Cardiovascular study enabled ensemble learning of 10-year risk of future coronary heart disease (CHD)

**Abstract**

10-year risk of future coronary heart disease (CHD) refers to the probability that a person will develop coronary heart disease within the next ten years. Accurate identification of 10-year risk of future CHD enables healthcare providers to implement preventive strategies, including lifestyle changes, medications, and regular monitoring, thus greatly reducing the risk of CHD. However, there are currently few ensemble-based computational methods available for predicting 10-year risk of future coronary heart disease (CHD). In this study, we introduce an innovative ensemble approach, successfully enabling the computational evaluation of 10-year risk of future coronary heart disease (CHD). We conducted a comprehensive model selection process that involved multiple basic machine learning algorithms and deep learning architectures such as recurrent neural networks, convolutional neural networks and Transformer-based models. Trained with scaled 6 nominal variables (e.g. male, education and so on) one-hot encoded and 8 continuous variables (e.g. age, BMI and so on), the performance of individual models achieved an AUROC between 0.552 and 0.717. To further enhance prediction accuracy and robustness, we then constructed the ensembles of these individual models with different combinations. The best performance attained by our ensemble models reached an AUROC of 0.715. Shapley additive explanations were conducted to explain the significant contributions of specific genomic feature, providing an insight of most important variables for CHD risk prediction.

**Keywords**

Coronary heart disease (CHD); Ensemble learning; One-hot encoding; Recurrent neural networks; Convolutional neural networks; 1 dimensional transformer encoder

**1 Introduction**

[coronary heart disease (CHD) 介绍一段]

[computational ways to do the prediction of CHD 的相关工作的介绍]

Here, we took advantage of model-based ensemble deep learning and incorporated 7 nominal variables (e.g. male, education and so on) one-hot encoded and 8 continuous variables (e.g. age, BMI and so on), to develop a computational evaluation method for 10-year risk of future coronary heart disease (CHD). Our research comprehensively investigated the mainstream machine learning algorithms (SVM, Logistic Regression, XGBoost) and deep learning algorithms (CNN, ResNet, LSTM, Transformer). Among them, the LSTM, inspired by traditional recurrent neural networks (RNNs), achieved an AUROC around 0.711, which is the highest among the algorithms. Furthermore, by incorporating ensembles of multiple algorithms, our proposed methods finally achieved an AUROC of 0.715. **Figure 1** provides a detailed schematic diagram of LSTM and the ensemble models. To interpret proposed models, Shapley value analysis was conducted to analyze how different features contribute to the overall predictions and provide insights into most important variables for CHD risk prediction.

**2 Methods**

**2.1 Benchmark dataset**

To develop our 10-year risk of future CHD prediction framework, medical records from 644 CHD-positive patients (dataset P) and 3594 CHD-negative patients were obtained from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. Since there is a serious bias between positive and negative data, the negative data (dataset N) was randomly selected from these medical records of CHD-negative patients with 1:1 P-to-N ratio. The dataset was randomly split into the training and testing part with a ratio of 4:1. For the training part, it was divided into training set and validation set when applying the 5-fold cross validation.

**2.2 Medical records**

The 15 medical records include both demographic, behavioral and medical risk factors. For demographic risk factors, sex (stored in ‘male’) and age (stored in ‘age’) are taken into consideration. For behavioral risk factors, whether or not the patient is a current smoker (stored in ‘Current Smoker’) is examined and even more specific, the number of cigarettes that the person smoked on average in one day (stored in ‘cigsPerDay’) was recorded. Medical risk factors can be categorized into two conditions: history or current. For history medical risk factors, we check whether or not the patient was on blood pressure medication (stored in ‘BP Meds’), had previously had a stroke (stored in ‘prevalent Stroke’), the patient was hypertensive (stored in ‘prevalent Hyp’) and the patient had diabetes (stored in ‘diabetes’). For current medical risk factors, we examined the total cholesterol level (stored in ‘totChol’), systolic and diastolic blood pressure (stored in ‘sys BP’ and ‘dia BP’), body mass index (stored in ‘BMI’), heart rate (stored in ‘heartRate’), glucose level (stored in ‘glucose’) and the target for prediction which is 10 year risk of coronary heart disease CHD (stored in ‘TenYearCHD’).

**2.2 Data preprocessing**

Above 15 variables can be categorized into 6 nominal variables (e.g. male, education and so on) one-hot encoded and 8 continuous variables (e.g. age, BMI and so on) for training features and 1 nominal variables (‘TenYearCHD’) for labelling. To eliminate false ordinality in numeric categories, one-hot encoding strategy was applied to those 6 nominal variables, thus preventing the model incorrectly interpret these category labels as representing some sort of order or magnitude. To prevent an unfair feature examination, we maintain consistent data scale using **formula 1** towards the 14 training features across all sets. Without consistent scaling, a higher coefficient might not necessarily imply higher importance, as it could be reflecting the scale of the feature rather than its actual influence.

where is the original value of a feature, is the mean of the feature, is the standard deviation of the feature and is the standardized value.

**2.3 Model training and ensemble models development overview**

We evaluated three traditional machine learning methods: SVM, Logistic Regression, and XGBoost, alongside four deep learning methods: Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Transformer-Encoder, and ResNet (Residual block + CNN). Initially, these methods were trained and evaluated using these scaled 15 variables with nominal variables encoded in one-hot format. This approach was employed using a benchmark dataset with a balanced positive-negative ratio of 1:1. The dataset was split into a 4:1 ratio for training and testing, respectively, with five-fold cross-validation utilized during training.

Subsequently, the top five performing methods were selected to construct ensemble models, where the final result was obtained by aggregating the predicted values from each model. Five different combinations of these selected methods were evaluated and compared. The same testing dataset was utilized for evaluating each individual combination.

**2.4 LSTM**

Long Short-Term Memory Network (LSTM) is a specialized variant of Recurrent Neural Network (RNN), which has been widely used in NLP tasks such as text categorization, machine translation and language models. LSTMs effectively addressed the challenge of vanishing or exploding gradients encountered by conventional RNNs when processing long series of data. In our study, the input of LSTM was set to be a 1-dimentional input vector, consisting of 24 dimensions for 16 one-hot encoded nominal variables and 8 continuous variables. The hidden size was set to be 256. We utilized two LSTM layers for model construction, followed by a linear layer to reduce the output to a single value. A sigmoid activation function was applied at the end to obtain a binary output. The LSTM model architecture is displayed in **Figure 1-D**. For training, we employed binary cross-entropy loss as the loss function and Adam optimizer with learning rate set to be 0.001. The epoch was set to be 50.

**2.5 ResNet**

ResNet is a model composed by convolutional blocks and residual structures to evaluate the risk of 10-year risk of future coronary heart disease (CHD). To be specific, both convolutional blocks were configured as 1-dimensional with a kernel size of 3, stride of 1, and padding of 1. Within the first convolutional block, the data passed through a convolutional layer, followed by a batch normalization and a ReLU activation. Subsequently, the processed data underwent another convolutional layer and a batch normalization. The number of channels of each convolutional layer is set to [8, 16]. The second block shared similar architecture as the first convolutional block. The number of channels of each convolutional layer is set to [32, 16]. Two residual blocks were implemented after two convolutional blocks, though a shortcut connection, adding the output from the previous layer to the output of the current layer. Next, the output is flattened and passed through dense layers, and fed into a sigmoid activation function for binary classification (suffer CHD within ten years Yes/No). Here is the formula **(2)** for sigmoid activation function, where z is the input in the last two neuron before activation.

The loss function employed in this ResNet is binary cross entropy (i.e. BCELoss):

where is the total number of samples. is the label of each individual sample, with a value of either 0 or 1. is the predicted probability that the sample belongs to class 1, as predicted by the model for each sample. denotes the natural logarithm. The loss can also be expressed as

The Adam optimizer was implemented with a starting learning rate of 0.001, which takes advantage of the momentum strength and adaptive learning rate simultaneously. The epoch for the training was set to be 100.

**2.6 Transformer-Encoder method**

Transformer is a prevalent Seq2seq model first proposed for neural machine translation and subsequently has been applied to many NLP tasks. It functioned through an Encoder-Decoder architecture. Since it has the potential to be adapted for various types of disease risk evaluation tasks, we involved it in our study. Here, we only employed the Encoder module of the Transformer model: our model is mainly comprised by three multi-head self-attention layers. In each multi-head attention layer, multiple self-attention heads are parallel attention structures operating on the input sequence. For each attention head, three linear transformations of the input embeddings are derived: Query (Q), Key (K), and Value (V). These vectors are then used to compute the attention scores as follows:

where is the column number of those equal-sized three matrixes, i.e. the matrix dimension. A softmax function is used to weight the corresponding value vectors. The concatenation of the resulting projection of this kind of attention head is then linearly transformed to finally produce the output of the multi-head attention layer. The model architecture is displayed in **Figure 1-D**.

Details of hyperparameters used in this study were set as follow. The input to this model is a vector that incorporating scaled 16 one-hot encoded nominal variables and 8 continuous variables, thus the total feature length is of 24. This is to make the input fit the model's dimensional requirement: the input 1-d vector was transformed into a 3×8 matrix. In each multi-head self-attention layer, the head number is set to be 4. The dimension of each Query, Key, and Value matrix is (8/4)×3. The number of Encoder layers is set to 3. The final output of the Encoder module is flattened and passed through a fully connected layer with an output dimension of 2 to predict whether a person will suffer CHD or not within 10 years or not in the future. The loss function employed in our Transformer Encoder is cross entropy. Stochastic Gradient Descent (SGD) is selected as the optimizer with a learning rate of 0.005 and without momentum mechanism. The epoch for training was 15.

**2.7 Ensemble Learning**

Ensemble learning is a technique that combines the power of multiple predictive models to improve overall performance. The ensemble model was selected as one of our comparison methods for three reasons including reduced overfitting, improved accuracy, and improved robustness. Ensemble model (**Figure 1-E**) considered multiple combinations of methods to give a comprehensive evaluation. Methods with high AUROC in cross validation set and independent test set were selected for constructing ensemble model. In machine learning methods, Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), eXtreme Gradient Boost (XGB) were selected, while in deep learning methods, Long-Short Term Memory Neural Network (LSTM) and Recurrent Neural Network (ResNet) were selected. Here we considered five combinations following their AUROC: 1. LR and LSTM (top 2); 2. LR, NB and LSTM (top 2 in ML and top 1 in DL); 3. LR, NB and ResNet (top 2 in ML and rank 2 in DL); 4. LR, NB, RF and LSTM (top 4); 5. LR, NB, RF, XGB and LSTM (top 5). Those ensemble models were compared with each other. The ensemble method aggregates predictions from these different models by taking an average of their output as a fusion result. Ideally, due to the combination of multiple models, the integrated model tends not to be overly dependent on specific features of the training data, thus reducing the risk of overfitting. Meanwhile, the integrated model is less sensitive to small changes and outliers in the data.

**2.8 Performance evaluation metrics**

The following evaluation metrics were applied. We used the Receiver Operating Characteristic (ROC) curve (sensitivity against 1-specificity) and the area under the ROC curve (AUROC). Besides AUROC, sensitivity (Sn), specificity (Sp), overall accuracy (ACC), F1 score, and Matthew's Correlation Coefficient (MCC) were also included. A 5-fold cross-validation was applied on 80% of the data as training datasets, while the rest of 20% were used as testing datasets for independent testing.

Among them, TP represents the number of true positives, while TN represents true negatives; FP stands for the number of false positives, and FN stands for the number of false negatives.

**3 Results**

**3.1 Single-model methods evaluation**

To develop our prediction method, we first evaluate the performance of different approaches for CHD risk prediction within 10 years using both machine learning and deep learning methods. Five different machine learning classifiers were included: Logistic regression (LR), Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and eXtreme Gradient Boost (XGB). Four deep learning methods were included: Convolutional Neural Network (CNN), Long-Short Term Memory (LSTM) and Residual Neural Network (ResNet) and 1 dimensional Transformer Encoder (Trans-Encoder). Performance evaluation of these methods has been summarized in **Table 1**. Additionally, we demonstrated a combined ROC figure for cross validation (**Figure 2a**) and independent test (**Figure 2b**). The results indicated that logistic regression and naïve bayes for machine learning approaches and LSTM for deep learning approaches are generally outperform other models.

**Table 1 Performance evaluation of single-model methods**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Methods & Modes** | | **Sn (%)** | **Sp (%)** | **ACC (%)** | **F1** | **MCC** | **AUROC** |
| **LR** | Cross Validation | 0.611 | 0.711 | 0.650 | 0.657 | 0.309 | 0.720 |
| Independent Test | 0.659 | 0.696 | 0.678 | 0.677 | 0.358 | 0.717 |
| **NB** | Cross Validation | 0.159 | 0.947 | 0.534 | 0.273 | 0.255 | 0.694 |
| Independent Test | 0.121 | 0.667 | 0.519 | 0.205 | 0.099 | 0.709 |
| **RF** | Cross Validation | 0.566 | 0.674 | 0.612 | 0.615 | 0.233 | 0.663 |
| Independent Test | 0.545 | 0.632 | 0.605 | 0.585 | 0.214 | 0.666 |
| **SVM** | Cross Validation | 0.575 | 0.625 | 0.578 | 0.599 | 0.155 | 0.653 |
| Independent Test | 0.553 | 0.664 | 0.628 | 0.603 | 0.262 | 0.658 |
| **XGB** | Cross Validation | 0.566 | 0.646 | 0.592 | 0.604 | 0.189 | 0.623 |
| Independent Test | 0.492 | 0.570 | 0.550 | 0.528 | 0.104 | 0.632 |
| **CNN** | Cross Validation | 0.959 | 0.028 | 0.471 | 0.632 | -0.036 | 0.505 |
| Independent Test | 0.962 | 0.016 | 0.500 | 0.663 | -0.067 | 0.587 |
| **LSTM** | Cross Validation | 0.724 | 0.592 | 0.655 | 0.667 | 0.319 | 0.711 |
| Independent Test | 0.742 | 0.603 | 0.674 | 0.700 | 0.349 | 0.709 |
| **ResNet** | Cross Validation | 0.561 | 0.565 | 0.563 | 0.550 | 0.126 | 0.565 |
| Independent Test | 0.455 | 0.619 | 0.535 | 0.500 | 0.075 | 0.559 |
| **Trans-Encoder** | Cross Validation | 0.468 | 0.551 | 0.512 | 0.477 | 0.019 | 0.513 |
| Independent Test | 0.606 | 0.421 | 0.516 | 0.561 | 0.027 | 0.498 |

图表

中度可信度描述已自动生成

**Figure 2 The ROC curves of single-model methods. (A)** The ROC curves of multiple approaches for predicting 10-year risk of future coronary heart disease (CHD) under the 5-fold cross-validation test. **(B)** The ROC curves of multiple approaches for predicting 10-year risk of future coronary heart disease (CHD) under the independent dataset test.

**3.2 Ensemble-model methods evaluation**

From **Table 1**, we can see that LR, NB, RF and XGB are the top 4 machine learning approaches with the best performance; and LSTM and ResNet are the top 2 deep learning approaches with the best performance. Next, we constructed ensemble models by applying different combinations of models from those 6 methods. The performance evaluation toward those methods in independent test is summarized in **Table 2** and a combined ROC figure for independent test is demonstrated in **Figure 3**. The ensemble method aggregated predictions from these different models by fusing their output to make a final decision. All ensemble methods demonstrated improvement in AUROC (0.698-0.717) compared to the average AUROC of individual methods (0.498-0.717). The ensemble model of LR and LSTM combination outperformed other combinations with an AUROC value of 0.717.

**Table 2 Performance evaluation of ensemble-model methods**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Ensemble Methods** | **Sn (%)** | **Sp (%)** | **ACC (%)** | **F1** | **MCC** | **AUROC** |
| LR + LSTM | 0.689 | 0.675 | 0.682 | 0.689 | 0.363 | 0.717 |
| LR + NB + LSTM | 0.189 | 0.913 | 0.543 | 0.297 | 0.147 | 0.715 |
| LR + NB + ResNet | 0.136 | 0.929 | 0.523 | 0.226 | 0.106 | 0.697 |
| LR + NB + RF + LSTM | 0.287 | 0.897 | 0.585 | 0.415 | 0.232 | 0.704 |
| LR + NB + RF + XGB + LSTM | 0.364 | 0.857 | 0.605 | 0.485 | 0.253 | 0.698 |



**Figure 3 The ROC curves of ensemble-model methods.** The figure shows ROC curves of ensemble models with different method combinations under the independent dataset test.

**3.3 Model interpretation**

Gaining insights into the key input features and the underlying mechanisms behind model decisions are crucial for making further improvements to the model. However, interpreting deep learning models can be challenging. To overcome this limitation, we adopted an approach to interpret the role of different risk factors using machine learning alternative, following a previously published work. The Shapley additive explanations was employed to assess the relative importance of each input feature in the prediction. In our study, we explained a model using a gradient explainer which relies on expected gradients (an extension of integrated gradients). It served as a more general approach than deep explainer and can be applied to any differentiable model, thus being selected in our work.

**4 Conclusion**

**5 Code availability**

The deep learning based approaches, including CNN, LSTM, ResNet, 1 dimensional Transformer-Encoder and all ensemble models, were implemented with Pytorch 2.1.0. Codes for model construction can be available at:

**6 References**