


Employing feature engineering strategies to improve the performance of machine learning algorithms on echocardiogram dataset

DIGITAL HEALTH
Volume 9: 1–15
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076231207589
journals.sagepub.com/home/dhj



Huang-Nan Huang¹, Hong-Ming Chen¹, Wei-Wen Lin^{2,3,4}, Chau-Jian Huang⁵,
Yung-Cheng Chen⁶, Yu-Huei Wang² and Chao-Tung Yang^{6,7} 

Abstract

Objectives: This study mainly uses machine learning (ML) to make predictions by inputting features during training and inference. The method of feature selection is an important factor affecting the accuracy of ML models, and the process includes data extraction, which is the collection of all data required for ML. It also needs to import the concept of feature engineering, namely, this study needs to label the raw data of the cardiac ultrasound dataset with one or more meaningful and informative labels so that the ML model can learn from it and predict more accurate target values. Therefore, this study will enhance the strategies of feature selection methods from the raw dataset, as well as the issue of data scrubbing.

Methods: In this study, the ultrasound dataset was cleaned and critical features were selected through data standardization, normalization, and missing features imputation in the field of feature engineering. The aim of data scrubbing was to retain and select critical features of the echocardiogram dataset while making the prediction of the ML algorithm more accurate.

Results: This paper mainly utilizes commonly used methods in feature engineering and finally selects four important feature values. With the ML algorithms available on the Azure platform, namely, Random Forest and CatBoost, a Voting Ensemble method is used as the training algorithm, and this study also uses visual tools to gain a clearer understanding of the raw data and to improve the accuracy of the predictive model.

Conclusion: This paper emphasizes feature engineering, specifically on the cleaning and analysis of missing values in the raw dataset of echocardiography and the identification of important critical features in the raw dataset. The Azure platform is used to predict patients with a history of heart disease (individuals who have been under surveillance in the past three years and those who haven't). Through data scrubbing and preprocessing methods in feature engineering, the model can more accurately predict the future occurrence of heart disease in patients.

Keywords

Precision medicine, feature selection, machine learning, data scrubbing, correlation matrix

Submission date: 9 April 2023; Acceptance date: 28 September 2023

¹Department of Applied Mathematics, Tunghai University, Taichung City

²Cardiovascular Center, Taichung Veterans General Hospital, Taichung City

³Department of PostBaccalaureate Medicine, National Chung Hsing University, Taichung

⁴Department of Life Science, Tunghai University, Taichung City

⁵Department of Information Management, ShuZen junior College of Medicine and Management, Kaohsiung City

⁶Department of Computer Science, Tunghai University, Taichung City

⁷Research Center for Smart Sustainable Circular Economy, Tunghai University, Taichung City

Corresponding authors:

Chao-Tung Yang, Department of Computer Science, Tunghai University, Research Center for Smart Sustainable Circular Economy, Tunghai University, Taichung City 407224.
Email: ctyang@thu.edu.tw

Wei-Wen Lin, Cardiovascular Center, Taichung Veterans General Hospital, Taichung Department of Post-Baccalaureate Medicine, National Chung Hsing University, Department of Life Science, Tunghai University, Taichung City 407224.
Email: weinlinecho@gmail.com



Introduction

This study uses the largest ultrasound dataset of heart disease in Asia for research. This is a huge challenge. Faced with such a huge dataset, it is actually very difficult for data analysts to clarify the internal structure and status of the data situation.

Therefore, at the beginning of the exploration of the dataset, we can only adopt the most conservative strategy, that is, from the most common sampling method, the most in-depth discussion, and then proceed to data cleaning, so we introduce various machine learning algorithms and ensemble learning concept to conduct research on ultrasound datasets.¹⁻⁹

At the beginning of this study, data preprocessing will be conducted. At the same time, it is necessary to select helpful features for input in machine learning algorithms and go on with model training. Generally, feature selection can be considered from two aspects: first, the diversity of features. If the divergence of a feature is not high, for example, the variance is close to zero, meaning that the difference in the sample on this feature is very small, this feature does not contribute much to the differentiation of the sample and can therefore be screened out, and second, it needs to understand the correlation between features and the target. Features that are highly correlated with the target should be selected as input features first because they contribute more to the predictive ability of the model.

In the field of feature engineering, feature selection is a critical step that can help us screen out features that contribute more to model training and prediction, improving the accuracy and stability of the model. In this study, the method of recursive feature elimination (RFE) was used to select important features. This feature selection algorithm¹⁰ is easy to configure and use and can effectively select features that are more correlated with the predicted target variable from the training dataset to improve the performance. Specifically, the method first trains with all features as input then removes the feature with the smallest importance and retrains a new model with the remaining features. This process is repeated until only one feature is left, and the feature importance ranking is recorded. Based on this ranking, the algorithms can obtain the accuracy of each feature subset and find the optimal number of feature selections, and through training, this study can also determine the quality of features, resulting in higher accuracy and stability.

These methods analyze the correlation between each feature and the target, select the most representative and predictive features as input, and thus improve the training effect of machine learning algorithms and models. In practical operation, this study used the *feature_selection* library in scikit-learn to perform feature selection. This package provides multiple methods for feature selection, which can be chosen according to specific situations. Feature selection is a crucial step in machine learning model training, as selecting the best features can help improve model accuracy and

generalization ability. After completing the data cleaning process for the cardiac ultrasound dataset, this study uploaded the accurate dataset to a machine learning model on the Azure platform for prediction. In addition, this paper divided the dataset into three treatment plans, including cardiac catheter ablation, ventricular defibrillator, and drug control, to train the machine learning model. Due to differences in age and

Table 1. Terms for features in the raw dataset.

Features	Terminology
LV (mm)	Left ventricular diastolic diameter
VS (mm)	Ventricular septal diastolic thickness
LVPW (mm)	Left ventricular posterior wall diastolic thickness
LA (mm)	Left atrial diameter
AO (mm)	Aortic diameter
TR_PG_Mean (mmHg)	Tricuspid regurgitation mean pressure gradient
LVEJValue	Left ventricular ejection fraction
RV (mm)	Right ventricular diastolic diameter
LVID (mm)	Left ventricular systolic diameter
PA (mm)	Pulmonary artery diameter
AR	Aortic regurgitation
AS	Aortic stenosis
MR	Mitral regurgitation
MS	Mitral stenosis
TR	Tricuspid regurgitation
PR	Pulmonary insufficiency
AR_Gr (mmHg)	Aortic reflux pressure gradient
MR_Gr (mmHg)	Mitral regurgitation pressure gradient
TR_Gr (mmHg)	Tricuspid regurgitation pressure gradient
PR_Gr (mmHg)	Pulmonary valve regurgitation pressure gradient
Prosthetic	Artificial heart valve

Table 2. Operations on features in the raw dataset.

Features	Approach
id	Delete
groupno	Predict target
index number	Preservation
billing date	Preservation
sex	Preservation
age	Preservation
LV (mm)	Preservation
VS (mm)	Preservation
LVPW (mm)	Preservation
LA (mm)	Preservation
AO (mm)	Preservation
TR_PG_Mean (mmHg)	Preservation
LVEJValue	Preservation
aid	Delete
date of birth	Delete
check date	Delete
RV (mm)	Preservation
LVID (mm)	Preservation
PA (mm)	Preservation
AR	Preservation
AS	Preservation
MR	Preservation
MS	Preservation
TR	Preservation
PR	Preservation
AR_Gr (mmHg)	Preservation
MR_Gr (mmHg)	Preservation

(continued)

Table 2. Continued.

Features	Approach
TR_Gr (mmHg)	Preservation
PR_Gr (mmHg)	Preservation
Prosthetic	Preservation
TR_PG_Max (mmHg)	Delete
TR_PG_Max (mmHg).1	Delete
LVEFLabel	Delete
TRLabel	Delete

physique in the ultrasound patient raw dataset, different prediction results are generated, so this study emphasizes the accuracy of the raw dataset to improve the judgment accuracy of the machine learning model and assist clinicians in improving diagnostic accuracy.¹¹

Materials and methods

This study lasted for more than three years, including the collection of data from heart disease patients and the attempt to use different machine learning algorithms to explore ultrasound datasets. Through experiments with different algorithms, we propose a set of effective strategies for picking out highly correlated features.^{12–15}

It is also mentioned that this study is limited to collecting all the features of the original dataset, which makes our research have its limitations, which also depends on the interpretation of the disease features that cause abnormal cardiac ultrasounds by medical experts. Therefore, at the beginning of the exploration of the original ultrasound dataset, there was a discussion with the cardiologist, that is, the current original ultrasound dataset has a complete collection of features that can cause cardiac ultrasound abnormalities and through different machine learning strategies. The selected key features^{16–21} were also discussed with professional cardiologists. The experts believed that the features selected by the machine learning strategy are highly correlated with heart diseases that cause ultrasound abnormalities.

The current ultrasound dataset is mainly related to the treatment plan. If other datasets are combined arbitrarily, the target to be predicted will diverge and become larger and more chaotic. In addition, our goal is only to focus on the heart disease tracking dataset of Taiwanese people. If additional datasets from other countries are added, the

Table 3. Data type of raw dataset features.

Features	Data type
Gender	int
Age	float 64
LV (mm)	float 64
VS (mm)	float 64
LVPW (mm)	float 64
LA (mm)	float 64
AO (mm)	float 64
TR_PG_Mean (mmHg)	float 64
LVEJValue	float 64
RV (mm)	float 64
LVID (mm)	float 64
PA (mm)	float 64
AR	int 64
AS	int 64
MR	int 64
MS	int 64
TR	int 64
PR	int 64
AR_Gr (mmHg)	int 64
MR_Gr (mmHg)	float 64
TR_Gr (mmHg)	float 64
PR_Gr (mmHg)	float 64
Prosthetic	int 64

Table 4. Numerical encoding.

Gender	Nonnumeric data conversion
Male	1
Female	0

Table 5. Handling missing values.

Features	Null numbers
Gender	0
Age	0
LV (mm)	0
VS (mm)	0
LVPW (mm)	0
LA (mm)	5
AO (mm)	5
TR_PG_Mean (mmHg)	20
LVEJValue	7
RV (mm)	57
LVID (mm)	48
PA (mm)	56
AR	0
AS	0
MR	0
MS	0
TR	0
PR	0
AR_Gr (mmHg)	0
MR_Gr (mmHg)	1
TR_Gr (mmHg)	1
PR_Gr (mmHg)	1
Prosthetic	0

treatment plan will be out of focus and not conducive to professional interpretation by cardiologists. The research materials of this study were mainly provided by the Department of Cardiology, Taichung Veterans General Hospital, and the ethics committee of its review committee had approved this study (IRB number: CE23277C).

Data preprocessing

The echocardiogram dataset used in this study often contains incomplete or defective data, so data scrubbing is a very important step. This section will introduce the process of data organization. First, we collected raw data on ultrasound results for 563 patients. Next, we compared the data of 36,994 ultrasound records with the same ID (patient ID) and obtained a total of 1306 records. Then, this study divided the patients into nine groups according to the definition of diseases and follow-up targets. After excluding the duplicates, this study selected 547 and 1212 diagnostic patient data, respectively. Before performing data cleaning, our research first discussed with the physicians what each feature represents, explored which feature values could be removed first, and explained why some data items have missing values or some features have a higher degree of correlation, in order to fully understand the data situation, which is helpful for subsequent scrubbing processing. In this section, we also provide abbreviations of medical terms for the raw dataset (Table 1).

Nonnumeric data conversion

Machine learning algorithms are a type of technology based on mathematical models that require computation based on numerical data (Table 2). However, the datasets we collect usually include text data, such as gender (male, female) and education level (elementary school, middle school, high school). Machine learning algorithms are a type of technology based on mathematical models that require computation based on numerical data. However, the datasets we collect usually include text data, such as gender (*male, female*) and education level (*elementary school, middle school, high school*), which need to be preprocessed before uploading to the Azure platform's machine learning algorithms for processing. Before preprocessing, our research need to convert text data into numerical data. An instance of

this is that gender can be converted to 0 or 1, where 0 represents male and 1 represents female; education level can be converted to 1, 2, or 3, where 1 represents elementary school, 2 represents middle school, and 3 represents high school. In addition, to ensure data accuracy and completeness, it's crucial to tackle issues such as missing values, duplicates, and outliers during the preprocessing stage. Once the preprocessing is done, the

Table 7. RFE feature importance ranking.

Features	Ranking
Gender	15
Age	3
LV (mm)	2
VS (mm)	8
LVPW (mm)	7
LA (mm)	1
AO (mm)	4
TR_PG_Mean (mmHg)	5
LVEJValue	6
RV (mm)	10
LVID (mm)	9
PA (mm)	11
AR	19
AS	17
MR	20
MS	18
TR	21
PR	22
AR_Gr (mmHg)	13
MR_Gr (mmHg)	14
TR_Gr (mmHg)	12
PR_Gr (mmHg)	16
Prosthetic	23

Table 6. Correlation strength.

Absolute value of correlation strength	Correlation coefficient
1.0 to 0.8	Much strong positive
0.8 to 0.6	Strong positive
0.6 to 0.4	Moderately positive
0.4 to 0.2	Weak positive
0.2 to 0.0	Much weak positive
0	No correlation

study can proceed to upload the numerical data to Azure platform's machine learning algorithms for further analysis (Tables 3 and 4).

Handling missing values

Data scrubbing and handling missing values are important steps in data analysis. These steps help ensure data

accuracy and completeness, thereby improving the effectiveness and credibility of data analysis. Data scrubbing typically involves the following steps: The first step is removing duplicate data. When there is duplicate data in the dataset, it can negatively impact data analysis, so it needs to be removed. The second step is handling missing values. Missing values often occur in datasets and can affect data analysis, so they need to be addressed. Common methods for handling missing values include

Table 8. Ranking of XGBoost algorithm.

Features	Ranking
Gender	2
Age	1
LV (mm)	8
VS (mm)	9
LVPW (mm)	14
LA (mm)	12
AO (mm)	3
TR_PG_Mean (mmHg)	5
LVEJValue	4
RV (mm)	13
LVID (mm)	15
PA (mm)	16
AR	20
AS	17
MR	21
MS	18
TR	22
PR	23
AR_Gr (mmHg)	10
MR_Gr (mmHg)	7
TR_Gr (mmHg)	6
PR_Gr (mmHg)	11
Prosthetic	19

Table 9. Ranking of random forest algorithm.

Features	Ranking
Gender	16
Age	2
LV (mm)	3
VS (mm)	7
LVPW (mm)	8
LA (mm)	1
AO (mm)	4
TR_PG_Mean (mmHg)	5
LVEJValue	6
RV (mm)	10
LVID (mm)	9
PA (mm)	11
AR	21
AS	17
MR	22
MS	18
TR	23
PR	20
AR_Gr (mmHg)	12
MR_Gr (mmHg)	14
TR_Gr (mmHg)	13
PR_Gr (mmHg)	15
Prosthetic	19

deletion, interpolation, and filling. The third step is handling outliers. Outliers refer to extreme or unusual values that may be present in the dataset and can also affect data analysis. Therefore, they need to be handled. Methods for handling outliers include deletion, replacement, and correction. The fourth step is handling inconsistent data, which may include inconsistent units or data formats. These inconsistencies can interfere with data analysis

and need to be addressed. Handling missing values is also an important part of data cleaning. Missing values refer to empty or unfilled values that may be present in the dataset and can also affect data analysis. Therefore, they need to be addressed. Common methods for handling missing values include deletion, interpolation, and filling. Deletion involves removing rows or columns containing missing values from the dataset. This method is

Table 10. Ranking of CatBoost algorithm.

Features	Ranking
Gender	10
Age	1
LV (mm)	6
VS (mm)	8
LVPW (mm)	7
LA (mm)	3
AO (mm)	2
TR_PG_Mean (mmHg)	4
LVEJValue	5
RV (mm)	9
LVID (mm)	11
PA (mm)	17
AR	19
AS	15
MR	20
MS	18
TR	21
PR	22
AR_Gr (mmHg)	13
MR_Gr (mmHg)	14
TR_Gr (mmHg)	12
PR_Gr (mmHg)	16
Prosthetic	23

Table 11. Ranking of LightGBM algorithm.

Features	Ranking
Gender	16
Age	4
LV (mm)	2
VS (mm)	7
LVPW (mm)	8
LA (mm)	1
AO (mm)	3
TR_PG_Mean (mmHg)	5
LVEJValue	6
RV (mm)	10
LVID (mm)	9
PA (mm)	11
AR	19
AS	17
MR	20
MS	18
TR	21
PR	22
AR_Gr (mmHg)	13
MR_Gr (mmHg)	14
TR_Gr (mmHg)	12
PR_Gr (mmHg)	15
Prosthetic	23

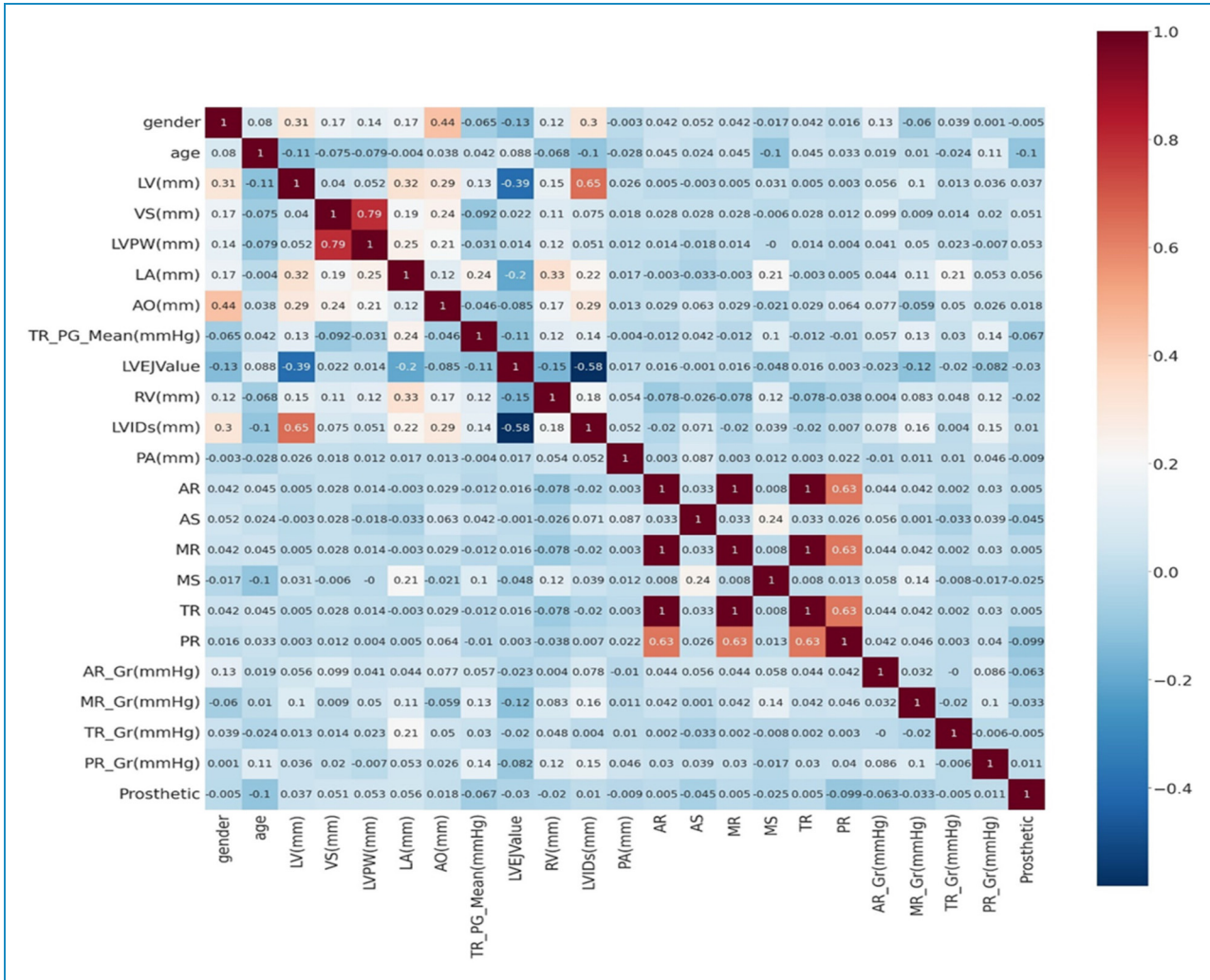


Figure 1. Correlation matrix of the raw dataset. The figure presented depicts a grid displaying the correlation coefficient between two specific features. It is worth noting that the correlation coefficient measures the strength of the linear relationship between the two variables. Therefore, the higher the correlation coefficient, the greater the magnitude of the correlation between the two features. In addition to the correlation coefficient, this study also utilizes a visualization technique to further illustrate the degree of correlation between the two features. This technique involves shading the corresponding grid cells, where the intensity of the shading is directly proportional to the correlation coefficient. In other words, the darker the color of the cell, the higher the degree of correlation between the two features. Overall, this approach provides a comprehensive way to understand the relationship between the features being analyzed. By combining statistical measures such as the correlation coefficient with visual aids, researchers can better interpret the data and draw more accurate conclusions. This method has the potential to be applied in various fields beyond the scope of this study, making it a valuable tool in data analysis and visualization.

relatively simple but can reduce the amount of data, potentially affecting the results of data analysis (Table 5).

Correlation matrix

In this research, understanding the significance of the correlation coefficient is crucial for identifying essential features required for training machine learning algorithms.¹⁷ The correlation coefficient represents the correlation between two groups of features. The larger the absolute value of the correlation coefficient, the higher the

correlation between the two groups, and 0 means that the two groups of features are not related.²²⁻²⁴

The common formula for correlation matrix calculation is as follows:

$$r(x, y) = \frac{COV(x, y)}{S_x S_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

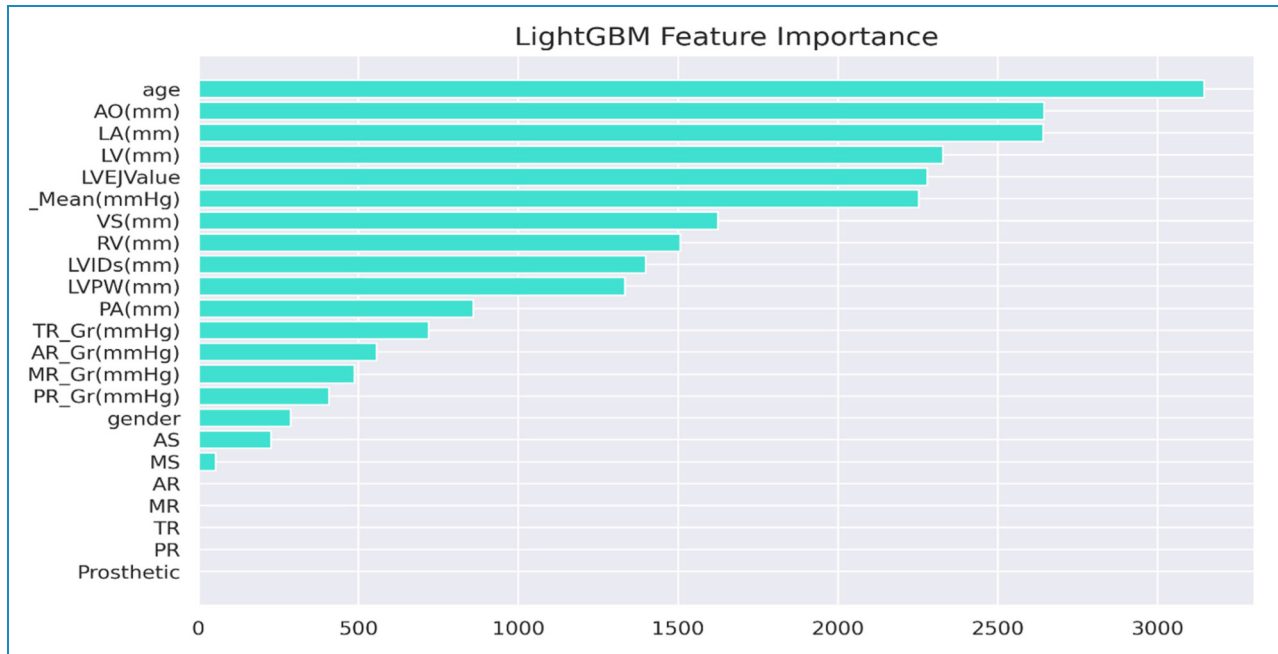


Figure 2. LightGBM feature importance ranking. The figure presented in this study contains a horizontal axis in green, which represents the ranking of each feature being analyzed. It is important to note that the methodology used in this study did not involve selecting specific features, which means that the ranking values depicted on the axis may be relatively high. However, this approach offers a comprehensive way to understand the relationships between the different features and how they impact the results being analyzed.

The matrix mainly discusses the linear relationship between two variables, and its value is between -1 and 1 . According to the correlation coefficient analysis in the formula, it is mainly used to explore the linear relationship between two continuous variables (x, y). If the absolute value of the correlation coefficient between two variables is large, it indicates that the degree of mutual covariation is large. Generally, if there is a positive correlation between two variables, then when x increases, y will also increase. Conversely, if there is a negative correlation between two variables, then when x increase, y will decrease accordingly.

Through the matrix, it will help this research to find that some features are not very relevant to the establishment of the model, and these features can be removed. Visual representation will help this research reduce the amount of features, improve the accuracy of the machine learning model, and at the same time improve the generality of the model and reduce the risk of overfitting. The relationship between the correlation strength and the coefficient is as follows (Table 6).

In this study, it can be observed from the matrix that the correlation coefficient between LVIDS and LV is 0.65 , indicating a strong positive correlation. Moreover, the correlation coefficient between LVIDS and LVEJValue is -0.58 , indicating a moderately positive correlation. Additionally, the correlation coefficient between AR and PR is 0.63 , which is the same as that between MR and PR and TR and PR, indicating a

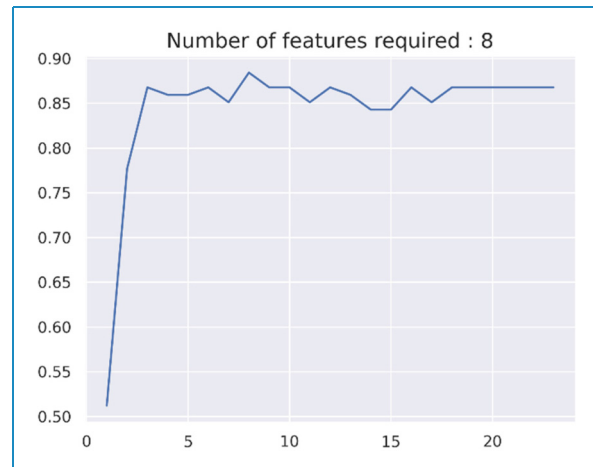


Figure 3. Implementing RFE algorithm. The figure presented in this study is a visual representation of the relationship between the accuracy of an algorithm and the 23 features that were analyzed. The Y-axis denotes the accuracy of the algorithm after executing the critical features, while the X-axis represents the 23 features in their respective order. It is worth noting that the critical features were identified through a rigorous selection process and were deemed to be the most important in determining the accuracy of the algorithm. This approach not only increases the accuracy of the algorithm but also helps to reduce the computational load and improve efficiency.

strong positive correlation as well. However, when selecting feature values, it's important to avoid selecting highly dependent variables, as the new feature may be derived

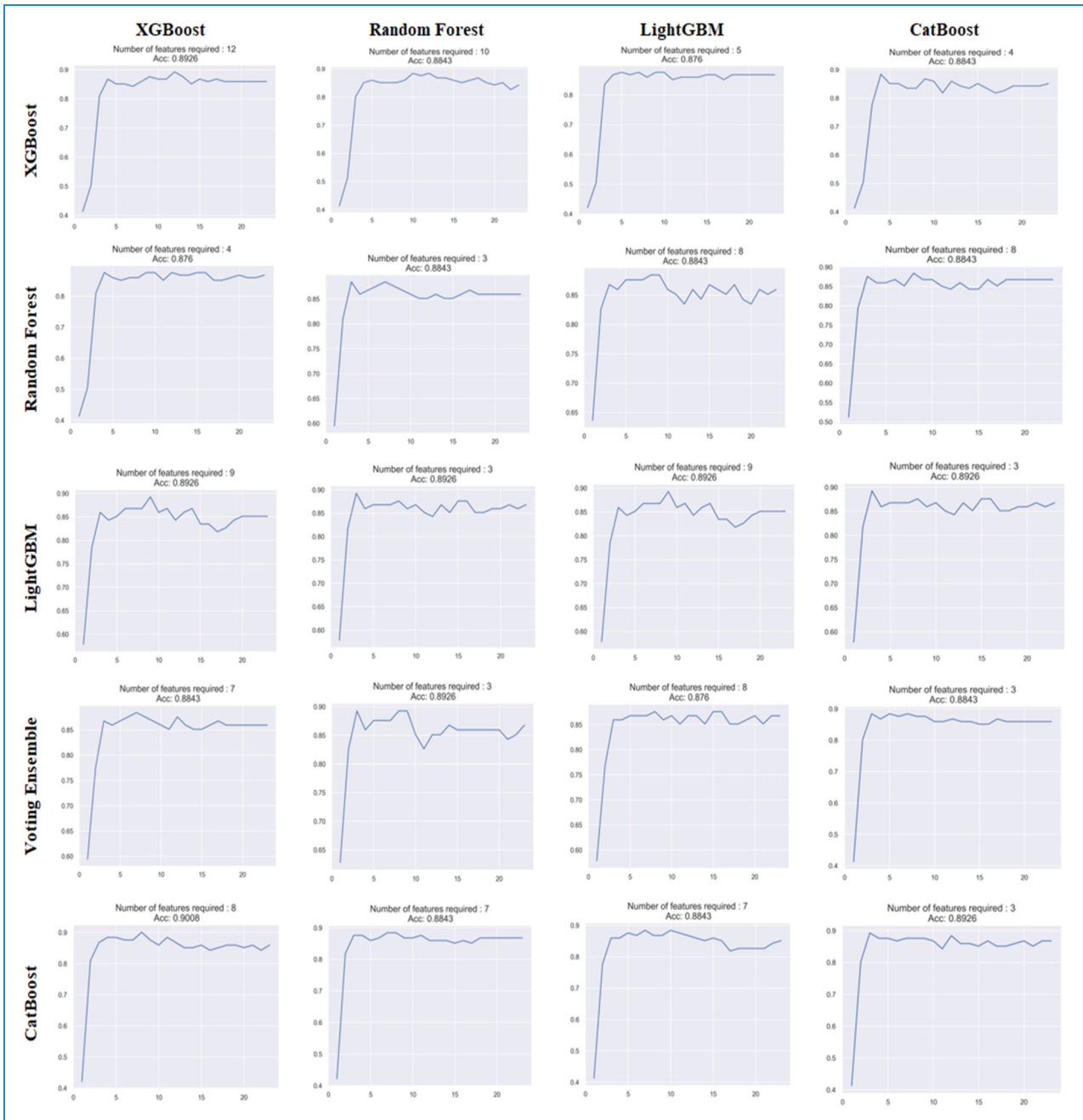


Figure 4. The number of features and the accuracy of each algorithm. The figure presented in this study is a valuable tool for visualizing the results of various algorithms used in the feature selection process. It highlights the features that were selected by each algorithm, as well as the ranking results of each algorithm after execution. By examining the results of multiple algorithms, researchers can gain a more comprehensive understanding of the features that are most impactful for the problem being analyzed. This approach not only increases the accuracy of the analysis but also helps to identify patterns and trends that may have been missed by using a single algorithm.

from existing feature data, resulting in a high correlation between the new and old features. In such cases, the two features may contain identical information. Consequently, when such features are included in the prediction algorithm, they do not contribute significantly to the model's predictive ability.^{25–28}

Critical feature selection in the first stage

This section tests the selected features using five supervised learning models to find the best accuracy and number of features. This study chose to use the original cardiac ultrasound dataset, which has been processed with StandardScaler and contains 1212 data points, because it

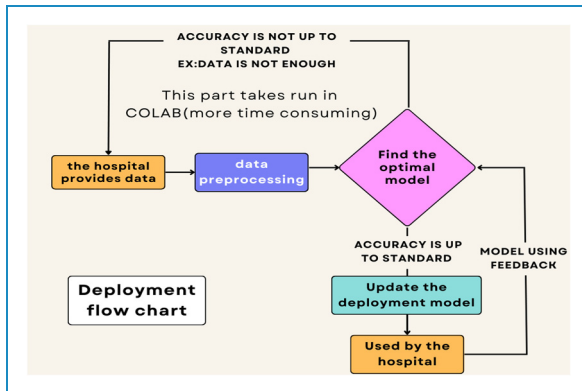


Figure 5. System operation scenario. In this study, the development process of the system is thoroughly explained, highlighting the key stages of the project and the methodologies used to achieve the desired outcome. One of the most notable aspects of the study is the use of Google Colaboratory's cloud virtual platform for calculation comparison. This innovative approach allowed the researchers to compare the computational efficiency and accuracy of different algorithms and make data-driven decisions about which methodologies to deploy on our user interface. The use of cloud-based platforms for data analysis has become increasingly popular in recent years, due to the scalability, flexibility, and cost-effectiveness they offer. By leveraging the power of the cloud, researchers can analyze large datasets more quickly and efficiently and collaborate with colleagues around the world in real time. In addition, cloud-based platforms like Google Colaboratory provide a wide range of tools and resources for data analysis, making it easier for researchers to test and compare different methodologies and identify the most effective approach for their specific research question.

has higher accuracy and more features. Since the Voting Ensemble package does not have the `.coef_` or `.feature_importances_` attributes, it cannot be used for feature selection, so this section is only used for prediction models. In this section,²⁹ we can use the `.feature_importances_` attribute in the scikit-learn package, as well as the RFE package, to know the importance ranking of each feature, but we cannot know the changes in accuracy. Therefore, this study will use the following algorithm to run out the ranking of features in order to find critical factors.³⁰

The reason why different algorithms are emphasized here is that we didn't know anything about such a huge dataset at the beginning, even the structure and state of the data, so the strategy we adopted at the beginning was to clean the data first and how to sample strategy. For the training strategy of machine learning, how to sample from the raw dataset is also an important issue. Therefore, we first adopt the strategy of integrated learning, and integrated learning is mainly divided into two sampling methods:

- (a) Bagging: The method of sampling from the raw data set here is to take and put back. The most common algorithm is Random Forest, and the training strategy is to train multiple classifiers.
- (b) Boosting: After sampling, it is thrown directly into the model training, and a new classifier is generated to evaluate the importance of each sample.

Therefore, in order to understand the performance of the algorithms in different sampling strategies, we explore the cardiac ultrasound datasets for several common algorithms,



Figure 6. Operating system interface. The welcome page of a website or application is often the first point of contact for users, and as such, it is essential that it is well-designed and easy to use. In this study, the welcome page is designed with a click button, which allows the operator to interact with the system in an intuitive and straightforward way.

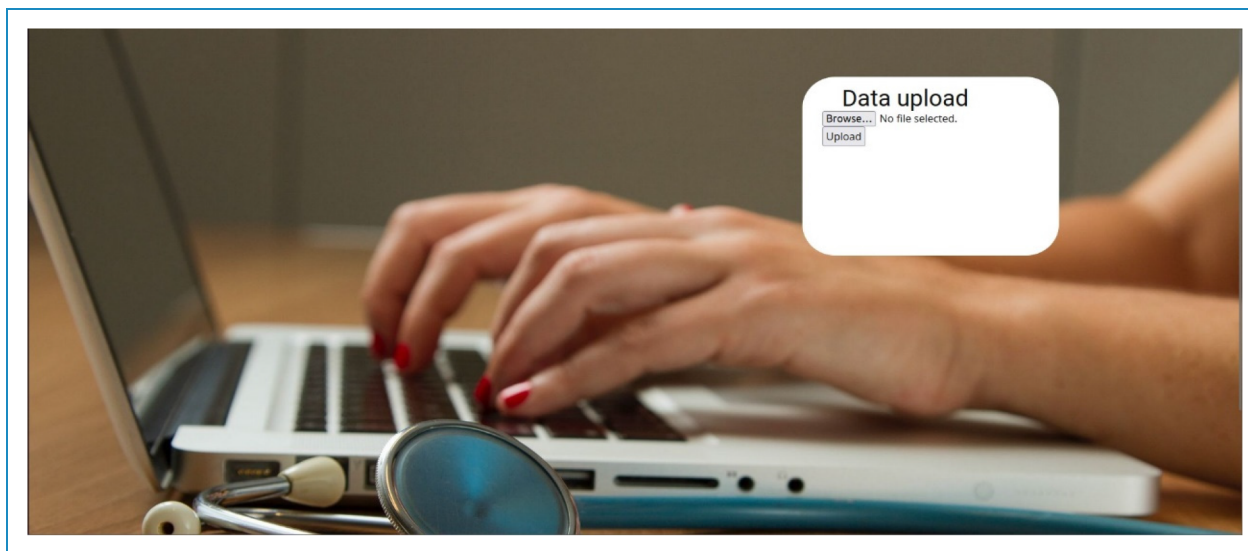


Figure 7. File upload interface.

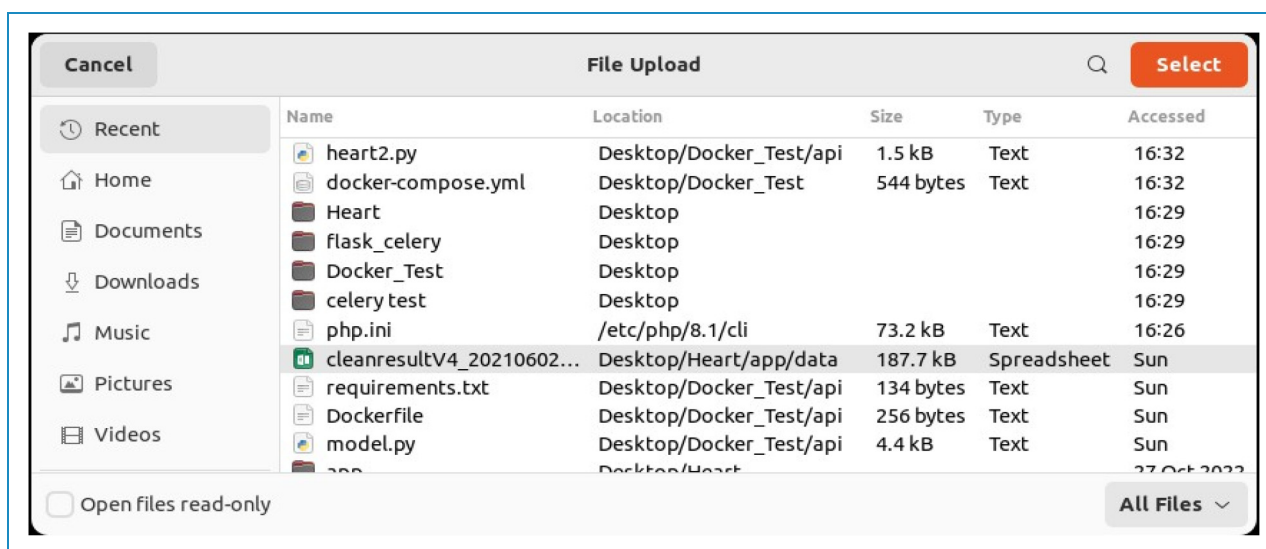


Figure 8. Select log file. When uploading a record file in Excel format, it is essential to ensure that the file does not have any encoding problems. Encoding problems can occur when the file is saved in a format that is not compatible with the system or software being used, resulting in errors or corrupted data. To avoid encoding problems, it is recommended to first check the format of the record file before uploading it. This can be done by opening the file in Excel or a similar program and checking the file properties or format settings. If there are any issues with the file format or encoding, they can be identified and corrected before uploading the file.

and list the rankings so that we can know the important features selected after different sampling methods (Table 7).

Due to the unavailability of relevant packages on the Azure platform, the research team designed a solution to use the RFE package for ranking feature importance and combining the feature names with their rankings to obtain a “list” of important feature rankings. Through these lists, we can easily understand the importance rankings deletion list generated by the four algorithms, as shown in Tables 8–11.

Critical feature selection in the second stage

In the second stage of this study, feature selection^{25,26,29,31,32} is used to help select the most representative features for building models, thereby improving model accuracy and interpretability. However, the number of features is neither necessarily the more, the better, nor is it necessarily the less, the better, as too many features can increase computation and noise, while too few features may overlook important information. Therefore, to address



Figure 9. System prediction result.

this issue, this study uses the RFE algorithm to implement feature selection by iteratively removing the least important features to achieve the effect of feature selection. In addition to using packages, this study also adds a list to record accuracy to display changes in accuracy and find the optimal number of feature selections. During this process, it is important to note that when features are gradually deleted in the loop, repeated cross-validation is needed to ensure that the importance of features is adequately evaluated. Furthermore, it is also important to note that the feature selection algorithm may encounter some problems due to overfitting or underfitting, so adjustment and optimization are required in practical applications (Figures 1–3).

Comparison of different algorithms

Figure 4 shows the impact of feature quantity on accuracy when using XGBoost, Random Forest, LightGBM, and CatBoost as feature selection models and XGBoost, Random Forest, LightGBM, Voting Ensemble, and CatBoost as prediction models in supervised machine learning. Specifically, we used XG Boost, Random Forest, LightGBM, and CatBoost as feature selection models and XG Boost, Random Forest, LightGBM, Voting Ensemble, and CatBoost as prediction models to investigate the effect of feature quantity on accuracy. These results help us understand the impact of feature quantity on machine learning model performance. It can be observed from the figures that different feature selection and prediction models have different requirements for feature quantity. Overall, appropriate feature quantity is crucial for improving machine learning model performance, and both too

many or too few features can have a negative impact. Therefore, in practical applications, we adjust the feature quantity according to the specific situation to obtain the best machine learning model performance.

System development

Our research team has developed a web interface^{34,35} aimed at providing a convenient and user-friendly tool to aid doctors in the diagnosis.^{10,36} The web interface can accept raw ultrasound record files uploaded by doctors and generate prediction results and other relevant information through our model running in the backend. The model combination used in this study includes feature selection by Random Forest or CatBoost and a Voting Ensemble as the prediction model. Through repeated trials, we selected the optimal model combination to ensure the highest prediction accuracy. During the development process, we used various technologies and tools, such as Python, scikit-learn, and Flask Web framework to achieve our development goals, and ultimately deployed the system on the AWS platform (Figures 5–9).

The commendable performance of the system^{37–39} is evident from the experimental results, where it operated efficiently within a CPU processing time of 132 s. The accuracy, a remarkable 0.8926, underscores the system's ability to generate accurate predictions. Other notable accuracy results include a macroscopic accuracy of 0.9468, a microscopic accuracy of 0.8926, and a weighted accuracy of 0.8968. Taken together, these metrics validate the predictive effectiveness and precision of the system.

Doctors who use this web interface^{40–44} can easily diagnose and predict heart disease in patients and

receive useful feedback and information. The web interface displays the optimal number of features, accuracy, and which features are best, enabling doctors to better understand how the model operates and make better diagnoses and predictions.³³ In addition, the web interface also provides other useful features, such as editing and saving prediction results.

Conclusions and future work

This study uses machine learning techniques to analyze ultrasound datasets as a key approach. Notably, this dataset is the largest cardiac data repository in Asia, specifically tailored for individuals of Asian descent, particularly the Taiwanese population. An effective selection strategy was developed by intentionally incorporating these demographic characteristics into the ultrasound data collection. This strategy has the potential to guide individuals in other countries to identify key determinants when processing similar disease data in the coming years. As such, this study represents an important milestone in the advancement of medical interventions for heart disease.

In the next stage of our research, we will use eigenvalues and corresponding labeled symptoms to make predictions, so we will try different types of deep learning models for disease research, because if only the original medical image is labeled, the amount of information is insufficient, and it is not very helpful for classification or prediction results. If we integrate the selected key features and labeled information and then use the deep learning model to train or adjust parameters (e.g., ventricular insufficiency, result in abnormal heart sound), it will effectively improve the classification results or the accuracy of prediction, which will be a great breakthrough in the diagnosis and interpretation of doctors.

Acknowledgements: Our research team would like to express our sincere gratitude to the National Science and Technology Council for providing us with the essential resources and support that enabled us to successfully complete this study and obtain valuable experience and results. We would also like to acknowledge the Taichung Veterans General Hospital for their project 110DHA0500853, which provided us with practical applications and experience in the field of clinical medicine. The support from these entities and individuals played a crucial role in the completion of our research, and we extend our heartfelt appreciation to all who have contributed to this study.

Contributorship: Huang-Nan Huang and Hong-Ming Chen, two professors from the Department of Mathematics, Tunghai University, provided the concept of the model of machine learning algorithms as a strategy for analyzing datasets, while Wei-Wen Lin provided the clinical experience for cardiac interpretation. Lecturer Chau-Jian Huang provided data cleaning strategies and practical operations, Yung-Cheng Chen was responsible for the establishment of models and the implementation of various algorithms and data analysis,

Yu-Huei Wang radiologist provided the interpretation of each feature in the cardiac dataset, and Chao-Tung Yang was responsible for the integration and opinion integration of the medical team and the computer science team.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: The cardiac ultrasound dataset for this study was mainly provided by the Department of Cardiology, Taichung Veterans General Hospital, and the study was approved by the ethics committee (IRB number: CE23277C).

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was made possible with the generous research grants from the National Science and Technology Council. The project numbers are NSTC110-2221-E-029-020- MY3, NSTC1121-2621-M-029-004, and NSTC1121-2622-E-029-003.

ORCID iD: Chao-Tung Yang  <https://orcid.org/0000-0002-9579-4426>

Patient consent: The ultrasound dataset used in this study is the huge data collected by the Department of Cardiology, Taichung Veterans General Hospital, for three years. Before the data were collected, the patients were informed that the data would be used in the study without violating their personal privacy. The purpose of the study is medical research based on public welfare and nonprofit, and the consent of these patients was obtained.

References

1. Liu H, Jiang T and Guo X. Empowering or exploiting? The role of persuasive technology in promoting sustainable consumption behavior. *Social Media + Society* 2023; 9: 1–12.
2. Sotirakopoulos P and Driessen S. The performative politics of social media influencers: a case study of #EndSARS. *Social Media + Society* 2023; 9: 1–12.
3. Zheng Y and Yang H. The effects of social media on political polarization: evidence from a natural experiment in China. *Social Media + Society* 2023; 9: 1–13.
4. Kim H and Lee H. An analysis of consumer behavior on Instagram shopping: the role of social comparison and perceived value. *Social Media + Society* 2022; 8: 1–13.
5. Ju S and Kim J. Social media use and academic performance among college students: a longitudinal study. *Social Media + Society* 2022; 8: 1–11.
6. Chen Y and Liu Y. The effects of social media-based health campaigns on public knowledge and behavior: a meta-analysis. *Social Media + Society* 2022; 8: 1–11.
7. Li X, Yang J and Zhan J. Exploring the factors influencing online health information seeking behavior among college students in China. *Social Media + Society* 2022; 8: 1–11.

8. Zuo M and Zhang M. The influence of different message framing on individuals' donation intention in online crowd-funding. *Social Media + Society* 2022; 8: 1–10.
9. Wu X, Wu H and Xie Y. Understanding the mechanisms behind user churn in social media: a mixed-methods study. *Social Media + Society* 2022; 8: 1–12.
10. Han J, Lu L and Shang L. Effects of positive psychological interventions on posttraumatic stress disorder: a systematic review and meta-analysis. *Digital Health* 2021; 8: 1–11.
11. Li Q, Zhang H, Chen Y, et al. Effects of regular singing intervention on psychological well-being and physical health in Chinese adults: a randomized controlled trial. *Digital Health* 2021; 8: 1–10.
12. Onan A and Korukoğlu S. A feature selection model based on genetic rank aggregation for text sentiment classification. *J Inf Sci* 2016; 43: 25–38.
13. Onan A, Korukoğlu S and Bulut H. A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Inf Process Manag* 2017; 53: 814–833.
14. Onan A. Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency Comput: Practice Exp* 2020; 33: e5909.
15. Onan A. Mining opinions from instructor evaluation reviews: a deep learning approach. *Comput Appl Eng Educ* 2019; 28: 117–138.
16. Onan A. Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach. *Comput Appl Eng Educ* 2020; 29: 572–589.
17. Onan A. An ensemble scheme based on language function analysis and feature engineering for text genre classification. *J Inf Sci* 2016; 44: 28–47.
18. Onan A and Alp Toçoğlu M. A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification. *IEEE Access* 2021; 9: 7701–7722.
19. Onan A. Topic-enriched word embeddings for sarcasm identification. In: *Computer science on-line conference*. Cham: Springer, 2019, pp.293–304. https://doi.org/10.1007/978-3-030-19807-7_29
20. Onan A. Biomedical text categorization based on ensemble pruning and optimized topic modelling. *Comput Math Methods Med* 2018. <https://doi.org/10.1155/2018/2497471>
21. Onan A, Bulut H and Korukoglu S. An improved ant algorithm with LDA-based representation for text document clustering. *J Inf Sci* 2017; 43: 275–292.
22. Lee S, Lee SY and Hong S. Health-related internet use among older adults in South Korea: a systematic review and meta-analysis. *Digital Health* 2021; 8: 1–10.
23. Li Q, Wu Y and Cai Y. Effects of exercise intervention on cognitive function in older adults: a systematic review and meta-analysis. *Digital Health* 2021; 7: 1–16.
24. Liu Y and Wang Y. The impact of online learning on medical students' academic performance: a systematic review and meta-analysis. *Digital Health* 2021; 8: 1–8.
25. Zhang J, Li Y and Lü L. Online social support and psychological well-being during the COVID-19 pandemic: a longitudinal study in China. *Social Media + Society* 2020; 6: 1–12.
26. Onan A. Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification. *J King Saud Univ-Comput Inf Sci* 2022. <https://doi.org/10.1016/j.jksuci.2022.02.025>
27. Onan A. Consensus clustering-based undersampling approach to imbalanced learning. *Sci Program* 2019. <https://doi.org/10.1155/2019/5901087>
28. Onan A, Korukoğlu S and Bulut H. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Syst Appl* 2016; 57: 232–247.
29. Lee SH and Koo C. What factors influence the adoption of electric vehicles? An empirical study of South Korean consumers. *Digit Journalism Social Media* 2021; 2: 259–268.
30. Onan A. Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. *IEEE Access* 2019; 7: 145614–145633.
31. Ransbotham S, Grance T and Koenig M. The cybersecurity workforce gap: how to address it and improve security. *Digit Journalism Social Media* 2021; 2: 339–353.
32. Zhao Y, Zhao Y and Lu Y. A systematic review of virtual reality interventions for dementia care. *Digit Health* 2020; 6: 1–17.
33. Teixeira R and Salavisa I. Building communication in online education: an analysis of asynchronous video messages. *Digit Journalism Social Media* 2021; 2: 175–186.
34. Joo HJ, Cho YJ, Lee SJ, et al. The effectiveness of immersive virtual reality in mental health treatment: a systematic review and meta-analysis. *Digital Health* 2021; 7: 1–15.
35. Jin M and Shin W. Making the best of your time online: the effects of temporal framing on online behavior. *Social Media + Society* 2021; 7: 1–11.
36. Huang W, Xu C and Liu Y. A systematic review of wearable sensor-based systems for fall detection. *Digital Health* 2021; 7: 1–17.
37. Palma AC and Loureiro LM. Antecedents and outcomes of e-health literacy: a systematic review. *Digital Health* 2021; 7: 1–25.
38. Wang H and Zhang X. The relationship between social media use and depression in Chinese adolescents: a cross-sectional study. *Digital Health* 2021; 7: 1–12.
39. Chen L, Chen R and Zhang H. Application and prospect of medical artificial intelligence in China. *Digital Health* 2021; 7: –9.
40. Milne-Ives M, van Velthoven MH, Meinert E, et al. Mobile apps for mental health self-management: a systematic review and meta-analysis. *Digital Health* 2021; 7: 1–17.
41. Park S, Cho M and Kang M. Use of virtual reality technology for pain management: a systematic review. *Digital Health* 2021; 7: 1–15.
42. Zhang H, Chen Y and Li Y. Effects of regular singing intervention on psychological well-being and physical health in Chinese adults: a randomized controlled trial. *Digital Health* 2021; 8: 1–10.
43. Chen J, Zhang Y and Li X. The influence of interpersonal trust on customer loyalty in online marketplaces: a social exchange perspective. *Digit Journalism Social Media* 2021; 2: 293–306.
44. Srinivasan S and Krishnaswamy KN. Predictive modeling for safety and productivity in an iron ore mine using sensor data: a case study. *Digit Journalism Social Media* 2021; 2: 187–196.