ції стандартного та оптимізованого криптоалгоритмів AES. Програмна реалізація алгоритму AES, яка базувалась на офіційній специфікації стандарту шифрування, була модифікована для зменшення часу обробки даних з умовою збереження криптографічної стійкості шифру. Загальні принципи запропонованого методу полягають у перетворенні усіх двовимірних масивів на одномірні, додаванні допоміжних таблиць для операцій ShiftRows та MixColumns, об'єднанні операцій зі схожими принципами опрацювання елементів. Результати моделювання показали, що модифікована реалізація алгоритму AES демонструє скорочення часу обробки до 50% при шифруванні та до 75% при дешифруванні даних у порівнянні з відомими результатами.

Ключові слова: *криптографічна обробка даних, симетричний блоковий шифр, оптимізація обробки даних.*

Отримано: 03.09.2024

UDC 004.9; 004.8
DOI: 10.32626/2308-5916.2024-25.88-96

Myhajlo Sydoruk*,
Solomiia Liaskovska*,,** PhD

*Lviv Polytechnic National University, Lviv,
**Kingston University, London, United Kingdom

AN ENSEMBLE METHOD FOR THE FRAUD DETECTION IN TRANSACTIONS

In today's world, bank fraud has become one of the significant threats to the financial stability and security of clients of financial institutions. The development of technologies, in particular in the field of machine learning, opens up wide opportunities for building effective systems for detecting and preventing fraud in the banking sector [1, 2].

Detecting fraudulent transactions is an important task that requires thoughtful and technological solutions. One of these methods is the use of machine learning approaches and methods.

This paper proposes the use of an ensemble method that combines several machine learning models at once. This approach will reduce the probability of false positives and increase classification accuracy. In addition, for the optimal operation of the model, preprocessing of the data will be carried out, in particular, their normalization, balancing of classes, as well as the selection of features. During the research, it is important not only to achieve high accuracy, but also to reduce as much as possible the number of fraudulent transactions that will be mistakenly classified as normal [3]. This is related to the business requirements of the banking sector, as each such transaction causes losses to the system's reputation, as well as direct financial losses.

Within the framework of the study, it is substantiated that the use of this approach gives better classification results than single models due to the compensation of the shortcomings of each of them. The choice of this approach is also due to high practicality, compatibility with financial systems, as well as ease of integration.

This paper analyzes the proposed model, its advantages and disadvantages in comparison with analogues. The ensemble method helps to combine the advantages of simple models and reduce the impact of their shortcomings on the final result. In general, the choice of software should depend on the technical requirements of the project and to obtain better results, different models and approaches should be analyzed.

Keywords: *bank fraud, classification, classification, random forest, linear regression, decision tree, neural networks.*

Introduction. Solving the task of detecting fraudulent transactions in the banking sector requires a comprehensive and systematic approach, as fraud can take various forms and evolve depending on context and technological innovations.

Fraud detection may require analyzing large volumes of data from various sources, including customer transaction history, internal and external user behavior, device data, and geolocation. Various machine learning algorithms, such as classification, clustering, and association rules, can be used to solve the problem. It is important to choose the algorithm that best matches the specific type of data and the nature of the fraud.

For immediate response to suspicious activities, it is crucial to have a monitoring system that can quickly react to potential threats. Fraud detection may require analyzing not only the transaction itself but also contextual information such as the customer's previous transactions, typical behavior, etc.

One of the main challenges in developing a program for detecting fraudulent transactions is the large number of imbalanced datasets. Many of them contain a large number of non-fraudulent transactions and only a few that were not initiated by the cardholder. Therefore, the following points should be emphasized during the work:

Data heterogeneity: Transaction data can be heterogeneous in various parameters such as transaction amount, type of payment card, geographical location, etc., which can complicate the classification process.

Personal data and privacy: The processing and analysis of customers' personal data for fraud detection must take into account the privacy and confidentiality of this data and comply with relevant legal norms and regulations.

Classification model accuracy: The model may incorrectly classify transactions as fraudulent or legitimate, leading to false positives (when legitimate transactions are mistakenly identified as fraudulent) or false negatives (when fraudulent transactions are mistakenly identified as legitimate).

Adaptability to new fraud types: Fraudsters continuously improve their methods, so the classification model may fail to account for new types of fraud if the data is not sufficiently representative.

Despite all the aforementioned challenges, a model built on machine learning principles can accurately detect suspicious transactions.

Formulation of the problem.

The object of the research is classification algorithms, including machine learning models.

The subject of the research is a model for detecting fraudulent transactions based on a limited dataset.

The aim of the work is to create a model that can accurately and quickly determine, based on information about a bank transaction, whether it is safe or if there is a risk that the card data has been compromised by fraudsters.

Methods. Fraud detection in transactions involves data collection, processing, model selection and tuning, as well as evaluation and deployment.

Data should be collected from various sources, such as bank transactions, log files, customer data, and purchase history. These data should include timestamps, transaction amounts, transaction locations, types of transactions, and information about the payer and recipient. Access to internal databases may be required for this purpose.

A crucial step is processing the obtained datasets. We must remove duplicates, fill in missing data, and address anomalies. Afterward, the data should be normalized and clustered if necessary. Categorical data must be encoded into numerical values. Creating new informative features can significantly improve the model's performance. Alongside this, it is important to select the most relevant features. New features could include, for example, the average amount or the number of transactions made by a particular user within a day or other time period. Fraudulent transactions typically make up less than 1% of all payments, and as a result, the collected datasets will be imbalanced. Working with such data can be inefficient. This can be addressed using resampling methods, such as oversampling (e.g., SMOTE) or under-sampling [4-6]. After this, it is necessary to select a method and create a machine learning model. Since the task is a type of classification problem, logistic regression, decision trees, Random Forest, gradient boosting, and deep learning can be highlighted as possible solutions. Each of these models has its own advantages and disadvantages, and the final choice depends on specific requirements and conditions. Once the model is trained, it should be evaluated using appropriate metrics. The primary metrics include accuracy (Accuracy), recall (Recall), precision (Precision), F1-score, and ROC-AUC. These metrics will help assess how well the model identifies fraudulent transactions and distinguishes them from legitimate ones [7].

Logistic regression is the most common method for solving classification tasks, including fraud detection in transactions. It is based on modeling the probability that an object belongs to one of two classes by using a linear combination of input features. Logistic regression employs the sigmoid function to transform the linear combination of features into the probability that a transaction is fraudulent. Logistic regression models the probability that an observation belongs to a certain class. The formula for logistic regression is as follows:

$$P(y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)}},$$

where $P(y=1|X)$ – the probability that the target variable y equals 1 (for example, «fraudulent transaction») given the features X, β_0 – this is the intercept term (intercept), β_1, $\beta_2,\ldots,\beta_n$ – these are the regression coefficients corresponding to the influence of the features X_1, $X_2,\ldots,X_n$ – on the probability [6-8].

Decision tree is an effective tool for solving classification tasks, including fraud detection in transactions. The main idea is to build a model that divides the feature space into subspaces using rules based on feature values and makes decisions at each node of the tree. A decision tree consists of nodes, branches, and leaves, where each internal node represents a check on a feature, each branch represents the outcome of that check, and each leaf represents the final class or value (in the case of regression). The process of building a decision tree involves the iterative selection of the best feature to split the data at each step.

The root node of the decision tree selects the best feature for splitting the data using metrics. The splitting process continues recursively until one of the stopping criteria is reached: maximum tree depth, minimum number of samples required for a split, or minimum number of samples in a leaf. After building the decision tree, the model is evaluated using a test dataset.

To improve model performance, it is important to tune the hyperparameters of the decision tree, such as maximum tree depth, minimum number of samples required for a split, and minimum number of samples in a leaf. This can be done using cross-validation or grid search methods.

Decision tree is prone to overfitting, especially if they are deep and have many nodes. This can lead to the model performing well on training data but poorly generalizing to new data. Decision trees can also be unstable, as small changes in the data can lead to significant changes in the structure of the tree.

Neural networks are a powerful tool for solving classification tasks, including fraud detection in transactions. The main idea of neural net-

works lies in their ability to automatically detect complex patterns and relationships in data, making them effective for tasks involving high levels of complexity and dimensionality.

A neural network consists of several layers of neurons: an input layer, one or more hidden layers, and an output layer. The input layer receives the data, the hidden layers perform the bulk of the work by detecting patterns, and the output layer provides the classification result. Each neuron in a layer is connected to all neurons in the next layer through weights that are learned during the training of the network. Activation functions, such as ReLU (Rectified Linear Unit) or sigmoid, are used to introduce non-linearity and improve learning [7-9].

The training process of a neural network involves data preparation. This includes cleaning the data, normalizing numerical features, encoding categorical variables, and possibly using techniques to handle imbalanced class distributions. The next step is to choose the model architecture, which can range from simple models with one or two hidden layers to complex deep neural networks with many layers. Once the architecture is defined, the model is trained using the backpropagation algorithm and optimizers such as Adam or SGD (Stochastic Gradient Descent). Training involves feeding data into the network, calculating errors, updating weights based on gradients, and repeating this process until desired results are achieved. After training, the model is evaluated using a test dataset with various metrics.

Data Processing and Model Architecture Building. Fraudulent transaction datasets are typically class-imbalanced. Valid transactions usually make up over 95%, which complicates model training. In such cases, synthetic data generation should be employed. The main idea of this approach is to interpolate between existing samples and create new ones based on this interpolation. One of the simple ensemble methods to use is logistic regression, as it has a small number of parameters, which reduces the risk of overfitting, and its results are easy for specialists to interpret and understand why a particular transaction was flagged as fraudulent.

Neural networks are capable of detecting complex dependencies that simpler models may not uncover. They are also robust to noise and incorrect data, which allows them to produce results that may differ from other ensemble models. In this task, where the sequence of transactions matters, this can be addressed by adding the time of the last transaction and its status as additional attributes to the dataset. This approach allows the use of standard deep neural networks (DNNs) instead of recurrent neural networks (RNNs), leading to faster model training. Compared to regression models, neural networks approach data analysis from a different perspective, which reduces the risk of simultaneous errors in both models. The results of the models should be combined to obtain a final assessment. A good model for this task would be one

based on decision trees or random forests. For this task, a decision tree is preferable due to its simplicity and interpretability. Random forests, on the other hand, are prone to overfitting on simpler datasets.

Results and discussion. For this task, a stacking ensemble method is suitable. Stacking (stacking ensemble) is an ensemble technique that combines predictions from multiple models to improve overall performance. The core idea is to use an additional model trained on the predictions of base models to make the final decision, rather than using simple aggregation methods like voting or averaging.

Thus, the training dataset will be divided between the logistic regression model and the neural network. This approach requires a sufficiently large dataset due to the risk of overfitting. Both models will output the probability that a given transaction is fraudulent. The decision tree will then produce the final result based on these two probabilities and the input transaction data. This helps to reduce the risk of overfitting in the final model. Figure 1 demonstrates the architecture of the neural network

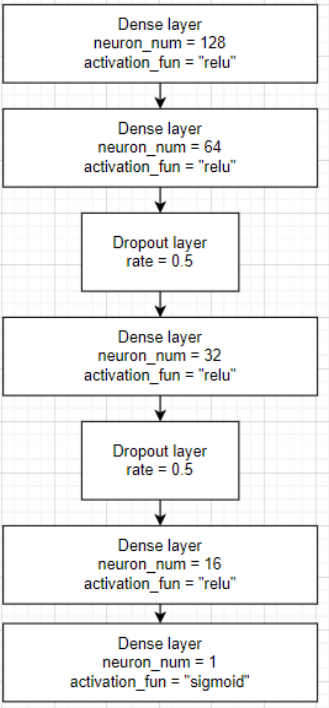

Fig. 1. *Architecture of a Deep Neural Network for Binary Classification*

The first layer is always the input layer. The number of neurons in this layer corresponds to the number of features being analyzed. Following

this, hidden layers and dropout layers are added to prevent overfitting by removing some connections. Three hidden layers should be included, each with a different number of neurons. The final layer will be the output layer with 1 neuron due to the nature of the binary classification task. The best activation function for this task is the sigmoid function, as it converts the output to a range of [0, 1], allowing it to be interpreted as a probability.

Figure 2 demonstrates the Architecture of the Ensemble Method.

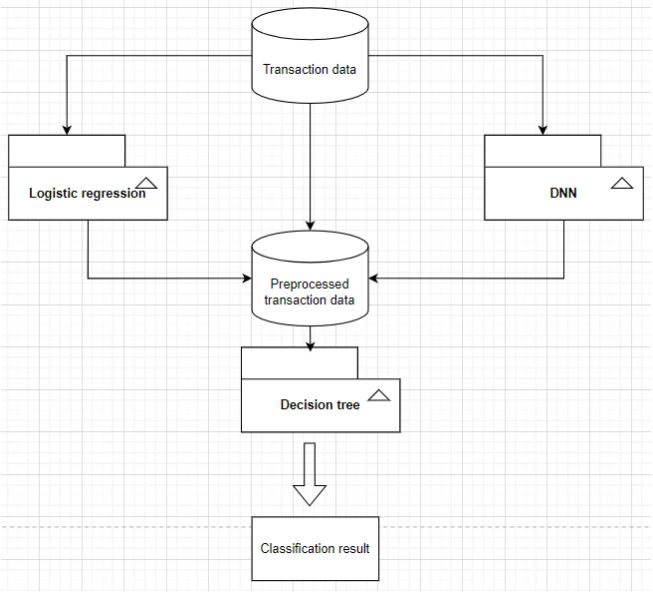

Fig. 2. *Architecture of the Ensemble Model for Binary Classification*

Experiments and Results Analysis. The final testing of the proposed architecture was conducted on a dataset consisting of 2,000 transaction records. Of these, 5%, or 100 instances, were identified as fraudulent. The results of the experiment, including accuracy values and the confusion matrix, are recorded in Table 1.

Table 1

The results of the experiment, including accuracy values and the confusion matrix

	Accuracy	**True positive**	**True negative**	**False positive**	**False negative**
Logistic Regression	90.2%	1780	24	26	170
Decision Tree	81%	1604	17	33	346
Neural Network	93.4%	1827	39	11	123
Ensemble Method	96.9%	1892	46	4	58

As we can see, the use of the ensemble method resulted in higher accuracy; more importantly, the final model significantly reduced the number of fraudulent transactions that were incorrectly identified as valid.

Conclusion. This paper explored the application of ensemble methods for detecting fraudulent transactions, which is one of the most complex and critical tasks in financial analytics and security. The study examined ensemble methods, as well as logistic regression, random forests, decision trees, and neural networks. Specifically, it constructed a model architecture based on these methods.

The research found that the optimal combination of models is a stacking ensemble, where logistic regression and neural networks analyze the stream of input data, while the decision tree, leveraging its different approach to identifying dependencies, provides the final result. This model will analyze a wide range of transaction patterns and, as a result, have increased accuracy.

Decision tree is typically prone to overfitting, especially with unbalanced datasets. However, in the proposed model, it will function as a meta-model and effectively combine the predictions of other models. Additionally, the decision tree will not work with raw unprocessed data, which positively affects the risk of overfitting.

The drawbacks of this approach include the complexity of tuning and the need for a large dataset for analysis. An effective stacking ensemble requires time for proper model tuning, and small datasets may lead to decreased performance.

Acknowledgment. The British Academy's Researchers at Risk Fellowships Programme supportsthis research.

References:

1. Islam M. A., Uddin M. A., Aryal S., Stea G. An ensemble learning approach for anomaly detection in credit card data with imbalanced and overlapped classes. *Journal of Information Security and Applications.* 2023. Vol. 78. URL: https://www.sciencedirect.com/science/article/pii/S2214212623002028.
2. Abdul Salam M., Fouad K. M., Elbably D. L., Elsayed S. M. Federated learning model for credit card fraud detection with data balancing techniques. *Neural Computing and Applications.* 2024. Vol. 36, No. 11. P. 6231-6256. URL: https://link.springer.com/article/10.1007/s00521-023-09410-2.
3. Almarshad F. A., Gashgari G. A., Alzahrani A. I. A. Generative adversarial networks-based novel approach for fraud detection for the european cardholders 2013 dataset. *IEEE Access.* 2023. Vol. 11. P. 107348-107368. URL: https://ieeexplore.ieee.org/document/10265011.
4. Документація Tensorflow. URL: https://www.tensorflow.org/api_docs.
5. Ансамблеві моделі для класифікації. URL: https://www.kaggle.com/-code/emstrakhov/ensembles-in-machine-learning.
6. Hancock J. T., Bauder R. A., Wang H., Khoshgoftaar T. M. Explainable machine learning models for medicare fraud detection. *Journal of Big Data.* 2023. Vol. 10, No. 1.

7. Alsayaydeh J. A. J., Aziz A., Rahman A. I. A. et al. Development of programmable home security using GSM system for early prevention. *ARPN Journal of Engineering and Applied Sciences.* 2021. Vol. 16. No. 1. P. 88-97.
8. Fedorchenko I., Oliinyk A., Alsayaydeh J. A. J. et al. Modified genetic algorithm to determine the location of the distribution power supply networks in the city. 2020. DOI: 10.5281/zenodo.5163692.
9. Shakhovska N., Liaskovsky D., Augousti A., Liaskovska S., Martyn Y. Design and Deployment of Data Developer Toolkit in Cloud Manufacturing Environments. *CEUR-WS.* 2024. Vol. 3699. P. 47-56.

АНСАМБЛЕВИЙ МЕТОД ДЛЯ ВИЯВЛЕННЯ ШАХРАЙСТВА В ТРАНЗАКЦІЯХ

У сучасному світі банківське шахрайство стало однією зі значущих загроз фінансовій стабільності та безпеці клієнтів фінансових установ. Розвиток технологій, зокрема в галузі машинного навчання, відкриває широкі можливості для побудови ефективних систем виявлення та запобігання шахрайству в банківській сфері.

Виявлення шахрайських транзакцій є важливим завданням, що потребує продуманих та технологічних рішень. Одним з таких методів є використання підходів та методів машинного навчання.

В даній роботі пропонується використання ансамблевого методу, який поєднує одразу кілька моделей машинного навчання. Такий підхід дозволить зменшити ймовірність помилкових спрацювань та підвищити точність класифікації. Окрім цього для оптимальної роботи моделі буде проведений препроцесинг даних, зокрема їх нормалізація, балансування класів а також вибір ознак. В ході дослідження важливо не лише досягти високої точності, а й якомога сильніше зменшити кількість шахрайських транзакцій, що будуть помилково класифіковані як нормальні. Це пов'язано з бізнес вимогами банківської сфери, оскільки кожна така транзакція завдає втрат репутації системи а також безпосередньо фінансових збитків.

У рамках дослідження обгрунтовано, що використання даного підходу дає кращі результати класифікації, ніж одиночні моделі завдяки компенсації недоліків кожної з них. Вибір даного підходу зумовлений також високою практичністю, сумісністю з фінансовими системами а також простотою інтеграції.

В даній роботі проведено аналіз запропонованої моделі, її переваги та недоліки у порівнянні з аналогами. Ансамблевий метод допомагає поєднати переваги простих моделей та зменшити вплив їхніх недоліків на кінцевий результат. В загальному, вибір програмного забезпечення повинен залежати від технічних вимог проекту і для отримання кращих результатів слід аналізувати різні моделі та підходи.

Ключові слова: *банківське шахрайство, класифікація, випадковий ліс, лінійна регресія, дерево рішень, нейронні мережі.*

Отримано: 31.08.2024