LEVERAGING SMOTE AND RANDOM FOREST FOR IMPROVED CREDIT CARD FRAUD DETECTION

Maddala Ruchita[1], Maridu Bhargavi[2], Maddala Rakshita[3], Bellamkonda Chaitanya Nandini[4], and Irfan Aziz[5]

[1,2,3,4,5]Department of CSE, Vignan's Foundation for Science, Technology and Research, Vadlamudi, Guntur, Andhra Pradesh, India

[1]ruchita.2801@gmail.com

[2]bhargaviformal@gmail.com

[3]rakshita.2801@gmail.com

[4]bellamkondachaitanyanandini@gmail.com

[5]azizirfan387@gmail.com

Abstract

The accelerating speed of digital transactions has made credit card theft a serious concern to customers and banks alike. Traditional means of fraud detection are no longer as strong as today's more sophisticated methods. This paper presents a summary of various machine learning techniques applied in credit card fraud detection. Several algorithms are highlighted, including logistic regression, random forests, and boosting methods such as gradient boosting and LightGBM. These algorithms were selected due to their capacity to manage big datasets and identify patterns unique to fraud. According to the results, the best performance was obtained by random forests, which achieved 99.30% accuracy, outperforming all other methods. Gradient boosting and logistic regression were also competitive, reaching 98.5% accuracy. This study offers recommendations for improving fraud detection techniques while showcasing the efficacy of machine learning in this regard.

Key words: Fraud detection, accuracy, random forest, machine learning, logistic regression, credit card fraud, and boosting algorithms.

1 Introduction

Credit card fraud refers to the unauthorized gain of cardholder information through telephone scams, text message phishing, or hacking of the Internet. Most of this is usually done with the help of software tools to execute illegal

transactions. As a result, credit card transactions are now sent through a verification module to prevent such fraudulent acts. In the process of verification, the details of the transaction are analyzed to check if there are any fraudulent signs. If detected as fraudulent, the transaction is straightaway blocked by the verification module [1].

Traditional rule-based fraud detection systems are not very effective, as they cannot keep up with new, hidden fraud schemes. For this reason, machine learning techniques are becoming more popular because they can learn from fraud tactics that change over time. By identifying hidden patterns in transaction data, these algorithms provide a potent substitute, simplifying the process of distinguishing between legitimate and fraudulent transactions.

Handling the problem of uneven class distribution is one of the main obstacles in the use of machine learning to credit card fraud detection.Fraud makes up less than 1% of the dataset, making it difficult for most algorithms to correctly identify fraud. There are different ways to handle this issue. For example, SMOTE can create synthetic fraud samples to balance the dataset. Additionally, ensemble methods like Random Forest, Gradient Boosting, XGBoost, and AdaBoost show great potential to improve the fraud detection rate.

In this paper, we compare these algorithms, including Logistic Regression, Random Forest, and several boosting algorithms. We also discuss how feature selection techniques affect performance and evaluate the models using metrics like precision, recall, and F1-score. This study aims to solve the problems of imbalanced datasets by using a combination of strong algorithms to improve detection accuracy for credit card fraud and reduce false positives.

2 Literature Survey

Kumar et al. [2] conducted a comprehensive study on credit card fraud detection, exploring various techniques and methodologies aimed at addressing the complexities of fraudulent activities in financial transactions. Their research emphasizes the significance of machine learning (ML) algorithms in identifying fraudulent transactions through the analysis of historical data.

Alarfaj et al. [3] explored the detection of credit card fraud by utilizing a blend of cutting-edge machine learning and deep learning techniques. Their work demonstrates the effectiveness of various algorithms in accurately detecting fraudulent transactions. The authors conducted extensive experiments, comparing multiple approaches to identify the best-performing models. They concluded that integrating traditional machine learning methods with advanced deep learning techniques significantly enhances detection capabilities, making it a robust solution for combating credit card fraud.

Negi et al. [4] presented a study focused on detecting credit card fraud by using both deep learning and machine learning methods. Their research highlights the importance of applying advanced algorithms to boost the accuracy of fraud detection models. The researchers conducted experiments that demonstrated the effectiveness of their framework in accurately identifying fraudulent transactions.

Sailusha et al. [5] conducted a comprehensive study on credit card fraud

detection, focusing on machine learning algorithms such as Random Forest and Adaboost. Their research emphasizes the significance of these algorithms in classifying fraudulent and non-fraudulent transactions through accuracy, precision, recall, and F1-score evaluations.

3 Methodology

The dataset used in this research is the Credit Card Fraud Detection dataset, consisting of 284,807 transaction records in total. However, the dataset only contains 492 fraudulent transactions (Class 1). This certainly adds complexity to predictive modeling, as majority class models tend to bias toward the majority class, in this case, the legitimate transactions.

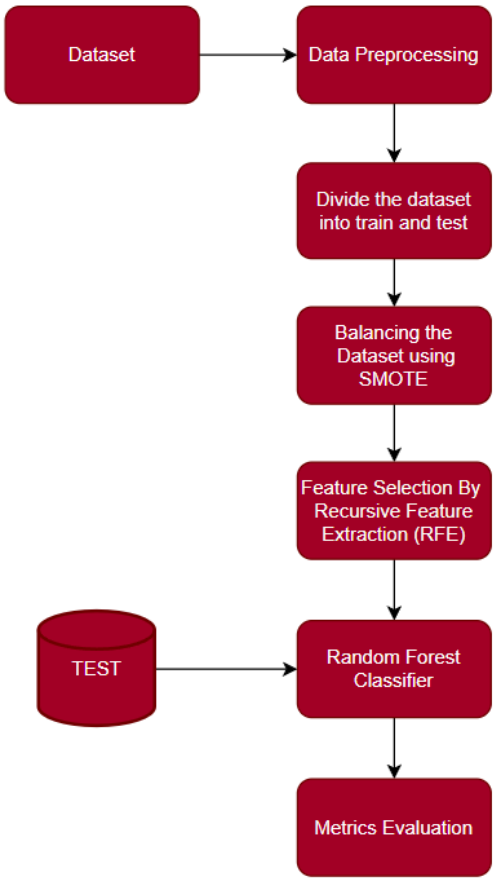


Figure 1: Proposed model flowchart

The proposed methodology for credit card fraud detection involves several critical steps in developing a robust and efficient model. The methodology highlights the following key components:

The **Algorithm 1** of the proposed model, as illustrated below which includes data preprocessing, feature selection through Recursive Feature Elimination (RFE), and training multiple machine learning algorithms such as Logistic Regression, Random Forest, and XGBoost. A Voting Classifier is then employed to combine the predictions of these models, enhancing the detection of fraudulent transactions.

Algorithm 1 Credit Card Fraud Detection Model Process

Input: Dataset
$D = \{(x_i, y_i)\}_{i=1}^{n}$, where x_i are features and $y_i \in \{0, 1\}$ is the label (0 = legitimate, 1 = fraud)
Output: Best-performing model with metrics
Step 1: Preprocessing
- Remove duplicate entries from dataset D:

$$D = \{(x_i, y_i) \mid (x_i, y_i) is unique\} \tag{1}$$

- Split D into train (80%) and test (20%) sets:

$$D_{train}, D_{test} = split(D, 0.8, 0.2) \tag{2}$$

Step 2: Balancing the dataset using SMOTE
- Apply SMOTE to handle class imbalance on the training data:

$$D_{train}^{balanced} = SMOTE(D_{train}) \tag{3}$$

Step 3: Feature Selection using RFE
- Use Recursive Feature Elimination (RFE) to extract k important features:

$$F_k = RFE(D_{train}^{balanced}, k) \tag{4}$$

Step 4: Apply machine learning models
- Train the following models on F_k:
M = { Logistic Regression, Random Forest, Gradient Boosting, XGBoost, AdaBoost, CatBoost, LightGBM, Voting Low, Voting High }
Step 5: Model Evaluation
- For each model $m \in M$, evaluate using the following metrics:Accuracy Precision Recall F1-Score

- Random Forest provides the best performance for these metrics
Return Best model with metrics

3.1 Data Load and Initial Exploration

The data has been imported into a Pandas DataFrame so that its structure, value types, and any missing values may be examined first. A general overview using the `info()` method provided information related to the dataset; confirmation about the presence of null values was done using the `isnull().sum()`

method.

3.2 Class Distribution Analysis

The target variable, `Class`, was analyzed to gain insight into the distribution between legitimate and fraudulent transactions. Value counts showed a significant imbalance, with most transactions being legitimate. Additionally, the statistical summary of the `Amount` feature will be inspected individually for the two classes to capture the stark contrasts in transaction amounts.

3.3 Splitting the Dataset

The dataset was first separated into independent variables (X) and the target variable (Y). Afterward, the dataset was split into two parts, with 80% allocated for training and the remaining 20% set aside for testing. This approach guarantees that the model has ample data for learning during training while still preserving enough data for an unbiased assessment.

3.4 Handling Class Imbalance

SMOTE was used to bolster the training data by creating synthetic instances for the minority class—more precisely, the fraudulent transactions—in order to address the class imbalance. Because of the imbalanced dataset, this method aids in balancing the class distribution and reducing model bias. By looking at the distribution following resampling, it was possible to determine that SMOTE was effective in providing a more fair representation of the classes after balancing.

3.5 Feature Selection

After applying oversampling, feature selection techniques are utilized to determine the most significant features for fraud detection. This step enhances model performance by reducing dimensionality and focusing on the features that contribute the most to predicting fraudulent transactions. Methods such as Recursive Feature Elimination (RFE) are employed to identify the best features according to their predictive potential. The process allows better accuracy of the model and provides a clear understanding of the significant factors that contribute to fraudulent cards in credit card transactions.

3.6 Proposed Machine Learning Algorithms

We developed six machine learning models: Random Forest, Logistic Regression, and five boosting algorithms, namely Gradient Boosting, XGBoost, AdaBoost, CatBoost, and LightGBM. Each model contributes significantly to enhancing the accuracy of detecting fraudulent credit card transactions. Descriptions of these models are provided below.

3.6.1 Logistic Regression

Logistic Regression classifies transactions as fraudulent or legitimate by using the logistic function, which is defined as follows:

$$P(Fraud|X) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n)}} \tag{5}$$

In this equation, $P(Fraud|X)$ represents the probability of a transaction being fraudulent, α_0 indicates the intercept, while $\alpha_1, \alpha_2, \ldots, \alpha_n$ are the coefficients corresponding to the features $X_1, X_2, \ldots, X_n$. This straightforward structure facilitates the learning of complex relationships between features and fraud probability, ultimately enhancing the accuracy of the fraud detection system.

3.6.2 Random Forest

Random Forest generates many decision trees and predicts the mode of their classifications for detection purposes. The final prediction can be expressed as:

$$\hat{Z} = mode(R_1(X), R_2(X), \ldots, R_k(X)) \tag{6}$$

where $R_i(X)$ represents the predictions from the separate trees and k is the total number of trees. Because it can capture intricate correlations between characteristics and handle high-dimensional datasets with efficiency, this model is well-suited for fraud detection.

3.6.3 Gradient Boosting

Gradient Boosting builds a sequence of models in which each new model is fit to correctly classify the misclassifications of the previous models. The output generated by the model after iteration m can be expressed as:

$$F_m(Z) = F_{m-1}(Z) + \eta h_m(Z) \tag{7}$$

where $F_{m-1}(Z)$ denotes the prediction made by the preceding model, η represents the learning rate, and $h_m(Z)$ signifies the new weak learner introduced at iteration m. This enhancement refines the model's predictions to better identify fraudulent transactions.

3.6.4 XGBoost

XGBoost is a highly efficient and scalable variant of gradient boosting that utilizes regularization techniques to reduce the risk of overfitting. The function is formulated as:

$$\mathcal{L} = \sum_{i=1}^{m} l(t_i, \hat{t}_i) + \sum_{j=1}^{J} \Omega(h_j) \tag{8}$$

In this mathematical equation, l represents the loss function, t_i denotes the actual value, $\hat{t}_i$ indicates the predicted value, and Ω signifies the regularization term associated with each function h_j. XGBoost excels in managing large datasets and capturing complex patterns in credit card fraud detection.

3.6.5 AdaBoost

AdaBoost integrates several weak classifiers to create a robust classifier. The prediction can be formulated as:

$$G(X) = \sum_{k=1}^{K} \beta_k f_k(X) \tag{9}$$

In this equation, β_k indicates the weight assigned to the k-th weak classifier $f_k(X)$, and K represents the total number of weak classifiers. This approach focuses on challenging cases to enhance the model's accuracy in identifying fraudulent transactions.

3.6.6 LightGBM

LightGBM achieves efficiency and scalability even for large datasets. The objective function for LightGBM can be described as:

$$\mathcal{F} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{j=1}^{J} \Omega(f_j) \tag{10}$$

This is similar to XGBoost, where $\mathcal{F}$ is the overall loss function. LightGBM enhances the speed and accuracy of a model while reducing memory usage, making it suitable for real-time applications such as credit card fraud detection.

By using these models, the individual advantages of each model are combined to improve the overall dependability and accuracy of fraud detection systems.

4 Metrics

The performance of our ensemble model is measured according to several essential metrics, which are spelled out in the section that follows:

4.1 Confusion Matrix

A confusion matrix is essential in machine learning as it evaluates the effectiveness of classification models. It compares the predicted classes with the actual classes, summarizing the model's effectiveness by presenting counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This information can yield insights into the model's capacity to differentiate between legitimate and fraudulent transactions.

4.2 Precision

Precision quantifies the frequency of accurately anticipated positive cases that actually occur. The ratio of genuine positive instances to all cases that have been labeled as such is known as precision. Formulaically, precision is expressed as:

$$Precision = \frac{A}{A+B} \tag{11}$$

4.3 Recall

Recall is an evaluation of how good the model is in picking out all the actual positives from the total positive count. It represents how well it can classify frauds. The recall formula is stated to be:

$$Recall = \frac{A}{A+C} \tag{12}$$

4.4 F1-Score

The harmonic mean of recall and accuracy, or the F1-score, provides a fair comparison of the two measurements. When class distributions are unbalanced, it is highly helpful. The following formula is used to get the F1-score:

$$F1Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{13}$$

4.5 Accuracy

The effectiveness of a classification model in identifying things is measured by its accuracy. It may be expressed as follows: It is defined as the ratio of successfully predicted instances to all occurrences in the dataset.

$$Accuracy = \frac{A+D}{A+D+B+C} \tag{14}$$

Variable Definitions
A = Correctly Predicted Positives (TP)
B = Incorrectly Predicted Positives (FP)
C = Missed Positives (FN)
D = Correctly Predicted Negatives (TN)

4.6 Evaluation of Proposed Model vs. Existing Model

In order to improve performance measures beyond accuracy, we examined a variety of machine learning methods for credit card fraud detection in our study.

The current model from the base study largely tested Logistic Regression, Random Forest, and Naive Bayes. The findings showed that:

- Naive Bayes achieved an accuracy of 99.30%.
- Random Forest and Logistic Regression both recorded an accuracy of 98.5% [2].

In contrast, our proposed model demonstrated superior performance metrics as follows:

- With a precision of 1.00, recall of 0.87, and an F1-score of 0.81, the Random Forest model has an accuracy of 99.93%.
- Other models, such as Gradient Boosting and XGBoost, also performed admirably with accuracies around 99.87%.
- Our Voting Classifier (Soft) achieved an accuracy of 99.83%, demonstrating the efficacy of ensemble methods in fraud detection.

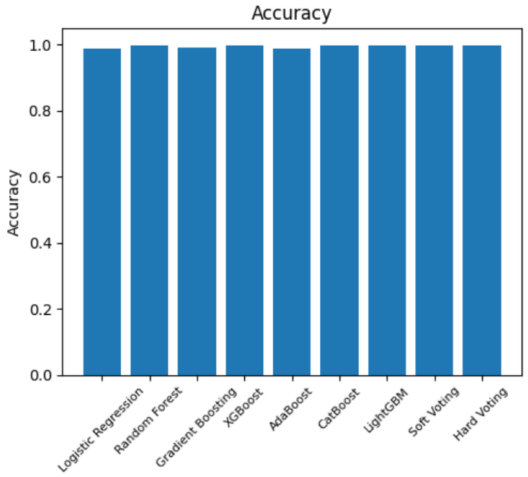
Table 1: Comparison of Performance Metrics

Model	**Accuracy**
Existing Model (Naive Bayes)	99.30%
Existing Model (Random Forest)	98.5%
Proposed Model (Random Forest)	**99.93%**
Proposed Model (Voting Classifier)	99.83%

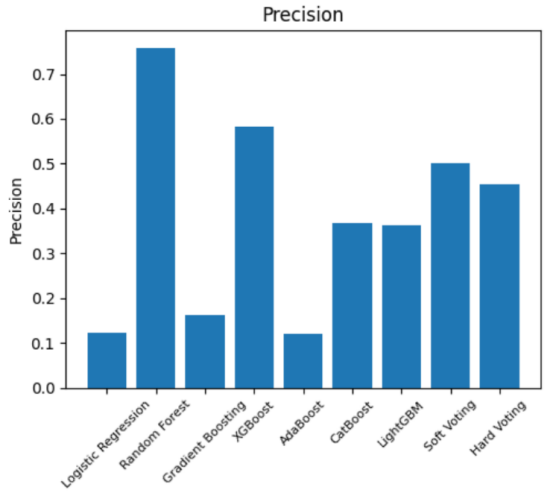
5 Results

In this part, we present the results achieved after processing the suggested machine learning techniques for fraud detection against a credit card. For this purpose, we evaluate the performance of each model against all metrics described above, namely precision, recall, F1-score, and accuracy. The outcomes are thoroughly examined to offer valuable perspectives on the efficacy of the used models.

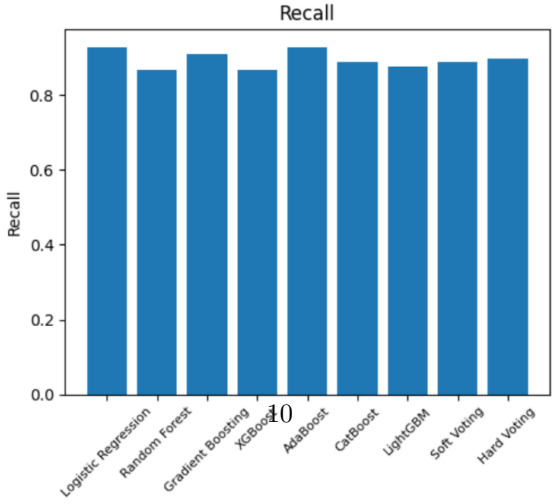
Table 2 summarizes the performance of various machine learning algorithms toward credit card fraud detection. The conclusion reached is that ensemble models, particularly the Random Forest and Voting Classifiers, outperform traditional algorithms like Logistic Regression and AdaBoost in terms of the metrics presented.



(a) Accuracy



(b) Precision



(c) Recall

Figure 2: Performance Metrics

Table 2: Performance Metrics of the Proposed Models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.99	1.00	0.93	0.22
Random Forest	**0.9993**	**1.00**	**0.87**	**0.81**
Gradient Boosting	0.9987	1.00	0.87	0.70
XGBoost	0.9987	1.00	0.87	0.70
AdaBoost	0.9882	1.00	0.93	0.21
CatBoost	0.9972	1.00	0.89	0.52
LightGBM	0.9971	1.00	0.88	0.51
Voting Classifier (Soft)	0.9983	1.00	0.89	0.64
Voting Classifier (Hard)	0.9980	1.00	0.90	0.60

It was the most accurate model and had strong precision and recall values, thereby proving that this model catches fraudulent transactions effectively. In fact, this shows the output in the base paper, which identifies Random Forest as a top performer for fraud detection tasks.

A Voting Classifier with both soft and hard voting balanced the resulting precision and recall well. Soft voting especially proved to have better precision than hard voting, which reveals the desirable effect of prediction aggregation by multiple models to strengthen the overall performance of the model.

While Logistic Regression and AdaBoost provided a great rate in terms of recall, they suffered from very low precision values that led to many false positives. The weakness brought about by this case necessitates more powerful methods to avoid false alarms in real-life application scenarios.

Such analysis establishes the fact that ensemble methods are crucial for improving fraud detection capabilities. Since these models can function using multiple classifiers and combine them, this builds better generalization and robustness in unseen data.

Overall findings show that traditional models can recognize fraudulent transactions; however, ensemble techniques significantly enhance precision and recall, making ensemble a better choice for credit card fraud detection.

6 Conclusion

In this study, we analyzed various machine learning algorithms and evaluated their performance based on accuracy, precision, recall, and F1-score in the context of credit card fraud detection. The final results confirm that ensemble methods, in particular Random Forest and Voting Classifiers, surpass classic models such as Logistic Regression and AdaBoost in detecting fraudulent transactions.

The best performance is related to the Random Forest model, as it achieved the highest accuracy along a strong balance between recall and precision, favoring this method as one of the top algorithms for detecting fraud. However, training Voting Classifiers showed further improvement in performance, as the combination of the outputs of multiple models into one was especially useful for tuning the trade-off between precision and recall.

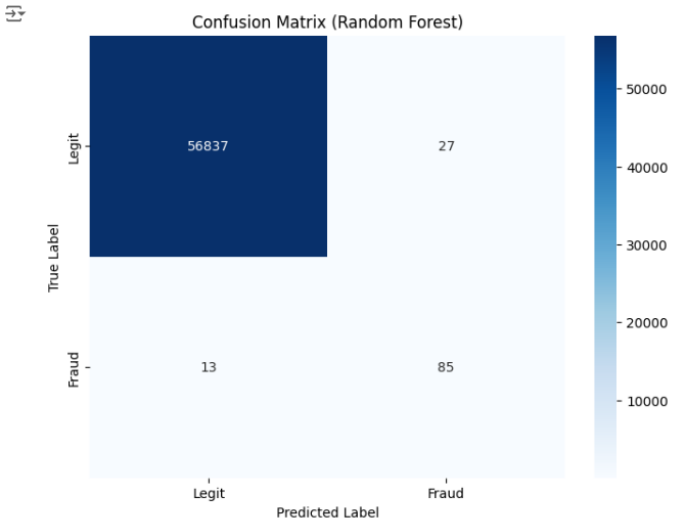


Figure 3: Confusion Matrix of Random Forest

Advanced techniques have been applied: SMOTE, which addresses imbalanced datasets, and Recursive Feature Elimination (RFE) to select features. The above-mentioned techniques have significantly contributed to the overall effectiveness of the models deployed. The study emphasizes the dire need for a strong ensemble approach that could find complex patterns of fraudulent activities with better reliability and minimal false positives.

Future research will consider other ensemble techniques that may be employed on real-time data analysis to further boost the adaptability and effectiveness of fraud detection models. The results have confirmed that an integrated approach with different machine learning algorithms applied may lead to much more effective solutions in combating credit card fraud.

7 References

[1] Bhakta, S. S., Ghosh, S., & Sadhukhan, B. (2023). Credit Card Fraud Detection Using Machine Learning: A Comparative Study of Ensemble Learning Algorithms. In *2023 9th International Conference on Smart Computing and Communications (ICSCC)* (pp. 296-301). Kochi, Kerala, India. doi: 10.1109/ICSCC59169.2023.10335075.

[2] Kumar, A., Poojitha, M. V., Anuhya, T., Srinivas, K., & Bhargavi, M. (2024). Credit Card Fraud Detection. In *2024 8th International Conference on Inventive Systems and Control (ICISC)* (pp. 79-82). Coimbatore, India. doi: 10.1109/ICISC62624.2024.00020.

[3] Alarfaj, F. K., Malik, I., Khan, H. U., Almusallam, N., Ramzan, M., & Ahmed, M. (2022). Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms. *IEEE Access*, 10, 39700-39715. doi: 10.1109/ACCESS.2022.3166891.

[4] Negi, S., Das, S. K., & Bodh, R. (2022). Credit Card Fraud Detection using Deep and Machine Learning. In *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 455-461). Salem, India. doi: 10.1109/ICAAIC53929.2022.9792941.

[5] R. Sailusha, V. Gnaneswar, R. Ramesh and G. R. Rao, "Credit Card Fraud Detection Using Machine Learning," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 1264-1270, doi: 10.1109/ICICCS48265.2020.9121114.