frontiers in Physiology

SYSTEMATIC REVIEW
published: 14 January 2022
doi: 10.3389/fphys.2021.811661

Practical Lessons on 12-Lead ECG Classification: Meta-Analysis of Methods From PhysioNet/Computing in Cardiology Challenge 2020

Shenda Hong[1,2*†], *Wenrui Zhang*[3†], *Chenxi Sun*[4,5], *Yuxi Zhou*[6,7] *and Hongyan Li*[4,5*]

[1] National Institute of Health Data Science, Peking University, Beijing, China, [2] Institute of Medical Technology, Peking University Health Science Center, Beijing, China, [3] Department of Mathematics, National University of Singapore, Singapore, Singapore, [4] Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China, [5] School of Electronics Engineering and Computer Science, Peking University, Beijing, China, [6] School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, China, [7] RIIT, TNList, Department of Computer Science and Technology, Tsinghua University, Beijing, China

OPEN ACCESS

Edited by:
Kuanquan Wang,
Harbin Institute of Technology, China

Reviewed by:
Runnan He,
Peng Cheng Laboratory, China
Varun Gupta,
KIET Group of Institutions, India

***Correspondence:**
Shenda Hong
hongshenda@pku.edu.cn
Hongyan Li
leehy@pku.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: 09 November 2021
Accepted: 21 December 2021
Published: 14 January 2022

Citation:
Hong S, Zhang W, Sun C, Zhou Y and Li H (2022) Practical Lessons on 12-Lead ECG Classification: Meta-Analysis of Methods From PhysioNet/Computing in Cardiology Challenge 2020. Front. Physiol. 12:811661. doi: 10.3389/fphys.2021.811661

Cardiovascular diseases (CVDs) are one of the most fatal disease groups worldwide. Electrocardiogram (ECG) is a widely used tool for automatically detecting cardiac abnormalities, thereby helping to control and manage CVDs. To encourage more multidisciplinary researches, PhysioNet/Computing in Cardiology Challenge 2020 (Challenge 2020) provided a public platform involving multi-center databases and automatic evaluations for ECG classification tasks. As a result, 41 teams successfully submitted their solutions and were qualified for rankings. Although Challenge 2020 was a success, there has been no in-depth methodological meta-analysis of these solutions, making it difficult for researchers to benefit from the solutions and results. In this study, we aim to systematically review the 41 solutions in terms of data processing, feature engineering, model architecture, and training strategy. For each perspective, we visualize and statistically analyze the effectiveness of the common techniques, and discuss the methodological advantages and disadvantages. Finally, we summarize five practical lessons based on the aforementioned analysis: (1) Data augmentation should be employed and adapted to specific scenarios; (2) Combining different features can improve performance; (3) A hybrid design of different types of deep neural networks (DNNs) is better than using a single type; (4) The use of end-to-end architectures should depend on the task being solved; (5) Multiple models are better than one. We expect that our meta-analysis will help accelerate the research related to ECG classification based on machine-learning models.

Keywords: electrocardiogram, machine learning, deep learning, classification, practical lessons, physionet challenge, meta-analysis

1. INTRODUCTION

Cardiovascular diseases are one of the leading causes of death worldwide (Virani et al., 2021). Electrocardiogram (ECG) is the most representative and important non-invasive tool for diagnosing cardiac abnormalities (Kligfield, 2002). The effectiveness of using a standard 12-lead ECG for the diagnosis of various cardiac arrhythmias and other diseases has been proven in several

studies (Kligfield et al., 2007). Owing to the predictability of ECG for short-term and long-term mortality risks (Raghunath et al., 2020), accurate and timely detection of cardiac abnormalities based on 12-lead ECG can significantly help save people's lives (Virani et al., 2021). However, manual interpretation of ECG is time-consuming, and different cardiologists may disagree on complicated cases (Hannun et al., 2019; Ribeiro et al., 2019).

In recent years, machine-learning methods have been employed to rapidly detect cardiac abnormalities in 12-lead ECGs (Ye et al., 2010; Jambukia et al., 2015; Minchole et al., 2019; Al-Zaiti et al., 2020). Newly emerging deep-learning models have further achieved comparable performance to clinical cardiologists on many ECG analysis tasks (Hannun et al., 2019; Hong et al., 2020b; Sinnecker, 2020; Elul et al., 2021; Somani et al., 2021), such as cardiovascular management (Fu et al., 2021; Siontis et al., 2021) and arrhythmia/disease detection (Attia et al., 2019; Erdenebayar et al., 2019; He et al., 2019; Hong et al., 2019a,b, 2020a; Zhou et al., 2019; Raghunath et al., 2020; Ribeiro et al., 2020). However, as high-quality real-world ECG data is difficult to acquire, most deep-learning models are designed to detect only a small fraction of cardiac arrhythmias, owing to the limitations of the datasets.

PhysioNet/Computing in Cardiology Challenge 2020 (Challenge 2020) provided high-quality 12-lead ECG data obtained from multiple centers with a large set of cardiac abnormalities (Goldberger et al., 2000; Alday et al., 2020; PHY, 2020; Raghunath et al., 2020). The aim of Challenge 2020 was to identify clinical diagnoses from 12-lead ECG recordings, providing an opportunity to employ various advanced methods to address clinically important questions that are either unsolved or not well-solved (Alday et al., 2020). The datasets for Challenge 2020 were sourced from multiple medical centers worldwide. As shown in **Table 1**, all the datasets contain recordings, diagnostic codes, and demographic data. There are 66,361 ECG recordings, and the number of diagnostic classes is 111. As shown in **Figure 1**, 27 diagnoses are included to evaluate the methods by using an evaluation metric designed by Challenge 2020. This evaluation metric assigns different weights to different classes based on the harmfulness of misdiagnosis in the clinic. The unnormalized challenge score is the summation of the element-wise dots of the confusion matrix and a given reward matrix.

Many well-designed methods were proposed in Challenge 2020. To obtain a comprehensive understanding of how these methods benefit automated ECG interpretation, a more systematic analysis is needed to compare the differences and similarities among them. Thus, in this study, we conduct a meta-analysis of the 41 methods that qualified to be in the final rankings. We analyze the methods in terms of five aspects: data processing, feature engineering, machine-learning models, training strategy, and applications to the real world (see **Figure 2**). Through our meta-analysis, we gather the details of the five aforementioned aspects and conduct the Mann-Whitney *U*-test to verify the effectiveness of the methods. Finally, we discuss the reasons for the effectiveness or ineffectiveness of the methods and summarize five practical lessons that can be applied in real-world scenarios or scholarly research.

Our main practical lessons are the following:

1. Data augmentation should be employed and adapted to specific scenarios.
2. Combining different features can improve performance.
3. A hybrid design of different types of deep neural networks (DNNs) is better than using a single type.
4. The use of end-to-end architectures should depend on the task being solved.
5. Multiple models are better than one.

Abbreviations: Challenge 2020, PhysioNet/Computing in Cardiology Challenge 2020; CinC, Computing in Cardiology; CNN, Convolutional Neural Network; DNN, Deep Neural Network; ECG, Electrocardiogram; RNN, Recurrent Neural Network.

2. METHOD

2.1. Search Strategy and Inclusion Criteria

In Challenge 2020, 70 teams successfully implemented their methods on the platform's test data. We conduct our analysis for the 41 teams that qualified to be on the final rankings of the Computing in Cardiology (CinC) conference[1]. The reasons for the disqualification of the other 29 teams are the following: the method did not work on the hidden set, the team failed to submit a preprint or a final article on time, or the team was absent in CinC.

2.2. Data Extraction

To investigate the techniques applied by each team, we considered five aspects of the methods that formed a solution pipeline (see **Figure 2**): *data preprocessing*, *feature engineering*, *machine-learning models*, *training strategy*, and *applications to the real world*. **Table 2** presents these five aspects.

We confirmed whether a team used a specific technique in their solution by using a three-step reading and checking strategy. First, each reviewer carefully read the full text of 41 papers and extracted data for a single aspect. The data includes whether or how the teams employ techniques involved in the aspect. If a technology is not mentioned in a paper, we assumed that the corresponding team did not use that technology. All the results were gathered together and summarized in a spreadsheet file. Second, each reviewer checked the whole spreadsheet and added comments on what they disagreed with. Finally, all the reviewers discussed the disagreements and corrected the mistakes in the spreadsheet. Thus, we reached the final spreadsheet, and this spreadsheet can be found at https://github.com/hsd1503/cinc2020_meta.

2.3. Analytic Approach

Our analytic approach consists of three main steps. First, for each technique mentioned in **Table 2**, we calculated the usage percentage of the method of the 41 teams. Then, we collected official scores on the test set of each team, and grouped teams based on whether they employed a specific technique. Finally, we statistically analyzed whether these techniques are useful for

[1]https://www.cinc.org/archives/2020/

TABLE 1 | Overview of databases used in Challenge 2020.

Database	Total Patients	Recordings in Training set	Recordings in Validation set	Recordings in Test set	Total Recordings
CPSC (Liu et al., 2018)	9,458	10,330	1,463	1,463	13,256
INCART (Tihonenko et al., 2008)	32	74	0	0	74
PTB (Bousseljot et al., 1995; Wagner et al., 2020)	19,175	22,353	0	0	22,353
G12EC (G12, 2020)	15,742	10,344	5,167	5,167	20,678
Undisclosed	Unknown	0	0	10,000	10,000
Total	Unknown	43,101	6,630	16,630	66,361

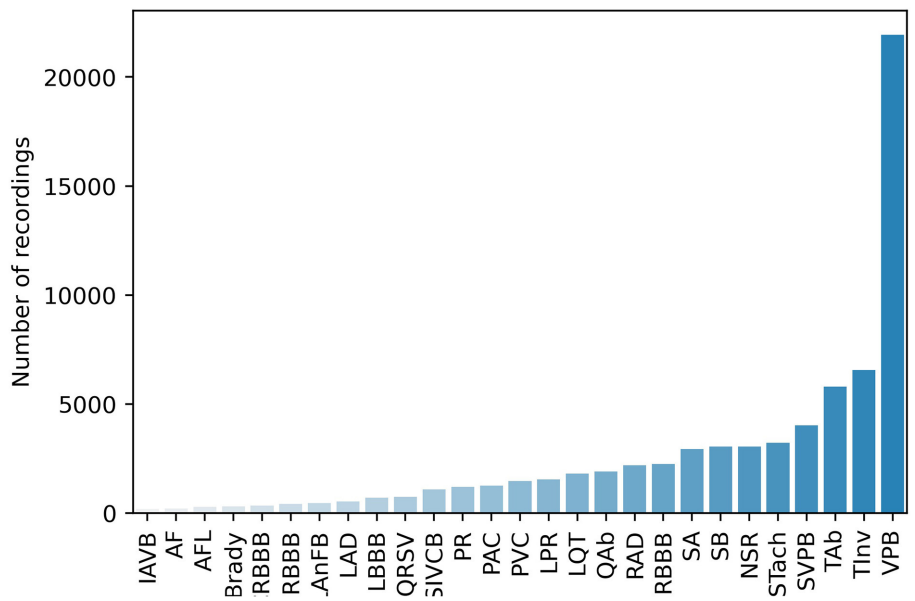

FIGURE 1 | Number of recordings of each scored diagnosis.

ECG classification. The commonly used student t-test requests that the data follows the normal distribution. However, the distribution is unknown. So we adopt the Mann Whitney U-test, a more general and also widely used statistical test method. We conducted the Mann-Whitney U-test (Mann and Whitney, 1947) using SciPy library version 1.6.2[2] and Python version 3.8.8 for each technique. An alternative hypothesis is that the treated technique can improve the performance of the model. We combined two groups, sorted them in ascending order, and assigned ranks for samples (the smallest sample is set as 1, the second smallest sample as 2, and so on). We calculated the sum of the ranks of the two groups referred to as R_1 and R_2. The U-statistics are computed as

$$U_i = R_i - \frac{n_i(n_i+1)}{2}, i = 1, 2 \tag{1}$$

where n_i is the number of samples in the i-th group. Then, we let $U = 3U_1$ because our alternative hypothesis is that the values of group 1 are statistically larger than those of group 2. The Z-statistics are computed as

$$Z = \frac{U - \frac{n_1 \times n_2}{2} - 0.5}{\sqrt{\frac{n_1 \times n_2}{12} \times ((n_1 + n_2 + 1) - \frac{tie}{(n_1+n_2)\times(n_1+n_2-1)})}}, \tag{2}$$

$$tie = \sum_{i=1}^{n_1} count(group_{1i}), \tag{3}$$

where $count(group_{1i})$ represents the number of values in groups 1 and 2, equal to the i-th value in group 1. The p-value is

$$p = P(x > Z), x \sim N(0, 1). \tag{4}$$

To better visualize the results, we drew box plots for each technique. The box figures show groups for the median, upper quartile, lower quartile, outliers, and

[2]https://www.scipy.org

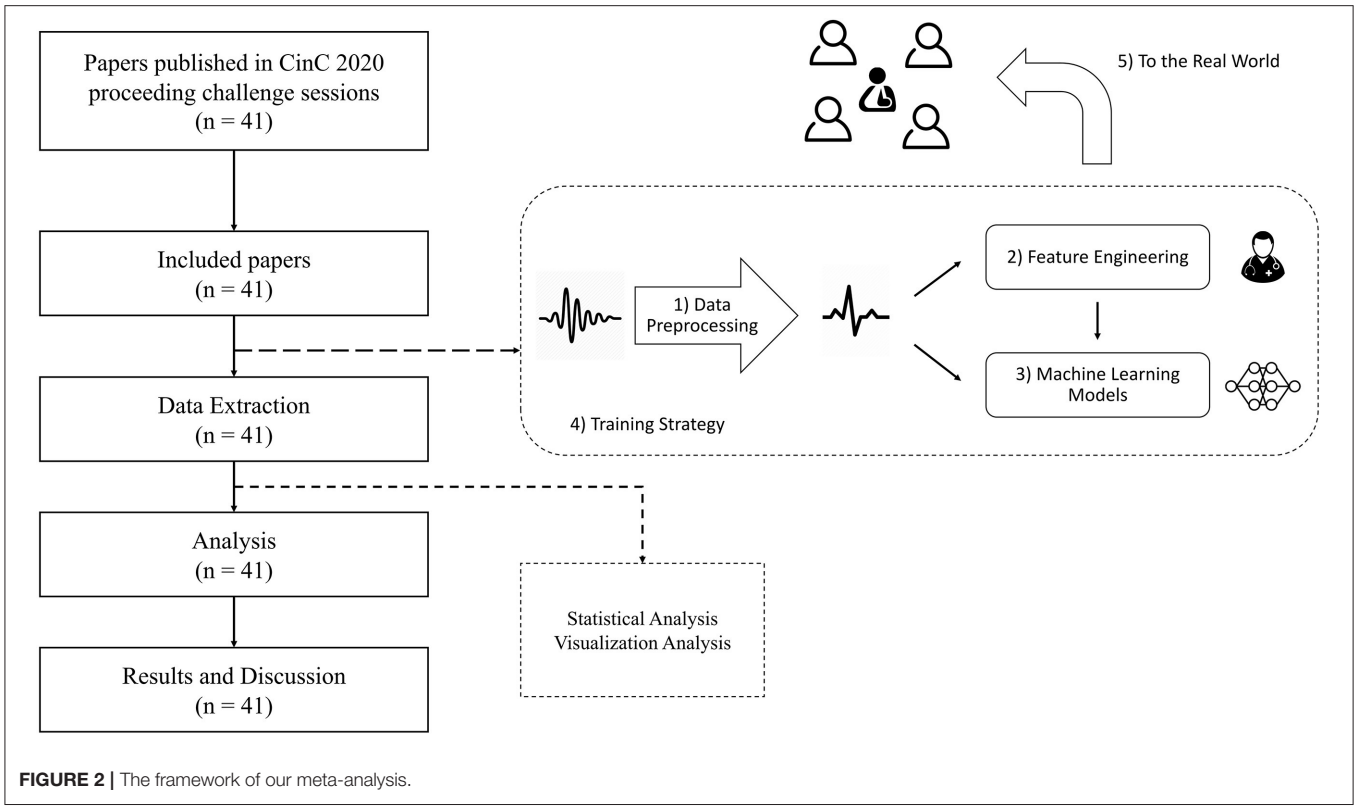

FIGURE 2 | The framework of our meta-analysis.

TABLE 2 | Details of employed techniques.

Aspect	Inclusion	Usage(%)	# in top-10 methods	*p*-value
Data preprocessing	Signal processing	95.12	10	N.A.
	Data augmentation	31.70	6	0.071
	Imbalance handling	53.66	7	0.252
Feature engineering	Hand features	36.59	0	0.983
	Demographic features	29.27	5	0.109
Machine-learning models	Deep neural network	82.93	10	0.116
	Convolutional neural network	82.93	10	0.116
	Recurrent neural network/transformer	31.71	4	0.317
	Attention	24.39	6	0.006
Training strategy	Model ensemble	36.59	4	0.878
	End-to-End	80.49	10	0.139
	Multi-binary classification	58.54	10	0.002
Applications to the real world	Post-processing	2.38	1	N.A.
	Interpretability	4.76	0	N.A.
	Unknown classes and unseen patients	0	0	N.A.

N.A. means that the hypothesis test is not conducted.

range of official scores on the test set. In addition, we discussed and explained why some techniques are beneficial and explore practical lessons from the methods in Challenge 2020.

3. RESULTS

3.1. Overview

The overall meta-analysis results are listed in **Table 2**. We can observe that some techniques are used by the majority of the teams. The results indicate that ECG classification is a complex process that includes multiple techniques. Among these techniques, signal processing, DNNs, convolutional neural networks (CNNs), end-to-end and multi-binary classifications are used by all of the top 10 teams. In addition, we have several significant findings: 1) deep-learning methods were more popular than traditional methods in Challenge 2020; 2) all the teams that employed deep-learning methods used CNNs; and 3) none of the top-10 teams used hand-labeled features (except demographic features); they all adopted end-to-end models instead.

3.2. Data Preprocessing

In this section, we focus on three components of data preprocessing: *signal processing*, *data augmentation*, and *imbalance handling*.

3.2.1. Signal Processing

Signal processing is the most common technique used for ECG classification. We did not attempt to verify the effect of signal processing because different teams set different sampling rates and window sizes and applied various methods. Instead, we summarize and discuss the most common signal processing techniques used in Challenge 2020: resampling, resizing, filtering, and normalization.

Resampling aims to eliminate the differences in the sampling rates among the different input samples. This is necessary because varied sampling-rate inputs degrade the classification models. In the real world, ECG recordings are collected from various medical devices with different sampling rates. Training machine-learning models with this type of data is difficult because their data distributions are inconsistent. This problem can be solved by interpolating the data to a unified sampling rate. Resizing is often realized by cutting signals into a fixed length (known as the window size). Resizing also aims to satisfy another common training request: that the length of the training samples should be the same. Filtering, usually by using band-pass filters, is applied to denoise raw signals. This prevents the model from being disturbed by noise, and this can usually improve performance. Normalization standardizes the signals to a normal distribution or even distribution by transforming signal values in the range of [0, 1] or [-1, 1]. Data distributions can be unified and the influence of noise and outliers can be alleviated through normalization. Other signal processing techniques such as zero-padding (Natarajan et al., 2020), median filters (Hsu et al., 2020), and wavelet transformation denoising (Zhu et al., 2020) can also be used.

3.2.2. Data Augmentation

Data augmentation is an efficient tool for increasing the size and enhancing the quality of the training data. It mainly aims to generate more data covering unseen input spaces (Wen et al., 2020). Data augmentation can make the model more robust by enlarging the size and adding noise or causing transformation. This is also an effective method to avoid overfitting. Common data augmentation methods in Challenge 2020 included the introduction of external data (Bos et al., 2020; Zhu et al., 2020), addition of noise (Chen et al., 2020; Weber et al., 2020), and random cropping (Duan et al., 2020a; Weber et al., 2020). All these methods enlarged the size of the training data. However, when augmentation is performed, the extent of augmentation (such as the stride of the sliding window augmentation) must be considered. Augmenting too much may destroy the distribution of data and cause failure in learning common patterns in data.

In **Figure 3**, we can see that data augmentation is intuitively beneficial in Challenge 2020. All descriptive statistics are larger when data augmentation is performed. The p-value of the Mann-Whitney U-test is 0.07, which is slightly larger than 0.05 without data augmentation. As the sample size is small, we believe that the alternative hypothesis holds.

3.2.3. Imbalance Handling

The training data in Challenge 2020 suffer from heavy class imbalance (as shown in **Figure 1**), which results in predictions being biased toward the majority classes. This is because the training samples of the majority class dominate in the training phase, and they bias the model objectives so that it is easier to obtain higher overall accuracy. In addition, classes with minority sample sizes are more difficult to learn. Even when a classification model is successfully trained, it would very likely become an over-fitted model. Therefore, solving this problem also significantly affects model performance. As shown in **Figure 3**, handling class imbalance can improve the performance of the models.

In Challenge 2020, teams attempted to overcome this problem in two main ways: threshold optimization (Chen et al., 2020; Fayyazifar et al., 2020; Zhao et al., 2020) and weighted loss (Bos et al., 2020; Min et al., 2020). Threshold optimization aims to select the appropriate thresholds corresponding to each class; this has proven to be feasible (Kang et al., 2019). This method is based on models preferring to output a high probability for major classes; thus, setting a low threshold for minor classes can help alleviate this problem. Loss weights are assigned for each class, and the weighted loss forces each class to contribute equally to training the model. In addition, over-sampling (Zisou et al., 2020), down-sampling (Hsu et al., 2020), and other methods have been employed in Challenge 2020.

3.3. Feature Engineering

In this section, we examine how the teams choose or engineer features for model inputs in terms of two aspects: *hand features* and *demographic features*.

3.3.1. Hand Features

In our analysis, we regard hand features as features extracted through non-machine-learning methods, while not simply selecting raw features such as age and sex. Hand features can be further divided into temporal and frequent features.

1. **Temporal Features**: Temporal features are related to the morphological characteristics of ECG waves. The extraction of temporal features consists of two steps: wave detection

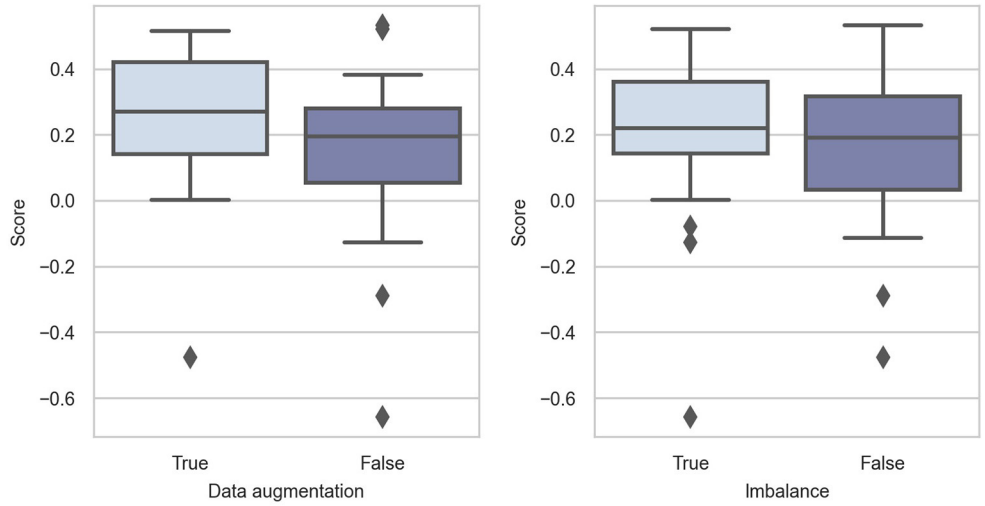

FIGURE 3 | Box-plots of score distributions of data preprocessing techniques.

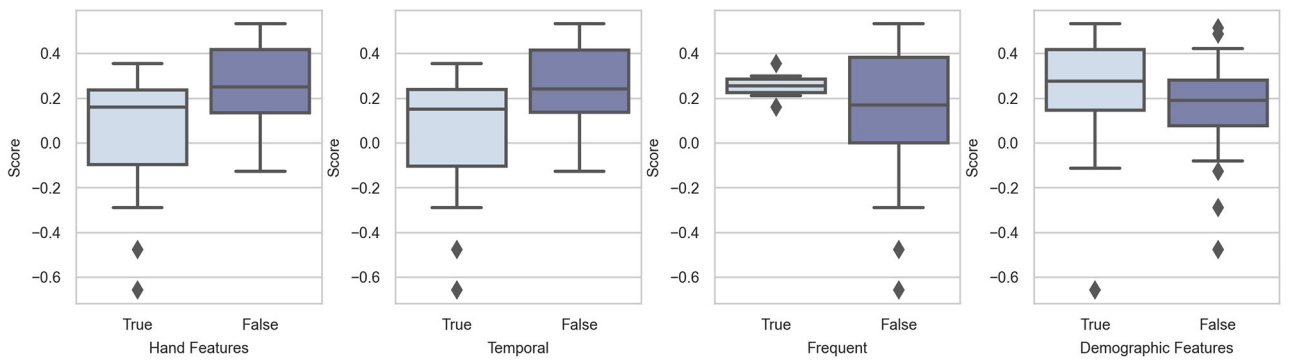

FIGURE 4 | Box-plots of score distributions of feature engineering techniques.

and measurement computing. In Challenge 2020, teams usually employed traditional waves detection methods, such as P-wave, QRS-complex, and T-wave, and then explicitly computed ECG measurements as feature vectors, such as P-wave duration, PR interval, QRS duration, and ST slope. The details of the temporal features can be found in Hong et al. (2019b). These ECG-specific features have proven to be effective for the diagnosis of cardiac diseases.

2. **Frequent Features**: Frequency domain is also an important part of ECG hand features. Thus, some teams extracted features focusing on the frequency spectrum, excluding temporal information. The frequency domain helps to inspect signals from a different view rather than only from the temporal domain. For example, the frequency bands of 0.67–5 Hz, 1–7 Hz, and 10–50 Hz are commonly considered as the dominant components of P-wave, T-wave, and QRS-complex, respectively.

We conducted the Mann-Whitney *U*-test to verify whether adding hand features is beneficial. However, the results were not satisfactory as per our expectations. On the contrary, the results showed that hand features have negative effects (p = 0.983). This may be because of the model architecture. Among the 15 teams that added hand features, 7 abandoned deep-learning methods and adopted traditional machine-learning methods, such as XGBoost (Wong et al., 2020). The model's inferiority may influence the results of the difference between adding and not adding hand features. We also conducted a hypothesis test on temporal features and frequent features, and the resulting p-values were 0.984 and 0.128, respectively. The results of the hypothesis test showed that temporal features are not helpful in improving the performance of the models. The addition of frequent features can yield better prediction results than that of temporal features. **Figure 4** also supports our statistical results.

3.3.2. Demographic Features

Demographic features, such as age and sex, have proven to be useful in ECG classification. Some cardiac diseases occur more frequently in specific patient subgroups. For example, ventricular fibrillation is predominantly observed in the aged (Iwami et al., 2003). Recent studies have shown that the difference between the chronological age and DNN-estimated age can be used as a predictor of mortality (Ladejobi et al., 2021; Lima et al., 2021). Although the features extracted from raw signals may include information related to these hand features, explicitly taking these hand features as input of models can help the model learn more knowledge than raw signals.

The statistical and graphic results proved our hypothesis that demographic features can help models make more accurate predictions. The *p*-value of the hypothesis test was 0.109. This means that adding demographic features is likely to be beneficial. As shown in **Figure 4**, the scores are relatively higher in the group with demographic features. In addition, we observe that the first and second teams and three other top-10 teams input demographic features to their models. We can, therefore, conclude that demographic features are helpful in the context of Challenge 2020.

3.4. Machine-Learning Models

In this section, we focus on the model architectures employed in Challenge 2020. We determined whether the teams used *basic machine-learning methods* or *DNNs*. We classified *DNNs* into three categories: *CNNs*, *recurrent neural networks (RNNs)/transformers*, and *attention mechanisms*.

3.4.1. Basic Machine-Learning Methods

The basic machine-learning methods are all machine-learning techniques excluding DNNs, such as rule-based models and decision tree models. The most notable advantage of these models is that they are relatively easy to use compared with DNNs. Thus, these models can achieve good performance with less data, shorter training times, and lower computation resources. However, most of the time spent on traditional machine-learning methods is to extract features manually, requiring more intervention by specialists. In addition, the ECG data provided in Challenge 2020 were sufficient to support a more complex model (DNN architectures). Conventional methods may fit a large amount of data. In Challenge 2020, several teams adopted basic machine-learning methods, such as XGBoost (Uguz et al., 2020), random forest (Ignacio et al., 2020), and rule-based models (Smisek et al., 2020), whereas others combined these traditional methods with DNNs (Duan et al., 2020b; Zisou et al., 2020).

3.4.2. DNNs

With the development of deep learning, we observe that most teams preferred to use DNN architectures. The prominent advantage of DNNs is that explicit feature extraction by human experts is not necessary, as features are automatically extracted by DNNs based on powerful learning ability and flexible design (Hong et al., 2020b). Related studies have shown that features extracted by DNNs are more informative (having higher importance scores than a random forest classifier) than hand features (Hong et al., 2017). The performance of deep-learning methods is also higher than that of traditional methods on many tasks, such as atrial fibrillation detection from single-lead ECG (Clifford et al., 2017) and sleep staging (Ghassemi et al., 2018). Therefore, the use of appropriate DNN architectures is of great significance.

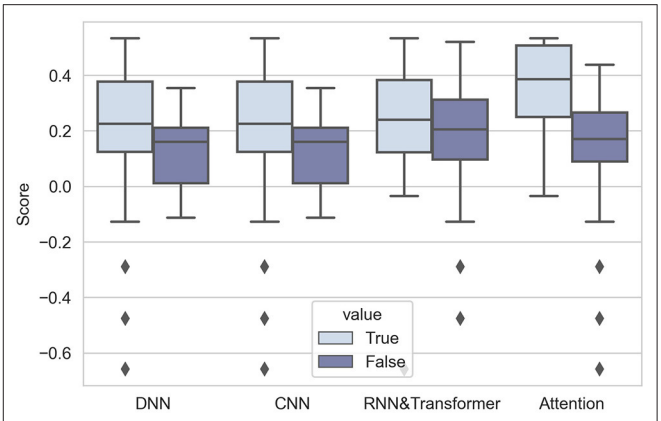

FIGURE 5 | Box-plots of score distributions of machine-learning models.

As expected, the performance of DNNs was comparatively better in Challenge 2020. The 10 highest-ranking teams used DNNs, proving the popularity and effectiveness of DNNs. The first box figure in **Figure 5** shows the performance of the DNN and non-DNN models, with the DNN models exhibiting higher scores.

The analysis of CNNs, RNNs or transformers, and attention mechanisms is presented as follows:

1. **CNNs**: CNN is one of the most popular DNN architecture that has been widely used in computer vision, signal processing, and natural language processing. The essential of the "convolutional" operation is local connectivity between two adjacent neural network layers, which makes it focus on the locality features while also reducing the model parameters (easier training). Such networks can automatically extract hierarchical representations relying on stacked trainable small convolutional filters (kernels). These filters can efficiently extract local representations and can reduce the complexity of models by sharing the same parameters in each layer. It is demonstrated that CNNs can capture more details in 12-lead ECG signals (Baloglu et al., 2019), so CNN is a proper choice as a feature extractor.

   It is notable that all DNNs used in Challenge 2020 include CNNs. Most of them employed a popular CNN architecture named ResNet (Residual Networks) (He et al., 2016a). The core component of ResNet is skip connections, which aims to solve the optimization degradation problem in the back-propagation process (as the network depth increases, accuracy gets saturated and degrades) (He et al., 2016b). In Challenge 2020, the results are in accord with the general point of view–using CNNs can significantly improve the

performance, as shown in the second group of box-plots of **Figure 5**.

2. **RNNs/transformers**: In addition to CNN, RNNs and Transformer (Vaswani et al., 2017) are also widely used DNNs, especially for sequential data, such as time series, event sequences and natural language (Hong et al., 2020b). RNNs take the output from the previous step as input and iteratively update hidden states and memory. Transformer adds attention to sequential modeling and allows sequences to be parallel processed. Different from CNNs, RNNs and transformers mainly focus on temporal dependency rather than local representation. Another advantage is that RNNs and Transformer can handle inputs of various lengths, which is also sometimes necessary for time series data.

   Some teams combine two kinds of architectures in Challenge 2020 (Fayyazifar et al., 2020; Hasani et al., 2020; Natarajan et al., 2020; Oppelt et al., 2020) by applying an RNN or Transformer on representations obtained by CNNs. This is commonly preferred for long ECG signals, because combining two kinds of DNNs can both extract local features and summarize features along the time dimension to obtain global representations. From the third box figure in **Figure 5**, we can see that RNNs and Transformers can help improve the performance of models.
3. **Attention mechanism**: Because of the emergence of the Transformer, attention becomes a widely used mechanism in DNN architectures. The attention mechanism is essentially a kind of weighted sum, and we categorize it into two classes: position-wise attention and channel-wise attention. In detail, we see Transformer as position-wise attention, because Transformer assigns different weights for features extracted from different time points. In addition, we see squeeze-and-excitation block (Hu et al., 2018) as channel-wise attention, because SE block produces weights for each channel of input features. These two kinds of attention mechanisms both have characteristics of plug and play, which means they can easily be combined with DNN models. By applying attention, models can focus on key time steps of long time series (position-wise), or more informative channels (channel-wise attention).

   The results are notable: 4 highest-ranking teams all add the attention mechanism to their models, showing the prevalence of attention. The result of the Mann-Whitney *U*-test also proves that attention can improve the performance of models (p-value is 0.0059, less than 0.01). The effect of using attention is intuitively shown in the fourth box figure in **Figure 5**.

3.5. Training Strategy

In this section, we analyze three aspects of the model: *model ensemble*, *end-to-end*, and *multi-binary classification*.

3.5.1. End-to-End

The end-to-end model takes raw data as input and outputs the target directly, without considering how the features are generated or what they represent. During the process of training end-to-end models, less supervision is required, making it more applicable in the real world. Non-end-to-end models divide the whole task into several sub-tasks, indicating that different sub-tasks may not be consistent and the gap between them may result in non-optimal performance. In addition, when the model is divided into multiple parts, the errors of each part may accumulate and propagate into the next stage. In contrast, end-to-end training can provide more space for models to adjust themselves depending on the input data, making models fit the data better. However, the interpretability of the end-to-end model is always a critical question, especially for medical purposes. Without knowing how the model makes decisions, the results may be unreasonable to be accepted by clinicians and difficult to verify.

In Challenge 2020, the top-10 teams adopted end-to-end models, showing the popularity of such models. As shown in **Figure 6**, these models perform considerably better than non-end-to-end models. The effect and popularity of end-to-end models in Challenge 2020 were related to DNNs, because most DNNs are structured in this manner.

3.5.2. Multi-Binary Classification

A multi-label classification problem can be solved as a multi-class problem directly or a combination of multiple binary classification problems. In detail, multi-binary classification means training a binary classifier for each class to decide whether the sample belongs to this class. This means that the predictive possibility of each class is independent. This is advantageous for training because the multi-class task can be divided into several simple binary classification tasks. However, this neglects the relationship between different diseases, which may have negative effects. In contrast, the output of the multi-class problem is only a vector representing the predictive possibilities, and the sum of these is 1. This is a relatively difficult task compared with the multi-binary classification.

Whether using multi-binary classification significantly influences the performance. The p-value of the Mann-Whitney *U*-test is 0.0018, indicating that using multi-binary classification can significantly improve the performance of the classifiers. Consequently, we believe that the relationship between diseases is not very important. For such a difficult multi-class task, multi-binary classification can reduce the difficulty of training and help achieve better performance.

3.5.3. Model Ensemble

A model ensemble is a learning paradigm that combines multiple learners to improve the overall performance. The commonly used ensemble methods include *bagging* (average predictions or votes for one prediction) and *boosting* (weighted bagging). The core concept of bagging is to average the predictions of several models or make predictions according to the majority vote. Boosting can be regarded as a type of weighted bagging because the classifiers are assigned different weights. Bagging and boosting can improve the performance of the ensemble model by reducing the error caused by the variance and bias, respectively. The motivations behind the ensemble used in Challenge 2020 were mainly to combine models designed for different features or to enhance the ensemble model by combining several models. In addition,

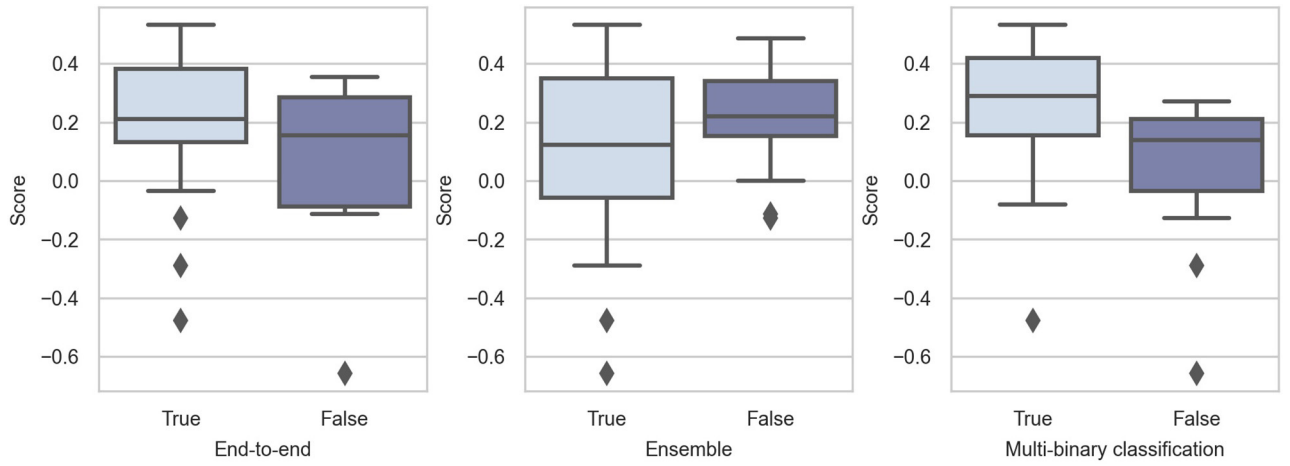

FIGURE 6 | Box-plots of score distributions of training strategy techniques.

bagging was the most commonly used ensemble method in Challenge 2020.

We notice that the three highest-ranking teams used the model ensemble (Natarajan et al., 2020; Zhao et al., 2020; Zhu et al., 2020), but only 14 out of 41 teams employed this strategy. The results are not expected, and we believe that this is because the model ensemble can help improve the single model, whereas it is less meaningful to compare among different teams.

3.6. Applications to the Real World

In this section, we consider some techniques that are necessary in real-world scenario but rare in Challenge 2020, which are *post-processing*, *interpretability*, *unknown classes* and *unseen patients*.

3.6.1. Post-processing

We exclude ensemble methods and threshold optimization in this section because they are mentioned in the previous sections. Except for these two types of post-processing techniques, we find that only one team performed hard sample mining (Chen et al., 2020) as a post-processing technique. Based on the same idea as hard sample mining, some techniques (Orphanidou et al., 2014) can be used to detect and remove low-quality ECG segments (hard samples). The summarized results for high-quality ECG segments are more believable. Finally, the interactions between labels can be considered to post-process the predictions. For example, the reward matrix in Challenge 2020 (see Figure 2 in Alday et al., 2020) indicates that class labels are correlated with each other. In this situation, predicting one label might help predict another correlated label.

3.6.2. Interpretability

The lack of model interpretability is a critical problem for machine-learning models, especially for deep-learning-based models. In Challenge 2020, only two teams (Raipal et al., 2020; Żyliński and Cybulski, 2020) showed feature importance in interpreting the model. The factors that lead to model predictions are unknown for clinicians. Here, we discuss two potential directions for improving the interpretability of the models.

Uncertainty represents how "certain" a model is of each prediction it generates. Although it is difficult to obtain statistical guarantees on performance (which requires true data distribution), estimating the level of uncertainty of predictions is more important than improving accuracy for clinicians (Tonekaboni et al., 2019). A common method for estimating the uncertainty of DNNs with dropout is Monte Carlo dropout (Gal and Ghahramani, 2016). This technique uses dropout during inference and applies the model on the same input multiple times to sample many outputs. In the real world, uncertainty can help clinicians to determine the degree of model reliability, as high uncertainty in ECG classification strongly corresponds with a low diagnostic agreement with the interpretation of the cardiologist (Vranken et al., 2021).

The relative importance is visualized, as clinicians view ECG signals as figures rather than numbers, unlike what deep-learning engineers do. Consequently, highlighting the important component of an ECG segment is vital for the interpretability of the models. The relative importance of different components obtained by models should be evaluated to examine the evidence of the results in a way that cardiologists can understand. Thus, the models can "explain" their predictions, while identifying more details that may be neglected by humans (Elul et al., 2021). The methods for achieving this goal include spectro-temporal attention (Elul et al., 2021) and layer-wise relevance propagation (Binder et al., 2016). These methods emphasize the more important part of the ECG signal on figures to help humans understand what the models care about the most.

In summary, interpretability is necessary for ECG signals in real-world scenarios, and it requires more attention. Researchers can attempt to explain why the models produce their predictions, and this can prompt the real-world application of automated ECG interpretation.

3.6.3. Unknown Classes and Unseen Patients

The classification model is trained on a limited range of datasets, but it is used on an unlimited range of data in the real world. A team in Challenge 2020 considered the differences among databases and employed domain adversarial training (Hasani et al., 2020). However, this team neglected individual-level differences. In a real-world scenario, there are numerous unknown classes and unseen patients. To solve this problem, the models for automated ECG interpretation should have the ability to quickly adapt to unknown classes and unseen patients.

For unknown classes, the model needs to 1) automatically detect whether there is an unknown class and 2) rapidly adapt to the unknown class. To achieve the first goal, we can decouple the multi-class classification task into multiple binary classification tasks and add new classification heads to meet the new rhythm types. If all the existing prediction heads output "False," it indicates that we have met an unknown rhythm type. Other techniques related to the open world (Bendale and Boult, 2015) are also useful. To achieve the second goal, an effective method is to build a separate task-specific simple machine-learning model on top of existing engineered features, such re-training only the final fully connected layer in DNNs, while maintaining the other weights.

A more difficult problem is the gap in the data distributions among different patients. This gap is caused by not only the physiological differences but also other factors such as medical devices and data storage formats (Elul et al., 2021). In this situation, the model trained on existing data might not work equally well on unseen patients. There are many noteworthy attempts to overcome this problem, such as meta-learning to find a set of easily generalized initial parameters (Banluesombatkul et al., 2020) and employing regularization on the loss function (Elul et al., 2021). These methods can benefit the performance on unseen patients and may be helpful on existing patients.

Thus, a good model is not the best on the training set, but the best on the unseen dataset. Thus, how to tackle the "unseen" problems is the key for machine-learning models.

4. DISCUSSION

In this section, we summarize and discuss the five most influential and interesting practical points based on previous results.

4.1. Data Augmentation Should Be Employed and Adapted to Specific Scenarios

It is universally accepted that increasing the amount of training data contributes to the improvement of deep-learning-based models. However, high-quality labeled data are limited in 12-lead ECG classification tasks. Data augmentation by generating synthetic patterns is a model-agnostic solution to this problem.

In addition to cropping, introducing external data, and adding noise, methods based on random transformations, such as flipping, window warping, and masking, are commonly used for the augmentation of time-series data. However, we notice that no method based on the time-frequency domain or frequency domain alone was used for Challenge 2020. In recent years, data augmentation from these two perspectives has drawn considerable attention in many fields (Lee et al., 2019; Park et al., 2019; Gao et al., 2020), including ECG classification tasks. Moreover, handling the severe class imbalance problem in ECG through data augmentation can be a future research direction.

Furthermore, choosing the most appropriate augmentation method remains a challenge. Although the authors in Iwana and Uchida (2021) discussed the advantages and disadvantages of various methods and offered suggestions for using different time-series data types, the effectiveness of various augmentation methods is still based on empirical experiences and experiments.

4.2. Combining Different Features Can Improve Performance

To fully utilize expert knowledge and metadata beyond raw signals, traditional features (not from deep-learning models) are commonly applied for tasks in the medical field (Supratak et al., 2016; Hong et al., 2017, 2019b). In Challenge 2020, teams not only used demographic features such as sex and age but also extracted signal-specific features using traditional methods. In terms of integration, most teams combined traditional-method-based features and deep-learning-based features using simple concatenation. In this way, models can learn extra information from traditional features and retain the generalization ability of deep features (Cheng et al., 2016; Natarajan et al., 2020).

However, most teams neglected a simple technique: hand feature interaction. DeepFM (Guo et al., 2017) provides an accessible solution to this problem, by adding an interaction technique to wide and deep architectures (Cheng et al., 2016). In addition, how to achieve "feature fusion" is a potential direction for better combining the two types of features. The outer product is another universally employed method (Gao et al., 2016; Yu et al., 2017).

4.3. A Hybrid Design of Different Types of DNNs Is Better Than Using a Single Type

First, deep-learning models prevailed in Challenge 2020. As shown in **Table 2**, 82.93% of the teams select DNNs as their models or part of their models. Some teams that use relatively simple models, such as rule-based models (Smisek et al., 2020), achieve good scores. We believe that DNNs exhibit a higher performance only when the model is suitable and the data are well-preprocessed. In addition, combining raw data and domain knowledge is a critical problem in deep-learning-based methods.

Second, the choice of the DNN type is essential. In Challenge 2020, CNN-based models were dominant: all DNNs were CNNs or extracted features from CNNs, indicating that CNNs may be a better choice when latent representations are extracted from raw ECG signals. However, RNNs or transformers can also be applied to discover the temporal dependency of the representations obtained by CNNs. This was a common way to combine CNNs and RNNs in Challenge 2020 (Hasani et al., 2020; Natarajan et al., 2020; Oppelt et al., 2020) and was adopted by three of the top-5 teams.

Finally, we want to emphasize the attention mechanism because of its significant performance improvement. The top-3 teams in Challenge 2020 used the attention mechanism, which can be classified as channel-wise attention (squeeze-and-excitation) and location-wise attention (transformers). The former assigns different weights to each channel (channels can be implicit in each layer of DNNs), and the latter assigns different weights to the representation vectors in each time step. Overall, both types of attention make models learn to recognize more useful information. We can easily add the attention mechanism to the designed models as a plugin block, and this has proven to be beneficial, as described in Section 3.4.2.

When an appropriate model for ECG tasks is constructed, the advantages of different base learners should be combined to design the most powerful model.

4.4. The Use of End-to-End Architectures Should Depend on the Task Being Solved

End-to-end models are becoming increasingly popular owing to the development of DNNs because they do not require significant manual interference, reducing the cost and time consumed in automated ECG interpretation.

However, although end-to-end models are attractive, some limitations exist. First, the performance of each part of the entire model cannot be quantified. Thus, each component of the whole model is designed empirically without any separable and measurable feedback, thereby causing difficulty in the modification of the model. Second, end-to-end models are much slower to be trained compared with decomposition methods because the gradients are noisier and less informative, as demonstrated theoretically and empirically (Shalev-Shwartz et al., 2017). Third, end-to-end models are less flexible because we cannot process the features generated in the middle layers.

Overall, it remains unclear whether to use end-to-end models, depending on the scenario and domain knowledge.

4.5. Multiple Models Are Better Than One

The model ensemble is a model-agnostic and efficient paradigm for improving the performance of a single model. A method from Challenge 2017 showed that the ensemble classifier outperformed single models (Hong et al., 2017). The most common ensemble methods included bagging (bootstrap aggregation), boosting, and stacking.

- Bagging is an ensemble method that trains base learners from different bootstrap samples (subsampling with replacement for the training data). Bagging is more efficient because the base models can be trained in parallel. We regard most ensemble models in Challenge 2020 as being obtained by bagging because they are trained in parallel. However, strictly speaking, they are different from bagging because their sampling methods include not only bootstrap but also other methods.
- Boosting is a family of methods that train models in order, with each base learner relying on the last one. For example, the incorrectly classified samples from the last base learner may be assigned a higher weight in the current training process to emphasize its importance. This is generally employed in decision tree models, such as AdaBoost (Freund and Schapire, 1997) and XGBoost (Chen and Guestrin, 2016).
- Stacking trains first-level learners by using training data and then takes the output from the first-level learners together with the training labels to train a second-level learner. In this way, all first-level learners are combined, and the second-level learner produces the prediction.

There is no conclusion about which ensemble method has the best performance among the three most common ensemble models. However, it is incorrect that more base learners lead to a better performance (Zhou et al., 2002). In other words, composing an ensemble with a part of the base learners instead of the whole set is more appropriate.

5. CONCLUSION

In this study, we collected 41 methods used in Challenge 2020 and conducted a meta-analysis on them, focusing on the aspects of data preprocessing, feature engineering, machine-learning models, training strategy, and applications to the real world. We statistically analyzed and visualized the effectiveness of each technique. We then discussed the advantages and disadvantages of the techniques in terms of the aforementioned aspects. Finally, we summarized five practical lessons based on the analysis, providing practical and instructive experiences in cardiac disease classification tasks based on ECG.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

SH: conceptualization. SH and WZ: methodology and writing-original draft. WZ: visualization. All authors contributed to data extraction, data interpretation, reviewing, editing, and approval of the final version.

FUNDING

This work was supported by the National Natural Science Foundation of China (nos. 62102008 and 62172018) and the National Key Research and Development Program of China under grant no. 2021YFE0205300.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2021.811661/full#supplementary-material

REFERENCES

Żyliński, M., and Cybulski, G. (2020). "Selected features for classification of 12-lead ecgs," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Alday, E. A. P., Gu, A., Shah, A. J., Robichaux, C., Wong, A.-K. I., Liu, C., et al. (2020). Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. *Physiol. Meas.* 41, 124003. doi: 10.1101/2020.08.11.20172601

Al-Zaiti, S., Besomi, L., Bouzid, Z., Faramand, Z., Frisch, S., Martin-Gill, C., et al. (2020). Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. *Nat. Commun.* 11, 1–10. doi: 10.1038/s41467-020-17804-2

Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., Asirvatham, S. J., Deshmukh, A. J., Gersh, B. J., et al. (2019). An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* 394, 861–867. doi: 10.1016/S0140-6736(19)31721-0

Baloglu, U. B., Talo, M., Yildirim, O., Tan, R. S., and Acharya, U. R. (2019). Classification of myocardial infarction with multi-lead ecg signals and deep cnn. *Pattern Recognit. Lett.* 122, 23–30. doi: 10.1016/j.patrec.2019.02.016

Banluesombatkul, N., Ouppaphan, P., Leelaarporn, P., Lakhan, P., Chaitusaney, B., Jaimchariya, N., et al. (2020). Metasleeplearner: a pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning. *IEEE J. Biomed. Health Inform.* 25, 1949–1963. doi: 10.1109/JBHI.2020.3037693

Bendale, A., and Boult, T. (2015). "Towards open world recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA: IEEE), 1893–1902.

Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., and Samek, W. (2016). "Layer-wise relevance propagation for neural networks with local renormalization layers," in *International Conference on Artificial Neural Networks* (Barcelona: Springer), 63–71.

Bos, M. N., van de Leur, R. R., Vranken, J. F., Gupta, D. K., van der Harst, P., Doevendans, P. A., et al. (2020). "Automated comprehensive interpretation of 12-lead electrocardiograms using pre-trained exponentially dilated causal convolutional neural networks," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Bousseljot, R., Kreiseler, D., and Schnabel, A. (1995). *Nutzung der EKG-Signaldatenbank CARDIODAT der PTBuber das Internet*. (New York, NY: Walter de Gruyter).

Chen, J., Chen, T., Xiao, B., Bi, X., Wang, Y., Duan, H., et al. (2020). "Se-ecgnet: multi-scale se-net for multi-lead ecg data," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA), 785–794.

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., et al. (2016). "Wide deep learning for recommender systems," in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, DLRS 2016* (New York, NY: Association for Computing Machinery), 7–10.

Clifford, G. D., Liu, C., Moody, B., Lehman, L. H., Silva, I., Li, Q., et al. (2017). "Af classification from a short single lead ecg recording: the physionet/computing in cardiology challenge 2017," in *2017 Computing in Cardiology (CinC)* (Rennes: IEEE), 1–4.

Duan, R., He, X., and Ouyang, Z. (2020a). "Madnn: a multi-scale attention deep neural network for arrhythmia classification," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Duan, R., He, X., and Ouyang, Z. (2020b). "Madnn: a multi-scale attention deep neural network for arrhythmia classification," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Elul, Y., Rosenberg, A. A., Schuster, A., Bronstein, A. M., and Yaniv, Y. (2021). Meeting the unmet needs of clinicians from ai systems showcased for cardiology with deep-learning-based ecg analysis. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2020620118. doi: 10.1073/pnas.2020620118

Erdenebayar, U., Kim, Y. J., Park, J.-U., Joo, E. Y., and Lee, K.-J. (2019). Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram. *Comput. Methods Programs Biomed.* 180:105001. doi: 10.1016/j.cmpb.2019.105001

Fayyazifar, N., Ahderom, S., Suter, D., Maiorana, A., and Dwivedi, G. (2020). "Impact of neural architecture design on cardiac abnormality classification using 12-lead ECG signals," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Freund, Y., and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55, 119–139. doi: 10.1006/jcss.1997.1504

Fu, Z., Hong, S., Zhang, R., and Du, S. (2021). Artificial-intelligence-enhanced mobile system for cardiovascular health management. *Sensors* 21, 773. doi: 10.3390/s21030773

G12 (2020). *Georgia 12-Lead ECG Challenge Database*. Available online at: https://www.kaggle.com/bjoernjostein/georgia-12lead-ecg-challenge-database/metadata

Gal, Y., and Ghahramani, Z. (2016). "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning* (Amsterdam: PMLR), 1050–1059.

Gao, J., Song, X., Wen, Q., Wang, P., Sun, L., and Xu, H. (2020). Robusttad: Robust time series anomaly detection via decomposition and convolutional neural networks. *arXiv preprint* arXiv:2002.09545.

Gao, Y., Beijbom, O., Zhang, N., and Darrell, T. (2016). "Compact bilinear pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 317–326.

Ghassemi, M. M., Moody, B. E., Lehman, L.-W. H., Song, C., Li, Q., Sun, H., et al. (2018). You snooze, you win: the physionet/computing in cardiology challenge 2018. *Comput. Cardiol.* 45, 1–4. doi: 10.22489/CinC.2018.049

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., et al. (2000). Physiobank, physiotoolkit, and physionet : Components of a new research resource for complex physiologic signals. *Circulation* 101, E215–E220. doi: 10.1161/01.CIR.101.23.e215

Guo, H., Tang, R., Ye, Y., Li, Z., and He, X. (2017). Deepfm: a factorization-machine based neural network for ctr prediction. *arXiv preprint* arXiv:1703.04247. doi: 10.24963/ijcai.2017/239

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., et al. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* 25, 65–69. doi: 10.1038/s41591-018-0268-3

Hasani, H., Bitarafan, A., and Baghshah, M. S. (2020). "Classification of 12-lead ecg signals with adversarial multi-source domain generalization," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 770–778.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). "Identity mappings in deep residual networks," in *European Conference on Computer Vision* (Amsterdam: Springer), 630–645.

He, R., Liu, Y., Wang, K., Zhao, N., Yuan, Y., Li, Q., et al. (2019). Automatic cardiac arrhythmia classification using combination of deep residual network and bidirectional LSTM. *IEEE Access* 7, 102119–102135. doi: 10.1109/ACCESS.2019.2931500

Hong, S., Wu, M., Zhou, Y., Wang, Q., Shang, J., Li, H., et al. (2017). "Encase: an ensemble classifier for ecg classification using expert features and deep neural networks," in *2017 Computing in cardiology (CinC)* (Rennes: IEEE), 1–4.

Hong, S., Xiao, C., Ma, T., Li, H., and Sun, J. (2019a). "MINA: multilevel knowledge-guided attention for modeling electrocardiography signals," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019* (Macao), 5888–5894.

Hong, S., Xu, Y., Khare, A., Priambada, S., Maher, K., Aljiffry, A., et al. (2020a). "Holmes: health online model ensemble serving for deep learning models in intensive care units," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Virtual Event, CA), 1614–1624.

Hong, S., Zhou, Y., Shang, J., Xiao, C., and Sun, J. (2020b). Opportunities and challenges of deep learning methods for electrocardiogram data: a systematic review. *Comput. Biol. Med.* 122:103801. doi: 10.1016/j.compbiomed.2020.103801

Hong, S., Zhou, Y., Wu, M., Shang, J., Wang, Q., Li, H., et al. (2019b). Combining deep neural networks and engineered features for cardiac

arrhythmia detection from ECG recordings. *Physiol. Meas.* 40, 054009. doi: 10.1088/1361-6579/ab15a2

Hsu, P.-Y., Hsu, P.-H., Lee, T.-H., and Liu, H.-L. (2020). "Multi-label arrhythmia classification from 12-lead electrocardiograms," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 7132–7141.

Ignacio, P. S., Bulauan, J.-A., and Manzanares, J. R. (2020). "A topology informed random forest classifier for ecg classification," in *2020 Computing in Cardiology* (Rimini,: IEEE), 1–4.

Iwami, T., Hiraide, A., Nakanishi, N., Hayashi, Y., Nishiuchi, T., Yukioka, H., et al. (2003). Age and sex analyses of out-of-hospital cardiac arrest in osaka, japan. *Resuscitation* 57, 145–152. doi: 10.1016/S0300-9572(03)00035-2

Iwana, B. K., and Uchida, S. (2021). An empirical survey of data augmentation for time series classification with neural networks. *PLoS ONE* 16:e0254841. doi: 10.1371/journal.pone.0254841

Jambukia, S. H., Dabhi, V. K., and Prajapati, H. B. (2015). "Classification of ECG signals using machine learning techniques: A survey," in *2015 International Conference on Advances in Computer Engineering and Applications*, 714–721. doi: 10.1109/ICACEA.2015.7164783

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., et al. (2019). Decoupling representation and classifier for long-tailed recognition. *arXiv preprint* arXiv:1910.09217.

Kligfield, P. (2002). The centennial of the einthoven electrocardiogram. *J. Electrocardiol.* 35, 123–129. doi: 10.1054/jelc.2002.37169

Kligfield, P., Gettes, L. S., Bailey, J. J., Childers, R., Deal, B. J., Hancock, E. W., et al. (2007). Recommendations for the standardization and interpretation of the electrocardiogram: part i: the electrocardiogram and its technology a scientific statement from the american heart association electrocardiography and arrhythmias committee, council on clinical cardiology; the american college of cardiology foundation; and the heart rhythm society endorsed by the international society for computerized electrocardiology. *J. Am. Coll. Cardiol.* 49, 1109–1127. doi: 10.1016/j.jacc.2007.01.024

Ladejobi, A. O., Medina-Inojosa, J. R., Shelly Cohen, M., Attia, Z. I., Scott, C. G., LeBrasseur, N. K., et al. (2021). The 12-lead electrocardiogram as a biomarker of biological age. *Eur. Heart J. Digital Health.* 2, 379–389. doi: 10.1093/ehjdh/ztab043

Lee, T. E. K., Kuah, Y., Leo, K.-H., Sanei, S., Chew, E., and Zhao, L. (2019). "Surrogate rehabilitative time series data for image-based deep learning," in *2019 27th European Signal Processing Conference (EUSIPCO)* (A Coruna: IEEE), 1–5.

Lima, E. M., Ribeiro, A. H., Paix ao, G. M., Ribeiro, M. H., Pinto Filho, M. M., Gomes, P. R., et al. (2021). Deep neural network estimated electrocardiographic-age as a mortality predictor. *Nat. Commun.* 12:5117. doi: 10.1038/s41467-021-25351-7

Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., et al. (2018). An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *J. Med. Imaging Health Inform.* 8, 1368–1373. doi: 10.1166/jmihi.2018.2442

Mann, H. B., and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* 18, 50–60. doi: 10.1214/aoms/1177730491

Min, S., Choi, H.-S., Han, H., Seo, M., Kim, J.-K., Park, J., et al. (2020). "Bag of tricks for electrocardiogram classification with deep neural networks," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Minchole, A., Camps, J., Lyon, A., and Rodriguez, B. (2019). Machine learning in the electrocardiogram. *J. Electrocardiol.* 57, S61–S64.

Natarajan, A., Chang, Y., Mariani, S., Rahman, A., Boverman, G., Vij, S., et al. (2020). "A wide and deep transformer neural network for 12-lead ecg classification," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Oppelt, M. P., Riehl, M., Kemeth, F. P., and Steffan, J. (2020). Combining scatter transform and deep neural networks for multilabel electrocardiogram signal classification. In 2020 Computing in Cardiology, pages 1-4. IEEE. doi: 10.22489/CinC.2020.133

Orphanidou, C., Bonnici, T., Charlton, P., Clifton, D., Vallance, D., and Tarassenko, L. (2014). Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring. *IEEE J. Biomed. Health Inform.* 19, 832–838. doi: 10.1109/JBHI.2014.2338351

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., et al. (2019). Specaugment: a simple data augmentation method for automatic speech recognition. *arXiv preprint* arXiv:1904.08779. doi: 10.21437/Interspeech.2019-2680

PHY (2020). *Physionet/Computing in Cardiology Challenge 2020.* Available online at: https://physionetchallenges.github.io/2020/

Raghunath, S., Cerna, A. E. U., Jing, L., Stough, J., Hartzel, D. N., Leader, J. B., et al. (2020). Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat. Med.* 26, 886–891. doi: 10.1038/s41591-020-0870-z

Raipal, H., Sas, M., Lockwood, C., Joakim, R., Peters, N. S., and Falkenberg, M. (2020). "Interpretable xgboost based classification of 12-lead ecgs applying information theory measures from neuroscience," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Ribeiro, A. H., Ribeiro, M. H., Paix ao, G. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., et al. (2020). Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nat. Commun.* 11, 1–9. doi: 10.1038/s41467-020-15432-4

Ribeiro, A. H., Ribeiro, M. H., Paixo, G., Oliveira, D. M., Gomes, P. R., Canazart, J. A., et al. (2019). Automatic diagnosis of the 12-lead ecg using a deep neural network. 11:1760. doi: 10.1038/s41467-020-16172-1

Shalev-Shwartz, S., Shamir, O., and Shammah, S. (2017). "Failures of gradient-based deep learning," in *International Conference on Machine Learning* (Sydney, NSW: PMLR), 3067–3075.

Sinnecker, D. (2020). A deep neural network trained to interpret results from electrocardiograms: better than physicians? *Lancet Digital Health* 2, e332–e333. doi: 10.1016/S2589-7500(20)30136-9

Siontis, K. C., Noseworthy, P. A., Attia, Z. I., and Friedman, P. A. (2021). Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat. Rev. Cardiol.* 18, 465–478. doi: 10.1038/s41569-020-00503-2

Smisek, R., Nemcova, A., Marsanova, L., Smital, L., Vitek, M., and Kozumplik, J. (2020). "Cardiac pathologies detection and classification in 12-lead ECG," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Somani, S., Russak, A. J., Richter, F., Zhao, S., Vaid, A., Chaudhry, F., et al. (2021). Deep learning and the electrocardiogram: review of the current state-of-the-art. *Europace* 23, 1179–1191. doi: 10.1093/europace/euaa377

Supratak, A., Wu, C., Dong, H., Sun, K., and Guo, Y. (2016). "Survey on feature extraction and applications of biosignals," in *Machine Learning for Health Informatics* (Springer), 161–182.

Tihonenko, V., Khaustov, A., Ivanov, S., Rivin, A., and Yakushenko, E. (2008). *St Petersburg Incart 12-Lead Arrhythmia Database.* PhysioBank, PhysioToolkit, and PhysioNet.

Tonekaboni, S., Joshi, S., McCradden, M. D., and Goldenberg, A. (2019). "What clinicians want: contextualizing explainable machine learning for clinical end use," in *Machine Learning for Healthcare Conference* (Ann Arbor: PMLR), 359–380.

Uguz, D. U., Berief, F., Leonhardt, S., and Antink, C. H. (2020). "Classification of 12-lead ECGS using gradient boosting on features acquired with domain-specific and domain-agnostic methods," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 5998–6008.

Virani, S. S., Alonso, A., Aparicio, H. J., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., et al. (2021). Heart disease and stroke statistics#x2014;2021 update. *Circulation* 143, e254–e743. doi: 10.1161/CIR.0000000000000950

Vranken, J. F., van de Leur, R. R., Gupta, D. K., Juarez Orozco, L. E., Hassink, R. J., van der Harst, P., et al. (2021). Uncertainty estimation for deep learning-based automated analysis of 12-lead electrocardiograms. *Eur. Heart J. Digital Health.* 2, 401–415. doi: 10.1093/ehjdh/ztab045

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., et al. (2020). Ptb-xl, a large publicly available electrocardiography dataset. *Sci. Data* 7, 1–15. doi: 10.1038/s41597-020-0495-6

Weber, L., Gaiduk, M., Scherz, W. D., and Seepold, R. (2020). "Cardiac abnormality detection in 12-lead ECGS with deep convolutional neural networks using data augmentation," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., et al. (2020). Time series data augmentation for deep learning: a survey. *arXiv preprint* arXiv:2002.12478. doi: 10.24963/ijcai.2021/631

Wong, A. W., Sun, W., Kalmady, S. V., Kaul, P., and Hindle, A. (2020). "Multilabel 12-lead electrocardiogram classification using gradient boosting tree ensemble," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Ye, C., Coimbra, M. T., and Kumar, B. V. (2010). Arrhythmia detection and classification using morphological and dynamic features of ecg signals. *Ann. Int. Conf. IEEE Eng. Med. Biol.* 2010, 1918–1921.

Yu, Z., Yu, J., Fan, J., and Tao, D. (2017). "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 1821–1830.

Zhao, Z., Fang, H., Relton, S. D., Yan, R., Liu, Y., Li, Z., et al. (2020). "Adaptive lead weighted resnet trained with different duration signals for classifying 12-lead ECGs," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Zhou, Y., Hong, S., Shang, J., Wu, M., Wang, Q., Li, H., et al. (2019). "K-margin-based residual-convolution-recurrent neural network for atrial fibrillation detection," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (Macao).

Zhou, Z.-H., Wu, J., and Tang, W. (2002). Ensembling neural networks: many could be better than all. *Artif. Intell.* 137, 239–263. doi: 10.1016/S0004-3702(02)00190-X

Zhu, Z., Wang, H., Zhao, T., Guo, Y., Xu, Z., Liu, Z., et al. (2020). "Classification of cardiac abnormalities from ecg signals using se-resnet," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Zisou, C., Sochopoulos, A., and Kitsios, K. (2020). "Convolutional recurrent neural network and lightgbm ensemble model for 12-lead ecg classification," in *2020 Computing in Cardiology* (Rimini: IEEE), 1–4.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.