Database	Records	Leads	Duration	Annotations	Source	Year	Papers
MIT-BIH Arrhythmia Database [1]	47	2	30 minutes	Beat Level, rhythm level. (annotations are keeping updating, please refer to https://archive.physionet.org/physiobank/annotations.shtml)	Boston's Beth Israel Hospital	1975-1979	54
PhysioNet Computing in Cardiology Challenge 2017 [2]	8528 train, 3658 test	1	30 seconds	Rhythm level: normal, AF, other, noise	AliveCor healthcare device	2017	33
PTB Diagnostic ECG Database [3]	549	15	Several minutes	Rhythm level: Myocardial infarction, Cardiomyopathy/Heart failure, Bundle branch block, Dysrhythmia, Myocardial hypertrophy, Valvular heart disease, Myocarditis, Miscellaneous, Healthy controls	National Metrology Institute of Germany	1995	16
MIT-BIH Atrial Fibrillation Database [4]	25	2	10 hours	Rhythm level: AF, atrial flutter (AFL), AV junctional rhythm, and others	Boston's Beth Israel Hospital	1983	8
2018 China Physiological Signal Challenge [5]	6877 train, 2954 test	12	15 seconds	Rhythm level: AF, I-AVB, LBBB, RBBB, PAC, PVC, STD, STE	11 hospitals	2018	7
QT Database [6]	105	2	15 minutes	Onset, peak, and end markers for P, QRS, T, and U waves	Compiled from several existing databases	1997	6
MIT-BIH Normal Sinus Rhythm Database [7]	18	2	24 hours	Beat Level: normal	Boston's Beth Israel Hospital	N.A.	5
St Petersburg INCART 12-lead Arrhythmia Database [8]	75	12	30 minutes	Rhythm level: Acute MI, Transient ischemic attack (angina pectoris), Prior MI, Coronary artery disease with hypertension, Sinus node dysfunction, Supraventricular ectopy, Atrial fibrillation or SVTA, WPW, AV block, Bundle branch block	St. Petersburg Institute of Cardiological Technics	2003	3
MIT-BIH Malignant Ventricular Ectopy Database [9]	22	2	30 minutes	ventricular tachycardia, ventricular flutter, and ventricular fibrillation	Compiled from two separate databases	1986	3
CU Ventricular Tachyarrhythmia Database [10]	35	1	8 minutes	ventricular tachycardia, ventricular flutter, and ventricular fibrillation	Creighton University Cardiac Center	1986	3

Table 2
Summary of databases.

leads (V1 to V4) and no abnormalities will be detected by a single lead. Examples of 15 lead ECG data from medical devices and single lead ECG data from healthcare devices are presented in Figures 5 and 6, respectively.

- Duration. Short-term ECG data (less than several minutes) and long-term ECG data can complement each other. Short-term ECG data is cheaper and easier to collect. Many cardiac diseases can be detected based on short-term ECG data, so such data represent the primary diagnostic tool in outpatient departments. However, long-term ECG can help to detect diseases with intermittent symptoms, such as paroxysmal ventricular fibrillation (VF) and AF.
- Annotations. Annotations include ECG measurement annotations (onset, peak, and end markers for P-, QRS-, T-, and U-waves), beat-level annotations (PAC, PVC, etc.), and rhythm-level annotations (covers both beat-level annotations and other diseases such as AF and VF). Annotation requires huge effort by medical experts.

The following databases were used by many of the selected papers:

- The MIT-BIH Arrhythmia Database [123] (54 papers) consists of 48 half-hour ECG records from 47 subjects at Boston's Beth Israel Hospital (now the Beth Israel Deaconess Medical Center). Each ECG data sequence has an 11 bit resolution over a 10 mV range with a sampling frequency of 360 Hz. This dataset is fully annotated with both beat-level and rhythm-level diagnoses.
- The PhysioNet Computing in Cardiology Challenge 2017 dataset [30] (33 papers) contains 8,528 de-identified ECG recordings with durations ranging from 9 s to just over 60 s that were sampled at 300 Hz by an AliveCor healthcare device. Among these recordings, 5154 are normal, 717 recordings are AF, 2,557 recordings are others, and 46 recordings are noise. Additionally, 3,658 test recordings are private for scoring. This dataset was collected by healthcare devices.
- The PTB Diagnostic ECG Database [15] (16 papers) contains 549 15 channel ECG records from 290 subjects. The sampling rate reaches as high as 10 kHz. Among these subjects, 216 have one of eight types of heart disease and 52 are healthy control, while 22 are unknown.
- The MIT-BIH Atrial Fibrillation Database [122] (8 papers) includes 25 10 h long-term 2 lead ECG recordings with a sampling rate of 250 Hz for human subjects with AF (mostly paroxysmal). The original recordings were collected at Boston's Beth Israel Hospital using ambula-

[1] https://physionet.org/content/mitdb/1.0.0/
[2] https://www.physionet.org/content/challenge-2017/1.0.0/
[3] https://physionet.org/content/ptbdb/1.0.0/
[4] https://physionet.org/content/afdb/1.0.0/
[5] http://2018.icbeb.org/Challenge.html
[6] https://physionet.org/content/qtdb/1.0.0/
[7] https://physionet.org/content/nsrdb/1.0.0/
[8] https://physionet.org/content/incartdb/1.0.0/
[9] https://physionet.org/content/vfdb/1.0.0/
[10] https://physionet.org/content/cudb/1.0.0/

Method	Model	F_{1N}	F_{1A}	F_{1O}	F_{1P}	F_{1NAOP}	F_{1NAO}
[173]	RNN + Expert Features	0.9030	0.8547	0.7366	0.5622	0.7641	0.8314
[57]	CNN	0.91	0.84	0.74	NA	NA	0.83
[141]	CNN + Expert Features	0.9151	0.8247	0.7437	NA	NA	0.8278
[67]	CRNN + Expert Features	0.9117	0.8128	0.7505	0.5671	0.7605	0.8250
[167]	CNN + Expert Features	0.9142	0.8153	0.7370	NA	NA	0.8222
[231]	CRNN	0.9090	0.8221	0.7319	0.5676	0.7577	0.8210
[187]	CRNN	0.9028	0.8221	0.7324	NA	NA	0.8191
[200]	CNN	0.9031	0.8203	0.7310	0.5251	0.7449	0.8181
[137]	CNN + Expert Features	0.9056	0.8284	0.7204	NA	NA	0.8181

Table 3
Comparison of deep learning methods on the PhysioNet Computing in Cardiology Challenge 2017 dataset. We only include methods with an F_1 score over 0.8 reported for the hidden test set.

tory ECG recorders with a 0.1 to 40 Hz recording bandwidth.

- 2018 The China Physiological Signal Challenge dataset [109] (seven papers) contains 6,877 (3178 female, 3699 male) 12 lead ECG recordings with durations ranging from 6 s to just over 60 s, which were collected at 11 hospitals with a sampling rate of 500 Hz. Among these recordings, 918 are normal, 1,098 recordings are AF, 704 are first-degree atrioventricular block, 207 recordings are left-bundle branch block (LBBB), 1,695 are right-bundle branch block (RBBB), 556 are PAC, 672 recordings are PVC, 825 are ST segment depression, and 202 are ST segment elevation. Additionally, 2,954 test recordings are private for scoring.

A summary of the deep learning methods tested on the PhysioNet Computing in Cardiology Challenge 2017 dataset is provided in Table 3. We only include methods with an F_1 score over 0.8 reported for the hidden test set. One can find the official leaderboards for the PhysioNet Computing in Cardiology Challenge 2017 at https://physionet.org/content/challenge-2017/1.0.0/, 2018 China Physiological Signal Challenge at http://2018.icbeb.org/Challenge.html, and PhysioNet Computing in Cardiology Challenge 2020 at https://physionetchallenges.github.io/2020/.

4. Discussion of Opportunities and Challenges

In this section, we will discuss the current challenges and problems related to deep learning based on ECG data. Additionally, potential opportunities are also identified in the context of these challenges and problems.

4.1. Data Collection

As shown in Table 2, there is no standard regarding collection procedures. Different studies have used various numbers of leads, durations, sources (subject backgrounds), etc. This makes it difficult to compare results between different datasets fairly. Additionally, high-quality data and annotations are difficult to acquire, so many current works are still using the MIT-BIH Arrhythmia Database, which was collected over 40 y ago. The most recent single-lead PhysioNet Computing in Cardiology Challenge 2017 and 12 lead 2018 China Physiological Signal Challenge used high-quality data, but they both focused on short-term ECG recordings. Researchers would welcome a new high-quality long-term ECG dataset with annotations and such a dataset would certainly inspire new innovative studies.

4.2. Interpretability

Deep learning models are often considered to be black-box models because they typically have many model parameters or complex model architectures, which makes it difficult for a human to understand why a particular result is generated by such a model. This challenge is much more severe in the medical domain because diagnoses without any explanation are not acceptable for medical experts.

There have been a few works focusing on enhancing the interpretability of ECG deep learning methods. For example, some works have [99, 69] explicitly added interpretable expert features to deep learning models that can be used for partial interpretation. Others have used multi-level attention weights [68] or attribute scores [168] to generate salient maps based on raw ECG data. There have also been several works [163, 69] on generating lower-dimensional embeddings using t-distributed stochastic neighbor embeddings [115] to derive interpretable results.

There are two worthwhile research directions regarding interpretability. The first is how to interpret complex deep learning models using relatively simple models. For example, one can first construct a black-box deep learning model for a specific task, and then construct a separate interpretable simple model that matches the deep learning model's predictions, and then interpret prediction results based on the simple model [149, 150]. The second one is how to construct an interpretable deep model directly. For example, when designing a deep model architecture, one can borrow neuron connection concepts from tree-based models [183] or add attention mechanisms to hidden layers [28, 68] because such mechanisms can be more easily understood by humans.

4.3. Efficiency

Because deep models are complex, it is difficult to deploy large models on portable healthcare devices, which is a major obstacle to applying deep learning models in real-world applications. In this context, one promising research direction is the model compression technique. For example, knowledge distillation is commonly used to transform large and powerful models into simpler models with a minor decrease in accuracy [64]. Additionally, one can use quantization, weight sharing, and careful coding of network weights [56] to compress a large model.

4.4. Integration with Traditional Methods

Most existing deep learning models are trained in an end-to-end manner, making them difficult to integrate with traditional expert-feature-based methods after model training is completed [135].

There are two main research directions for tackling this issue. The first is to use existing expert knowledge to design DNN architectures [70]. For example, the authors of [68] proposed guiding multilevel attention weights using expert