Journal of Biomedical Research

Predicting postoperative atrial fibrillation: An explainable deep learning approach

Harneet Kaur, Haiquan Chen, Jorge Catrip, Nathaen Weitzel, Hong Liu

Cite this article as:
Harneet Kaur, Haiquan Chen, Jorge Catrip, Nathaen Weitzel, Hong Liu. Predicting postoperative atrial fibrillation: An explainable deep learning approach[J]. *Journal of Biomedical Research*, In press. doi: 10.7555/JBR.39.20250387

View online: https://doi.org/10.7555/JBR.39.20250387

Available online at www.jbr-pub.org.cn

Open Access at PubMed Central

Journal of Biomedical Research, 2026 40(0): 1–6

Letter

JBR

Predicting postoperative atrial fibrillation: An explainable deep learning approach

Dear Editor,

Postoperative atrial fibrillation (POAF) is the most common complication after cardiac surgery. POAF causes a reduction in ventricular filling and atrial pooling, increases the risk of stroke threefold, and leads to other complications such as hypotension, pulmonary edema, acute kidney injury, heart failure, and encephalopathy[1–2]. In the present study, we developed TransAF, an accurate yet explainable transformer-powered deep learning system to predict the onset of POAF after cardiac surgery. By leveraging the transformer architecture[3–4] (the state-of-the-art deep learning technique for learning from heterogeneous, tabular data), TransAF represents a key step forward in bridging the gap between state-of-the-art deep learning research and the needs of the cardiac surgery community. The performance of TransAF and eight baseline predictive models was evaluated for comparison in terms of sensitivity, F1, and AUROC (the area under the receiver operating characteristic curve). Extensive experiments showed that TransAF outperformed all the state-of-the-art predictive methods with up to 46% and 15% improvements in terms of sensitivity and F1 scores, respectively. In addition to improved prediction performance, TransAF provides feature interaction analysis, feature attribution analysis, and feature importance analysis to support early intervention to optimize patient care by leveraging the latest gradient-based XAI (explainable artificial intelligence) algorithms. Specifically, to the best of our knowledge, TransAF constitutes the first attempt to explore the explainability of the POAF prediction results by revealing the complex, pairwise interactions among risk factors affecting POAF incidence through learning effective attention patterns using the transformer architecture[3–4].

We created a dataset from an institutional Society of Thoracic Surgeons (STS) database (records with missing values were excluded), which consisted of 2 068 patient records with 149 features, of which 1 587 were non-POAF patients and 481 were POAF patients. Patients with previously documented AF were excluded. Since our goal was to predict the onset of POAF after cardiac surgery, we excluded all postoperative features, which led to 82 features (77 categorical and 5 continuous). The definitions for all features are provided in ***Supplementary Tables 1*** and ***2***. Demographic, clinical, and surgical characteristics of the dataset are summarized in ***Supplementary Tables 3*** and ***4***.

As shown in ***Fig. 1A***, TransAF consists of three components: (1) a categorical and numerical feature encoder, (2) a self-attention embedding transformation, and (3) a co-attention embedding fusion. (1) *Categorical and numerical feature encoder*. For each input value, we used a neural network with one linear layer to extract its dense vector representation by treating each different input value as a unique token[4]. (2) *Self-attention embedding transformation.* We used the self-attention mechanism to refine the extracted embeddings to their contextual representations to learn the hierarchical dependencies within the numerical and categorical features[3]. Specifically, we created two separate stacks of self-attention transformer blocks, one for the categorical features and the other for the numerical features. Each self-attention block is composed of two components: multi-headed self-attention and a feed-forward network[5]. (3) *Co-attention embedding fusion*. We employed the co-attention mechanism to capture the complex categorical-numerical

Received: 08 September 2025; Revised: 15 January 2026; Accepted: 19 January 2026;

CLC number: R541.7, Document code: B

The authors reported no conflict of interests. Hong Liu is a Associate Editor-in-Chief of this journal and was excluded from all editorial decisions. The manuscript was handled independently, with peer review conducted by at least two external reviewers.

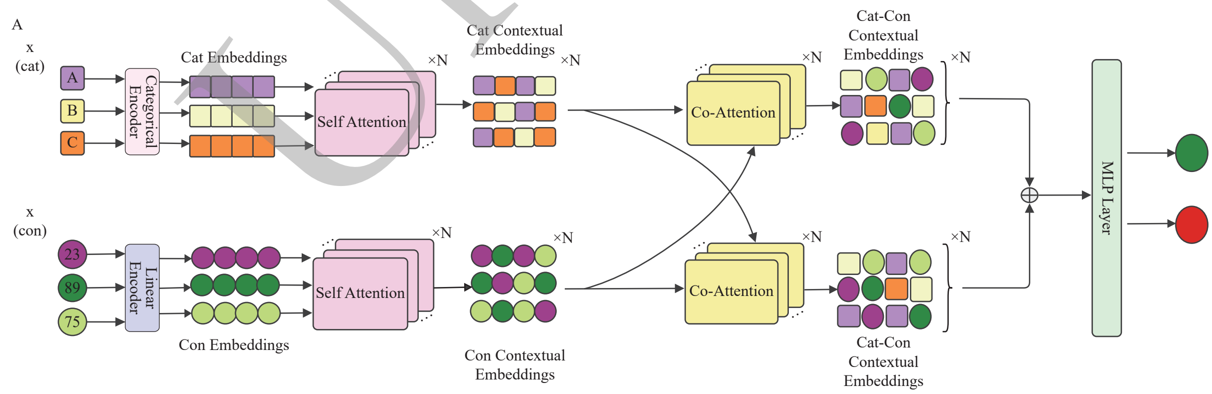
https://doi.org/10.7555/JBR.39.20250387

interactions by fusing the refined contextual representations for the numerical and categorical features[6–7]. Specifically, we used two separate stacks of cross-attention transformer blocks, one for the categorical features and the other for the numerical features. The co-attention design allows our model to effectively learn the complex interactions between continuous and categorical features. The output embeddings from the multi-headed co-attention blocks were then concatenated and passed to a feed-forward network with multiple fully connected layers, followed by an output layer with the softmax activation function as the classification head.

For performance comparison, we implemented the following models using Scikit-learn, XGBoost, and PyTorch. (1) SVM-R: a support vector machine with the kernel and C value set to the radial basis function (RBF) and 10 000[8–9]. (2) SVM-P: a support vector machine with kernel and C value set to the polynomial function and 10 000[8–9]. (3) RF: a random forest model with 100 trees. (4) GBDT: a gradient-boosted decision tree model with 100 trees and a learning rate of 0.05[8–9]. (5) FCNN: a feed-forward neural network including three fully-connected layers[8–9]. (6) ResNet: a residual neural network including nine ResBlocks and two LSTM layers[10]. (7) TransAF-S: an ablated version of TransAF where the co-attention blocks are removed to verify the effectiveness of the co-attention embedding fusion. (8) TransAF-C: an ablated version of TransAF where the self-attention blocks are removed to verify the effectiveness of the self-attention embedding transformation. (9) TransAF: the complete version of TransAF, which includes one fully-connected layer as categorical and numerical feature encoder, eight self-attention blocks (each block with eight heads), and one co-attention block, followed by one fully connected output layer, with the embedding size and batch size set to 64 and 128, respectively. The binary cross-entropy loss with the Adam optimizer and a learning rate of 0.000 01 was used. We performed the 5-fold cross-validation and reported the average metric scores on the positive label (POAF) for all the models. We also applied random oversampling by duplicating POAF instances in the training set for class balancing. An NVIDIA A100 GPU with 12 GB of RAM was used to train all the models. For all the deep learning models, *i.e.*, FCNN, ResNet, TransAF-S, TransAF-C, and TransAF, early stopping was used to prevent overfitting.

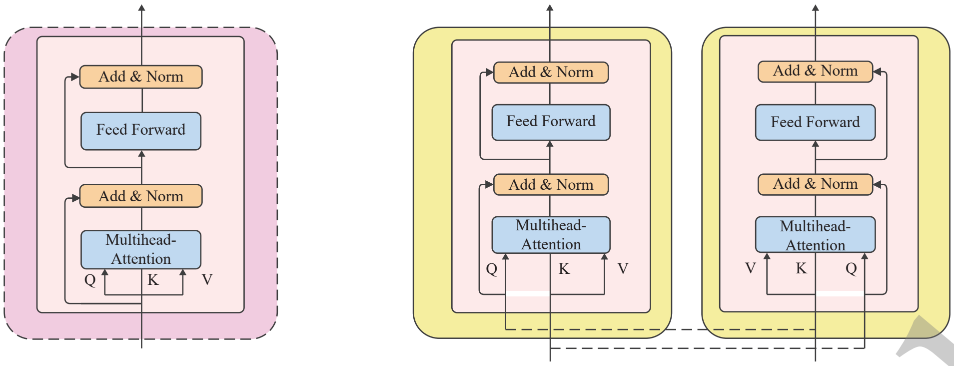
Supplementary Table 5 summarizes the performance results of all the models in terms of the sensitivity, F1, and AUROC scores on the positive label (POAF). As shown in ***Supplementary Table 5***, TransAF yielded a sensitivity score of 0.70, an F1 score of 0.38, and an AUROC score of 0.56, outperforming all baseline models. Notably, TransAF outperformed all other models in terms of sensitivity with an improvement of up to 46%. Specifically, SVM-R, SVM-P, RF, GBDT, FCNN, ResNet, and TransAF achieved a sensitivity score of 0.32, 0.24, 0.24, 0.29, 0.31, 0.24, and 0.70, respectively. In terms of the F1 score, TransAF also outperformed all baseline models with up to 15% performance improvement. Specifically, SVM-R, SVM-P, RF, GBDT, FCNN, ResNet, and TransAF achieved an F1 score of 0.29, 0.23, 0.24, 0.27, 0.31, 0.27, and 0.38, respectively. Furthermore, TransAF-C and TransAF-S underperformed TransAF in terms of sensitivity, F1, and AUROC scores. TransAF-C returned a sensitivity score of 0.47 and an F1 score of 0.33, while TransAF-S returned a sensitivity score of 0.33 and an F1 score of 0.29. This result verifies the effectiveness of the self-attention embedding transformation and the co-attention embedding fusion in TransAF.

To analyze feature interactions, we randomly selected two correctly predicted patients (one POAF patient and one non-POAF patient) from our test set.

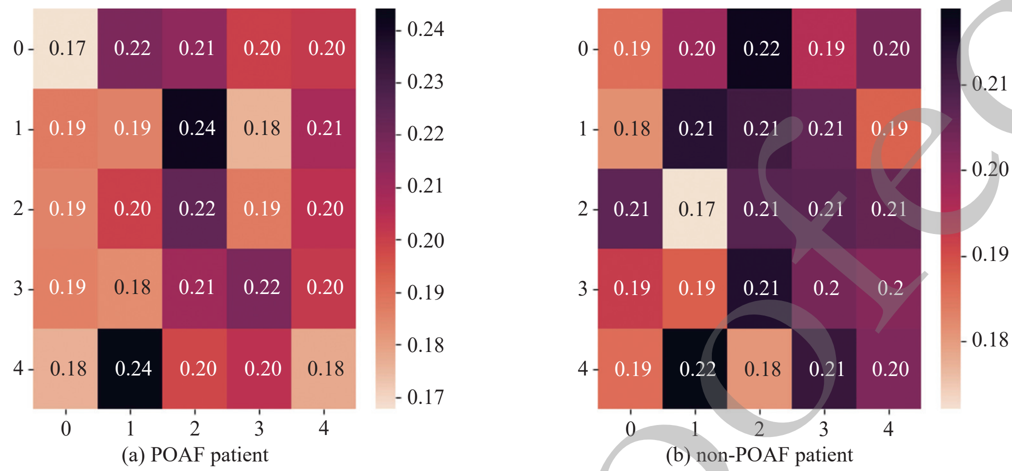


(Continued)

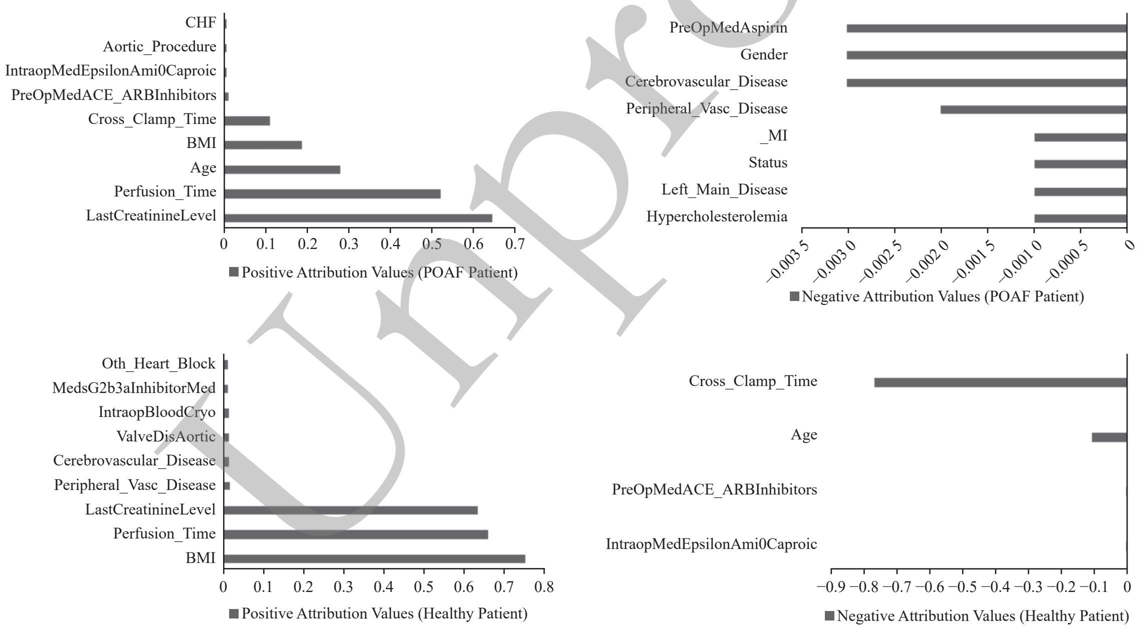
(Continued)



B



C



(Continued)

(Continued)

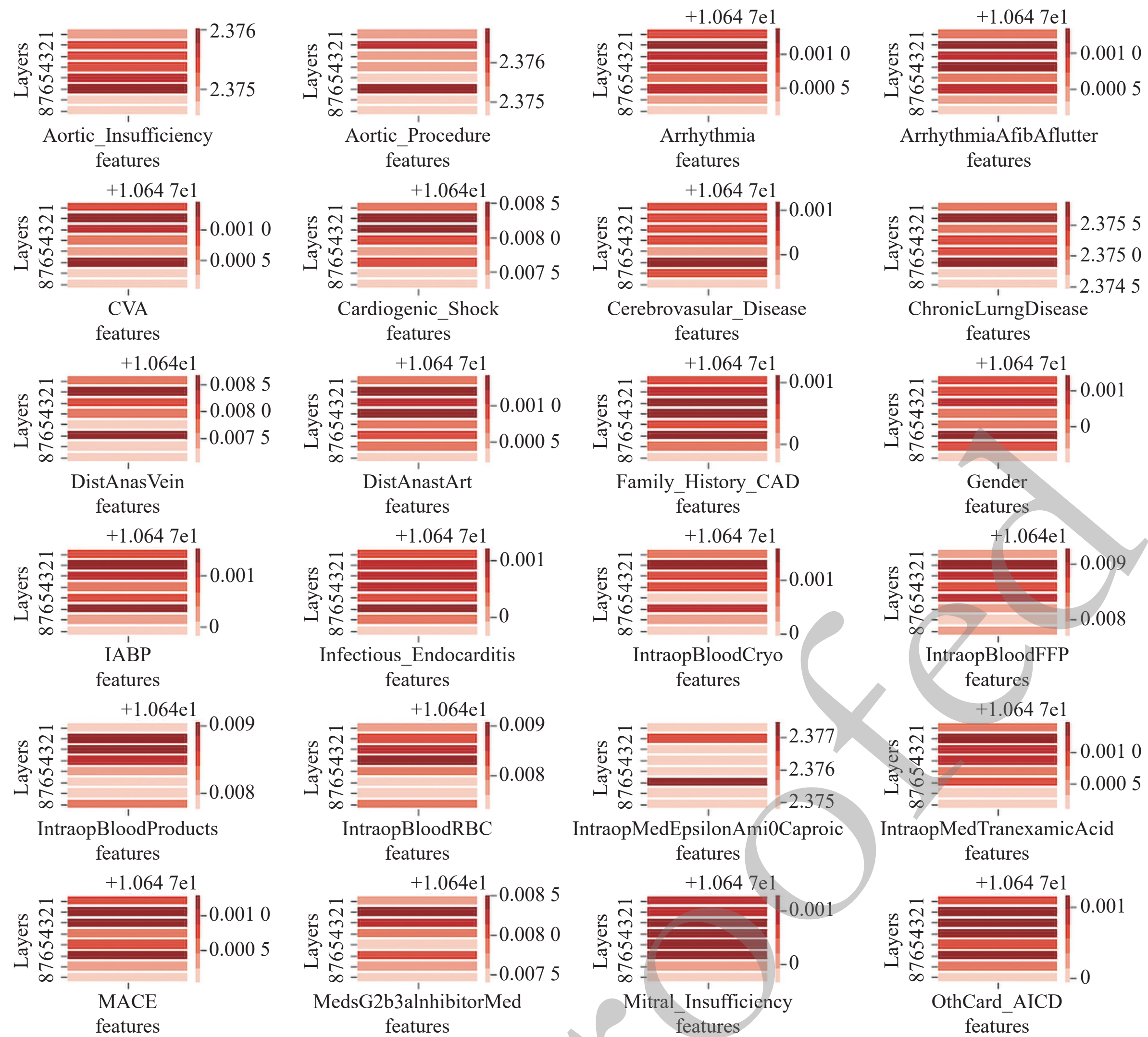


Fig. 1 **Postoperative atrial fibrillation (POAF) feature interaction and attribution analysis.** A: Architecture of TransAF (a); self-attention embedding transformation (b); co-attention embedding fusion (c). B: Feature interaction analysis. Heatmaps of the attention scores learned in the first self-attention layer among the 5 numerical features for the selected POAF patients (a) and the selected non-POAF patients (b). C: Feature attribution analysis. The features with the highest (a) and the lowest (b) attribution scores for the selected POAF patients. The features with the highest (c) and the lowest (d) attribution scores for the selected non-POAF patients. For (a) and (b), the target label was set to POAF (label 1), where a positive attribution score means the risk factor pushes the prediction up, increasing the probability for the model to output POAF. For (c) and (d), the target label was set to non-POAF (label 0), where a positive attribution score means the risk factor pushes the prediction down, increasing the probability for the model to output non-POAF. D: Illustration of how the attribution score of each risk factor evolved through the model layers for the selected POAF patients, where a positive attribution score means the risk factor value pushes the prediction up, increasing the probability for the model to output POAF.

Supplementary Fig. 1 shows the attention scores learned in the first self-attention block among the 77 categorical features for the selected POAF patient. Any off-diagonal entry (the darker the color, the higher the value) indicates the magnitude of the combined effect of two different risk factors on the corresponding POAF incidence. In *Supplementary Fig. 1*, features 1 (aortic stenosis), 2 (aortic insufficiency), 3 (aortic procedure), 7 (congestive heart failure (CHF)), 11 (chronic lung disease), 28 (intraoperative epsilon-aminocaproic acid), 45 (preoperative aspirin), 61 (prior heart failure), and 71 (aortic valve diseases) exhibit stronger interactions with each other than with the other features. *Supplementary Fig. 2* shows the attention scores learned in the first self-attention block among the 77 categorical features for the selected non-POAF patient. In *Supplementary Fig. 2*, feature 0 (number of coronary vessels corrected) shows stronger interactions with features 1 (aortic stenosis), 3 (aortic procedure), 6 (coronary artery bypass graft [CABG]), 7 (CHF), 15 (distal arterial anastomosis), 17 (gender), 18 (hypercholesterolemia), 19 (hypertension), 20 (intra-aortic balloon pump [IABP]), 26 (intraoperative blood products), 27 (intraoperative red blood cell), 28 (intraoperative epsilon-aminocaproic acid), 30 (intraoperative dexmedetomidine), 40 (preoperative antiplatelets), 45 (preoperative aspirin), 49 (preoperative lipid lowering medications), 52 (preoperative beta blockers), 65 (smoking), 70 (valve), and 75 (myocardial infarction) than with other features. ***Fig. 1B*** shows the heatmaps of the attention scores learned in the first self-attention block among the five numerical features for the selected POAF patient and the non-POAF patient, respectively. In ***Fig. 1B*** (a), for the POAF patient, features 1 (body mass index [BMI]), 2 (preoperative last creatinine level), and 4 (perfusion time) interact with each other with higher attention scores than with other features. In ***Fig. 1B*** (b), for the non-POAF patient, features 0 (age) and 2 (preoperative last creatinine level) interact more closely with each other than with other features.

We further performed a feature attribution analysis. We calculated the attribution scores of each risk factor for the two selected POAF and non-POAF patients by leveraging the latest gradient-based XAI (explainable artificial intelligence) algorithms. Attribution scores are proportional to the contribution of a feature value to a model's prediction, indicating how much a risk factor helps push the model prediction up (positive) or down (negative). Attributions across all features add up to the model's final prediction. ***Fig. 1C*** (a) and ***Fig. 1C*** (b) show the top features with the highest and lowest attribution scores at the input layer for the selected POAF patient, where age, perfusion time, last creatinine level, cross-clamp time, and BMI were the features with the highest attribution scores leading the model to predict POAF. On the other hand, ***Fig. 1C*** (c) and ***Fig. 1C*** (d) show the top features with the highest and lowest attribution scores at the input layer for the selected non-POAF patient, where age, perfusion time, and BMI were the features with the highest attribution scores leading the model to predict non-POAF, whereas cross-clamp time and last creatinine level were the two features with the lowest attribution scores. ***Fig. 1D*** shows how the attribution scores of various risk factors evolved through all the self-attention layers within TransAF for the selected POAF patient.

We next performed a feature importance analysis. ***Supplementary Table 6*** shows the averaged attribution scores of the risk factors at the input layer returned by TransAF across the POAF patients and non-POAF patients in our dataset, respectively. As shown in ***Supplementary Table 6***, for POAF patients, age, last creatinine level, cross-clamp time, and perfusion time had the highest averaged attributions in predicting POAF, while BMI negatively contributed to a POAF prediction. On the other hand, for non-POAF patients, age, BMI, history of renal failure, intra-arrtic ballon pump (IABP), and left main disease had the highest averaged attributions in predicting non-POAF, while cross-clamp time, perfusion time, and last creatinine level negatively contributed to a non-POAF prediction. Notably, some risk factors, such as age, last creatinine level, cross-clamp time, perfusion time, and BMI, contributed significantly (either positively or negatively) to both POAF and non-POAF predictions, indicating that they are strong predictors that can be used to predict both the presence and absence of POAF. For example, in older patients, ages typically contribute positively to the presence of POAF, whereas in younger patients, ages contribute positively to the absence of POAF.

In summary, we developed TransAF, an accurate yet explainable transformer-powered deep learning system to predict the onset of POAF after cardiac surgery. TransAF represents a key step forward in bridging the gap between state-of-the-art deep learning research and the needs of the cardiac surgery community. Our study suggests that the deep learning transformer architecture can be effectively leveraged to predict POAF and reveal the complex interactions among risk factors to support early and timely intervention in clinical practice, thereby optimizing preoperative, intraoperative, and postoperative care. However, there are also some limitations to this study. Since our study was a single-site retrospective study, the model's performance was limited by several factors: (1) to enable rigorous validation on the model's predictive performance, we excluded all postoperative features from model training; (2) our dataset lacked continuous electrocardiogram monitoring data; and (3) the number of POAF cases and the available clinical features were limited in our single-site STS Adult Cardiac Surgery database (ACSD). Our future research agenda includes extending and validating this study from a single site to multiple sites by using the national STS ACSD database and including all preoperative demographics, risk factors, medications, intraoperative variables, new-onset POAF, and other outcome data.

This work was supported in part by the Department of Anesthesiology and Pain Medicine at the University of California Davis Health, NIH grant UL1 TR000002 awarded to the University of California Davis Health, and the Research and Creative Activity Award of California State University, Sacramento.

Yours sincerely,

Harneet Kaur1, Haiquan Chen[1,✉], Jorge Catrip[2], Nathaen Weitzel[3], Hong Liu[3,✉]

[1]Department of Computer Science, California State University,
Sacramento, CA 95819-6021,
USA;
[2]Department of Surgery, University of California Davis Health, Sacramento, CA 95817,
USA;
[3]Department of Anesthesiology and Pain Medicine, University of California Davis Health, Sacramento, CA 95817,
USA.
✉Corresponding authors: Hong Liu, E-mail: hualiu@ucdavis.edu; Haiquan Chen, E-mail: haiquan.chen@csus.edu.

References

[1] Andrade JG, Aguilar M, Atzema C, et al. The 2020 Canadian Cardiovascular Society/Canadian Heart Rhythm Society comprehensive guidelines for the management of atrial fibrillation[J]. *Can J Cardiol*, 2020, 36(12): 1847–1948.

[2] Maisel WH, Kuntz KM, Reimold SC, et al. Risk of initiating antiarrhythmic drug therapy for atrial fibrillation in patients admitted to a university hospital[J]. *Ann Intern Med*, 1997, 127(4): 281–284.

[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc. , 2017: 6000–6010.

[4] Gorishniy Y, Rubachev I, Khrulkov V, et al. Revisiting deep learning models for tabular data[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc. , 2021: 1447.

[5] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 770–778.

[6] Liu R, Huang ZA, Hu Y, et al. Spatial-temporal co-attention learning for diagnosis of mental disorders from resting-state fMRI data[J]. *IEEE Trans Neural Netw Learn Syst*, 2024, 35(8): 10591–10605.

[7] Wu Y, Zhan P, Zhang Y, et al. Multimodal fusion with co-attention networks for fake news detection[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Kerrville: Association for Computational Linguistics, 2021: 2560–2569.

[8] Lu Y, Chen Q, Zhang H, et al. Machine learning models of postoperative atrial fibrillation prediction after cardiac surgery[J]. *J Cardiothorac Vasc Anesth*, 2023, 37(3): 360–366.

[9] Parise O, Parise G, Vaidyanathan A, et al. Machine learning to identify patients at risk of developing new-onset atrial fibrillation after coronary artery bypass[J]. *J Cardiovasc Dev Dis*, 2023, 10(2): 82.

[10] Tohyama T, Ide T, Ikeda M, et al. Deep learning of ECG for the prediction of postoperative atrial fibrillation[J]. *Circ Arrhythm Electrophysiol*, 2023, 16(2): e011579.