Journal of Public Health & Environment (2024)

Journal of Public Health & Environment

journal homepage: www.journal-phe.online

Research article

Arrhythmia recognition based on integrated learning of multi-model fusion Stacking

Xiangchen Ju[1#], Kai Yang[2#], Dongxiong Liu[3#], Heyu Gao[4#], Huiquan Li[5#], Zetong Zhang[5#]

1. School of Mathematics and Science, TaiZhou University, TaiZhou, China
2. School of Electronic & Information Engineering, Nanjing University of Information Science & Technology, Nanjing, China
3. School of Materials Science and Engineering, Northeastern University, Shenyang, Liaoning Province, China
4. Artificial Intelligence College, Shenyang Aerospace University, Liaoning Province, Shenyang, China
5. Business of College, Linyi University, Linyi, Shandong Province, China

Statement: [#]These authors contributed equally to this work and should be considered co-first authors

ABSTRACT

Current monitoring systems employing single model algorithms face challenges in accurately recognizing and alerting real-time arrhythmic events in patients with severe cardiac conditions. To address this issue and improve the detection accuracy of electrocardiogram (ECG) monitoring systems, this paper introduces a novel arrhythmia recognition model based on Stacking ensemble learning. This model integrates multiple base learners, including LightGBM, XGBoost, Random Forest, SVM, and Logistic Regression, and optimizes them using Gradient Boosting as the meta-learner. Hyperparameters were fine-tuned through grid search, and nested cross-validation was employed to train the model, ensuring robust predictive performance. The detection results indicate that the Stacking ensemble model significantly outperforms single models in both accuracy and stability, offering substantial practical application value for clinical ECG monitoring and diagnosis.

Key words: Arrhythmia，Multi-model fusion，Stacking ensemble learning，Grid search，Nested cross-validation

1. Introduction

The electrophysiological activity of the heart underpins cardiac contraction through signal transmission in myocardial fibers, detectable via surface electrocardiograms (ECGs), essential for cardiac disease diagnosis. This paper suggests deploying a real-time alert classification algorithm on ECGs to promptly identify arrhythmias, enhancing monitoring. Advances in AI, notably machine learning and deep learning, have bolstered cardiac monitoring. Employing LightGBM, XGBoost, and Random Forest within a Stacking ensemble framework, this study constructs a multi-model fusion for arrhythmia detection, validated through cross-validation and performance evaluation.

2. Model construction and experimental data

2.1. Construction of fusion model based on Stacking

This paper utilizes a Stacking strategy for multi-model fusion classification of arrhythmia detection data, beginning with data preprocessing, feature engineering, and rule division. Base learners include LightGBM, XGBoost, Random Forest, SVM, and Logistic Regression, optimized through grid search to fine-tune hyperparameters. Nested cross-validation trains the base learners, enhancing accuracy and generalizability. The base learners' predictions feed into a meta-classifier trained with Gradient Boosting to further increase prediction accuracy. Comprehensive model performance assessment using accuracy, recall, precision, F1 scores, confusion matrices, and learning curves confirms that the Stacking approach improves model generalizability and predictive accuracy[1].

2.2. Experimental data

This paper employs the "MIT-BIH Malignant Ventricular Ectopy Database" to construct a database of short ECG records, each containing 2-second segments of ECG signals, for evaluating arrhythmia detection algorithms. The database provides data of various temporal lengths related to cardiac arrhythmias, facilitating the testing and comparison of different arrhythmia detection algorithms.

3. Data preprocessing

This paper extracts features from 2-second ECG segments and employs a

six-level risk severity as the target encoding for classification prediction using a multi-model fusion via Stacking ensemble learning.

3.1. Feature selection

The paper calculates descriptive statistics such as maximum, minimum, mean, variance, count, median, range, standard deviation, and coefficient of variation for each 2-second ECG data segment using Python. These statistics comprehensively extract the characteristics of the data segments, as shown in Table 1.

Table 1. Data of each descriptive statistic

Maximum	Minimum	Mean	…	Extreme difference	Standard deviation	Coefficient of variation
15604	8.46E-07	68.88754	…	15604	860.2865	12.48827
9910.7	8.42E-06	42.34318	…	9910.7	557.5368	13.1671
…	…	…	…	…	…	…
4985.3	3.91E-05	49.85327	…	4985.3	376.0898	7.543933
1109.3	8.98E-08	12.36728	…	1109.3	82.59823	6.678771
…	…	…	…	…	…	…
15329	2.12E-08	68.34061	…	15329	843.2414	12.3388
6006.7	1.48E-06	42.70623	…	6006.7	384.0467	8.992757
…	…	…	…	…	…	…
5178.6	6.04E-07	42.97359	…	5178.6	329.8395	7.675401
1647.7	7.94E-05	21.439	…	1647.7	153.988	7.182608
…	…	…	…	…	…	…
322.86	2.92E-07	2.733945	…	322.86	19.5441	7.148681
887.06	1.71E-06	16.58329	…	887.06	92.80475	5.596282
…	…	…	…	…	…	…
573.82	1.69E-06	23.83287	…	573.82	71.46161	2.998447
460.45	1.12E-07	23.97056	…	460.45	70.73555	2.950935

3.2. Determination of target variables

The paper defines six major categories of cardiac arrhythmias as the target variables. It uses Python to apply ordinal encoding, mapping each category to an integer that reflects a hierarchy among the categories[2]. The integer values represent the severity levels of arrhythmia categories, with encoding from 0 to 5, decreasing in order of severity, as illustrated in Table 2.

Table 2. Encoding of target variables

Coding	Class
0	Life-threatening arrhythmia requiring immediate medical attention
1	Life-threatening arrhythmia
2	Life-threatening ventricular arrhythmias

3	Potentially dangerous ventricular arrhythmias
4	Supraventricular arrhythmia
5	No significant risk or normal sinus rhythm

3.3. Handling of outliers and missing values

The paper conducts statistical analysis of the samples using Python, initially employing box plots for outlier detection and management. Figure 1 and Figure 2 displays box plots of certain features both before and after the removal of outliers, illustrating the effect of this preprocessing step on the data distribution[3].

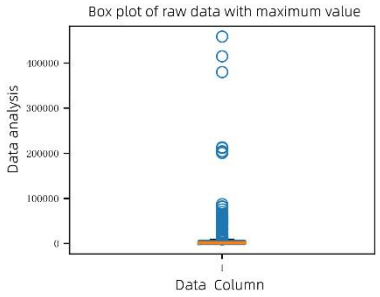

Figure 1. Original data box type diagram

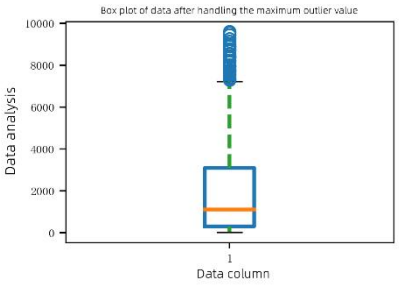

Figure 2 Data box diagram after outlier processing

After removing outliers, fill in the missing values of the KNN algorithm, where the n-highbors parameter is set to 5.

4.Feature Engineering

4.1. Yeo-Johnson

In order to improve the generalization ability of the model and prevent overfitting, when the absolute value of skewness exceeds 0.05, the data does not conform to the normal distribution, and it is necessary to perform Yeo-Johnson transformation to improve the distribution form. Figure 3 and Figure 4 shows the density histogram before and after partial feature conversion.

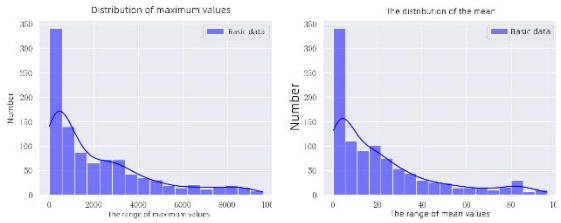

Figure 3. Histogram distribution of raw data

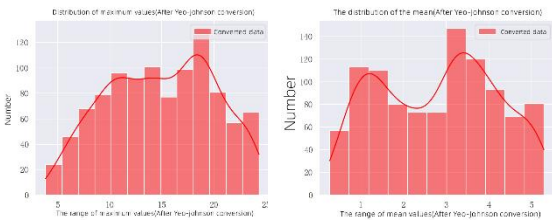

Figure 4. Histogram after Yeo-Johnson conversion

4.2. Feature screening

Initially, the paper uses the Pearson correlation coefficient to assess linear relationships between features, with coefficients ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation). To reduce redundancy, typically only one feature from a set of highly correlated features is retained. Further, the study employs the XGBoost algorithm and Python's SHAP library to calculate feature importance, which indicates the impact of each feature on the target variable. Features with high importance values are prioritized. These methods help optimize the structure and performance of the model[4]. Figure 5 and Figure 6 are correlation analysis and SHAP value calculation.

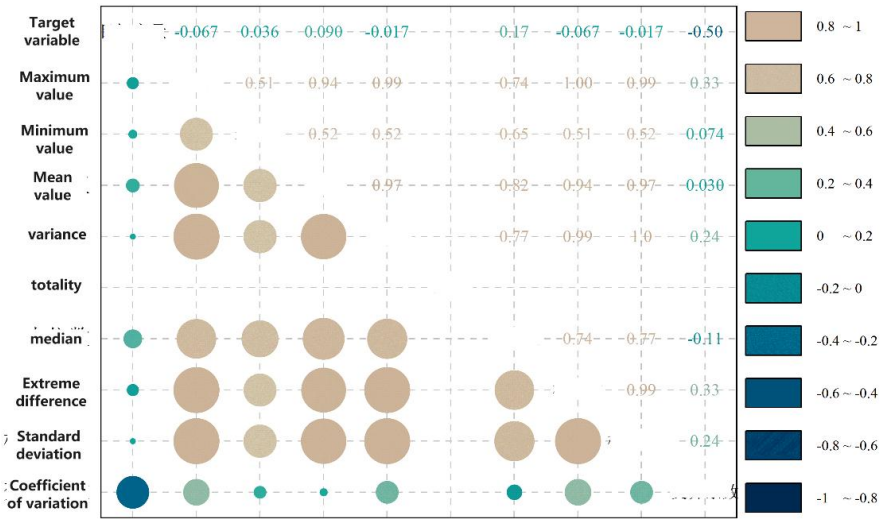

Figure 5. Feature correlation coefficient matrix heat map

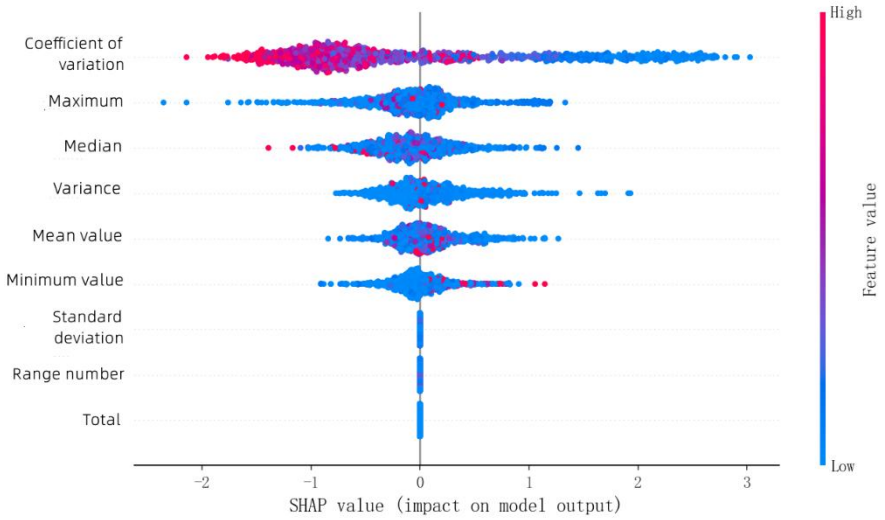

Figure 6. Importance of features

The analysis reveals that the feature "count" has a zero correlation with other variables and the lowest feature importance in XGBoost, while "range" shows high correlation with "maximum," "mean," "variance," and "standard deviation" and also low importance in XGBoost[5]. Therefore, both features are removed to retain the remaining features. Using the selected features for model training leads to more accurate and interpretable results.

5. Model training

5.1. Determination of hyperparameters

This study optimized the hyperparameters of various base models through grid search. Specifically, the LightGBM model parameters optimized included the number of leaves ('num_leaves'), learning rate ('learning_rate'), maximum depth of the tree ('max_depth'), minimum number of samples in leaf nodes ('min_child_samples') [6], and the data sampling ratio ('subsample'). For the XGBoost model, adjustments were made to the maximum depth of the tree ('max_depth'), learning rate ('learning_rate'), minimum loss reduction required for node splitting ('gamma'), data sampling ratio ('subsample'), and column sampling ratio ('colsample_bytree'). The Random Forest model's parameters optimized were the number of trees ('n_estimators'), maximum depth ('max_depth'), minimum number of samples required to split an internal node ('min_samples_split'), and the minimum number of samples required at a leaf node ('min_samples_leaf'). The Support Vector Machine model considered the penalty coefficient ('C'), influence of the kernel function ('gamma'), and kernel type ('kernel')

[7]. The Logistic Regression model optimized the penalty type ('penalty'), the reciprocal of regularization strength ('C'), and the optimization algorithm ('solver') [8]. The GridSearchCV explored all parameter combinations, and the best-performing combinations were selected via nested cross-validation, detailed in Table 3 as the optimal hyperparameter results.

Table 3. Optimal parameter combination

	Hyperparameter	Optimization result
LightGBM	num_leaves	31
	learning_rate	0.05
	max_depth	-1
	min_child_samples	20
	subsample	0.5
XGBoost	max_depth	3
	learning_rate	0.01
	gamma	0
	subsample	0.5
	colsample_bytree	0.5
Random Forest	n_estimators	200
	max_depth	3
	min_samples_split	2
	min_samples_leaf	1
SVM	C	1
	gamma	0.1
	kernel	linear
Logistic Regression	penalty	12
	C	1
	solver	liblinear

5.2. Test of model

In this paper, the evaluation indexes of each model were calculated by using the five-fold cross-validation[9], as shown in Table 4.

Table 4. Cross-validation scores

	Accuracy rate	Recall rate	Precision rate	F1
LightGBM	0.9608	0.9633	0.9650	0.9619
XGBoost	0.9829	0.9823	0.9875	0.9836
Random Forest	0.9659	0.9642	0.9689	0.9651
SVM	0.9435	0.9436	0.9479	0.9429
Logistic Regression	0.9492	0.9483	0.9521	0.9484
Stacking	0.9959	0.9950	0.9989	0.9955

In addition, the confusion matrix of each model test set is drawn, which provides a detailed view of the model training process and intuitively shows the classification

performance of the model[10]. As shown in Figure 7.

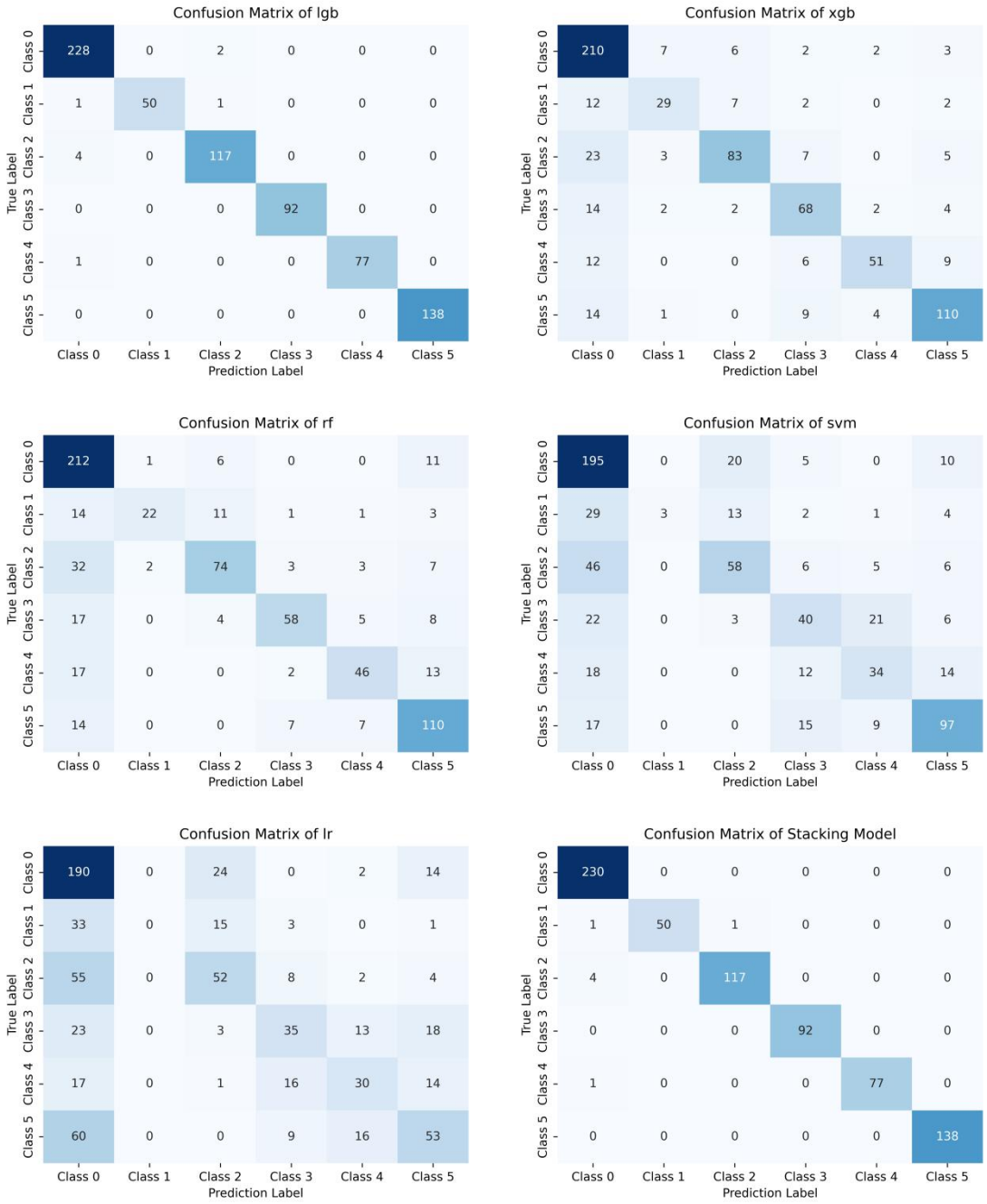

Figure 7 Confusion matrix

In summary, the Stacking fusion model is significantly higher than other single models.

6. Conclusion

The paper developed an arrhythmia detection model based on Stacking that exceeds the accuracy and stability of single models, showing significant improvements over existing research in terms of accuracy and robustness. This

ensemble model provides a more precise and reliable solution for arrhythmia detection. In the future, efforts will focus on optimizing the model by exploring advanced feature selection and ensemble strategies to enhance its generalizability. Additionally, the model's effectiveness will be validated through broader practical applications, particularly in the healthcare sector. Plans also include expanding the research scope to incorporate a more diverse range of arrhythmia detection algorithms, aiming to improve diagnostic accuracy and promote optimized patient care and treatment strategies.

References

[1] DE LUNA A B, COUMEL P, LECLERCQ J F. Ambulatory sudden cardiac death: Mechanisms of production of fatal arrhythmia on the basis of data from 157 cases. American Heart Journal, 1989, 117(1): 151-159. https://doi.org/10.1016/0002-8703(89)90670-4.

[2] LOWN B, WOLF M. Approaches to Sudden Death from Coronary Heart Disease. Circulation, 1971, 44(1): 130-142. https://doi.org/10.1161/01.CIR.44.1.130.

[3] THOMAS BIGGER J. Identification of patients at high risk for sudden cardiac death. The American Journal of Cardiology, 1984, 54(9): 3D-8D. https://doi.org/10.1016/S0002-9149(84)80277-5.

[4] Automated identification of shockable and non-shockable life-threatening ventricular arrhythmias using convolutional neural network. Future Generation Computer Systems, 2018, 79: 952-959. https://doi.org/10.1016/j.future.2017.08.039.

[5] GLIKSON M, NIELSEN J C, KRONBORG M B, Etc. 2021 ESC Guidelines on cardiac pacing and cardiac resynchronization therapy: Developed by the Task Force on cardiac pacing and cardiac resynchronization therapy of the European Society of Cardiology (ESC) With the special contribution of the European Heart Rhythm Association (EHRA). EP Europace, 2022, 24(1): 71-164. https://doi.org/10.1093/europace/euab232.

[6] Kumar Ashish, Tomar Harshit, Mehla Virender Kumar, et al. Stationary wavelet transform based ECGsignal denoising method. ISA Transactions, 2020, 114: 251-256. https://doi.org/10.1016/j.isatra.2020.12.029.

[7] Boda Somaraju, Mahadevappa Manjunatha, Dutta Pranab Kumar. A hybrid method for removal ofpower line interference and baseline wander in ECG signals

using EMD and EWT. BiomedicalSignal Processing and Control, 2021, 67: 102466. https://doi.org/10.1016/j.bspc.2021. 102466.

[8] Patro K K, Jaya Manmadha Rao M, Jadav A, et al. Noise removal in long-term ECG signals usingEMD-Based threshold method. Data Engineering and Communication Technology. Singapore:Springer, 2021: 461-469. https://doi.org/10.1007/978-981-16-0081-4_48.

[9] Bakshi A, Panigrahy M, Das J K. FPGA Based Digital Filters Design to Remove Noise from ECGSignal[A]. 2021 IEEE International Symposium on Smart Electronic Systems, Jaipur :IEEE, 2021:236-239. https://doi.org/10.1109/iSES52644.2021.00061.

[10]Liu W, Du Y. An Improved ECG Denoising Algorithm Based on Wavelet-scale Correlation Coefficients[A]. 2021 IEEE 15th International Conference on Anti-counterfeiting, Security, and Identification (ASID)[C], Xiamen: IEEE, 2021: 53-56. https://doi.org/10.1109/ASID52932.2021.9651689.