International Journal on Advanced Science Engineering Information Technology

Vol.12 (2022) No. 4
ISSN: 2088-5334

Re-implementation of Convolutional Neural Network for Arrhythmia Detection

Muhammad Ilham Rizqyawan [a,b,*], Yahdi Siradj [a], M Faizal Amri [b], Agus Pratondo [a]

[a] *Department of Multimedia Engineering Technology, Telkom University, Bandung, 40257, Indonesia*
[b] *Instrumentation Development Center, Indonesian Institute of Sciences, Bandung, 40135, Indonesia*
Corresponding author: [*]muha176@lipi.go.id

***Abstract*— Arrhythmia is an irregular heartbeat that may cause serious problems such as cardiac arrest and heart failure if left untreated. A dozen of studies have been conducted to make an automated arrhythmia detector. The classification approach uses a simple rule-based model, traditional machine learning, to a modern deep-learning technique. However, comparing an arrhythmia classifier performance is not an easy task. There are several different datasets, classification standards, data splitting schemes, and metrics. To assess the real performance of the developed models, it is important to train and evaluate the model in a standardized method such as the result score can become standard too. In this study, a set of CNN models from Acharya were re-implemented by re-training and re-evaluating it in a more standardized method. The model uses a raw ECG waveform with 260 samples around the QRS peaks and classifies it into five arrhythmia classes. The experiment was conducted using three configurations, using both intra-patient and inter-patient schemes. The experimental results show good performance for the intra-patient scheme but not for the inter-patient. There is a reduction of sensitivity and precision in the intra-patient scheme using a standardized method in this study compared to the original paper. This result indicates biased results caused by the oversampled test data in the original paper. In addition to the intra-patient result, the inter-patient result is also provided for a standardized comparison to other works in the future.**

***Keywords*— Arrhythmia classification; convolutional neural network; electrocardiography; heartbeat classification.**

Manuscript received 9 Oct. 2020; revised 21 May 2021; accepted 22 Jun. 2021. Date of publication 31 Aug. 2022.

I. INTRODUCTION

There have been many methods and approaches for automatic ECG signal detection and classification over the last decade, thanks to open-access ECG databases such as MIT-BIH [1]. In the development of P, QRS, and T wave detection, there are many types of detection: signal derivatives and digital filters [2], wavelets [3]–[5], and learning algorithms[6]–[8] approaches. The segmentation of the ECG signal, particularly the QRS complex, is said to be very close to the optimal and leaves small room for improvements[9].

There are also many approaches to arrhythmia classification. The method with the traditional machine learning algorithm usually also focuses on feature extraction methods. The most traditional feature in arrhythmia detection is the R-R interval and its derivatives [10], [11]. The other feature extraction technique widely used for the task is wavelet[12]–[16] since it is believed to be the best feature extraction method for arrhythmia detection[9]. For the classifier, the conventional techniques employed for the task is Linear Discriminant (LD) [17], Fuzzy Logic [18]–[20], Decision Tree-based methods [21]–[23], Support Vector Machine (SVM) [16], [24], [25], and Artificial Neural Network (ANN)[10], [11], [15]. The feature extraction is often not needed by a more modern algorithm such as reservoir computing[26] and deep learning-based technique [27]–[30]. They use raw data as the algorithm can extract the feature automatically, although some methods combine the raw data with traditional features[31], [32] to get a better result.

Unfortunately, measuring and comparing arrhythmia classifier performance is not an easy task[9]. There are different arrhythmia classification standards, data splitting, and metrics (especially for imbalance problems). For example, in using the dataset, the majority uses MIT-BIH, while Alfaras *et.al*.[32] combines MIT-BIH with the AHA dataset and Hannun *et.al*.[29] uses their dataset. In terms of data splitting protocol, the majority of the studies use general random sampling. The testing data can contain data from the same patients in the training data (intra-patient). Some studies

[31], [33] use the patient-specific protocol, some studies [26], [30] use the inter-patient protocol, while only a few studies provide both the intra and inter-patient protocol such as [28].

As few authors use the same standard, it is difficult to compare methods fairly. One of the solutions for this challenge is to re-implement the method published in literature with a more standardized protocol, including the inter-patient protocol[34]. By presenting the re-implemented result in the standards scheme, the method can be compared fairly.

Convolutional Neural Network (CNN) architecture from Acharya *et al*.[27] is one of the most cited papers for arrhythmia detection. It uses 1D CNN to classify the heartbeat with oversampling to deal with the imbalance problem. A separate feature extraction step usually incorporated in the traditional machine learning approach is not needed with CNN. Unfortunately, the result is difficult to compare as-is with other methods due to several reasons. The first reason is that the data were oversampled before splitting into training data and testing data in the cross-validation. It means both training data and testing data contained synthetic data from oversampling; thus, the accuracy of the results is questioned [28]. The second reason is that they did not provide the result from the inter-patient protocol.

In this paper, the CNN architecture from Acharya *et al*.[27] was re-implemented with several changes that address the problem mentioned in the previous paragraph. This way, the result's score will be more standards, and the architecture can be fairly compared with other methods.

II. Materials and Method

A. Overview

In this experiment, we re-implemented the Acharya architecture by training a set of CNN models with three configurations. The general outline of the experiment was as follows: data loading and preparation, data splitting, oversampling, model training, dan evaluation, as shown in Fig. 1 below.

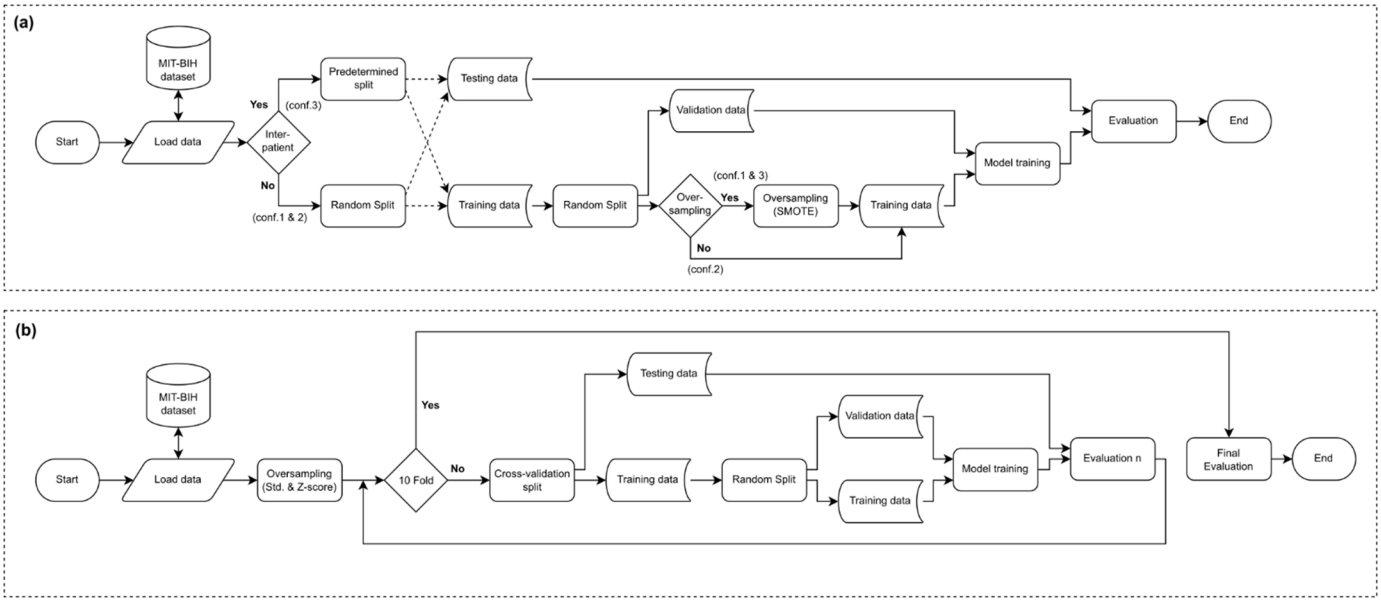


Fig. 1 Flowchart of the research stages: a) this paper, b) original paper [27]

In the data splitting stage, we prepare the data used for training, validation, and testing. This stage is important as the type of protocol (intra-patient or inter-patient) is mainly characterized by splitting the data. The next stage is oversampling. This stage is employed to overcome the imbalance problem of the arrhythmia dataset. In this experiment, only the training data were oversampled. We also trained a model without oversampling to investigate the effect of oversampling in this architecture.

The last stages are model training and evaluation. The training stage is the stage where the network is learning from the data by adjusting the weights in such a way as to minimize the error. After the model is done with the training, we need to test it to predict the data it has never seen before. The result is displayed in a confusion matrix, and from there, we can calculate other metrics to evaluate the model fully. The flowchart of the experiment in this study can be seen in Fig. 1. The details of the dataset, preparation, configurations, and evaluation metrics are described in the following sub-sections.

In data loading and preparation, we load the arrhythmia dataset (MIT-BIH), extract it, and transform it to the shape required by the architecture. Instead of recording our data, we used the widely available dataset for two major reasons. The first one is that we cannot do the annotation ourselves to decide the heartbeat class. Many cardiologists would need to manually annotate hundreds of thousands of heartbeats required to train the deep network. The second one to fairly assess the model's performance with other studies, we need to use a standardized dataset, at least as the testing data.

B. Dataset

Data from the MIT-BIH arrhythmia dataset [1] were used for this research. This dataset contains approximately 110.000 annotations of heartbeat with its arrhythmia label from 48 records, with each record length being 30 minutes long. Each

record contains two channels (lead II and V1) with a sampling rate of 360 Hz, 11-bit resolution over a 10mV range.

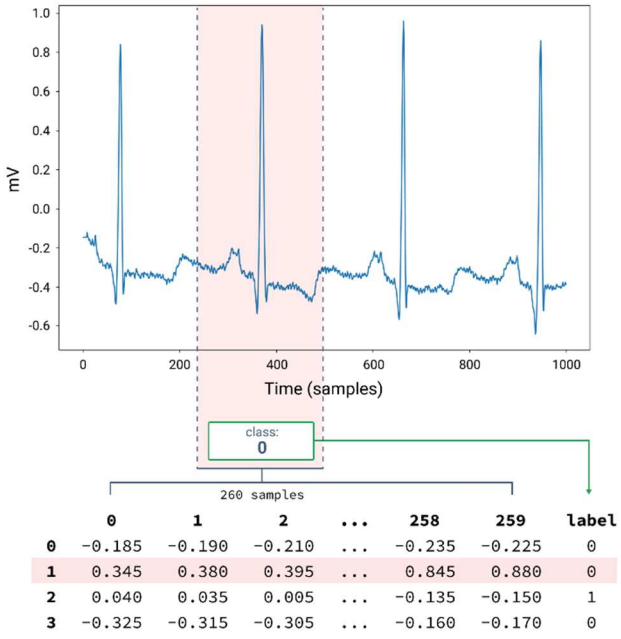
The data were classified into five classes according to the AAMI practice report recommendation [35]: Normal (N), Supraventricular ectopic beat (SVEB), Ventricular ectopic beat (VEB), Fusion beat (F), and Unknown beat (Q). In addition to the classification, four records of 48 records were excluded in this work to follow recommendations from AAMI.

C. CNN Architecture

CNN architecture from Acharya *et al.*[27] was used in this research. It contains three layers of 1D convolutional and max-pooling pairs as a feature extractor and is followed by three layers of fully connected layers as the classifier. The details of the architecture can be seen in Fig. 3. The Stochastic Gradient Descent (SGD) with Nesterov Momentum was used as the optimizer for the training stage. The learning rate was set at 0.003 with a momentum of 0.7.

D. Data Preparation

The original dataset contained ECG waveform data with two channels, the location of the QRS complex for each heartbeat, and the annotations of the heartbeat. For the classification, the data was prepared as follows: each heartbeat is considered as a row data, with its annotation as the label column. For the features, we use 260 samples of raw signal centered at R-peaks. The data was simply extracted by locating the R-peaks provided by the dataset and then extracting 130 samples before and after the peaks. The data preparation process can be seen in Fig. 2.



	0	1	2	...	258	259	label
0	-0.185	-0.190	-0.210	...	-0.235	-0.225	0
1	0.345	0.380	0.395	...	0.845	0.880	0
2	0.040	0.035	0.005	...	-0.135	-0.150	1
3	-0.325	-0.315	-0.305	...	-0.160	-0.170	0

Fig. 2 Data Preparation

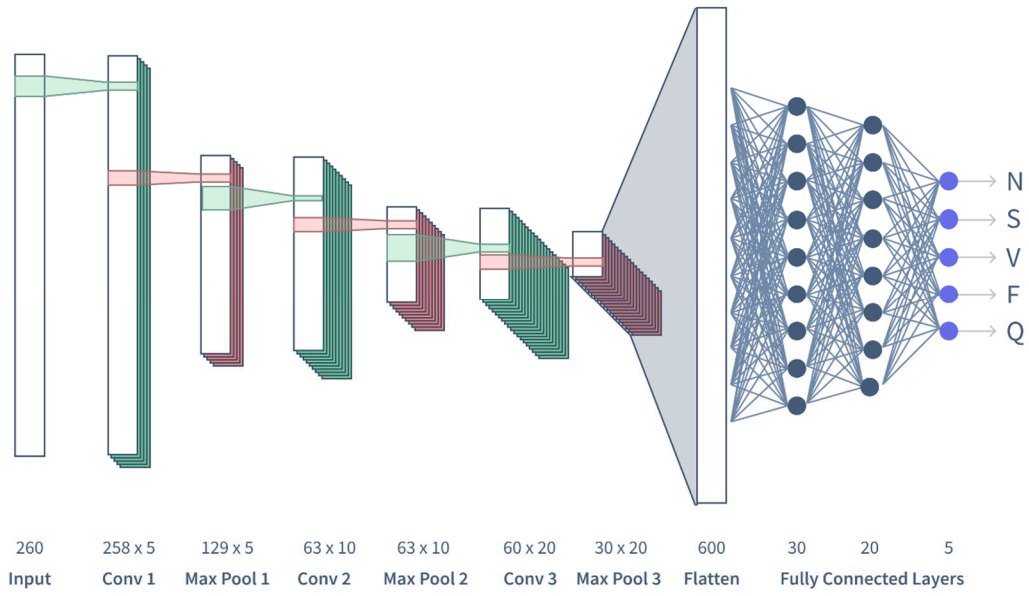


Fig. 3 Convolutional Neural Network architecture from Acharya[27] used in this study

E. Data-splitting Schemes

Supervised learning task such as arrhythmia classification relies on labeled data. The splitting scheme of these data into training and testing data has a big impact on how the model generalizes. There are three major data-splitting schemes in ECG arrhythmia classification: intra-patient, inter-patient, and patient-specific schemes.

1) Intra-patient: The majority of the study on arrhythmia classification uses an intra-patient scheme. The validation and testing data were sampled randomly from the original data. In other words, this scheme allows testing data and training data to have heartbeats from the same patient. This scheme is simple to implement, and in numbers, the models generally have better result scores than models with the inter-patient scheme. However, the models with this scheme are not good for classifying data from patients it has never seen before. To make the model comply with the new patient, the expert must manually annotate a portion of heartbeats from this new patient and re-train the model with the additional data[36].

2) Inter-patient: To address the generalization problem found in the intra-patient scheme, the inter-patient scheme was developed. In this scheme, we make sure there are no records from the same patient in training data and testing data, as opposed to the intra-patient scheme in the former configurations. The most widely adopted protocol for the inter-patient scheme is the protocol from de Chazal[34]. Forty-four records from the MIT-BIH dataset were split into two sets, namely DS1 and DS2, with DS1 used for training and DS2 for testing. Although this scheme was believed to be the most realistic scenario[9], the challenge of this task is much higher than the intra-patient scheme[9], [36]. Therefore, it is not uncommon for a model with high accuracy in the former scheme to get a lower score when it is applied to this scheme.

3) Patient-specific: Instead of trying to be better in generalization as the inter-patient scheme, this scheme goes in the opposite direction by focusing on adjusting the model to the specific patient. It means that every patient has their model specifically trained for them[31]. The model is trained by using two kinds of data. The first one is local data, which is the data from the specific patient. Moreover, the latter is global data, which is the data from many other patients. We cannot say that this scheme is bad, as it has an application in a certain scenario. But as the main objective of this study is to get a model that generalizes, the patient-specific scheme was not implemented in this study.

F. Experiment Configurations

The procedure for data preparation and data splitting was made in such a way to make the results can be analyzed more objectively by following the best-practice and standardized procedure. To achieve it, several changes in the training procedure were made in this work. The first change was in the preprocessing stage. The sampling of signal used for the input was not changed, but to minimize processing cost, there was no filtering and normalization used in this experiment as in the original paper.

The second change was the data splitting and oversampling. The data were oversampled in the original paper before it was split into training data and validation data/testing data. It means both training data and validation/testing data contained synthetic data from oversampling; thus, the accuracy of the results is questioned[28]. So in this experiment, only training data was oversampled using the synthetic minority over-sampling technique (SMOTE) [37]. Three configurations, as shown in TABLE .

TABLE I
CONFIGURATIONS

	Conf. 1	**Conf. 2**	**Conf. 3**
Total original heartbeat		100,733	
Data split scheme	Intra-patient	Intra-patient	Inter-patient
Oversampling	Yes, training data	No	Yes, training data
Training data	270,330	60,439	183,325
Validation data	20,147	20,147	10,205
Testing data	20,147	20,147	49,712

The configuration was created to assess the model from several perspectives. The first configuration was the most similar to the original paper, except for several points described in the previous paragraphs. The validation and testing data were sampled randomly from the original data. This scheme is known as the intra-patient scheme by allowing testing data and training data to have heartbeats from the same patient. The second configuration was the same as the first config, except there was no oversampling being made. This change was adopted to investigate the effectiveness of the oversampling.

The last configuration took a different approach in the form of a data separation protocol. The inter-patient protocol from de Chazal [34] was adopted for this configuration. Instead of using random sampling, the training and testing data were split using a predetermined list (DS1 and DS2). The list was proposed by [34] to make sure there are no data from the same patient in both training and testing.

G. Evaluation Metrics

Several standard metrics were used to measure the performance of the models: accuracy (acc), sensitivity (se), specificity (sp), precision (ppv), and f1-score (f1). The metrics were calculated by:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Se = \frac{TP}{TP+FN} \quad (2)$$

$$Spe = \frac{TN}{TN+FP} \quad (3)$$

$$Ppv = \frac{TP}{TP+FP} \quad (4)$$

$$F1 = 2 * \frac{Se*Ppv}{Se+Ppv} \quad (5)$$

True Positive (TP) is correctly predicted positive class, while True Negative (TN) is correctly predicted, negative class. In other words, the classifier is correctly recognizing the positive sample as positive (TP) and the negative sample as negative (TN). The classifier naturally wants a high TP and TN, while minimizing False Positive (FP) and False Negative (FN).

In addition to common metrics, other metrics are included, namely, Diagnostic Odds Ratio [38] (DOR) and Cohen Kappa score[39]. DOR is widely used in the medical community as a single indicator to measure diagnostic performance. DOR is denoted by:

$$DOR = \frac{TP}{FN} / \frac{FP}{TN} \quad (6)$$

The value of DOR ranges from 0 to infinity, but generally if the value is below 1. It means that the test does not discriminate between patients with and without the disorder. The next metric is the Cohen Kappa score. It is a measure for inter-rater agreement for categorical scales. Kappa score is denoted by:

$$Kappa = \frac{po-pe}{1-pe} \quad (7)$$

with po and pe that can be calculated with:

$$po = \frac{TP+TN}{TP+FP+TN+FN} \quad (8)$$

$$pe = (TN+FP)*(TN+FN)+(FN+TP)*(FP+TP) \quad (9)$$

Kappa score maximum value is 1, and according to Landis and Koch [40] a score < 0 is poor, 0.00 – 0.2 is slight, 0.21 –

0.4 is fair, 0.41 – 0.6 is moderate, 0.61 – 0.8 is substantial, and 0.81 – 1 is almost perfect agreement.

III. Result and Discussion

The model was implemented using Tensorflow 2 with Keras as the high-level API. Tensorflow was chosen because of the wide array of deployment options: from fully powered server or workstation to internet browser, mobile phone, and IoT device. This is important for us as we also investigate the possibility of deploying a real-time arrhythmia classifier in low computing platforms as mobile phone and edge devices[41].

The models were trained in Google Colab with Intel(R) Xeon(R) CPU @ 2.30GHz, Tesla K80 GPU, and 13GB of RAM. Each configuration's training was finished in approximately 40 minutes for 110 epochs. The history of accuracy and loss over epochs is shown in Fig. 4.

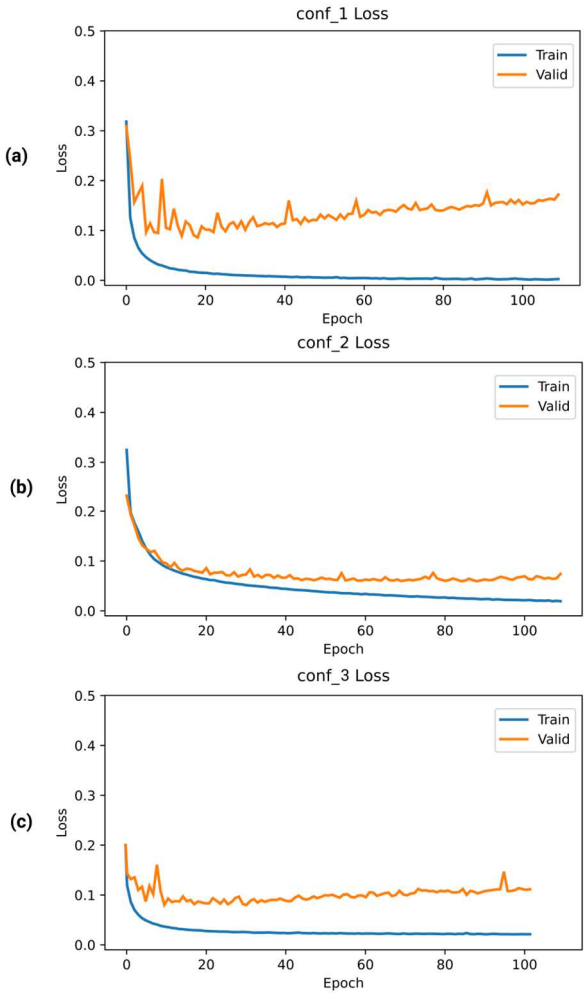


Fig. 4 Training loss history of: a) configuration 1, b) configuration 2, c) configuration 3

As can be seen in Fig. 4, the validation loss in configuration 1 is rising slightly, and the gap between validation loss and training loss becomes more significant over time. It indicates that there is overfitting in configuration 1. There is no raise observed in the validation loss of configuration 2, and the gap with the training loss is small compared to configuration 1. However, compared to configurations 1 and 3, the training loss in configuration 2 is still higher. It means that configuration 2 is underfit compared to other configurations. In general, configuration 3 shows a lower training loss at the early epoch than the previous configurations, but the validation loss is slightly rising over time, which also indicates overfitting, albeit not as high as configuration 1.

The final tests were done using test data to fully assess the model performance to see how the models predict the data it has never seen before. The result is shown in the confusion matrix of Fig. 5 and the calculated metrics in TABLE . In the final test, configuration 3 shows the worst performance in all metrics than the other configurations.

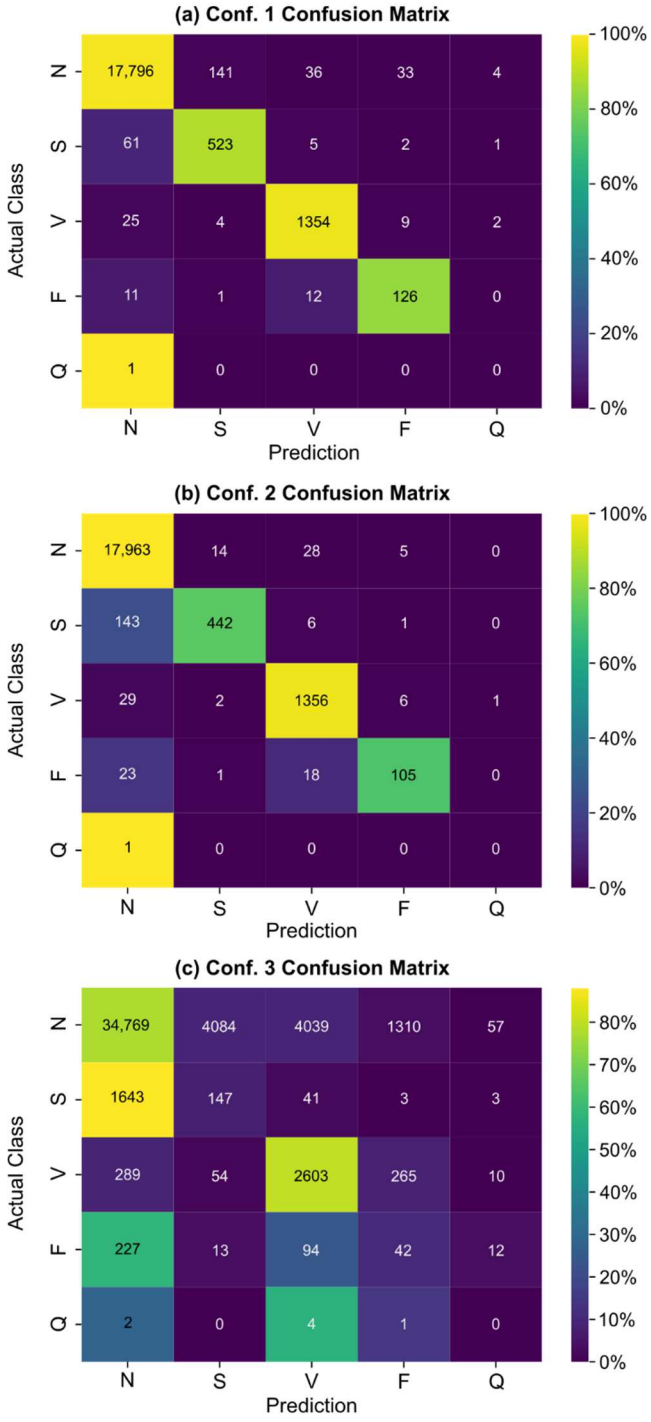


Fig. 5 Confusion matrix for: (a) configuration 1, (b) configuration 2, (c) configuration 3.

This result is expected since it is harder to get good results on the inter-patient scheme than the intra-patient scheme[9], [34]. The separation of the patient in training data and testing data and the ratio of training and testing data (TABLE) makes the inter-patient scheme more complicated. The real question is, how much worse the model performs in inter-patient compared to the intra-patient scheme?

From the data in TABLE , we can see that this model has a high score on the intra-patient scheme (conf. 1 & 2). This indicates that the model is suitable for the intra-patient scheme. But contrary to the previous configurations, on conf. 3, it only shows less than half the f1 score and less than a third of the kappa score compared to the intra-patient scheme. Thus, we can conclude that this CNN architecture is not suitable for the inter-patient scheme.

TABLE II
TEST RESULT METRICS

		Class				
		N	**S**	**V**	**F**	**Q**
Configuration 1	**acc**	0.985	0.989	0.995	0.997	1,000
	se	0.988	0.883	0.971	0.840	0.000
	spe	0.954	0.993	0.997	0.998	1,000
	ppv	0.995	0.782	0.962	0.741	0.000
	dor	1,730	1,008	11,943	2,381	0.000
	acc macro			0.993		
	f1			0.715		
	kappa			0.914		
Configuration 2	**acc**	0.988	0.992	0.996	0.997	1,000
	se	0.997	0.747	0.973	0.720	0.000
	spe	0.908	0.999	0.997	0.999	1,000
	ppv	0.989	0.963	0.963	0.900	0.000
	dor	3,785	3,387	12,833	4,283	0.000
	acc macro			0.994		
	f1			0.720		
	kappa			0.927		
Configuration 3	**acc**	0.766	0.883	0.904	0.961	0.998
	se	0.786	0.08	0.808	0.108	0.000
	spe	0.604	0.913	0.91	0.968	0.998
	ppv	0.941	0.034	0.384	0.026	0.000
	dor	5.581	0.916	42.657	3.670	0.000
	acc macro			0.902		
	f1			0.293		
	kappa			0.251		

Several studies have reported having an excellent performance for the latter case. Escalona-Moran *et al*.[26] used Reservoir Computing with Logistic Regression, and Mousavi *et al*. [28] combined CNN with Bidirectional Recurrent Neural Network (BiRNN) to get the results shown in TABLE .

Although the intra-patient scheme is considered not ideal for practical use, where the data to predict usually comes from the patient's record that is not used in the training data, it can be useful for other systems. One possible use case is the system where the model is inclusive to the patient of a particular hospital. In this system, the model will have an excellent performance on their patient, and only one model is needed for all the patients. The downside of this approach is that the model needs to be re-trained every time there is a new patient, and the model's performance might be decreasing inversely proportional to the number of samples.

The other possible—but slightly different—use case is a patient-specific classifier on personal devices[31], [33]. This scheme forms the data by combining specific patient data with globally aggregated data from all the patients. The downside of this approach is that every patient needs their own model, which will be expensive to train. Of course, further study is needed to investigate the application of this architecture in the patient-specific case.

TABLE III
INTER-PATIENT RESULT

		Class				
		N	**S**	**V**	**F**	**Q**
This paper (conf. 3)	**Acc**	76.56	88.25	90.35	96.13	99.82
	Se	78.56	8.00	80.81	10.82	0.00
	Sp	60.37	91.33	91.01	96.80	99.84
	Ppv	94.15	3.42	38.39	2.59	0.00
	Avg Acc			90.223		
	Avg Se			35.640		
	Avg Sp			87.869		
	Avg Ppv			27.709		
Mousavi *et al.* [28] (Inter-patient)	**Acc**	-	-	-	-	-
	Se	**99.68**	**88.94**	**99.94**	-	-
	Sp	**96.05**	**99.72**	**99.97**	-	-
	Ppv	**99.55**	**92.57**	**99.5**	-	-
	Avg Acc			-		
	Avg Se			**96.19**		
	Avg Sp			**98.58**		
	Avg Ppv			**97.21**		
Escalona-moran[26]	**Acc**	**96.28**	**96.28**	**99.71**	**99.91**	**99.99**
	Se	96.82	79.37	96.06	**92.26**	**57.14**
	Sp	91.89	96.93	**99.97**	**99.97**	**100**
	Ppv	98.98	49.8	99.49	**95.47**	**100**
	Avg Acc			**98.43**		
	Avg Se			84.33		
	Avg Sp			97.75		
	Avg Ppv			88.75		

The next question is the impact of the oversampling in the intra-patient scheme. In averaged metrics, the model trained without oversampling (configuration 2) shows a slightly better score. In general, both configurations 1 and 2 show mostly similar results. The difference in pattern is starting to show in the class-specific metrics. On average, configuration

1 has better sensitivity, while configuration 2 has better precision. In a high-risk context such as healthcare, where it is better to have more false positives in trade-off to have more false negatives, high sensitivity is preferred over high precision. In short, it is better to have falsely suspected arrhythmia instead of falsely ignored arrhythmia. This analysis suggests that the model with configuration 1 is preferred for the intra-patient scheme.

As the original paper generally followed the intra-patient scheme, the paper's comparison and the original paper are made with configuration 1. The difference between this paper (conf. 1) and the original paper is that in this paper, the final result was tested using test data, and there was no oversampling in the test data. In the original paper, the final result was tested using cross-validation, with oversampled data included. The difference in the result can be seen in TABLE .

TABLE IV
TEST RESULT'S DIFFERENCE WITH ORIGINAL PAPER IN INTRA-PATIENT SCHEME

		Class				
		N	**S**	**V**	**F**	**Q**
This Paper (conf. 1)	**Acc**	98.45	98.93	99.54	99.66	99.96
	Se	98.81	88.34	97.13	84	0
	Sp	95.41	99.25	99.72	99.78	99.97
	Ppv	99.45	78.18	96.23	74.12	0
	Avg Acc			99.31		
	Avg Se			73.66		
	Avg Sp			98.83		
	Avg Ppv			69.6		
Original Paper [27]	**Acc**	95.14	96.82	97.84	97.97	99.16
	Se	91.64	89.04	94.07	95.21	97.39
	Sp	96.01	98.77	98.78	98.66	99.61
	Ppv	85.17	94.76	95.09	94.69	98.41
	Avg Acc			97.39		
	Avg Se			93.47		
	Avg Sp			98.37		
	Avg Ppv			93.62		
Difference	**Acc**	3.31	2.11	1.7	1.69	0.8
	Se	7.17	-0.7	3.06	-11.21	-97.39
	Sp	-0.6	0.48	0.94	1.12	0.36
	Ppv	14.28	-16.58	1.14	-20.57	-98.41
	Avg Acc			1.92		
	Avg Se			-19.81		
	Avg Sp			0.46		
	Avg Ppv			-24.02		

In this result, we can see that there are significant difference scores in the average sensitivity and average precision, with a reduction of 21% in sensitivity and 25% in precision.

If we break down and see the specific class metrics, the highest gap comes from the F and Q class, wherein the pre-oversampled data, these classes are the two-top minority class among the others. This gap indicates that the original paper's test result is biased due to the use of oversampled data in the final test. Therefore, in this paper, we propose using the data in TABLE as a standardized score for Acharya's CNN architecture for arrhythmia detection, with configuration 1 as the score for the intra-patient scheme and configuration 3 as the score for the inter-patient scheme.

IV. CONCLUSION

In this study, a set of CNN models were trained and tested to investigate its performance in a more standardized method. The model shows good results for the intra-patient scheme but not for the inter-patient scheme. There is a reduction of sensitivity and precision in the intra-patient scheme using a standardized method in this study compared to the original paper, indicating biased results caused by the oversampled test data in the original paper. Therefore, this paper proposes the standardized score for Acharya's CNN architecture in arrhythmia detection, with configuration 1 as the score for the intra-patient scheme and configuration 3 as the score for the inter-patient scheme.

REFERENCES

[1] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, 2001, doi: 10.1109/51.932724.

[2] D. Pandit, L. Zhang, C. Liu, S. Chattopadhyay, N. Aslam, and C. P. Lim, "A lightweight QRS detector for single lead ECG signals using a max-min difference algorithm," *Comput. Methods Programs Biomed.*, vol. 144, pp. 61–75, Jun. 2017, doi: 10.1016/j.cmpb.2017.02.028.

[3] M. Rakshit and S. Das, "An efficient wavelet-based automated R-peaks detection method using Hilbert transform," *Biocybern. Biomed. Eng.*, vol. 37, no. 3, pp. 566–577, Jan. 2017, doi: 10.1016/j.bbe.2017.02.002.

[4] D. Berwal, A. Kumar, and Y. Kumar, "Design of high performance QRS complex detector for wearable healthcare devices using biorthogonal spline wavelet transform," *ISA Trans.*, vol. 81, pp. 222–230, Oct. 2018, doi: 10.1016/j.isatra.2018.08.002.

[5] A. Sharma, S. Patidar, A. Upadhyay, and U. Rajendra Acharya, "Accurate tunable-Q wavelet transform based method for QRS complex detection," *Comput. Electr. Eng.*, vol. 75, pp. 101–111, May 2019, doi: 10.1016/j.compeleceng.2019.01.025.

[6] J. S. Lee, S. J. Lee, M. Choi, M. Seo, and S. W. Kim, "QRS detection method based on fully convolutional networks for capacitive electrocardiogram," *Expert Syst. Appl.*, vol. 134, pp. 66–78, Nov. 2019, doi: 10.1016/j.eswa.2019.05.033.

[7] A. A. Suárez-León, C. Varon, R. Willems, S. Van Huffel, and C. R. Vázquez-Seisdedos, "T-wave end detection using neural networks and Support Vector Machines," *Comput. Biol. Med.*, vol. 96, pp. 116–127, May 2018, doi: 10.1016/j.compbiomed.2018.02.020.

[8] A. Malali, S. Hiriyannaiah, G. M. Siddesh, K. G. Srinivasa, and N. T. Sanjay, "Supervised ECG wave segmentation using convolutional LSTM," *ICT Express*, vol. 6, no. 3, pp. 166–169, Sep. 2020, doi: 10.1016/j.icte.2020.04.004.

[9] E. J. da S. Luz, W. R. Schwartz, G. Cámara-Chávez, and D. Menotti, "ECG-based heartbeat classification for arrhythmia detection: A survey," *Comput. Methods Programs Biomed.*, vol. 127, pp. 144–164, 2016, doi: 10.1016/j.cmpb.2015.12.008.

[10] F. Ma, J. Zhang, W. Liang, and J. Xue, "Automated Classification of Atrial Fibrillation Using Artificial Neural Network for Wearable Devices," *Math. Probl. Eng.*, vol. 2020, 2020, doi: 10.1155/2020/9159158.

[11] G. Sannino and G. De Pietro, "A deep learning approach for ECG-based heartbeat classification for arrhythmia detection," *Futur. Gener. Comput. Syst.*, vol. 86, pp. 446–455, Sep. 2018, doi: 10.1016/j.future.2018.03.057.

[12] Y. Jung and H. Kim, "Detection of PVC by using a wavelet-based statistical ECG monitoring procedure," *Biomed. Signal Process. Control*, vol. 36, pp. 176–182, Jul. 2017, doi: 10.1016/j.bspc.2017.03.023.

[13] M. Sharma, R. S. Tan, and U. R. Acharya, "Automated heartbeat classification and detection of arrhythmia using optimal orthogonal wavelet filters," *Informatics Med. Unlocked*, vol. 16, p. 100221, Jan. 2019, doi: 10.1016/j.imu.2019.100221.

[14] N. Sinha and A. Das, "Automatic diagnosis of cardiac arrhythmias based on three stage feature fusion and classification model using DWT," *Biomed. Signal Process. Control*, vol. 62, p. 102066, Sep. 2020, doi: 10.1016/j.bspc.2020.102066.

[15] R. Arvanaghi, S. Daneshvar, H. Seyedarabi, and A. Goshvarpour, "Fusion of ECG and ABP signals based on wavelet transform for cardiac arrhythmias classification," *Comput. Methods Programs Biomed.*, vol. 151, pp. 71–78, Nov. 2017, doi: 10.1016/j.cmpb.2017.08.013.

[16] C. K. Jha and M. H. Kolekar, "Cardiac arrhythmia classification using tunable Q-wavelet transform based features and support vector machine classifier," *Biomed. Signal Process. Control*, vol. 59, p. 101875, May 2020, doi: 10.1016/j.bspc.2020.101875.

[17] H. Zhou *et al.*, "A Novel Cardiac Arrhythmias Detection Approach for Real-Time Ambulatory ECG Diagnosis," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 10, Oct. 2017, doi: 10.1142/S0218001417580046.

[18] E. Ramirez, P. Melin, and G. Prado-Arechiga, "Hybrid model based on neural networks, type-1 and type-2 fuzzy systems for 2-lead cardiac arrhythmia classification," *Expert Syst. Appl.*, vol. 126, pp. 295–307, Jul. 2019, doi: 10.1016/j.eswa.2019.02.035.

[19] P. Kora, K. Meenakshi, K. Swaraja, A. Rajani, and M. Kafiul Islam, "Detection of Cardiac arrhythmia using fuzzy logic," *Informatics Med. Unlocked*, vol. 17, p. 100257, Jan. 2019, doi: 10.1016/j.imu.2019.100257.

[20] M. Lee, T. G. Song, and J. H. Lee, "Heartbeat classification using local transform pattern feature and hybrid neural fuzzy-logic system based on self-organizing map," *Biomed. Signal Process. Control*, vol. 57, p. 101690, Mar. 2020, doi: 10.1016/j.bspc.2019.101690.

[21] M. Mohanty, S. Sahoo, P. Biswal, and S. Sabut, "Efficient classification of ventricular arrhythmias using feature selection and C4.5 classifier," *Biomed. Signal Process. Control*, vol. 44, pp. 200–208, Jul. 2018, doi: 10.1016/j.bspc.2018.04.005.

[22] K. Gajowniczek, I. Grzegorczyk, and T. Ząbkowski, "Reducing false arrhythmia alarms using different methods of probability and class assignment in random forest learning methods," *Sensors (Switzerland)*, vol. 19, no. 7, p. 1588, Apr. 2019, doi: 10.3390/s19071588.

[23] K. Gajowniczek, I. Grzegorczyk, T. Ząbkowski, and C. Bajaj, "Weighted random forests to improve arrhythmia classification," *Electron.*, vol. 9, no. 1, p. 99, Jan. 2020, doi: 10.3390/electronics9010099.

[24] W. Yang, Y. Si, D. Wang, and B. Guo, "Automatic recognition of arrhythmia based on principal component analysis network and linear support vector machine," *Comput. Biol. Med.*, vol. 101, pp. 22–32, Oct. 2018, doi: 10.1016/j.compbiomed.2018.08.003.

[25] K. N. V. P. S. Rajesh and R. Dhuli, "Classification of ECG heartbeats using nonlinear decomposition methods and support vector machine," *Comput. Biol. Med.*, vol. 87, pp. 271–284, Aug. 2017, doi: 10.1016/j.compbiomed.2017.06.006.

[26] M. A. Escalona-Morán, M. C. Soriano, I. Fischer, and C. R. Mirasso, "Electrocardiogram classification using reservoir computing with logistic regression," *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 3, pp. 892–898, 2015, doi: 10.1109/JBHI.2014.2332001.

[27] U. R. Acharya *et al.*, "A deep convolutional neural network model to classify heartbeats," *Comput. Biol. Med.*, vol. 89, pp. 389–396, Oct. 2017, doi: 10.1016/j.compbiomed.2017.08.022.

[28] S. Mousavi and F. Afghah, "Inter- and Intra- Patient ECG Heartbeat Classification for Arrhythmia Detection: A Sequence to Sequence Deep Learning Approach," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, May 2019, vol. 2019-May, pp. 1308–1312, doi: 10.1109/ICASSP.2019.8683140.

[29] A. Y. Hannun *et al.*, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nat. Med.*, vol. 25, no. 1, pp. 65–69, 2019, doi: 10.1038/s41591-018-0268-3.

[30] L. Guo, G. Sim, and B. Matuszewski, "Inter-patient ECG classification with convolutional and recurrent neural networks," *Biocybern. Biomed. Eng.*, vol. 39, no. 3, pp. 868–879, Jul. 2019, doi: 10.1016/j.bbe.2019.06.001.

[31] S. Saadatnejad, M. Oveisi, and M. Hashemi, "LSTM-Based ECG Classification for Continuous Monitoring on Personal Wearable Devices," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 2, pp. 515–523, Feb. 2020, doi: 10.1109/JBHI.2019.2911367.

[32] M. Alfaras, M. C. Soriano, and S. Ortín, "A Fast Machine Learning Model for ECG-Based Heartbeat Classification and Arrhythmia Detection," *Front. Phys.*, vol. 7, p. 103, Jul. 2019, doi: 10.3389/fphy.2019.00103.

[33] J. P. Allam, S. Samantray, and S. Ari, "SpEC: A system for patient specific ECG beat classification using deep residual network," *Biocybern. Biomed. Eng.*, vol. 40, no. 4, pp. 1446–1457, Oct. 2020, doi: 10.1016/j.bbe.2020.08.001.

[34] P. De Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1196–1206, Jul. 2004, doi: 10.1109/TBME.2004.827359.

[35] ISO/ANSI/AAMI, "ANSI/AAMI/ISO EC57: Testing and reporting performance results of cardiac rhythm and ST-segment measurement algorithms," 2008.

[36] G. De Lannoy, D. François, J. Delbeke, and M. Verleysen, "Weighted conditional random fields for supervised interpatient heartbeat classification," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 1, pp. 241–247, Jan. 2012, doi: 10.1109/TBME.2011.2171037.

[37] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[38] A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. M. Bossuyt, "The diagnostic odds ratio: A single indicator of test performance," *J. Clin. Epidemiol.*, vol. 56, no. 11, pp. 1129–1135, Nov. 2003, doi: 10.1016/S0895-4356(03)00177-X.

[39] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104.

[40] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.

[41] M. I. Rizqyawan, A. Munandar, M. F. Amri, R. Korio Utoro, and A. Pratondo, "Quantized Convolutional Neural Network toward Real-time Arrhythmia Detection in Edge Device," in *Proceeding - 2020 International Conference on Radar, Antenna, Microwave, Electronics and Telecommunications, ICRAMET 2020*, Nov. 2020, pp. 234–239, doi: 10.1109/ICRAMET51080.2020.9298667.