Loan Approval Prediction using Machine Learning

|  |  |  |
| --- | --- | --- |
| Dr. Rama Krishna Eluri  Assistant Professor  Dept. of Computer Science and Engineering,  Narasaraopeta Engineering College  Narasaraopet, Andhra Pradesh,  India  ramakrishnaephd7@gmail.com | Battu Venkata Siva  Student  Dept. of Computer Science and Engineering  Narasaraopeta Engineering College  Narasaraopet, Andhra Pradesh,  India  [venkatasivabattu2002@gmail.com](mailto:venkatasivabattu2002@gmail.com) | Shaik Aariz Ahmed  Student  Dept. of Computer Science and Engineering  Narasaraopeta Engineering College  Narasaraopet, Andhra Pradesh, India  [aarizahmed2002@gmail.com](mailto:aarizahmed2002@gmail.com) |

*ABSTRACT-Loans are a crucial part of the modern world, and banks receive a significant portion of their profits from them. However, deciding whether to grant a loan to an applicant is a complex process that requires banks to consider many factors.*

*In this study, we suggest a machine learning-based method to streamline the loan acceptance prediction process. To determine whether or not a loan applicant's profile is relevant for approval, we employ effective machine learning algorithms. We base our predictions on important features. Additionally, we present a comparison study of various categorization methods to demonstrate how machine learning algorithms might enhance the loan approval procedure. Our results show that machine learning algorithms can significantly reduce the risk of loan defaults and improve the loan approval process. Moreover, to enhance prediction accuracy, we incorporate a voting ensemble technique into our methodology. This additional layer of analysis further refines our predictions, contributing to more reliable loan approval decisions.*

*KEYWORDS--Loan approval prediction, Loan default risk, Predictive modeling, Feature selection, SMOTE, Decision Tree,, Gradient Boost, Random Forest, Extra Tree, Comparative analysis, Voting Ensemble.*

# INTRODUCTION

One of the main functions of the banking sector is lending, which is crucial to maintaining the financial stability of a country. Banks receive a large amount of their interest revenuefrom loans, thus the loan approval procedure is quite important. Individual bank managers evaluate the risk of loan default and application eligibility, which might cause systemic disruptions that could affect the economy as a whole as well as possible financial losses for banks.

In this research, we use data-driven strategies to streamline the loan approval process. We start with the removal of null values [4], duplicates, and outliers from the data [2], cleaning it up, and examining any relationships between the variables. The purpose of feature selection is to find the most pertinent qualities. To address the disparity in class sizes, we employ the SMOTE [22]. We then use machine learning methods to forecast the results of loan approval, such as gradient boost, decision tree, random forest, and other techniques. Each algorithm's performance is assessed using various metrics. We also examine how each algorithm contributes to enhancing the loan approval procedure. Since the primary objective of this study is to employ data-driven methodologies to enhance the precision and effectiveness of the loan approval procedure, We do this with the implementation voting ensemble method.

# 2. LITERATURE SURVEY

Manjeet Kumar et al. [3] evaluated a number of classifiers, such as Light Gradient Boosting Machine (LGBM), Extra Trees, Random Forest, and Extreme Gradient Boosting (XGB) for the purpose of predicting bank loan default. Their study provides insightful information for financial institutions by highlighting the significance of debt income and work history in forecasting defaults

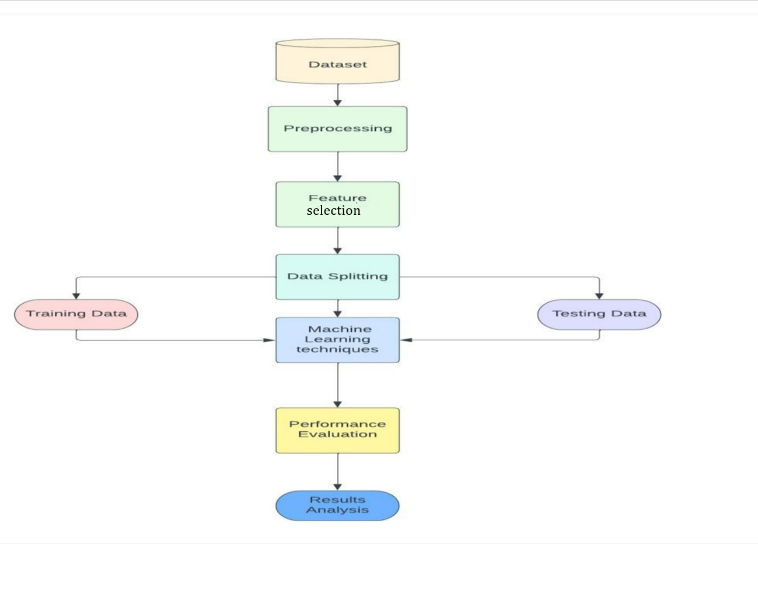
In the research, Mehul Madaan et al. [4] obtained 73% and 80% accuracy, respectively, in loan default prediction using Decision Trees and Random Forest algorithms. For financial organisations looking to enhance loan approval procedures and reduce credit risks, their study provides insightful information. Through the analysis of these algorithms' performance on a shared dataset, the research adds to the continuing investigation of machine learning applications in the banking industry.

Decision Trees and other machine learning models were used by Supriya et al. [2] to predict loan defaulters with an accuracy of 81.1%. Their research focused on data preparation methods, such as managing outliers and missing information, and resulted in a thorough examination of the characteristics that affect loan acceptance. The writers provided insightful information for improving credit risk assessment in the banking industry by highlighting the importance of variables like income level and credit history in loan sanctioning decisions.

Mahankali et al. [1] forecast loan approvals with an accuracy rate of 80.945% by using logistic regression. Their all-inclusive strategy comprises testing, model creation, and data pretreatment, offering a solid foundation for automated loan approval systems. This study provides useful information about the use of machine learning algorithms in banking settings and sets the standard for further research in the area. When comparing machine learning algorithms for forecasting bank loan risks, Alsaleem et al. [5] discovered that Multilayer Perceptron has the best accuracy (80%).

# 3. THE PROPOSED SYSTEM FOR LOAN APPROVAL PREDICTION

Our proposed model for loan approval prediction is structured around distinct phases and steps as shown in Fig.1 , each tailored to maximize accuracy, efficiency, interpretability, robustness, and fairness. The criteria as follows:

* Analysis of Dataset
* Visualization of Data
* Preprocessing Techniques
* Model Development

# Fig.1 Flow Chart

# *A. Analysis of Dataset*

Fig.2: Data Set

# We used the well-known Kaggle website, which hosts a variety of datasets, to find the datasets we needed for our predictive study. The dataset that we used is accessible [6] For our research to be successful, having access to these datasets shown in Fig.2 is essential for performing in-depth analysis and forecasts. The dataset consists of 13 columns and 614 entries.

# *B. Visualization of Data*

# As the target class to be predicted, the "Loan\_Status" column shows a class imbalance in the Fig.3, with roughly 68.7% of the entries labelled as "Y" (meaning loan approval) and the remaining 31.3% labelled as "N" (showing loan denial). The performance of several models, especially those that are sensitive to class distribution, may be impacted by this imbalance.

Fig3. Loan\_Status

# *C.* *Preprocessing Techniques*

The next stage after data visualisation is preprocessing, which gets the data ready for model training. The following are the steps.

## *Dealing Null Values:*

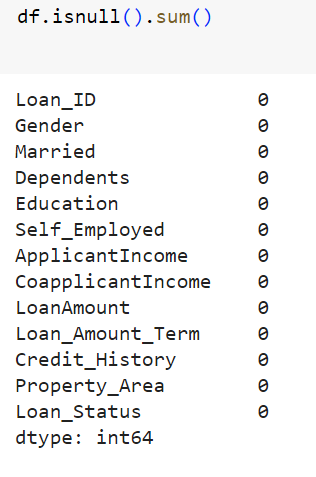
Reliability of the model and data quality depend on addressing null values [4]. By managing null values using methods like imputation or elimination, machine learning models become more accurate and the dataset's integrity is preserved. Potential biases are reduced by efficient null value management, allowing for more precise forecasts and perceptive analysis. The fig.4 illustrates how frequently data

Fig.4 Representing frequency of Null Values

are missing in various columns both before and after removal. It exposes deficiencies in important characteristics such as Gender, Dependents, Self\_Employed, and Credit\_History, for example.

* *One-hot Encoding:*

To ensure compatibility with numerical computations, one-hot encoding is essential for transforming categorical variables into a format that is appropriate for machine learning methods. Algorithms can successfully read and learn from categorical features by expressing categorical data as binary vectors. By doing this, the feature space is increased, maintaining the distinctive qualities of every category and improving.

## 

## Fig.5 Columns after One-hot encoding

## *Eliminating Data Outliers*

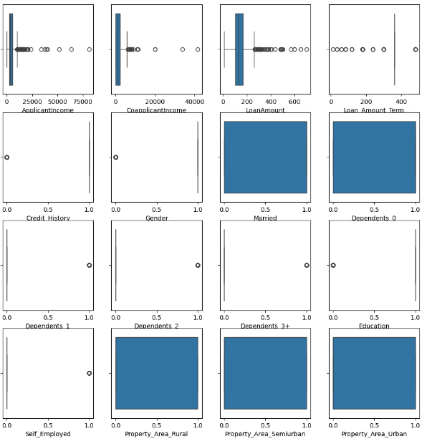
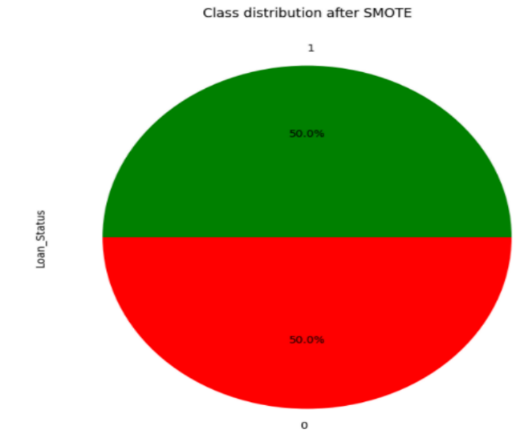


Fig 6: Visualizing data outliers

Handling outliers is essential to preserving the accuracy and integrity of machine learning models and statistical analysis [2]. Results can be severely skewed by outliers, which can also influence interpretations and provide false findings. Fig.6 Illustrating the presence of Outliers in the data. Since, the dataset had outliers that could have distorted the results of our investigation and the performance of the model. We utilised the quantile approach to solve this problem, identifying and eliminating extreme values by establishing thresholds based on quantiles.

* *Square Root Transformation:*

## Because it lessens the impact of extreme values, square root transformation is essential for stabilising variance, particularly in datasets with skewed distributions. It works by improving the symmetry and conformance of the data to normalcy assumptions, which can guarantee reliable inference and enhance the performance of specific statistical models.

## *Feature Selection:*

Feature selection is an essential process that helps machine learning models operate at their best by locating the most significant predictors for the target variable. We employ Recursive Feature Elimination (RFE) in our method, which is an iterative process of selecting features according to their importance ratings. The impact of each feature to the loan approval prediction job is carefully evaluated by integrating RFE with a logistic regression estimator. We are able to determine a subset of features shown in Fig 8, that minimises model complexity and maximises predicted accuracy through this iterative method.



Fig 7:selected features after RFE

## *SMOTE:*

When one class predominates over another in a situation like predicting loan acceptance, SMOTE (Synthetic Minority Over-sampling Technique) is an essential technique for resolving class imbalance in datasets[22]. As in Fig 9, the class imbalance bias is successfully reduced by using SMOTE, resulting in a more representative and trustworthy dataset that we can use to train our machine learning models. By using this method, the model's capacity to generalise across both groups is improved, which eventually leads to forecasts of loan acceptance that are more fair and accurate.

## Fig 8:Loan\_Status after SMOTE

# *D. Model Development*

To ensure their resilience and usefulness in a range of situations, machine learning models are constructed and evaluated using training datasets. The research employs the following methods: Random Forest, Decision Tree, Gradient Boosting, K-Nearest Neighbours (KNN), Extra Trees, and Logistic Regression. We use the implementation voting ensemble approach to improve the accuracy and efficiency of the loan approval process.

* Voting Ensemble Method:

A machine learning method called voting ensemble combines several models to provide predictions. A majority vote or an average is used to decide the final forecast, which is based on the weighted predictions of each model. When compared to separate models, it frequently results in increased accuracy and resilience.

y^​=argmaxj​i=1∑n​1(y^​i​=j) (1)

# 4. RESULTS & DISCUSSION

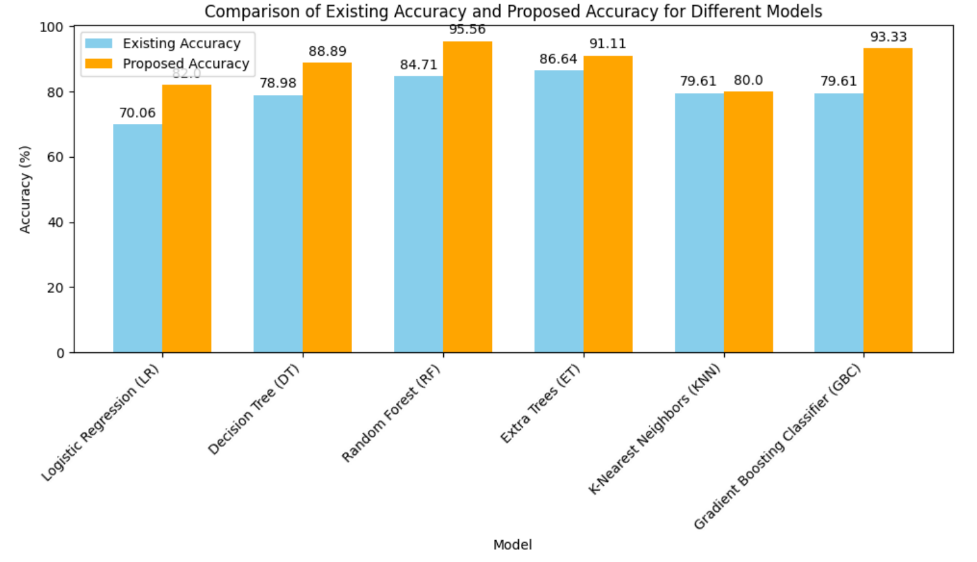
We found that the models that were used are Random Forest, Extra Trees, Decision Trees, Gradient Boosting, Logistic Regression, and K-Nearest Neighbors performed differently. Every algorithm performed differently, as shown in Fig 9 . In order to highlight, With an accuracy of 95.56%, Random Forest stood out among the group of models employed to forecast loan acceptance. It demonstrated resilience during both the training (98.325%) and testing (95.556%) stages of the process. Closely behind, Extra Trees Classifier showed great predictive strength with an accuracy of 91.11%, and Gradient Boosting, though doing effective, showed its effectiveness with an accuracy of 93.33%. On the other hand, the accuracies of the Decision Tree, K-Nearest Neighbours, and Logistic Regression models were much lower at 88.89%, 80.00%, and 82.22%, respectively. Also illustrated the comparitive representation of accurcies of models in the existing research [22] and current research.

Fig.9 Comparing Accuracies

* *Voting Ensemble Method:*

In order to increase the overall accuracy, we will integrate the predictions of several machine learning models in our next stage, which is the implementation of a voting ensemble approach.

After careful examination, Random Forest was clearly the best performer, with an astounding accuracy of 95.56%. In training and testing, this model demonstrated exceptional robustness, with accuracy values of 98.325% and 95.556%, respectively. With an accuracy of 91.11%, the Extra Trees Classifier demonstrated noteworthy predictive capability, trailing closely behind. Likewise, Gradient Boosting demonstrated its efficacy with a 93.33% accuracy rate.

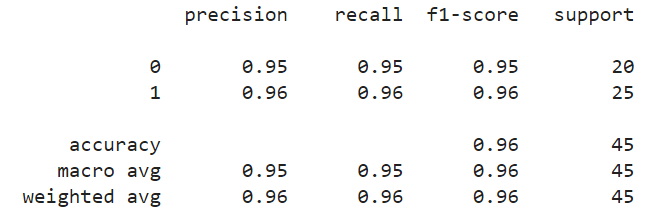
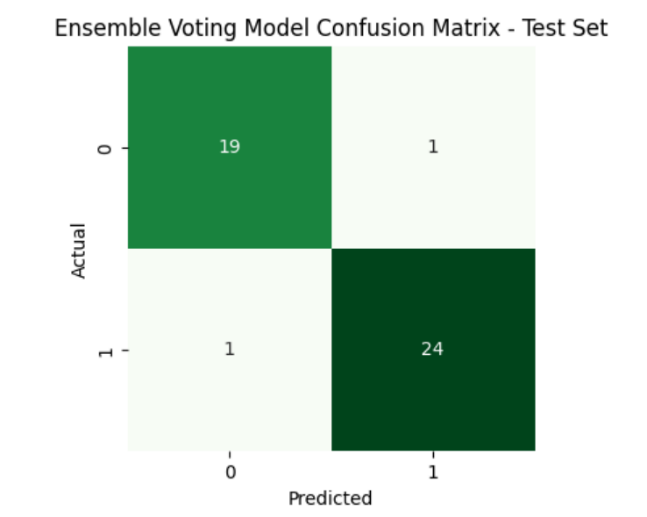
By combining Random Forest, Extra Trees Classifier, and Gradient Boosting, we will use ensemble learning to take advantage of the advantages of these high-performing models. In order to improve the accuracy of loan approval forecasts, this strategy seeks to use the combined predictive potential of models. Using the best-performing models and the ensemble approach, we were able to obtain an astounding accuracy of 95.55%. This suggests a notable improvement in the accuracy of loan approval forecasts.

Fig.10 (a) Metrics for Voting Ensemble Model

The confusion matrix's highlighted values suggest that the model is performing satisfactorily in classifying loan applications. The accuracy acquired with 24 true positives and 19 genuine negatives. With a precision of 95% for loans that have been granted, there is a high percentage of accurate forecasts among loans with this approval status. Furthermore, for loans that are authorised, recall is perfect and providing there are very few false negatives and false positives.

  
Fig.10 (b) Confusion Matrix of Voting Ensemble Model

# 5. CONCLUSION & FUTURE SCOPE

In summary, our initiative has shed light on the significant contribution that machine learning (ML) approaches have made to transforming the process of predicting loan approval. We have established a solid basis for our predictive models by carefully examining our dataset and utilising sophisticated preprocessing methods, such as managing null values, outliers, square root transformation, correlation analysis, feauture selection and SMOTE.

Going forward, our model achieved an amazing accuracy of 95.56%—serve as a testament to the ability of ML to provide incredibly precise predictions. This represents a notable improvement above the Existing accuracy of 87.26% from the previous base study and highlights the enormous progress made possible by our rigorous methodology.

Table1: Comparing Final Accuracies

|  |  |
| --- | --- |
|  | Accuracy(%) |
| Existing Accuracy | 87.26 |
| Proposed Accuracy | 95.56 |

Our research opens up a world of opportunities for improving loan approval prediction models in the future. Subsequent investigations may explore further into the field of deep learning, utilising neural networks capacity to address complex patterns in loan data. In addition, to guarantee flexibility in response to changing patterns and dynamics in loan application and repayment behaviour, ongoing model monitoring systems will be necessary. We are steadfast in our research to advance ML innovation and bring about revolutionary change in the field of financial decision-making processes as we move forward.

6. REFERENCES

1. Gopinath, Mahankali, K. Srinivas Shankar Maheep, and R. Sethuraman. 2021. “Customer Loan Approval Prediction Using Logistic Regression.” Advances in Parallel Computing. https://doi.org/10.3233/apc210103.
2. Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma,Namburi Vimala Kumari, k Vikash,“Loan Prediction by using Machine Learning Models”, International Journal of Engineering and Techniques.Volume 5 Issue 2, Mar-Apr 2019.
3. A. Uzair, T. Aziz, H. Ilyas, S. Asim, B. N. Kadhar, "An Empirical Study on Loan Default Prediction Models" Journal of Computational and Theoretical Nanoscience, Volume 16, Number 8, August 2019, pp. 3483-3488(6). DOI: https://doi.org/10.1166/jctn.2019.8312
4. M. Madaan et al. "Loan default prediction using decision trees and random forest: A comparative study" IOP Conf. Ser.: Mater. Sci. Eng. 2014. doi: 10.1088/1757-899X/1022/1/012042.
5. Alsaleem, M. Y., & Hasoon, S. O. (2020). Predicting bank loan risks using machine learning algorithms. AL-Rafidain J. Comput. Sci. Math., 14(1), 149–158.
6. Ramachandra, H. V., G. Balaraju, R. Divyashree, and Harish Patil. 2021. “Design and Simulation of Loan Approval Prediction Model Using AWS Platform.” 2021 International Conference on Emerging Smart Computing and Informatics (ESCI). https://doi.org/10.1109/esci50559.20219397049.

Data set used <https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset/.%20>.

1. Singh, Vishal, Ayushman Yadav, Rajat Awasthi, and Guide N. Partheeban. 2021. “Prediction ofModernized Loan Approval System Based on MachineLearning Approach.” 2021 International Conference on IntelligentTechnologies(CONIT).https://doi.org/10.1109/conit51480.2021.9498475.
2. S.m., Karthikeyan, S. M. Karthikeyan, and Pushpa Ravikumar. 2021. “A Comparative Analysis of Feature Selection for Loan Prediction Model.” InternationalJournalofComputerApplications.<https://doi.org/10.5120/ijca2021920992>.
3. Hassan, Amira Kamil Ibrahim and Ajith Abraham. "Modeling consumer loan default prediction using ensemble neural networks" , 2013 International Conference On Computing, Electrical And Electronic Engineering (ICCEEE). IEEE, 2013.
4. .Nitesh Pandey et al. (2022). "Loan Approval Prediction using Machine Learning Algorithms Approach.", IRJMETS, this paper achieved an accuracy of 78.3% using Logistic Regression, Decision Trees, and KNN. It discusses limitations in feature engineering and imbalanced datasets.
5. J. Tejaswini (2022). "Accurate Loan Approval Prediction Based on Machine Learning Approach.", IRJMETS, this paper achieved an accuracy of 88.1% using XGBoost. It emphasizes the importance of careful data pre-processing and feature engineering for optimal performance.
6. Anant Shinde et al. (2022). "Loan Prediction System Using Machine Learning." Published in the field of Finance by IEEE, this paper compares Logistic Regression, Decision Tree, and Random Forest models. It concludes that Decision Tree and Random Forest outperform Logistic Regression due to their non-linearity handling capabilities.
7. Dharavath Sai Kiran. Avula et al. (2023). "Loan Approval Prediction using Adversarial Training and Data Science." Within the domain of Finance, this paper discusses a model that produced low accuracy values and could only handle minimum-sized data.
8. Zhu L, Qiu D, Ergu D, Ying C, Liu K (2019) A study on predicting loan default based on the random forest algorithm. Procedia Comput Sci 162:503–513.
9. Nigmonov A, Shams S, Alam K (2022) Research Macroeconomic determinants of loan defaults: evidence from the U.S. peer-to-peer lending market. Res Int Bus Finan 59:101516.
10. Lee JW, Lee WK, Sohn SY (2021) Graph convolutional network-based credit default prediction utilizing three types of virtual distances among borrowers. Expert Syst Appl 168:114411.
11. Lim S-J, Thiel C, Sehm B, Deserno L, Lepsien J, Obleser J (2022) Distributed networks for auditory memory differentially contribute to recall precision. NeuroImage 256:119227.
12. Fontem B, Smith J (2019) Analysis of a chance-constrained new product risk model with multiple customer classes. Eur J Oper Res 272(3):999–1016 23. Bianco S, Mazzini D, Napoletano P, Schettin R (2019) Multitask painting categorization by deep multibranch neural network. Expert Syst Appl 135:90–101.
13. Wang L, Chen Y, Jiang H, Yao J (2020) Imbalanced credit risk evaluation based on multiple sampling, multiple kernel fuzzy self-organizing map and local accuracy ensemble. Appl Soft Comput 91:106262
14. Vuttipittayamongkol P, Elyan E, Petrovski A (2021) On the class overlap problem in imbalanced data classification. Knowl-Based Syst 212:106631 26. Papouskova M, Hajek P (2019) Two stage consumer credit risk modelling using heterogeneous ensemble learning. Decis Support Syst 118:33–45.
15. Ashofteh A, Bravo JM (2021) A conservative approach for online credit scoring. Expert Syst Appl 176:114835.

[22] N. Uddin, M.K. Uddin Ahamed, M.A. Uddin, Md. Manwarul Islam, Md. Alamin Talukder, and Sunil Aryal, "An ensemble machine learning based bank loan approval predictions system with a smart application," International Journal of Cognitive Computing in Engineering, vol. 4, pp. 327-339, 2023.http://doi.org/10.1016/j.ijcce.2023.09.001.