Enhancing Generalization in Arrhythmia Classification with Attention Mechanisms

Zhaojing Huang [†,*], Sarisha MacLachlan [†,*], Nhan Duy Truong [†], Omid Kavehei [†]

Abstract

Diagnosing cardiac arrhythmia through interpreting electrocardiogram (ECG) recordings is a challenging and time-consuming task, frequently resulting in inconsistent outcomes and misdiagnosis due to signal noise, interference, and comorbidities. To overcome these challenges, machine learning algorithms have been explored as potential solutions, with promising initial results. However, their lack of generalizability and explainability has hindered their widespread use in clinical settings.

This study focuses on evaluating and reproducing a popular Deep Neuron Network (DNN) model proposed by Ribeiro *et al* [19]. The performance of the model in classifying ECG recordings was found to be influenced by the characteristics of the training dataset, which was composed of different ECG recordings. Although the model exhibited strong generalizability with an F1 score of 0.87 when tested on the CPSC dataset, its performance was inconsistent when applied to the Shaoxing and Ningbo Hospital ECG dataset. To enhance the model's interpretability and performance, an attention layer was incorporated into the network, which improved its focus and resulted in an F_1 score of 0.87 from 0.83 trained on the same dataset.

1 Introduction

Cardiac arrhythmias are typically characterized as any deviation or alteration from an individual's typical heart rate pattern that cannot be justified physiologically. They often serve as an indication of an underlying cardiac ailment, and identifying them early on can help prevent serious clinical conditions like heart failure or stroke [15].

Cardiac arrhythmias are a fairly common medical problem affecting a significant portion of the global population. It is estimated that between 1.5 to 5% of people worldwide have some form of arrhythmia [9]. This means that millions of people are living with arrhythmias that can impact their quality of life and overall health. The incidence of arrhythmias tends to increase with age and is also influenced by other factors such as hypertension, diabetes, and pre-existing heart disease [5] [8]. This prevalence of arrhythmias can put a considerable strain on healthcare systems worldwide due to the resources needed to diagnose and treat these conditions [15]. Moreover, untreated or undiagnosed arrhythmias can lead to severe complications, further emphasizing the importance of early detection and proper management [12].

There are multiple clinical methods used nowadays to diagnose cardiac arrhythmias, which include examining the patient's medical history, physical assessment, and using specialized monitoring equipment. The electrocardiogram (ECG) is a widely utilized tool to measure the heart's electrical activity, and it was introduced by Waller in the early 1900s. It is considered essential in evaluating and diagnosing cardiovascular diseases [2]. This method offers a straightforward, low-cost, and non-intrusive approach to tracking heart signals. The most prevalent version is the standard 12-lead ECG (S12L-SCG), but there are also other types of ECGs that use different numbers of leads for various purposes.

Physicians frequently encounter challenges in interpreting ECG recordings because of the intricacy, unclear causes of arrhythmias, and their clinical associations. Depending solely on visually examining ECG recordings can result in incorrect diagnosis and categorization, which can have life-threatening outcomes [6]. Expertise at a high level is necessary to manually interpret ECGs, as identifying slight variations in amplitude and frequency is not easy to observe visually [4]. Furthermore, there is variation in interpretation among different observers and even for the same observer, and the task is both monotonous and requires a lot of time, even for highly experienced cardiologists [14]. Studies indicate that cardiolo-

[*]The two authors contributed equally to this paper.

[†]Z. Huang, S. MacLachlan, N.D. Truong and O. Kavehei are with the School of Biomedical Engineering, The University of Sydney, NSW 2006, Australia. (e-mail: {zhaojing.huang, smac0444, duy.truong, omid.kavehei}@sydney.edu.au).

gists can accurately recognize ECG abnormalities in the range of 53% to 96%, with as many as 33% of ECG readings containing some level of mistake, and up to 11% of cases leading to incorrect treatment [6]. As a result, the progress of digitalization in the healthcare sector has led to the introduction of computer-assisted ECG screening, which can aid medical professionals in identifying and diagnosing cardiovascular illnesses [6].

1.1 Background

The computer-aided diagnosis of cardiac conditions (CACD) can be a valuable resource for healthcare providers to obtain a second opinion and reduce diagnostic errors related to cardiac diagnose [10]. Typically, computer-aided cardiac diagnosis (CACD) systems involve four primary stages: preprocessing of the ECG signal, detecting heartbeats, extracting and selecting features, and constructing a classifier [1]. Despite advancements in computerized ECG interpretation, current systems still have a notable misdiagnosis rate, with an accuracy rate of 69.7% compared to 76.3% for cardiologists [6]. Traditional machine learning methods have made headway in tackling this issue, but they necessitate substantial expertise and preprocessing of the ECG signal for feature extraction that is done manually [17]. Additionally, the considerable variation in wave morphology among patients and the existence of noise make achieving a high degree of accuracy challenging [17]. The development of deep neural networks (DNNs) through progress in machine learning and artificial intelligence has resulted in notable achievements in image and speech recognition, and also holds promise for various healthcare and clinical applications.

Convolutional Neural Networks (CNNs), which are a type of deep learning algorithm, have effectively tackled complex image analysis tasks in various fields, including medical and non-medical applications [17]. These deep learning techniques entail supplying raw or minimally processed data into a network made up of numerous assembled mathematical equations or "hidden layers," which enables the network to automatically extract, choose, and categorize features without the need for manual feature extraction [17]. Although DNNs have demonstrated good performance in classifying arrhythmias, they are frequently confined to training on small, publically available datasets, need lengthy computation periods to tune the classifiers, and have not yet been verified in a clinical environment. By improving ECG interpretation, lowering misinterpretation, and minimising clinical treatment errors, a reliable, accurate, generalizable DNN with shown clinical performance has the potential to transform current practice in supporting clinicians in cardiac arrhythmia diagnosis.

Recent research has aimed to develop precise deep neural networks capable of accurately classifying cardiac arrhythmias from ECG tracings. However, the validation of the generalization of these models has been given insufficient attention in some publications. Ribeiro *et al* [19] constructed a DNN model that was trained on the most extensive 12 lead ECG dataset to date and achieved an impressive F1 score of over 80% and a specificity of more than 99% [19]. Nevertheless, this model has not been verified on other datasets to confirm its generalization ability. The Hybrid Deep CNN Model proposed by Ullah [21] and the Deep-Learning-Based Framework introduced by Jamil [11] have both attained an impressive F1 score of 0.99 and above. However, neither of these models has been cross-validated on other datasets, and their ability to generalize to new data has not been evaluated.

The objective of this study is to enhance the generalization performance of Ribeiro *et al* [19]'s model by conducting cross-dataset validation.

2 Prerequisite

2.1 DNN architecture

This study builds upon the model proposed by Ribeiro *et al* [19]. Specifically, we utilize their unidimensional residual neural network architecture which includes a 1-D convolutional layer, followed by batch normalization and a Relu activation function. The network then passes through four residual blocks before exiting with a dense layer, as depicted in Figure 1.

2.2 Attention Layer

Bahdanau *et al* [3] proposed an attention mechanism to address the bottleneck problem of fixed-length encoding vectors in sequence-to-sequence models. This problem is especially critical for longer or more complex sequences where the decoder's access to input information is limited. The attention mechanism enables the decoder to selectively focus on relevant input parts, improving its access to information at each time step. The attention mechanism can be broken down into three main computational steps:

Alignment scores are computed by the attention mechanism's alignment model, which takes encoded hidden states and the previous decoder

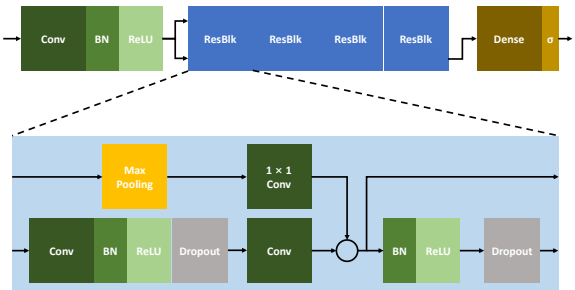

Figure 1: Ribeiro *et al* [19] developed a comprehensive DNN model for arrhythmia classification, which utilized multiple ResBlks in a specific configuration.

output. This model is represented by a Eq 1, typically a feedforward neural network.

$$e_{t,i} = a(s_{t-1}, h_i) \tag{1}$$

To calculate the **weights** as Eq 2, the attention mechanism applies a softmax function to the alignment scores that were computed earlier.

$$a_{t,i} = softmax(e_{t,i}) \tag{2}$$

The attention mechanism generates a distinct **context vector** for the decoder at each time step as Eq 3, which is determined by a weighted sum of all encoder hidden states.

$$c_t = \sum_{i=1}^{T} a_{t,i} h_i \tag{3}$$

3 Method

Using the model presented by Ribeiro *et al* [19] as the base model, this study aims to evaluate the cross-dataset generalization ability of the model, by investigating how training data characteristics impact the model's generalization performance. Additionally, this study proposes an attention mechanism to enhance the model's generalization ability.

3.1 Datasets

To evaluate the proposed method, this study primarily used the extensive ECG dataset collected by Telehealth Network Minas Gerais (TNMG) used by Ribeiro *et al* [19]. Additionally, the China Physiological Signal Challenge 2018 (CPSC) dataset [13] and the ECG database from Shaoxing and Ningbo Hospitals (SNH) were employed as validation datasets.

Telehealth Network Minas Gerais (TNMG)

The Telehealth Network Minas Gerais (TNMG) dataset, cited in the reference by Ribeiro *et al* [19], consists of 2,322,513 labeled S12L-ECG recordings with short duration, obtained from 1,676,384 unique patients. Between 2010 and 2016, the Telehealth Network of Minas Gerais, Brazil, gathered data using either a Tecnologia Eletronica Brasileira TEB ECGPC model or a Micromed Biotecnologia ErgoPC 13 model tele-electrocardiograph. Cardiologist annotations and the University of Glasgow ECG analysis program were used to label the recordings, resulting in a dataset containing a normal heart rhythm and six typical clinical arrhythmia categories. The recordings were sampled at 400 Hz, and zero-padding was used where necessary to ensure each of the 12 leads produced records of equal length, containing 4096 samples, which corresponds to approximately 10 seconds [19]. The six common clinical arrhythmia classes in the dataset is summarised in Table 1

This study utilized 1,048,575 distinct ECG 12-lead tracings from the TNMG dataset as part of its experimentation. The distribution of the dataset is depicted in Figures 2.

The TNMG dataset, a comprehensive collection of ECG recordings, offers a valuable resource for investigating the prevalence of ECG abnormalities in the general population. Figure 2 presents a striking proportion of normal ECG records in the TNMG dataset, with over 80% of the total records displaying no labelled abnormality. This finding is consistent with the theoretical expectation of a healthy population, as demonstrated in Table 1. Moreover, the distribution of ECG abnormalities across different age groups in the TNMG datas aligns with the known increased prevalence of arrhythmias in the elderly population. Notably, the number of patients with arrhythmias declines beyond the age of 75, possibly due to lower survival rates among the elderly population with arrhythmia. Overall, the present findings highlight the importance of utilizing large and diverse datasets such as TNMG to gain insights into the expected distribution of ECG abnormalities in the general population, while also identifying age as a crucial factor in the prevalence of certain cardiac pathology.

The analysis of arrhythmia classifications within the dataset demonstrates a relatively uniform distribution, with each type of arrhythmia occurring at comparable frequencies. Furthermore, this con-

Type	Acronyms	Description	Prevalence in Population
First Degree Atrioventricular Block	1dAVb	A PR interval that exceeds 200 ms on a surface electrocardiogram is an indication of first-degree atrioventricular block (AVB)	4% among elders [16]
Atrial Fibrillation	AF	Rapid and irregular heartbeat characterizes atrial fibrillation (AF), which is the most common type of arrhythmia	2% of European and North American population [22]
Left Bundle Branch Block	LBBB	It results in a sequence of activation in the right ventricle before the left ventricle, which leads to modifications in the left ventricle's perfusion, mechanics, and workload	0.032% among men and 0.037% among women [18]
Right Bundle Branch Block	RBBB	A condition that affects the electrical activity in the ventricles of the heart, causing a delay in the depolarization of the right ventricle due to disrupted transmission of signals in the His-Purkinje system	0.2% to 1.3% of general population [23]
Sinus Tachycardia	ST	It is a tachyarrhythmia characterized by a raised resting heart rate and an exaggerated heart rate response to slight exertion or alterations in body posture	1.16% of randomly selected individuals [20]
Sinus Bradycardia	SB	A condition where the heart's sinus node fires electrical impulses slower than normal, causing the heart rate to be slower than the usual resting rate	0.16% of elderly population [7]

Table 1: Classifications of Arrhythmia in Telehealth Network Minas Gerais (TNMG) Dataset

sistent distribution pattern is maintained across both genders, with no apparent gender-based variations observed. The age-based breakdown of arrhythmia classifications is also consistent with the general trend seen in the overall breakdown of age groups within the dataset. Notably, the majority of cases are concentrated within the middle age groups, underscoring the utility of this dataset for developing predictive models for arrhythmias. These findings offer valuable insights into the prevalence and distribution of arrhythmias in the general population and emphasize the importance of utilizing large and diverse datasets for developing robust predictive models for arrhythmia detection and classification.

China Physiological Signal Challenge 2018 (CPSC)

The CPSC training dataset is publicly accessible and includes an open-source collection of 6,877 12 lead ECG records. The recordings range in length from 6s to 60s and were sampled at a rate of 500Hz [13]. The ECGs are divided into eight different types of arrhythmia, with AF, 1dAVb, LBBB, and RBBB overlapping with the TNMG dataset. Additionally, ST-segment depression (STD), premature atrial contraction (PAC), ST-segment elevated (STE) and premature ventricular contraction (PVC) are also represented.

In order to align the CPSC dataset with the TNMG dataset introduced earlier, patients diagnosed with STD, PAC, STE, and PVC in their first labelled condition are excluded from our analysis. Additionally, the ECG tracings are resampled to 400Hz to ensure comparability with the TNMG dataset. The breakdown of the processed dataset is shown in Figure 3.

The dataset contains a high proportion of patients with arrhythmia, with approximately 75% of the total 4,622 unique ECG tracings exhibiting one or more types of arrhythmia. This high proportion is not a result of the reduction in dataset size, but rather an inherent characteristic of the original dataset. The dataset follows the general trend in the population with a higher frequency of arrhythmia among elderly patients. However, the distribution of patients across the four classifications of arrhythmia is uneven. The high proportion of patients with Right Bundle Branch Block (RBBB) is not reflective of the population and makes the dataset uneven for training algorithms.

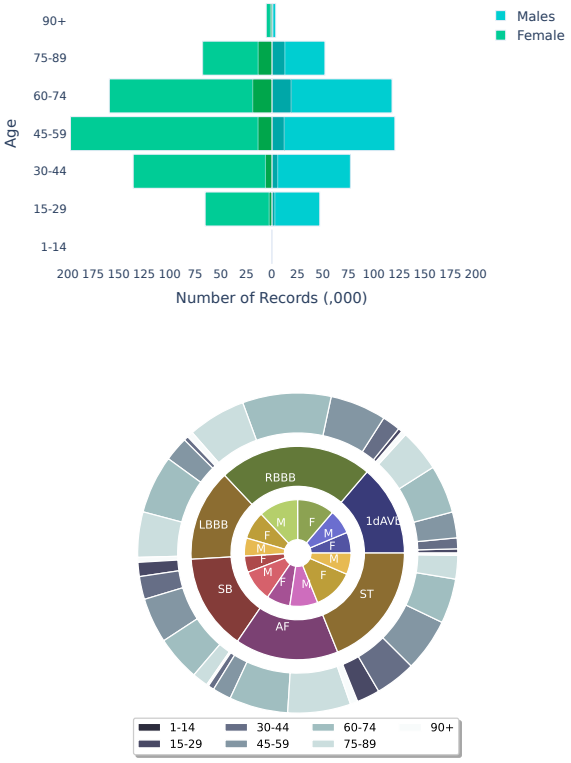

Figure 2: Upper: The TNMG dataset has been organized into specific age groups and genders. Additionally, a darker bar provides a breakdown of the data with Arrhythmia. Lower: Central donut chart provides a comprehensive depiction of the prevalence of various arrhythmias. The inner layer offers a nuanced breakdown by gender, while the outer layer offers a detailed breakdown by age group.

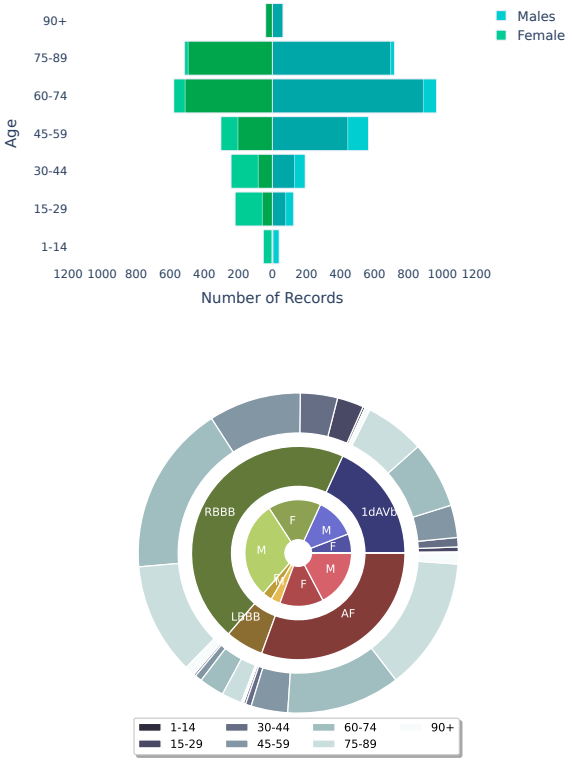

Figure 3: Upper: The processed CPSC dataset has been organized into specific age groups and genders. Additionally, a darker bar provides a breakdown of the data with Arrhythmia. Lower: Central donut chart provides a comprehensive depiction of the prevalence of various arrhythmias. The inner layer offers a nuanced breakdown by gender, while the outer layer offers a detailed breakdown by age group.

Shaoxing and Ningbo Hospitals (SNH)

The SNH dataset, which is open to the public, contains 45,152 12-lead ECG recordings gathered between 2013 and 2018 [24] [25]. These recordings are 10 seconds in length, and a licensed physician has assigned labels to each one indicating its association with one or more of the 63 arrhythmia types. In order to conform with the TNMG dataset, only ECG tracings that are categorized as normal or fall under the six classes existing in TNMG have been retained. Consequently, the total number of recordings has been decreased to 34,033. Furthermore, all tracings have been resampled at a rate of 400Hz. The processed dataset is presented in Figure 4.

The SNH dataset highlights a significant prevalence of arrhythmia among patients, with approximately 70% of the population presenting one or more types of arrhythmia. Upon closer examination of the patient classifications, the dataset reveals a significant skew towards SB, with more than half of the arrhythmia cases categorized as such. Such an imbalance towards SB in the dataset may pose a challenge when training a classification model for arrhythmia.

3.2 Subsets

To investigate how training data characteristics affect the model's generalization ability, this study selected subsets of ECG tracings from the TNMG dataset, using different selection criteria. The same model was trained on each of the subsets, and a cross-dataset performance analysis was conducted to examine the model's generalization capacity. The model is also compared to the model

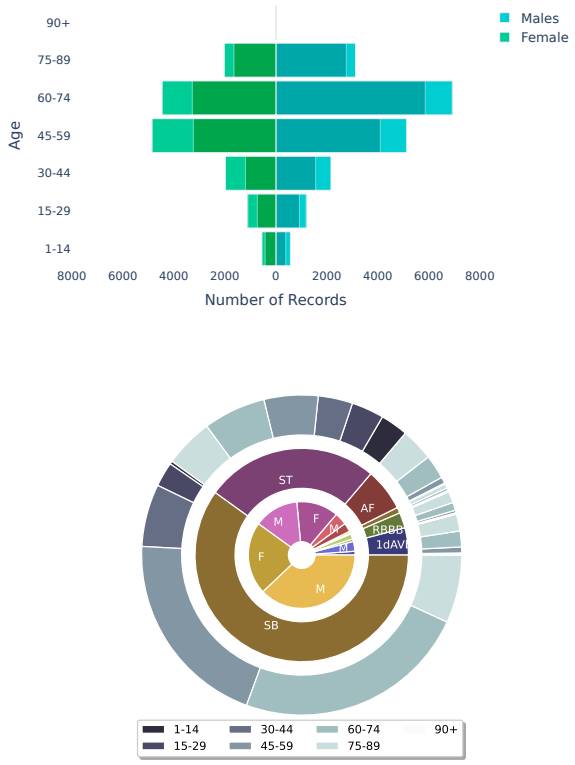

Figure 4: Upper: The processed SNH dataset has been organized into specific age groups and genders. Additionally, a darker bar provides a breakdown of the data with Arrhythmia. Lower: Central donut chart provides a comprehensive depiction of the prevalence of various arrhythmias. The inner layer offers a nuanced breakdown by gender, while the outer layer offers a detailed breakdown by age group.

trained in [19].

By using different selection criteria for the training data subsets, this study aims to provide insights into which characteristics of the training data have the most significant impact on the model's generalization ability. Furthermore, by analyzing the model's cross-dataset performance, this study can assess the degree to which the model can apply what it learned from the training data to new datasets. The findings from this study can inform future work on developing models with robust generalization capabilities in ECG analysis.

3.3 Attention Mechanism

Attention mechanisms have been shown to improve the performance of models [3], but this study goes beyond that and examines the effect of attention mechanisms on generalization. In this study, an additional attention layer is proposed to be added to the base model, which will also be trained on the six subsets of ECG tracings. The performance of this model with the attention mechanism will be compared to the performance of the base model without the attention mechanism, and their cross-dataset generalization abilities will be evaluated.

The proposed base model with the attention mechanism is illustrated in Figure 5. By including an attention mechanism, the model can selectively focus on the most informative regions of the ECG tracings, which could potentially improve the model's ability to generalize to new datasets. The findings from this study could provide insights into the benefits of incorporating attention mechanisms in ECG analysis models and how they impact the models' generalization capabilities.

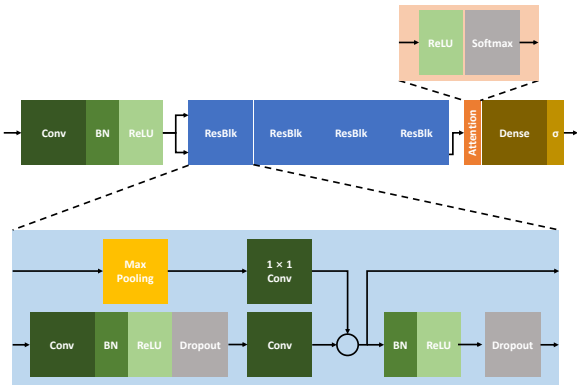

Figure 5: The attention layer that has been added to the base model integrates both ReLu and Softmax mechanisms, and it is located at the rear of the base architecture.

3.4 Evaluation Metrics

The models in this study are evaluated using the following performance metrics.

$$Specificity = \frac{TP}{TN + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

A higher score typically indicates better performance of the model. While F_1 score indicates the overall balanced performance of the model.

4 Experiment

4.1 Subsets

The objective of this study is to investigate the influence of distinct selection criteria for training data on the generalization capacity of a proposed model architecture. The model will be trained using four subsets of 12 lead ECG data from the TNMG dataset, each consisting of 21,000 traces, with varying selection criteria as presented in Table 2. The study's goal is to gain insight into how training data affects the model's generalization capability across different datasets.

Subsets	Selection Criteria
I	3,000 recordings from each of the six ECG classification types and 3,000 normal readings
II	Same selection criteria as approach I, but favor patients with multiple types of arrhythmia
III	Select the 1,500 oldest and youngest records for each ECG classification
IV	Randomly selected

Table 2: Subset Selection Criteria

As a result of some patients having more than one type of arrhythmia, there may be repeated selection of patients from the TNMG dataset. The remaining data required to reach the total of 21,000 records is randomly drawn from the dataset. These repeated selections may introduce discrepancies in the overall statistics. The statistics of the subsets are summarized in Table 3.

4.2 Training

The architecture of the base model shown in Figure 1 is trained on each of the subsets, and its performance is evaluated on a test dataset presented in [19]. The performance of these models is compared to that of a model trained in [19], and assess their cross-dataset generalization ability by testing them on the CPSC and SNH datasets.

Similarly, we train the proposed model with attention mechanism illustrated in Figure 5 on the subsets, and select the best performing model for evaluation of its cross-dataset generalization ability. The performance is compared to that of the base model, and its generalization capacity is evaluated.

To train our model, the hyperparameters that were previously used in [19] were deployed. These include 16 for kernel size, 64 for batch size, an initial learning rate of 0.001, the Adam optimizer, 0.8 for dropout rate, and a patience value of 9 epochs for early stopping.

By utilizing these hyperparameters, we aim to ensure consistency in our experiments and facilitate direct comparison with prior research. Such an approach enables a more rigorous evaluation of our proposed model architecture's performance in ECG signal classification tasks.

4.3 Model Testing

The models were evaluated on the test dataset described in [19], which consists of 827 unique 12-lead tracings of ECG. Further details regarding the breakdown of the test dataset can be found in Table 4.

To assess the generalization capacity of the models across different datasets, tests of the models were conducted on the CPSC and SNH datasets. The performance of the models was evaluated using the proposed metrics, and the results were discussed to determine their generalization performance.

5 Result

The experiment results are presented in the following two subsections. The first subsection highlights the findings of the base model architecture, while the second subsection examines the performance of the proposed model with attention mechanisms.

5.1 Base Model

The models trained on the subsets are tested on the test set of [19] and is compared to the result of the trained model in [19]. The perfomance matrics of the modles are presented in Table 5.

Then the trained models were evaluated on the CPSC dataset and their performance was compared to that of the model presented in [19], as shown in Table 6.

Ultimately, the models were assessed on the SNH dataset and their effectiveness was compared to that of the model introduced in [19], as illustrated in Table 7.

5.2 Boosted Model with Attention Mechanism

To comprehend the impact of the attention mechanism on the generalizability of the model, we trained the proposed model architecture depicted in Figure 5 on the four subsets. Its performance on the test set of [19] is presented in Table 8.

	I	II	III	IV
1dAVb	3,234	6,677	3,341	333
RBBB	4,320	7,020	3,819	589
LBBB	3,443	4,935	3,411	340
SB	3,367	3,624	3,324	357
AF	3,561	4,043	3,846	377
ST	3,299	3,219	3,395	455
Normal	3,000	3,000	3,000	18,793
Age Group				
90+				
75-89				
60-74				
45-59				
30-44				
15-29				

Table 3: Subset statistics with charts displaying different age groups with distinct bars for females and males. The charts organize the female bars on the left and the male bars on the right for visual clarity.

	Arrhythmia
1dAVb	28
RBBB	34
LBBB	30
SB	16
AF	13
ST	36
	Age Group
90+	
75-89	
60-74	
45-59	
30-44	
15-29	

Table 4: Test set statistics with charts displaying different age groups with distinct bars for females and males. The charts organize the female bars on the left and the male bars on the right for visual clarity.

	I	II	III	VI	[19]
P	0.814	0.768	0.801	0.791	0.924
R	0.868	0.926	0.893	0.817	0.935
S	0.994	0.992	0.992	0.994	0.997
F_1	0.839	0.831	0.834	0.797	0.926

Table 5: Assessment of model performance on TNMG test set

The top-performing model, incorporating the attention mechanism and achieving the highest F_1 score, was assessed on both the CPSC and SNH datasets. Its performance was compared to that of the TNMG model introduced in [19], and the findings are presented in Table 9.

	I	II	III	VI	[19]
P	0.929	0.883	0.901	0.795	0.910
R	0.844	0.841	0.861	0.759	0.837
S	0.980	0.967	0.976	0.946	0.981
F_1	0.883	0.860	0.879	0.770	0.867

Table 6: Assessment of model performance on CPSC dataset

	I	II	III	VI	[19]
P	0.479	0.512	0.483	0.149	0.479
R	0.502	0.549	0.527	1.000	0.526
S	0.852	0.819	0.822	0.000	0.815
F_1	0.465	0.504	0.488	0.221	0.482

Table 7: Assessment of model performance on SNH dataset

	I	II	III	VI
P	0.861	0.822	0.838	0.879
R	0.851	0.842	0.923	0.829
S	0.995	0.994	0.994	0.996
F_1	0.853	0.830	0.875	0.844

Table 8: The performance evaluation of the models with attention mechanism on TNMG test set

	Model with Attention		TNMG Model	
	CPSC	**SNH**	**CPSC**	**SNH**
P	0.904	0.496	0.901	0.479
R	0.856	0.573	0.861	0.526
S	0.970	0.810	0.976	0.815
F_1	0.879	0.522	0.879	0.482

Table 9: Comparative evaluation of model performance trained on TNMG subset III

6 Discussion

Based on the experimental results, it can be observed that the performance of a model is significantly influenced by the characteristics of the training data, even when the same model structure and amount of training data are used. The models trained on subsets I, II, and III outperformed the model trained on subset VI, primarily due to the balance of the training dataset. These subsets had a similar number of arrhythmia classifications, which contributed to their balanced nature. Furthermore, the results also indicate that a larger training dataset positively impacts the model's performance. When assessed on the TNMG test dataset, Ribeiro's model, which was trained on the complete TNMG dataset, demonstrated the highest level of performance compared to all other models.

The story changes when it comes to model generalization. As previously discussed, models trained on subsets I, II, and III still outperformed the model trained on subset VI, highlighting the significance of a balanced dataset. However, Ribeiro's model, which was trained on the complete TNMG dataset, did not perform better than the other models trained on smaller datasets in generalization. In terms of F_1 score, Ribeiro's model underperformed two models during testing on the CPSC and SNH datasets. This finding emphasizes the importance of data quality over data quantity in terms of generalization performance. The fact that a model trained on a smaller dataset can maintain its performance while consuming less power and time provides direction for future experimental designs. Researchers can focus on obtaining high-quality data rather than collecting large amounts of data, which can be time-consuming and costly. By using a smaller dataset, researchers can train models more efficiently and still achieve comparable performance, which can ultimately lead to more practical and effective solutions in various applications.

The final part of the experiment focuses on the impact of the attention mechanism on performance and generalization. The results clearly indicate that the attention mechanism enhances the performance of almost all subsets in terms of the F_1 score, with a particularly significant improvement observed in subset VI, which initially had the lowest F_1 score. This trend also translates into generalization, as evidenced by the comparison of the basic model and the model with the attention mechanism, both trained on subset III. The performance was notably enhanced, especially on the SNH dataset where the original performance was low. These findings demonstrate the potential of the attention mechanism to significantly improve model performance and generalization.

7 Conclusion

This study highlights how the quality of training data is crucial to the generalization performance of a model. A balanced training dataset can result in a smaller dataset achieving comparable performance to a larger one. Additionally, while attention mechanisms are known to enhance model performance, their ability to improve generalization has not been extensively studied. This research demonstrates that incorporating an attention mechanism can enhance the generalization performance of a model, especially when the original model's performance was poor.

This study is one component of an ongoing investigation into model generalization, and its findings will inform the design of future experiments. The results obtained in this stage of the research are valuable for guiding future research on this topic.

8 Acknowledgement

Zhaojing Huang would like to acknowledge support of the Research Training Program (RTP) provided by the Australia Government.

9 Code Availability

To obtain the code used in this paper, simply request it from the corresponding author, but be aware of any conditions or restrictions on its use.

10 Data Availability

This paper employs publicly accessible datasets such as CPSC and SNH, in addition to the TNMG dataset which can be obtained by requesting access from the dataset owner. The dataset is not publicly available, but access can be granted upon request with approval from the dataset owner.

References

[1] Miquel Alfaras, Miguel C Soriano, and Silvia Ortín. "A fast machine learning model for ECG-based heartbeat classification and arrhythmia detection". In: *Frontiers in Physics* (2019), p. 103.

[2] Majd AlGhatrif and Joseph Lindsay. "A brief review: history to understand fundamentals of electrocardiography". In: *Journal of community hospital internal medicine perspectives* 2.1 (2012), p. 14383.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[4] RR Bond et al. "Assessing computerized eye tracking technology for gaining insight into expert interpretation of the 12-lead electrocardiogram: an objective quantitative approach". In: *Journal of electrocardiology* 47.6 (2014), pp. 895–906.

[5] Judith Bouchardy et al. "Atrial arrhythmias in adults with congenital heart disease". In: *Circulation* 120.17 (2009), pp. 1679–1686.

[6] CJ Breen, GP Kelly, and WG Kernohan. "ECG interpretation skill acquisition: A review of learning, teaching and assessment". In: *Journal of electrocardiology* 73 (2022), pp. 125–128.

[7] Shuenn-Nan Chiu et al. "Long-term outcomes of pediatric sinus bradycardia". In: *The Journal of Pediatrics* 163.3 (2013), pp. 885–889.

[8] Elaine Chow et al. "Risk of cardiac arrhythmias during hypoglycemia in patients with type 2 diabetes and cardiovascular risk". In: *Diabetes* 63.5 (2014), pp. 1738–1747.

[9] Dhaval S Desai and Said Hajouli. "Arrhythmias". In: *StatPearls [Internet]*. StatPearls Publishing, 2022.

[10] Mohamed Hammad et al. "A multitier deep learning model for arrhythmia detection". In: *IEEE Transactions on Instrumentation and Measurement* 70 (2020), pp. 1–9.

[11] Sonain Jamil and MuhibUr Rahman. "A Novel Deep-Learning-Based Framework for the Classification of Cardiac Arrhythmia". In: *Journal of Imaging* 8.3 (2022), p. 70.

[12] Terese Lindberg et al. "Prevalence of unknown and untreated arrhythmias in an older outpatient population screened by wireless long-term recording ECG". In: *Clinical interventions in aging* (2016), pp. 1083–1090.

[13] Feifei Liu et al. "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection". In: *Journal of Medical Imaging and Health Informatics* 8.7 (2018), pp. 1368–1373.

[14] Roshan Joy Martis, U Rajendra Acharya, and Hojjat Adeli. "Current methods in electrocardiogram characterization". In: *Computers in biology and medicine* 48 (2014), pp. 133–149.

[15] Stanley Nattel et al. "New directions in cardiac arrhythmia management: present challenges and future solutions". In: *Canadian Journal of Cardiology* 30.12 (2014), S420–S430.

[16] Theodora Nikolaidou, Justin M Ghosh, and Andrew L Clark. "Outcomes related to first-degree atrioventricular block and therapeutic implications in patients with heart failure". In: *JACC: Clinical Electrophysiology* 2.2 (2016), pp. 181–192.

[17] Saman Parvaneh et al. "Cardiac arrhythmia detection using deep learning: A review". In: *Journal of electrocardiology* 57 (2019), S70–S74.

[18] Andrés R Pérez-Riera et al. "Left bundle branch block: Epidemiology, etiology, anatomic features, electrovectorcardiography, and classification proposal". In: *Annals of Noninvasive Electrocardiology* 24.2 (2019), e12572.

[19] Antônio H Ribeiro et al. "Automatic diagnosis of the 12-lead ECG using a deep neural network". In: *Nature communications* 11.1 (2020), p. 1760.

[20] Aino-Maija Still et al. "Prevalence, characteristics and natural course of inappropriate sinus tachycardia". In: *EP Europace* 7.2 (2005), pp. 104–112.

[21] Amin Ullah et al. "A hybrid deep CNN model for abnormal arrhythmia detection based on cardiac ECG signal". In: *Sensors* 21.3 (2021), p. 951.

[22] Zengwu Wang et al. "The disease burden of atrial fibrillation in China from a national cross-sectional survey". In: *The American journal of cardiology* 122.5 (2018), pp. 793–798.

[23] Yunyun Xiong et al. "The prognostic significance of right bundle branch block: a meta-analysis of prospective cohort studies". In: *Clinical Cardiology* 38.10 (2015), pp. 604–613.

[24] Jianwei Zheng et al. "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients". In: *Scientific data* 7.1 (2020), p. 48.

[25] Jianwei Zheng et al. "Optimal multi-stage arrhythmia classification approach". In: *Scientific reports* 10.1 (2020), p. 2898.