Unlocking Insights: Empirical Exploration of ECG Databases for Enhancing Arrhythmias Detection using Deep learning

Mohamed Ezzeldin A. Bashir[1], Akbar Khanan[1], Yasir Mohamed Abdulgadir[1], Abdul Hakim H. M. Mohamed[1], Senthilkumar Moorthy[1], and Keun Ho Ryu[2]

[1]*Department of Management Information Systems, College of Business Administration, A'Shariqiyah University, Ibra, Oman*

{mohamed.bashir, ,akbar.khanan, yasir.abdulgadir, abdulhakim.mohamed, senthilkumar.moorthy}@asu.edu.om

[2]*Data Science Laboratory, Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City 70000, Vietnam*

khryu@tdtu.edu.vn

Abstract

ECG databases play a crucial role in the development of the field of arrhythmia detection through deep learning techniques. With the increasing availability of diverse ECG datasets, an urgent need arises to comprehensively explore and analyze these databases. This paper highlights the importance of delving into ECG databases to uncover their unique features, challenges, and potential to improve the accuracy of arrhythmia detection. By examining the diverse characteristics of ECG datasets, researchers can identify the advantages and limitations of each database, Which leads to informed decisions when choosing the right datasets for specific research goals when utilizing deep learning techniques.

Keywords: *Electrocardiogram (ECG). Deep learning. and Arrhythmia.*

1. Introduction

The Electrocardiogram ECG signals are a very important medical method that can be employed by experts to extract very valuable information about the health condition of the heart. Therefore, to detect heart arrhythmia, which is the unusual heartbeat shown with a distinct shape in the ECG signal spotted by deflection on the P, QRS, and T waves, then some parameters are acquired, and an enormous finding is produced. Figure 1 illustrates the main ECG waves [1].

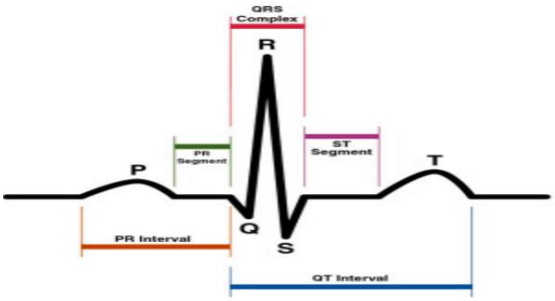

Figure 1 ECG Waves and Labels

In recent years, the discipline of medical diagnostics has undergone a transformative transformation with the integration of machine learning and deep learning technologies in the detection and classification of various cardiac arrhythmias. Accurate diagnosis of arrhythmias is extremely complicated to achieve. Traditional methods of detecting arrhythmias are often based on manual interpretation of ECG data by doctors, which can be time-consuming and subjective. However, the advent of machine learning and deep learning algorithms has revolutionized this process, enabling accurate detection of cardiac arrhythmias.

The role of databases in assessing the accuracy of arrhythmia detection algorithms is crucial. Datasets provide the empirical basis on which algorithms are tested, refined, and ultimately validated. High-quality and diverse datasets ensure rigorous testing, thereby verifying the ability of algorithms to deliver accurate results across a range of scenarios and

patient profiles [2]. In light of these considerations, this paper aims to delve into the myriad of arrhythmia datasets, their unique features, challenges, and implications for machine learning model performance. By doing so, we intend to bridge the gap between the diversity of the dataset and the optimal performance of the model, enabling the research community to harness the full potential of machine learning to detect cardiac arrhythmias. With this exploration, we aspire to lay the foundation for informed data set selection, tailored pretreatment approaches, and ultimately enhance the accuracy of arrhythmia detection.

2. Exploring arrhythmia datasets

In the arrhythmia detection research landscape, the availability of diverse datasets plays a pivotal role in shaping the effectiveness of machine learning models. These datasets originate from various sources, including medical institutions, research databases, and wearable devices. A notable feature of these datasets is their comprehensive representation of diverse patient profiles, covering age groups, genders, and clinical cases. Additionally, the datasets include a range of registration conditions, ranging from clinical settings to real-world scenarios, contributing to the richness and authenticity of the data annotations to these datasets, including classifications that are verified by experts, further enhancing their usefulness for training and validation purposes [3]. In the coming subsections, the main ECG databases exploration is going to be provided.

2.1. MIT-BIH arrhythmia database

In the field of arrhythmias detection with deep learning techniques, MIT-BIH's database is an important data source with widespread use and influence. This database was created by the Massachusetts Institute of Technology and Beth Israel Hospital. MIT-BIH's database is intended to serve as a benchmark for the evaluation of arrhythmia detection algorithms. It includes 48 ECG recordings each captured for 30 seconds at a sampling rate of 360 Hz. Because both normal sinus rhythms and other cardiac arrhythmias are captured, the dataset is ideal for developing and testing machine learning methods. The database contains 25 different types of cardiac arrhythmias, ensuring that cardiac abnormalities with fully representation. [4] This dataset is considered a useful tool for algorithm validation because the annotation process verifies the validity and reliability of the underlying truth labels. It has been used to create and assess several machine learning algorithms, from traditional methods to the most recent deep learning models [5].

2.2. Physio Net challenge database

Electrocardiograms (ECG), photoplethysmograms (PG), and other physiological signals are all included in the Physio Net challenge database. Each challenge is focused on a distinct medical issue, and the related database contains signal recordings with annotations that make it easier to create and test new algorithms. These databases give researchers access to actual data from numerous clinical and observational contexts, allowing them to investigate answers to pressing medical problems. Experts in the relevant medical domains thoroughly explain and validate the signals found in the Physio Net challenge database. These annotations offer precise ground of accurate classes, allowing researchers to assess the effectiveness of their algorithms. The database is a reliable source for algorithm validation and performance measurement thanks to a stringent annotation process that ensures the legitimacy and authenticity of the data. Modern algorithms for several medical applications, such as the identification of cardiac arrhythmias and the diagnosis of sleep apnea, have been made possible by using the Physio Net challenge database [6].

2.3. PTB diagnostic ECG database

The PTB diagnostic ECG database is a broadly used database by researchers in the field of using deep learning to monitor heart health. PTB database was a result of the great efforts of Physikalisch-Technische Bundesanstalt (BTP) in Germany. This database is considered a reliable source for

analyzing the heartbeat recorded in an electrocardiograph device. It is a crucial dataset for the creation and validation of the algorithm as it contains a variety of ECG recordings from both healthy people and people with a range of cardiac disorders. There are 549 records in all, each with an ECG recording of 15 lead samples at 1000 Hz. A variety of heart diseases, such as myocardial infarction, hypertrophy, and arrhythmias, are captured in these recordings. Additionally, each record in the database has annotations that pinpoint the locations of significant ECG waves. Experienced cardiologists provide in-depth explanations of the ECG recordings in the PTB diagnostic ECG database. These annotations offer important details regarding the temporal alignment of ECG waves, enabling researchers to precisely assess and validate their algorithms. Expert annotations guarantee the accuracy and consistency of the fundamental data, enhancing the database's value for algorithm evaluation. [7].

2.4. BTP-XL database

The BTP-XL database is a rich source for researchers in the field of heartbeat analysis and cardiac health monitoring. It provides a wide variety of ECG recordings from different people, including healthy groups and patients with different heart arrhythmias. More than 21,000 ECG recordings of 12 leads, each captured at 1000 Hz and lasting 10 seconds, are available in this database. A variety of cardiac disorders and arrhythmias are covered by these recordings, which are divided into various classes. ECG recordings depicting different cardiac arrhythmias, ischemia episodes, and other heart disorders are included in classes along with normal ECG recordings. Each ECG recording in the PTB-XL database has expert annotations that identify different heart abnormalities and rhythm disturbances. Accurate detection and classification of various heart diseases are made possible by these annotations, which are crucial for testing and training machine learning algorithms. The BTP-XL database's incorporation of thorough clinical and demographic data for each patient is one of its distinguishing qualities. Age, gender, medical history, and other pertinent information are included in this database, which puts ECG recordings in perspective and enables researchers to look for connections between patient traits and heart diseases [8].

2.5. AHA database

The American Heart Association (AHA) has created a database of arrhythmias and regular electrocardiograms (ECGs) that is available on a USB device and is divided into two series of neatly annotated edited recordings. The AHA utilizes the first series of recordings, to create arrhythmia detection procedures, then uses the second series of recordings, to validate the final procedure. The American Heart Association, the Board of Clinical Cardiology, the Committee on Electrocardiography, and Clinical Electrophysiology started working on this database in 1977. The funding came from the National Heart, Lung, and Blood Institute. Real cardiac patient recordings were provided to the study by the relevant institutions. Due to the annotated recordings that allow for the accurate identification of a particular cardiac arrhythmia, the beats and arrhythmias detection professionals refer to this database. This database is considered useful for technical training programs, and in-service training in hospitals, and medical schools besides those who are interested to utilize technology to analyze heartbeats. There are 154 recordings on the database overall, classified into eight kinds of arrhythmia. Each recording lasts for three hours, the last 30 of which were spent grading the rhythms. Timing data beat arrhythmia classification data, and 2- channel ECG data at 250 samples per second and 12 bits of resolution are all provided in this database [9].

2.6. INCART database

The INCART Database is a useful database for cardiovascular research, particularly in the area of electrocardiographic (ECG) analysis, commonly known as INCART PTB (INCART PreTerm Birth). The database contains a collection of ECG recordings made by both healthy people and people with a range of heart arrhythmias. It was used to create and evaluate algorithms for assessing ischemia, finding

arrhythmias, and other cardiac diagnostic procedures. It has numerous recordings of the 12-lead ECG that were sampled at 257 Hz with a resolution of 16 bits. The recordings are divided into many groups, such as those that depict distinct cardiac diseases like arrhythmias and myocardial infarction and normal recordings. For beats analysis and algorithm development, the database intends to provide a wide-ranging and comprehensive collection of ECG signals. Experts explain ECG recordings in the INCART database to provide fundamental accurate classifications for different ECG occurrences and abnormalities. Information about relevant features, such as CSR complexes and St chipsets, is included in the annotations, such annotations help to assess how well ECG signal analysis algorithms perform. The research community has utilized the INCART database extensively to create and validate algorithms for various cardiac diagnostic tasks [10].

2.7. The LUDB (Lund University Database)

LUDB (Lund University Database) is a collection of ECG waves and related annotations accessible to the public and is used for studies on signal processing techniques and arrhythmia identification. This database is provided by Swedish Lund University. More than 8000 ECG recordings, categorization annotations, and other information are included in LUDB, along with high-level diagnostic annotations, reference RR periods, and heart rate movements. According to the 12- recording device, the lead configurations of the ECG recordings in the database range from 1 to 12 leads and represent both the adult and pediatric populations. ECG recordings are annotated with classes that categorize rhythm following the association for the development of medical devices (AMI) standards. The database also includes lengthy recordings, such as one that lasts for a whole month and others that last for a week or more. Numerous arrhythmia detection and classification investigations have utilized LUDB database [11].

3. Related work

Shenda Hong et [12] explore 191 papers that utilized deep learning to detect heartbeats, but a very limited number of datasets are covered. The paper makes a summary of the source of data collection stating that most papers are from medical devices and a few from healthcare devices. Mentioned that the number of leads used to extract the ECG signal is 12 or less. also addressed the registration duration. Finally, the annotations include ECG measurement. The effort in this regard is very limited, as dealing with these databases in general without going into details makes the benefit limited, as well as including specialized databases to monitor a particular arrhythmia in the comparison like MIT-BIH Atrial Fibrillation Database can generate confusion.

E. Merdjanovska and A. Rashkovska [13] put appreciated effort into a review of 45 diverse General databases of ECG using the latest computational methods. The comparison was based on frequency, number of leads, length of recordings, and number of people participating in recording the ECG data. Except for the number of participants in recording ECG rhythms, the comparison's elements are relatively limited and of little use to experts who are creating applications based on deep learning and its algorithms.

Wasimuddin, M. et [14] followed the same way by reviewing many databases, but the number of classes added in the comparison was deemed new which has not been discussed before. Seeing the number of arrhythmias in the analysis and comparison reflects significance for those interested in applying the concepts of deep learning in this field. The paper did not address in detail what those classes are, nor did not address any accurate statistics for those classes, such as the number of repetitions of each class. Also, some efforts have discussed the importance of the number of leads, the duration of registration, how to register, and other things. [15] however, the literature despite its importance that covering this regard, there are some issues extremely necessary to be considered, such as the number and type of categories in each database, as well as the number of records representing the number of people who participated in recording ECG's

records. This, in addition to the number and type of features in each database, is of great importance when conducting learning processes in both supervised and unsupervised [16,17, 18].

6. Arrhythmias database comparison

In this part, we will review the most important databases that have been used to monitor heart health using deep learning principles and tools, by reviewing the most important components. Table 1 illustrates the details of these databases, like the frequency, number of leads, number of records, number of features, and number of arrhythmias.

Table 1. ECG databases details

Database	Frequency	No. Leads	No. Records	No. Features	No. Classes
MIT-BIH arrhythmia	360 Hz	2	48	34	25
Physio Net challenge	1000	12	8,528	1250	12
PTB diagnostic ECG	1000 Hz	15	549	87	9
BTP-XL	1000 Hz	12	21000	549	101
AHA	250	2	154	279	8
INCART	257	2	200	86	13
LUDB	250	12	800	220	5

The frequency is a very important incision for sampling the data recorded which represents pulses through an electrocardiograph, it is usually ranging from 250 to 1000 Hz. The larger this value, the more it is reflected on several things, the most important of which is the increase in detail across the waveform, and the reaction speed increases towards any change in pulses, which is reflected in determining the type of arrhythmia. All these are of great importance when designing any model based on machine learning to monitor heart health, and at the same time, this reflects some disadvantages, such as increasing the details represented by increasing the features that constitute an obstacle to classifying the type of arrhythmia, also creates an additional obstacle, which is the hardware limitation. Dealing with databases with a high sample rate poses a real challenge for researchers to absorb and deal with the huge number of features, and this is noted in Table 1 that whenever the frequency value increases, the features increase, and therefore the need to make an estimated effort during the pre-processing stage to reduce the huge number of features is requested. Also, the number of records in the database is very important, which positively affects many things, including better model accuracy, enhance robustness, better feature learning, and training convergence. However, it is important to note that there are dimension returns as the steady increase of records leads to unnecessary complexity, which negatively affects the performance of the model, especially if the database is huge in terms of the number of features and arrhythmias to be monitored, and this complexity negatively affects the computational process and the memory usage.

7. Conclusion

In the light of comparison between the most famous types of databases employed to monitor heart health according to applications based on machine learning and deep learning, and on the premise that data is the building block for the development of accurate systems that serve various needs in this field, we conclude that there are a huge number of databases with various characteristics related to frequency, number of records, and arrhythmias, and the number of features. All researchers in the field should be aware of employing efforts in the

right direction. More efforts should be expended in the future to study more databases and address several details such as the quality and number of categories to be monitored in each database and study the characteristics of each base in more detail.

10. References

[1] Serhani MA, T El Kassabi H, Ismail H, Nujum Navaz A. ECG monitoring systems: Review, architecture, processes, and key challenges. Sensors. 2020, pp.1796.

[2] Wasimuddin, M., Elleithy, K., Abuzneid, A.S., Faezipour, M. and Abuzaghleh, O., 2020. Stages-based ECG signal analysis from traditional signal processing to machine learning approaches: A survey. IEEE Access, 8, pp.177782-177803.

[3] Zhang Y, et al. A deep learning algorithm for detection of different types of arrhythmias using a wireless stethoscope system. IEEE Journal of Biomedical and Health Informatics, 2020, PP.1083-1090.

[4] Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. IEEE Engineering in Medicine and Biology Magazine, 2001, pp. 45-50.

[5] Goldberger AL, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation, 2000, pp. 215-220.

[6] Moody GB, Mark RG. The impact of the PhysioNet Challenge Database on medical signal processing research. IEEE Engineering in Medicine and Biology Magazine, 2001, pp.45-50.

[7] Struzik ZR, et al. Open access to large scale of ECG data – the PhysioNet/Computers in Cardiology Challenge 2003. Computers in Cardiology, 2003, pp.167-170.

[8] Goldberger AL, et al. PhysioNet: The PTB-XL ECG Data Set. IEEE Transactions on Biomedical Engineering, 2020, pp. 3123-3131.

[9] Strath SJ, Kaminsky LA, Ainsworth BE, Ekelund U, Freedson PS, Gary RA, Richardson CR, Smith DT, Swartz AM. Guide to the assessment of physical activity: clinical and research applications: a scientific statement from the American Heart Association. Circulation. 2013, pp.2259-79.

[10] Taddei A, Distante G, Emdin M, Pisani P, Moody GB, Zeelenberg C, Marchesi C. The European ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography. European Heart Journal, 1992, pp.1164-1172.

[11] Kiranyaz S, Ince T, Gabbouj M. Real-Time PatientSpecific ECG Classification by 1-D Convolutional Neural Networks, Computer, 2017, pp. 39-47.

[12] Hong, S., Zhou, Y., Shang, J., Xiao, C. and Sun, J. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. Computers in biology and medicine, 2020, pp.103801.

[13] Merdjanovska, E. and Rashkovska, A. Comprehensive survey of computational ECG analysis: Databases, methods and applications. Expert Systems with Applications, 2022, pp.117206.

[14] Wasimuddin, M., Elleithy, K., Abuzneid, A.S., Faezipour, M. and Abuzaghleh, O. Stages-based ECG signal analysis from traditional signal processing to machine learning approaches: A survey. IEEE Access, 20520, pp.177782- 177803.

[15] Bashir MEA, Mohamed A.H.H.M., Khanan A., Fattah F.A.M.A., Wang L., Ryu K.H. Cache Learning Method for Terrific Detection of Atrial Fibrillation. Advances in Intelligent Information Hiding and Multimedia Signal Processing 2021, pp 512-519.

[16] Bashir MEA, Ryu KS, Yun U, Ryu KH. Pro-Detection of Atrial Fibrillation using a Mixture of Experts. IEICE Transactions on Information and Systems, 2012, pp.2982–2990.

[17] Bashir MEA, Shon HS, Lee DG, Kim H, Ryu KH. Realtime automated cardiac health monitoring by combination of active learning and adaptive feature selection. KSII Trans Internet Inf Syst. 2013, pp.99–118.

[18] Bashir MEA, Lee DG, Li M, Bae JW, Shon HS, Cho MC, et al. Trigger learning and ecg parameter customization for remote cardiac clinical care information system. IEEE Trans Inf Technol Biomed. 2012, pp.561–71.