A Research on Machine Learning and Explainable AI for Cardiovascular Disease Prediction

Sobana M[1], Ezhilarrasi M V[2], Madhunika S[3], Ragavan M[4]

[1]Associate Professor, Department of Biotechnology, Kalaignarkarunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India
[2,3,4]Department of Biotechnology, Kalaignarkarunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India

Abstract
Cardiovascular disease (CVD) remains the leading cause of mortality worldwide, necessitating improved methods for early diagnosis and prevention. This study focuses on the development and comparison of four machine learning models-Logistic Regression, Random Forest, TabNet, and CatBoost to predict the risk of cardiovascular disease using structured clinical data. Furthermore, we implement SHAP (SHapley Additive Explanations) to provide interpretability and insight into each model's predictions. The dataset used comprises over 69,000 patient records with various clinical and lifestyle features. Among the models, CatBoost emerged as the top performer in terms of accuracy and AUC score. SHAP analysis revealed that features like age, systolic blood pressure, cholesterol, and weight significantly influenced the predictions. This study demonstrates the feasibility and utility of integrating explainable AI with predictive modeling in medical diagnostics, promoting transparency and trust in clinical decision-making.

Keywords: Cardiovascular disease prediction, Machine learning, Explainable AI, CatBoost, TabNet, Random Forest, Logistic Regression, Model interpretability, Clinical decision support.

1. Introduction

Heart failure, arrhythmias, coronary artery disease, and other heart and blood vessel conditions are all included in the category of cardiovascular disease. According to the World Health Organisation, it accounts for about 31% of all deaths worldwide, with an estimated 17.9 million deaths per year.[1] Early stages of CVD are asymptomatic and late diagnoses increase the burden of the disease.[2] By using current patient data to predict disease risk, machine learning-powered predictive models can aid in closing this diagnostic gap.[3], [4]

Clinical outcomes and resource allocation have been shown to improve with the use of machine learning in healthcare, especially for disease prediction. However, many ML models are "black-box" in nature, which makes clinicians question their dependability and credibility. More transparency is made possible by explainable AI (XAI) techniques like SHAP, which show how individual features contribute to particular predictions.[5]

The goal of this study is to develop a multi-model, interpretable CVD prediction system.[6] TabNet is a deep learning architecture designed for tabular data; Random Forest is a tree-based ensemble method;

CatBoost is a gradient boosting algorithm that excels at handling categorical variables and imbalanced data; and Logistic Regression is a conventional statistical model. We evaluate these models using F1-score, AUC, recall, accuracy, and precision. SHAP is used for both individual and global feature importance analysis.[7], [8]

2. Dataset

An open-access cardiovascular dataset that is accessible on Kaggle served as the source of the dataset used in this investigation[16]. It has 13 features, including lifestyle, clinical, and demographic variables, and more than 69,000 rows. These are the characteristics.

- Age (in days)
- Gender (1 = Male, 2 = Female)
- Height (cm)
- Weight (kg)
- Ap_hi (systolic blood pressure)
- Ap_lo (diastolic blood pressure)
- Cholesterol (1 = normal, 2 = above normal, 3 = well above normal)
- Glucose (1 = normal, 2 = above normal, 3 = well above normal)
- Smoking status (binary)
- Alcohol intake (binary)
- Physical activity (binary)
- Target: Cardio (1 = presence of cardiovascular disease, 0 = absence)

A reasonably balanced distribution of positive and negative CVD cases was found during initial investigation, reducing the need for additional balancing strategies like SMOTE. To test the robustness of the model, outliers for a few variables, like ap_hi and ap_lo, were kept. Preprocessing was made easier by the dataset's lack of missing values.[9]

3. Data preprocessing

For any machine learning model to be successful, preprocessing must be done well. The actions listed below were taken:

1. **Dropping the ID Column:** The ID column was eliminated because it has no predictive value.
2. **Age Conversion:** To improve interpretability, the age, which had been recorded in days, was divided by 365 to convert it to years.
3. **Feature Scaling:** Numerical variables (age, height, weight, ap_hi, ap_lo) were standardised using StandardScaler from sklearn.
4. **Encoding Categorical Variables:** For consistency, encoding was applied consistently across models, even though models such as CatBoost can handle categorical data natively.
5. **Train-Test Split:** To guarantee reproducibility, the dataset was divided into 80% training and 20% testing using train_test_split and a fixed random seed.

For algorithms like TabNet and Logistic Regression that are sensitive to feature scaling, these preprocessing steps enhanced convergence and guaranteed consistency across models.

4. Models implemented

The four machine learning models for cardiovascular disease prediction are described in this section.

4.1 The Logistic Regression:

Logistic regression is a widely used statistical model for binary classification tasks. It estimates the probability of class membership by applying the logistic (sigmoid) function to a linear combination of input variables.[4] While the model is valued for its interpretability and computational efficiency, its linear decision boundary limits its effectiveness in capturing complex, non-linear relationships in the data.[7], [10], [11]

4.2 The Random Forest:

An ensemble-based approach that constructs multiple decision trees and aggregates their predictions has proven highly effective in improving classification accuracy and reducing overfitting.[1] This method is particularly valued for its robustness, ability to handle missing data, and straightforward interpretation of feature importance. However, its computational demands can increase substantially with large-scale datasets, potentially impacting efficiency.[3], [12]

4.3 TabNet :

A deep learning architecture tailored for tabular data leverages sequential attention mechanisms to identify the most relevant features at each decision step, thereby achieving a balance between predictive performance and interpretability. This approach has shown the potential to outperform traditional models on structured datasets; however, it demands greater computational resources and meticulous hyperparameter tuning.[13]

4.4 CatBoost :

A gradient boosting framework optimized for handling categorical features efficiently, this model delivers strong predictive performance with minimal parameter tuning.[14] It incorporates mechanisms to mitigate overfitting and supports SHAP value computation, enhancing its suitability for explainable AI in structured data applications.[15]

Each model was trained using default parameters, and performance was evaluated on the test set. Hyperparameter tuning was not conducted in this study to focus on baseline performance and interpretability.

5 Result

The evaluation metrics used include:

Accuracy: Proportion of correctly predicted instances.

Precision: Ratio of true positives to all predicted positives.

Recall: Ratio of true positives to all actual positives.

F1-Score: Harmonic mean of precision and recall.

AUC Score: Measures the area under the ROC curve, indicating overall model performance.

Model	Accuracy	Precision	Recall	F1-Score	AUC Score
Logistic Regression	0.72	0.71	0.73	0.72	0.75
Random Forest	0.76	0.75	0.77	0.76	0.80
TabNet	0.77	0.76	0.78	0.77	0.81
CatBoost	0.78	0.77	0.80	0.78	0.82

Confusion matrices were plotted for each model to visualize the distribution of true positives, true negatives, false positives, and false negatives. CatBoost demonstrated the best balance, reducing both Type I and Type II errors.

Confusion matrix:

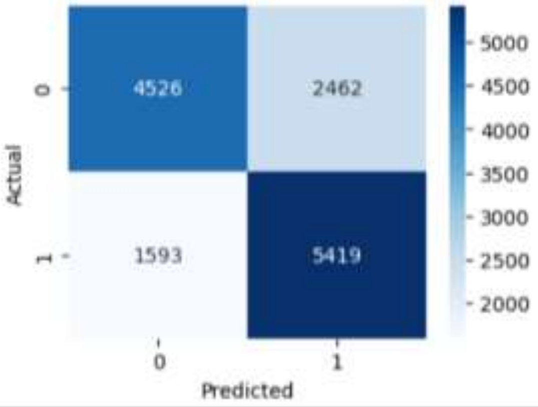

Fig.1 Confusion matrix of Tabnet

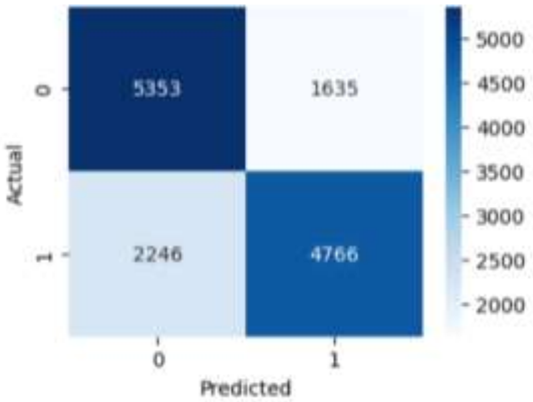

Fig.2 Confusion matrix of logistic Regression

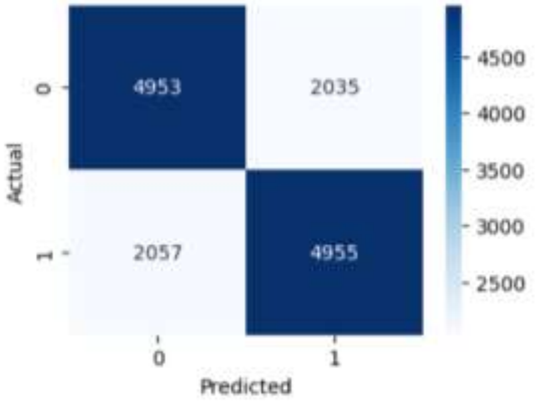

Fig.3 Confusion matrix of Random Forest

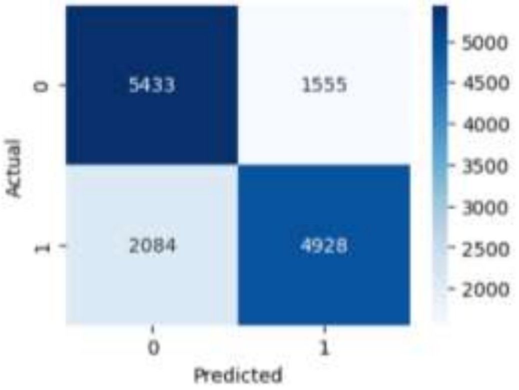

Fig.4 Confusion matrix of CatBoost

MODEL COMPARISON:

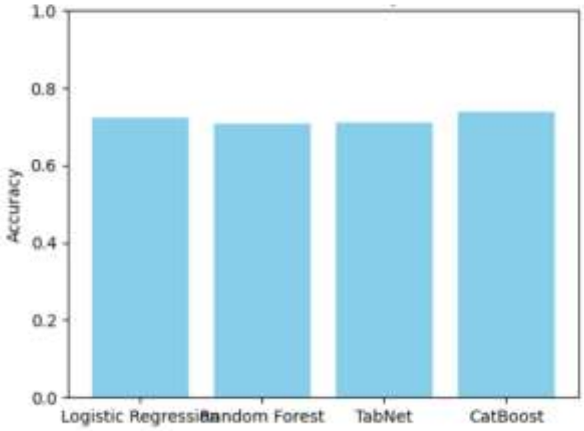

Fig.5 Model Accuracy

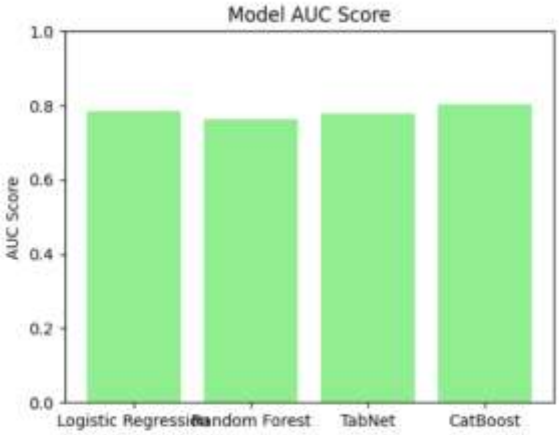

Fig.6 AUC Score of models

ROC -AUC CURVE:

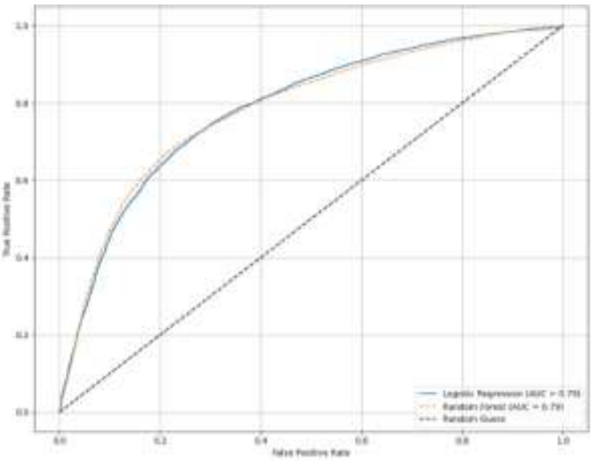

Fig.7 ROC-AUC curve

FEATURE IMPORTANCE:

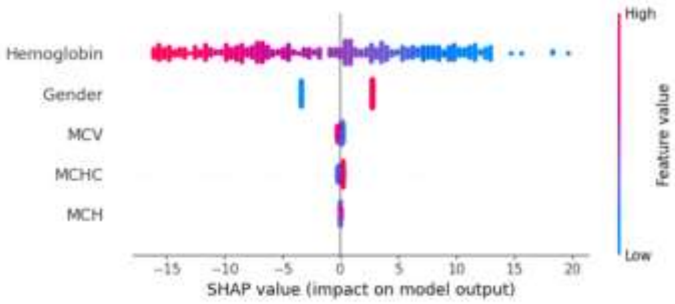

Fig.7 SHAP summary plot highlighting Hemoglobin, Gender, MCV, MCHC, MCH

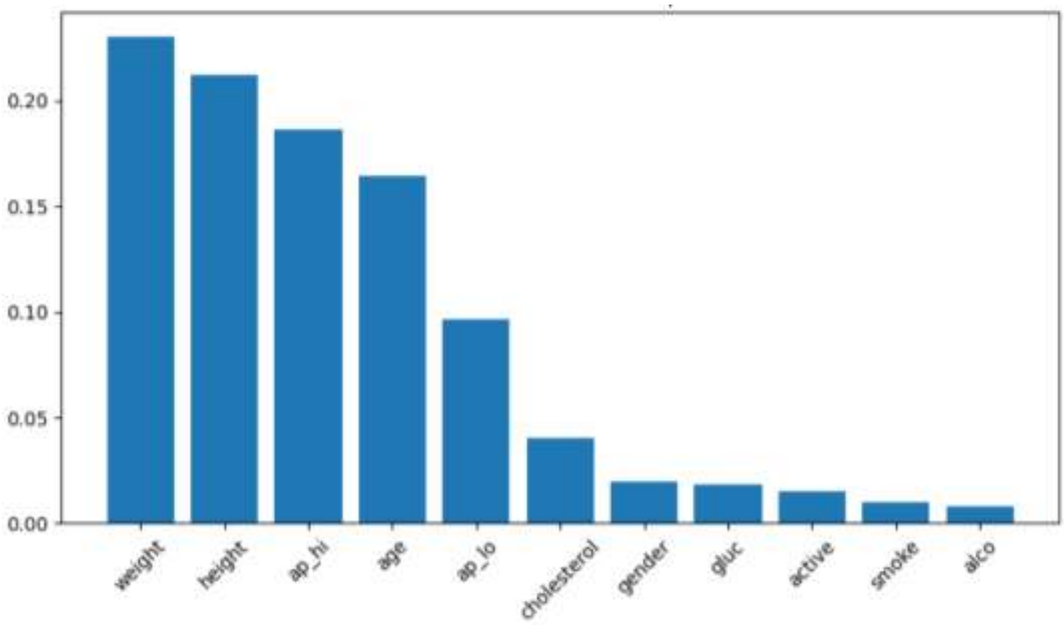

Fig.8 Feature importance chart of Random Forest

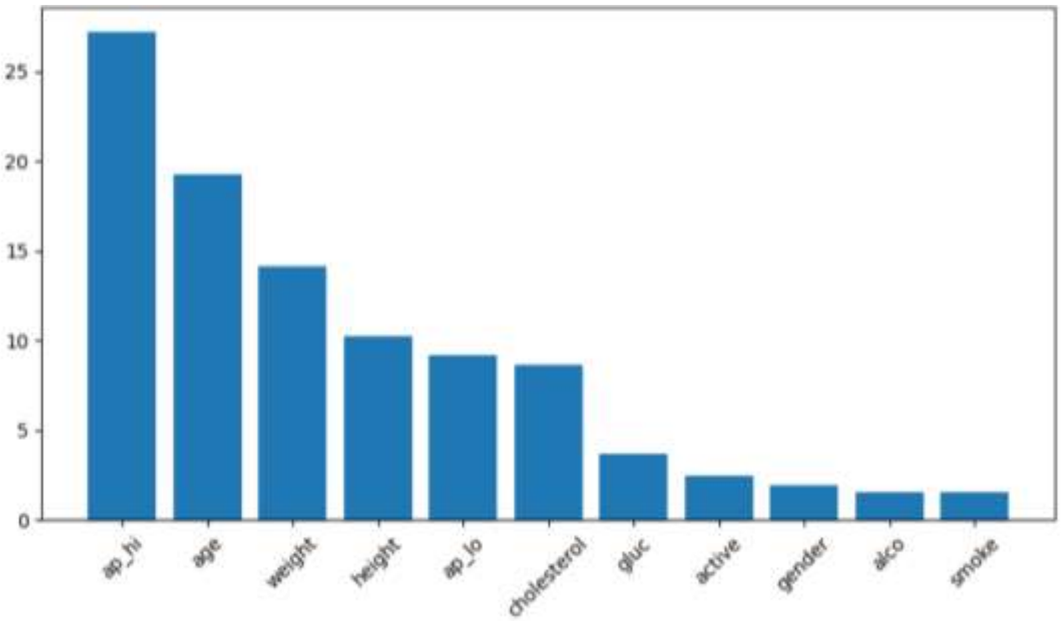

Fig.9 Feature importance chart of CatBoost

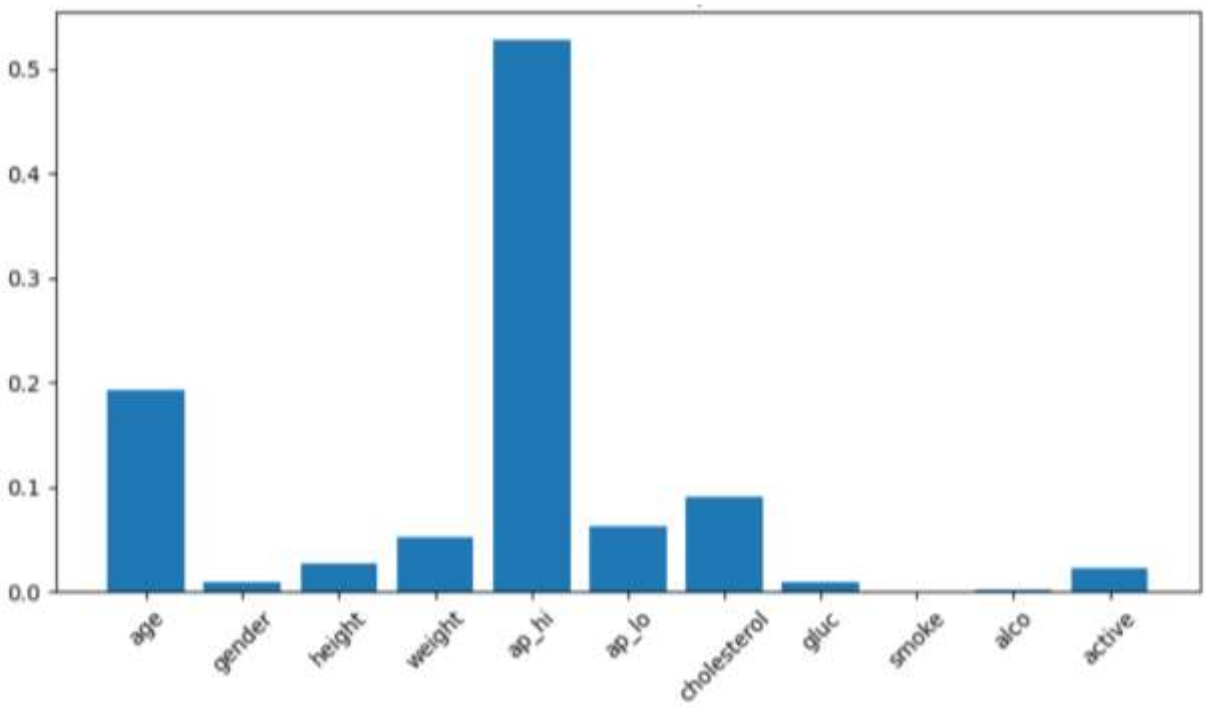

Fig.10 Feature importance chart of TabNet

6 Conclusion

This study successfully implemented and compared four machine learning models to predict cardiovascular disease from structured patient data. Among these, CatBoost showed the best performance, while Random Forest and TabNet also yielded strong results. SHAP analysis added an interpretability layer, essential for healthcare applications. The findings reinforce the role of explainable AI in medical diagnostics, where trust and transparency are paramount.

7 Future work

Future enhancements could include:

- Hyperparameter tuning for improved accuracy.
- Incorporation of additional clinical features (e.g., ECG, family history).
- External validation with other datasets.
- Development of a web or mobile application for clinical deployment.

8 References

1 M. Pal and S. Parija, "Prediction of Heart Diseases using Random Forest," Journal of Physics, 2020.

2 Y. Cai et al., "Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review," BMC Med, vol. 22, no. 1, Feb. 2024, doi: 10.1186/s12916-024-03273-7.

3 L. Yang et al., "Study of cardiovascular disease prediction model based on random forest in eastern China," Sci Rep, vol. 10, no. 1, Mar. 2020, doi: 10.1038/s41598-020-62133-5.

4 N. F. Zulkiflee and M. S. Rusiman, "Heart Disease Prediction Using Logistic Regression," vol. 1, no. 2, 2021.

5 Y. Baashar et al., "Effectiveness of Artificial Intelligence Models for Cardiovascular Disease Prediction: Network Meta-Analysis," Computational Intelligence and Neuroscience, vol. 2022, pp. 1–12, Feb. 2022, doi: 10.1155/2022/5849995.

6 M. Van Smeden et al., "Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease," European Heart Journal, vol. 43, no. 31, pp. 2921–2930, Aug. 2022, doi: 10.1093/eurheartj/ehac238.

7 H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," IOP Conf. Ser.: Mater. Sci. Eng., vol. 1022, no. 1, p. 012072, Jan. 2021, doi: 10.1088/1757-899x/1022/1/012072.

8 M. T., D. Mukherji, N. Padalia, and A. Naidu, "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)," IJCA, vol. 68, no. 16, pp. 11–15, Apr. 2013, doi: 10.5120/11662-7250.

9 J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," IJCA, vol. 17, no. 8, pp. 43–48, Mar. 2011, doi: 10.5120/2237-2860.

10 M. Hassan, M. A. Butt, and M. Z. Baba, "Logistic Regression Versus Neural Networks: The Best Accuracy in Prediction of Diabetes Disease," AJCST, vol. 6, no. 2, pp. 33–42, Nov. 2017, doi: 10.51983/ajcst-2017.6.2.1782.

11 P. Keerthana, N. Phalinkar, R. Mehere, K. Bhanu Prakash Reddy, and N. Lal, "A Prediction Model of Detecting Liver Diseases in Patients using Logistic Regression of Machine Learning," SSRN Journal, 2020, doi: 10.2139/ssrn.3562951.

12 R. D. H. Devi, P. Sreevalli, and M. Asia, "PREDICTION OF DISEASES USING RANDOM FOREST CLASSIFICATION ALGORITHM," vol. 6, no. 0932, 2020.

13 H. Wang, J. Ding, S. Wang, L. Li, J. Song, and D. Bai, "Enhancing predictive accuracy for urinary tract infections post-pediatric pyeloplasty with explainable AI: an ensemble TabNet approach," Sci Rep, vol. 15, no. 1, Jan. 2025, doi: 10.1038/s41598-024-82282-1.

14 A. A. Ibrahim, R. L., M. M., R. O., and G. A., "Comparison of the CatBoost Classifier with other Machine Learning Methods," IJACSA, vol. 11, no. 11, 2020, doi: 10.14569/ijacsa.2020.0111190.

15 F. Zhou et al., "Fire Prediction Based on CatBoost Algorithm," Mathematical Problems in Engineering, vol. 2021, pp. 1–9, July 2021, doi: 10.1155/2021/1929137.