Fuzzy Systems and Soft Computing
ISSN : 1819-4362

COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR FRAUD DETECTION IN BLOCKCHAIN

[1]Gajam Manoj Kumar, [2]Gande Sathvik, [3]Konakanchi Sairam, [4]J Crisil, [5]C.Vijay Kumar

[1,2,3,4] UG Scholar, Department of CSE (AI&ML)

[5]Assistant Professor, Department of CSE (AI&ML)

CMR Institute of Technology, Hyderabad, Telangana, India-501401

ABSTRACT:

Fraudulent transactions have a huge impact on the economy and trust of a blockchain network. Consensus algorithms like proof of work or proof of stake can verify the validity of the transaction but not the nature of the users involved in the transactions or those who verify the transactions. This makes a blockchain network still vulnerable to fraudulent activities. One of the ways to eliminate fraud is by using machine learning techniques. Machine learning can be of supervised or unsupervised nature. In this paper, we use various supervised machine learning techniques to check for fraudulent and legitimate transactions. We also provide an extensive comparative study of various supervised machine learning techniques like decision trees, Naive Bayes, logistic regression, multilayer perceptron, and so on for the above task .

INTRODUCTION:

The problem of detecting fraudulent transactions is being studied for a long time. Fraudulent transactions are harmful to the economy and discourage people from investing in bitcoins or even trusting other blockchain-based solutions. Fraudulent transactions are usually suspicious either in terms of participants involved in the transaction or the nature of the transaction. Members of a blockchain network want to detect Fraudulent transactions as soon as possible to prevent them from harming the blockchain network's community and integrity. Many Machine Learning techniques have been proposed to deal with this problem, some results appear to be quite promising [4], but there is no obvious superior method. This paper compares the performance of various supervised machine learning models like SVM, Decision Tree, Naive Bayes, Logistic Regression, and few deep learning models in detecting fraudulent transactions in a blockchain network. Such comparative study will help decide the best algorithm based on accuracy and computational speed trade-off. Our goal is to see which users and transactions have the highest probability of being involved in fraudulent transactions.

PROPOSED SYSTEM :

The workflow for detecting fraudulent activity is summarised in Figure 1. Essentially, after the Blockchain network has approved a transaction after all basic checks, our proposed system kicks in and does additional checks to detect if the

http://doi.org/10.36893/FSCC.2024.V09I01.109-120

transaction can be fraudulent. This approach makes sure that there is no extra overhead of even checking the transactions that the Blockchain network itself can easily invalidate.

The work done can be divided mainly into three phases:

1. Preprocessing phase

2. Building and training various models

3. Performance evaluation of all the models.

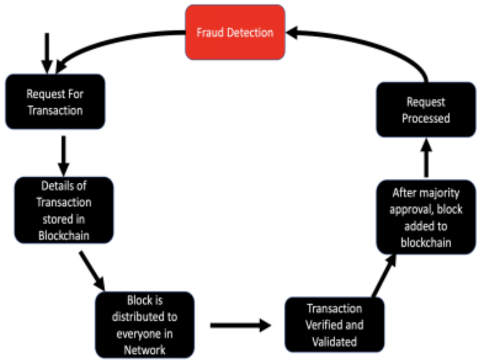

Fraud Detection
Request For Transaction
Details of Transaction stored in Blockchain
Block is distributed to everyone in Network
Transaction Verified and Validated
After majority approval, block added to blockchain
Request Processed

Fig. 1. Workflow of applying check for Fraud Detection

We preprocess using node-embedding in the network using the node2vec algorithm. Then, we read and convert the shorter version of concatenated rating dataset into a dataframe. Then, we create a function for the perception store features. This function extracts the features of a node using the ”source” and ”target” columns of the dataset. These features are then stored in a CSV file. We then run the node2vec algorithm in python and create a dictionary of nodes and corresponding embeddings. We also create a network edge list file and then reduce embeddings dimensionality for 2D projections. This dimensionnality reduction can be obtained using algorithms like t-SNE.

We then normalize the features extracted from the node2vec algorithm and create a file that contains the normalized values. We assign a score of 1 if the transaction is rated badly (fraud) and 0 otherwise. We then calculate the mean and standard deviation of the node features and save it to a CSV file. We then divide all our obtained data into train and test sets.

Phase II- Building various models, training and testing them.

We divide our data into train(0.8) and test(0.2) data. We then check the ratio of fraudulent and honest transactions in our train and test sets. We use machine following machine learning and deep learning models to predict if a transaction is fraudulent:

1.Logistic Regression: This is a simple linear classifier. Logistic regression works well for binary classification problems.

2. Multilayer Perceptron: Multilayer perceptron helps in separation data that cannot be classified using a linear classifier by introducing non linearity.

3. Naive Bayes: This model uses the Bayes theorem to calculate the probability of a transaction being fraudulent.

4. Adaboost: This is an ensemble learning method to boost the performance of binary classifiers.

5. Decision Tree: This classifier has a sequence of conditions and questions on data based on various features.

6. SVM: It uses a kernel method to transform the data in the dataset, and based on these transitions, it finds a boundary between all possible outputs.

7. Random Forest Classifier: This classifier fits a number of decision trees on small batches of the dataset.

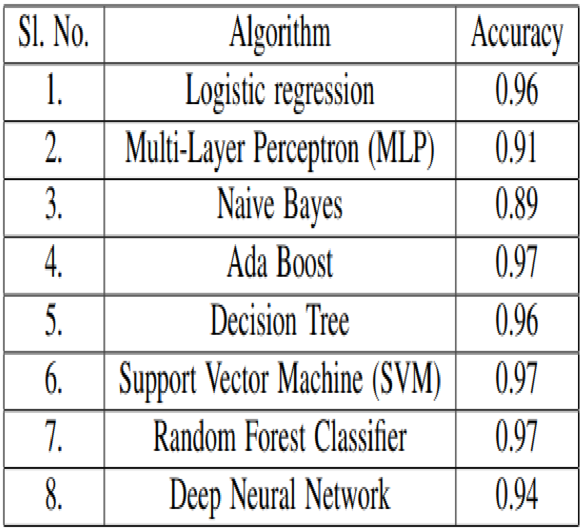
8. Neural Network: This model consists of six dense layers and four hidden layers. Relu and sigmoid were used as activation functions.

Phase III-Evaluation of models on test set

We evaluate all our classification models using bootstrap sampling. In machine learning, bootstrap sampling involves drawing sample data with replacement from the dataset to estimate a parameter. So we first choose the number of bootstrap samples. Then, we choose the sample size. Then, for each bootstrap sample, we draw a sample with chosen bootstrap size (with replacement) and test the sample's data. For this purpose, we use the accuracy metric, which is a standard metric used in machine learning problems. We then take the mean of all accuracies obtained in this fashion to evaluate the skill of our model.

EXITING SYSTEM:

We applied eight different supervised learning algorithms to the dataset. The dataset contains information about trust on different nodes or ratings given to them. This information is useful in detecting if a certain node's transaction can be fraudulent or not. The following table summarizes the accuracy obtained in each case.



Sl. No.	Algorithm	Accuracy
1.	Logistic regression	0.96
2.	Multi-Layer Perceptron (MLP)	0.91
3.	Naive Bayes	0.89
4.	Ada Boost	0.97
5.	Decision Tree	0.96
6.	Support Vector Machine (SVM)	0.97
7.	Random Forest Classifier	0.97
8.	Deep Neural Network	0.94

e observed that using Ada Boost, SVM, and Random Forest classifier gave the best results among the seven different algorithms. Also, since these algorithms already provide an accuracy of 97% we would like to build a fraud detector that will use the scores and decisions from the three algorithms together to decide if a transaction is fraudulent or not finally.

SYSTEM STUDY FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

SYSTEM DESIGN

UML DIAGRAMS :

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

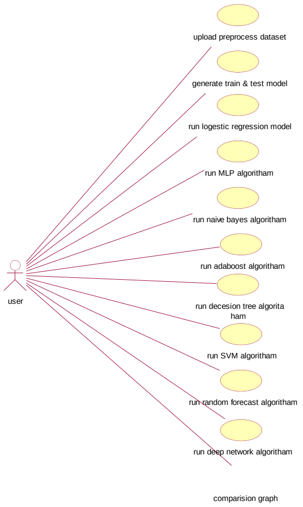
GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

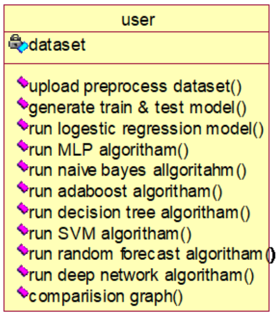
USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.


upload preprocess dataset
generate train & test model
run logestic regression model
run MLP algoritham
run naive bayes algoritham
run adaboost algoritham
run decesion tree algorita ham
run SVM algoritham
run random forecast algoritham
run deep network algoritham
comparision graph
user

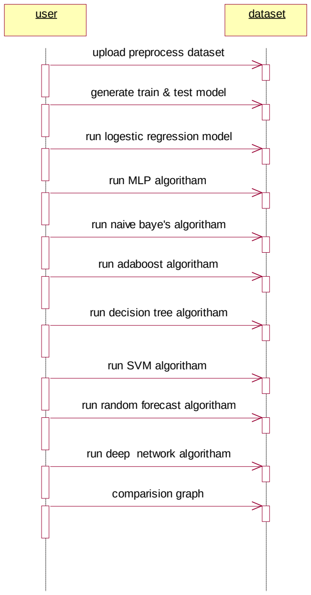
CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.


user
dataset
upload preprocess dataset()
generate train & test model()
run logestic regression model()
run MLP algoritham()
run naive bayes allgoritahm()
run adaboost algoritham()
run decision tree algoritham()
run SVM algoritham()
run random forecast algoritham()
run deep network algoritham()
compariision graph()

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

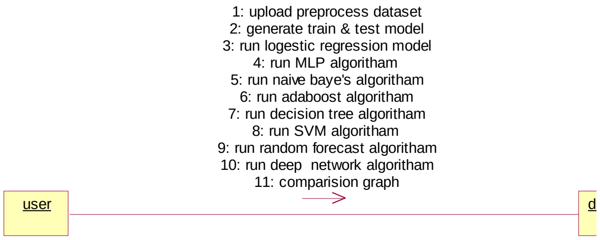

user
dataset
upload preprocess dataset
generate train & test model
run logestic regression model
run MLP algoritham
run naive baye's algoritham
run adaboost algoritham
run decision tree algoritham
run SVM algoritham
run random forecast algoritham
run deep network algoritham
comparision graph

COLLABRATION DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for

http://doi.org/10.36893/FSCC.2024.V09I01.109-120

choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.


1: upload preprocess dataset
2: generate train & test model
3: run logestic regression model
4: run MLP algoritham
5: run naive baye's algoritham
6: run adaboost algoritham
7: run decision tree algoritham
8: run SVM algoritham
9: run random forecast algoritham
10: run deep network algoritham
11: comparision graph
user
d

IMPLEMENTATION:

MODULES:

To implement this project we have designed following modules

Upload & Preprocess Dataset: button to upload and read dataset and then remove missing values

Generate Train & Test Model:button to get below output

In above screen we can see all data converted to numeric format and we can see total records found in dataset with total columns and then split dataset into train and test and now train and test data is ready and now click on each button to run all algorithms and get below output

we can see the performance or accuracy of each algorithm and below is the remaining algorithm accuracy

we can see accuracy of AdaBoost, Decision Tree and SVM and below is the accuracy of remaining algorithms

we can see random forest and Deep neural accuracy and in all algorithms Random Forest is giving better accuracy.

Comparison Graph:button to get below output

What is Machine Learning : -

Before we take a look at the details of various machine learning methods, let's start by looking at what machine learning is, and what it isn't. Machine learning is often categorized as a subfield of artificial intelligence, but I find that categorization can often be misleading at first brush. The study of machine learning certainly arose from research in this context, but in the data science application of machine learning methods, it's more helpful to think of machine learning as a means of *building models of data*.Fundamentally, machine learning involves building mathematical models to help understand data. "Learning" enters the fray when we give these models *tunable parameters* that can be adapted to observed data; in this way the program can be considered to be "learning" from the data. Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data. I'll leave to the reader the more philosophical digression regarding the extent to which this type of mathematical, model-based "learning" is similar to the "learning" exhibited by the human brain.Understanding the problem setting in machine learning is essential to using these tools effectively, and so we will start with some broad categorizations of the types of approaches we'll discuss here.

Categories Of Machine Leaning :-

At the most fundamental level, machine learning can be categorized into two main types: supervised learning and unsupervised learning.

Supervised learning involves somehow modeling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data. This is further subdivided into *classification* tasks and *regression* tasks: in classification, the labels are discrete categories, while in regression, the labels are continuous quantities. We will see examples of both types of supervised learning in the following section.

Unsupervised learning involves modeling the features of a dataset without reference to any label, and is often described as "letting the dataset speak for itself." These models include tasks such as *clustering* and *dimensionality reduction*. Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of the data. We will see examples of both types of unsupervised learning in the following section.

Need for Machine Learning

Human beings, at this moment, are the most intelligent and advanced species on earth because they can think, evaluate and solve complex problems. On the other side, AI is still in its initial stage and haven’t surpassed human intelligence in many aspects. Then the question is that what is the need to make machine learn? The most suitable reason for doing this is, “to make decisions, based on data, with efficiency and scale”.

Lately, organizations are investing heavily in newer technologies like Artificial Intelligence, Machine Learning and Deep Learning to get the key information from data to perform several real-world tasks and solve problems. We can call it data-driven decisions taken by machines, particularly to automate the process. These data-driven decisions can be used, instead of using programing logic, in the problems that cannot be programmed inherently. The fact is that we can’t do without human intelligence, but other aspect is that we all need to solve real-world problems with efficiency at a huge scale. That is why the need for machine learning arises.

Challenges in Machines Learning :-

While Machine Learning is rapidly evolving, making significant strides with cybersecurity and autonomous cars, this segment of AI as whole still has a long way to go. The reason behind is that ML has not been able to overcome number of challenges. The challenges that ML is facing currently are −

Quality of data − Having good-quality data for ML algorithms is one of the biggest challenges. Use of low-quality data leads to the problems related to data preprocessing and feature extraction.

Time-Consuming task − Another challenge faced by ML models is the consumption of time especially for data acquisition, feature extraction and retrieval.

Lack of specialist persons − As ML technology is still in its infancy stage, availability of expert resources is a tough job.

No clear objective for formulating business problems − Having no clear objective and well-defined goal for business problems is another key challenge for ML because this technology is not that mature yet.

Issue of overfitting & underfitting − If the model is overfitting or underfitting, it cannot be represented well for the problem.

Curse of dimensionality − Another challenge ML model faces is too many features of data points. This can be a real hindrance.

Difficulty in deployment − Complexity of the ML model makes it quite difficult to be deployed in real life.

Applications of Machines Learning :-

Machine Learning is the most rapidly growing technology and according to researchers we are in the golden year of AI and ML. It is used to solve many real-world complex problems which cannot be solved with traditional approach. Following are some real-world applications of ML

- Emotion analysis
- Sentiment analysis
- Error detection and prevention
- Weather forecasting and prediction
- Stock market analysis and forecasting
- Speech synthesis
- Speech recognition
- Customer segmentation
- Object recognition
- Fraud detection
- Fraud prevention
- Recommendation of products to customer in online shopping

CONCLUSION:

A method has been proposed for the detection of fraudulent transactions in a blockchain network using machine learning. In this method, various supervised learning approaches like support vector machines, decision trees, logistic regression, and dense neural networks were analyzed. A thorough comparative analysis of all the approaches is performed through accuracy. This work can be extended for the comparative study of unsupervised algorithms like clustering. In the future, we also plan to do an exhaustive study on fraudulent activities in a private blockchain

REFERENCES :

[1] Cai, Y., Zhu, D. Fraud detections for online businesses: a perspective from blockchain technology. Financ Innov 2, 20 (2016). https://doi.org/10.1186/s40854-016-0039-4

[2] Hyvarinen, H., Risius, M. & Friis, G. A Blockchain-Based Approach ¨ Towards Overcoming Financial Fraud in Public Sector Services. Bus Inf Syst Eng 59, 441–456 (2017). https://doi.org/10.1007/s12599-017-0502- 4

[3] Xu, J.J. Are blockchains immune to all malicious attacks?. Finance Innov 2, 25 (2016). https://doi.org/10.1186/s40854-016-0046-5

[4] Ostapowicz M., Zbikowski K. (2019) Detecting Fraudulent Accounts on ˙ Blockchain: A Supervised Approach. In: Cheng R., Mamoulis N., Sun Y., Huang X. (eds) Web Information Systems Engineering – WISE 2019. WISE 2020. Lecture Notes in Computer Science, vol 11881. Springer, Cham. https://doi.org/10.1007/978-3-030-34223-4 2

[5] Podgorelec, B., Turkanovic, M. and Karakati ´ c, S., 2020. A Machine ˇ Learning-Based Method for Automated Blockchain Transaction Signing Including Personalized Anomaly Detection. Sensors, 20(1), p.147.

[6] Farrugia S, Ellul J, Azzopardi G. Detection of illicit accounts over the Ethereum blockchain. Expert Systems with Applications. 2020 Jul 15;150:113318.

[7] Pham, Thai, and Steven Lee. ”Anomaly detection in bitcoin network using unsupervised learning methods.” arXiv preprint arXiv:1611.03941 (2016).

[8] Monamo, Patrick, Vukosi Marivate, and Bheki Twala. ”Unsupervised learning for robust Bitcoin fraud detection.” 2016 Information Security for South Africa (ISSA). IEEE, 2016.

[9] Shi, Fa-Bin, et al. ”Anomaly detection in Bitcoin market via price return analysis.” PloS one 14.6 (2019): e0218341.

[10] Li, Ji, et al. ”A Survey on Blockchain Anomaly Detection Using Data Mining Techniques.” International Conference on Blockchain and Trustworthy Systems. Springer, Singapore, 2019.

[11] P. N. Sureshbhai, P. Bhattacharya and S. Tanwar, ”KaRuNa: A Blockchain-Based Sentiment Analysis Framework for Fraud Cryptocurrency Schemes,” 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, 2020, pp. 1-6, doi: 10.1109/ICCWorkshops49005.2020.9145151.

[12] Brenig, Christian, and Gunter M ¨ uller. ”Economic analysis of cryptocur- ¨ rency backed money laundering.” (2015).

[13] Lorenz, Joana, et al. ”Machine learning methods to detect money laundering in the Bitcoin blockchain in the presence of label scarcity.” arXiv preprint arXiv:2005.14635 (2020).

[14] Bartoletti, Massimo, Barbara Pes, and Sergio Serusi. ”Data mining for detecting Bitcoin Ponzi schemes.” 2018 Crypto Valley Conference on Blockchain Technology (CVCBT). IEEE, 2018.

1. S.Dhanalakshmi,, M. Ganga Eswari1, “A Survey on Criminal Identification Using Multi Biometric Traits in Image Processing”, in International Journal of Computer Science and Engineering,(IJCSE), Volume 3,Issue:9, E-ISSN: 2347-2693, September 2015,PP 201-204
2. D.Vijasekar, S.Dhivya, amd, S.Dhanalakshmi,, “Survey on Detection of Glaucoma in Fundus Image by Segmentation and Classification”, International Journal of Engineering Research Technology(IJERT), Volume 4, Issue 9,ISSN:2278-0181, September 2015,PP 529-532
3. D.Vijayasekar, S.Dhivya, and S.Dhanalakshmi,, “Wiener Filter Operation on Blurred Images”, International Journal of Engineering Research Technology(IJERT), Volume 10, Special Issue 8,ISSN:2278-0181, September 2015,PP 197-200
4. M.Ganga Eswari, D.Vijayasekar, and, S.Dhanalakshmi,, “Criminal Identification Using Biometric Traits in

Image Processing", International Journal of Applied Engineering Research Technology(IJAER), ISSN 0973-4562 Vol. 10 No.85 (2015), October 2015,PP 265-268 (SCOPUS/Annexure-II Journals)

2. 15. K. P. Reddy, M. Satish, A. Prakash, S. M. Babu, P. P. Kumar and B. S. Devi, "Machine Learning Revolution in Early Disease Detection for Healthcare: Advancements, Challenges, and Future Prospects," 2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA),
3. 16. M. Satish, Prakash, S. M. Babu, P. P. Kumar, S. Devi and K. P. Reddy, "Artificial Intelligence (AI) and the Prediction of Climate Change Impacts," 2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), Hamburg, Germany, 2023, pp. 660-664, doi: 10.1109/ICCCMLA58983.2023. 10346636.
4. 17. Prakash, S. M. Babu, P. P. Kumar, S. Devi, K. P. Reddy and M. Satish, "Predicting Consumer Behaviour with Artificial Intelligence," *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, Hamburg, Germany, 2023 Hamburg, Germany, 2023, pp. 638-643, doi: 10.1109/ICCCMLA58983.2023. 10346963., pp. 687-692, doi: 10.1109/ICCCMLA58983.2023. 10346916.