Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 859−880
DOI 10.47065/bits.v8i2.10916
https://ejurnal.seminar-id.com/index.php/bits

An Interval-Informed Hybrid CNN-LSTM for Ten-Category ECG Beat and Event Classification

I Putu Bagus Erix Wijaya[1,2], Arda Ardiyansyah[1], Satria Mandala[1,*]

[1] Humic Engineering (Humic Eng), School of Computing, Telkom University, Bandung, Indonesia
[2] Fakultas Sosial, Informatika Kesehatan, Universitas Bhakti Kencana, Bandung, Indonesia
Emails: [1]putu.bagus@bku.ac.id, [2]ardaardiansyah24@gmail.com , [3,*]satriamandala@telkomuniversity.ac.id
Correspondence Author Email: satriamandala@telkomuniversity.ac.id
Received: **06/08/2026**, *Revised*: **22/08/2026**, *Accepted*: **04/09/2026**, *Published*: **08/09/2026**

Abstract−Automated electrocardiogram (ECG) classification remains challenging because waveform-based deep-learning models capture rich morphology, whereas clinically relevant conduction and timing intervals are not always represented explicitly. This study investigates whether a compact interval-informed representation can retain useful discriminative information for ten ECG beat and event categories. MIT-BIH Arrhythmia Database recordings were filtered and divided into seven-second segments. RR, PR, and QT intervals and QRS width were summarized using minimum, maximum, mean, median, skewness, and kurtosis, yielding 24 features. To prevent information leakage, the original samples were partitioned before oversampling; feature scaling and SMOTE were applied only to training data, while validation and held-out test data retained their original distributions. Leakage-controlled five-fold cross-validation yielded 92.15% ± 0.43% accuracy and 86.78% ± 0.68% macro-F1. On the untouched 826-sample test set, the model achieved 92.62% accuracy, 87.49% macro-F1, and 92.68% weighted F1. The most frequent bidirectional errors occurred between Normal and Premature Ventricular Contraction beats. These results support the use of interval-statistical features as a compact representation for multicategory ECG analysis, while showing that rare-class performance and fiducial reliability remain limiting factors for broader clinical use.

Keywords: Electrocardiogram; Hybrid CNN-LSTM; Interval-Informed Representation; Arrhythmia Classification; Statistical Features

1. INTRODUCTION

An electrocardiogram (ECG) records the electrical activity of the heart and supports the identification of rhythm disorders, conduction abnormalities, and beat-morphology changes. During manual interpretation, clinicians examine waveform patterns, beat-to-beat regularity, conduction intervals, and QRS-complex duration. This process becomes increasingly demanding for long ambulatory recordings, as a single record may contain thousands of beats with substantial morphological variability. The task is further complicated when categories share similar characteristics, as in Normal and Premature Ventricular Contraction (PVC) beats, or when signal artifacts resemble QRS complexes. These challenges motivate the development of automated methods that can provide more consistent ECG beat classification.

The MIT-BIH Arrhythmia Database is widely used for arrhythmia-classification research [1]. Hybrid deep-learning studies have combined CNNs with recurrent networks to learn from raw ECG beats and rhythm segments [2], [3]. Other work has incorporated attention-based spatio-temporal modeling and imbalance-aware training on hybrid CNN-LSTM networks [4], [5]. Lightweight 1D CNN-LSTM architectures and ensemble multi-model approaches have also been proposed for beat-wise classification [6], [7]. Further hybrid CNN-LSTM formulations have targeted general arrhythmia classification tasks [8], [9]. Bi-LSTM variants and CNN-LSTM models applied to signal-based arrhythmia contemplation have likewise been reported [10], [11] Attention-based hierarchical dual-structured RNN designs with dilated CNNs represent a further extension of this line of work [12]. More recent studies have introduced lightweight, explainable CNN-LSTM models alongside incorporated CNN-LSTM frameworks for ECG classification [13], [14]. CNN-LSTM-SE and ensemble CNN-BiLSTM techniques have further advanced heartbeat classification performance [15], [16]. Optimization-enhanced CNN-LSTM networks using Aquila optimization and hybrid CNN-BLSTM architectures have also been explored [17], [18]. Explainable attention-based hybrid CNN-GRU models and DenseNet-ABiLSTM frameworks extend these designs toward interpretability and multimodal signals [19], [20]. Hybrid CNN-LSTM models for IoT-based ECG monitoring and RanA-optimized deep CNN-BiLSTM frameworks represent continued refinement of this direction [21], [22]. Most recently, lightweight parallel CNN-LSTM networks with augmentation and joint CNN-Bi-LSTM-Transformer architectures with SHAP explanations have been introduced [23], [24]. An attention-based CNN-GRU model with explainable artificial intelligence has similarly been proposed for arrhythmia classification [25]. Although these approaches demonstrate the effectiveness of end-to-end or waveform-oriented learning, their inputs generally retain substantially more amplitude-level information than an interval-based representation.

The present study addresses a different question: whether a compact set of physiologically interpretable interval descriptors can retain useful discriminative information for a broad multicategory task. RR characterizes beat-to-beat regularity, PR reflects atrioventricular conduction, QRS width represents ventricular depolarization duration, and QT encompasses ventricular depolarization and repolarization [26]. These quantities are complementary, but they do not reproduce the complete morphology available in a raw ECG waveform. Accordingly, the proposed 24-feature representation should be viewed as a structured dimensionality reduction of selected timing and conduction characteristics rather than as a replacement for all waveform information.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 859−880
DOI 10.47065/bits.v8i2.10916
https://ejurnal.seminar-id.com/index.php/bits

Several previous ECG-classification studies have evaluated hybrid deep-learning models using fewer heartbeat categories and waveform-oriented inputs [2], [6]. Similar limitations characterize other formulations that emphasize lightweight or explainable architectures over categorical breadth [9], [13]. This pattern also extends to more recent optimization-based and explainable designs [19], [25] Such formulations do not necessarily include paced beats, fusion beats, conduction abnormalities, junctional escape beats, blocked atrial premature beats, and noncardiac QRS-like events within a single classification task. In particular, an isolated QRS-like artifact can resemble a genuine beat and is therefore relevant in the present study as an explicit event category rather than only as preprocessing noise.

To investigate this compact-representation setting, ECG recordings are divided into seven-second segments. Prior to interval extraction, signal quality is commonly improved using filter-based denoising techniques such as Butterworth and Chebyshev filters [27]. Accurate delineation of the QRS complex additionally relies on robust R-peak detection methods, including visibility-graph-based approaches [28]. Six descriptors — minimum, maximum, mean, median, skewness, and kurtosis — are then calculated for each of the four interval families, producing 24 input features. Statistical and interval-derived ECG features have previously been used to summarize physiologically relevant signal characteristics [26], [29]. However, the contribution of the present study is not a claim that these features reproduce full ECG morphology or universally outperform end-to-end models; rather, the study evaluates how effectively an explicit multi-interval statistical representation supports ten-category ECG beat and event classification.

The contributions of this study are threefold. First, it constructs a compact representation that combines distributional descriptors of RR, PR, QT, and QRS intervals while explicitly acknowledging that full waveform morphology is not retained. Second, it formulates a ten-category task covering Normal beats, arrhythmias, conduction abnormalities, paced beats, fusion beats, and isolated QRS-like artifacts. Third, it evaluates a hybrid CNN-LSTM using a leakage-controlled protocol in which class balancing is restricted to training data, with macro-averaged and class-specific metrics used to complement aggregate accuracy.

2. RESEARCH METHODOLOGY

2.1 Research Stages

The research stages are illustrated in Figure 1. ECG signals from the MIT-BIH Arrhythmia Database are filtered, divided into seven-second segments, and processed to derive RR, PR, and QT intervals and QRS width. Six statistical descriptors are calculated for each interval family, producing 24 features per segment. The original feature samples are then partitioned into development and held-out test sets before any oversampling is performed.

To prevent information leakage, class balancing is performed only after partitioning. The held-out test set is never used for feature scaling, SMOTE generation, hyperparameter selection, or model fitting. Within five-fold cross-validation on the development set, the scaler and SMOTE are fitted exclusively to the training portion of each fold; the corresponding validation fold remains untouched. After hyperparameter selection, a final model is trained on the complete development set using training-only scaling and SMOTE and is evaluated once on the original-distribution held-out test set.

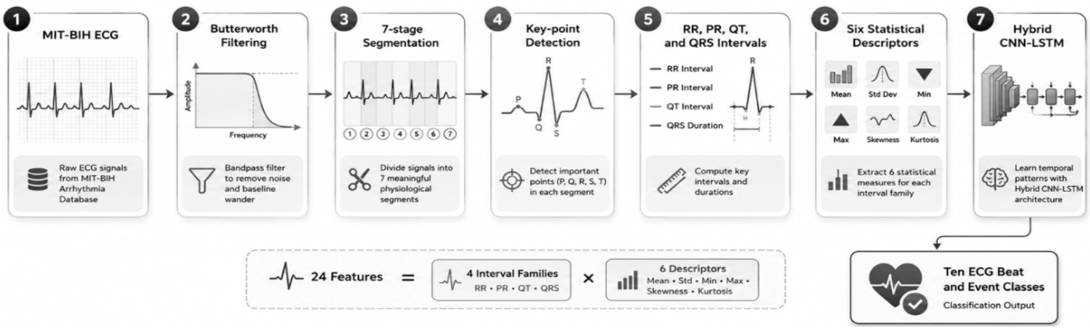

Figure 1. Stages of the interval-informed hybrid CNN-LSTM pipeline

As shown in Figure 1, the framework does not provide the complete ECG amplitude sequence directly to the model. Instead, each recording is transformed into an interval-statistical representation. This transformation is a central distinction of the study because the CNN-LSTM processes 24 physiologically meaningful descriptors rather than raw waveforms.

2.2 ECG Dataset and Beat/Event Categories

This study uses the MIT-BIH Arrhythmia Database available through PhysioNet [1]. The source database contains 48 two-channel ambulatory ECG recordings from 47 subjects. Each recording is approximately 30 minutes long and was sampled at 360 Hz. The database includes beat annotations used to construct the classification labels.

Segmentation and feature extraction produced 4,128 original labeled samples. Before applying SMOTE, these samples were stratified into a development set of 3,302 samples (80%) and a held-out test set of 826 samples (20%).

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 859−880
DOI 10.47065/bits.v8i2.10916
https://ejurnal.seminar-id.com/index.php/bits

The original class distribution and the resulting partition are reported in Table 4. SMOTE-generated samples are not counted as independent test observations and are created only inside the training data used for model development.

The ten ECG beat and event categories used in this study are summarized in Table 1.

Table 1. ECG beat and event categories

Code	Category
/	Paced Beat
A	Atrial Premature Contraction
F	Fusion of Ventricular and Normal Beat
L	Left Bundle Branch Block Beat
N	Normal Beat
R	Right Bundle Branch Block Beat
V	Premature Ventricular Contraction
j	Nodal or Junctional Escape Beat
x	Non-conducted P-wave or Blocked APB
—	Isolated QRS-like Artifact

As shown in Table 1, the ten categories comprise Paced Beat, Atrial Premature Contraction, Fusion of Ventricular and Normal Beat, Left Bundle Branch Block Beat, Normal Beat, Right Bundle Branch Block Beat, Premature Ventricular Contraction, Junctional Escape Beat, Non-conducted P-wave or Blocked APB, and Isolated QRS-like Artifact. The term "ECG beat and event categories" is used because Normal and artifact events are not arrhythmias.

2.3 ECG Preprocessing and Segmentation

Low-frequency components were attenuated using a fifth-order high-pass Butterworth filter. The Butterworth design was selected because it provides a smooth magnitude response within its passband [27]. The ECG recordings were then divided into seven-second segments. This duration was selected to include enough beats for interval-statistical calculation without producing an excessively long observation window.

Visibility-graph methods, specifically FastNVG and FastWHVG, were used for sample-accurate R-peak identification [28]. The detected R peaks served as temporal anchors for interval construction. RR intervals were calculated between consecutive R peaks. PR, QT, and QRS-width values were then obtained from the corresponding P-, Q-, S-, and T-related landmarks produced by the feature-extraction routine using conventional interval definitions: PR from P-wave onset to QRS onset, QRS width from QRS onset to QRS offset, and QT from QRS onset to the end of the T wave [26]. The same extraction procedure was applied consistently to all seven-second segments.

Fiducial localization was not evaluated as a separate prediction task in this study. Therefore, the reported classification performance reflects both the landmark-extraction stage and the downstream CNN-LSTM. Any displacement or missed detection of a required landmark changes the corresponding RR, PR, QT, or QRS measurement and propagates to the six derived descriptors (minimum, maximum, mean, median, skewness, and kurtosis). This dependency is considered when interpreting the results and limits direct generalization to signals for which fiducial delineation is unreliable.

2.4 Interval-Informed Statistical Feature Extraction

Each segment yielded four interval families: RR, PR, QT, and QRS width. Each family was summarized using the minimum, maximum, mean, median, skewness, and kurtosis [26], [29]. The mean describes central tendency and was calculated using Equation (1). The median provides a central measure that is less sensitive to extreme values, whereas the minimum and maximum define the range of interval variation within a segment.

$$\mu X = (1/n)\,\Sigma_{i=1}^{n}\, x_i \tag{1}$$

Skewness in Equation (2) quantifies the asymmetry of an interval distribution relative to its mean. This descriptor complements measures of central tendency because two segments may have similar means but different distributional shapes.

$$\mathrm{Skew}(X) = (1/n)\,\Sigma_{i=1}^{n}\,[(x_i - \mu X)/\sigma X]^3 \tag{2}$$

Kurtosis was calculated using Equation (3), with a subtraction of three to obtain excess kurtosis relative to a normal distribution. This descriptor characterizes the concentration of the distribution and the presence of extreme values.

$$\mathrm{Kurt}(X) = (1/n)\,\Sigma_{i=1}^{n}\,[(x_i - \mu X)/\sigma X]^4 - 3 \tag{3}$$

Because four interval families and six descriptors were used, each segment was represented by 24 features. All features were standardized using the Z-score in Equation (4) so that differences in feature scale would not dominate model optimization.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 859−880
DOI 10.47065/bits.v8i2.10916
https://ejurnal.seminar-id.com/index.php/bits

$$z = (x - \mu)/\sigma \quad (4)$$

The original class distribution was highly imbalanced. To avoid leakage, Z-score parameters were estimated only from the training data and then applied unchanged to the corresponding validation or held-out test data. SMOTE [30] was subsequently fitted only to the scaled training portion of each cross-validation fold. No synthetic sample was generated from, or introduced into, a validation fold or the held-out test set. For final model fitting, the same procedure was applied to the complete development set after hyperparameter selection, while the 826-sample test set remained untouched.

2.5 Hybrid CNN-LSTM Framework

The model configuration is presented in Table 2. The input consists of 24 statistical features arranged as a one-dimensional representation. Two Conv1D layers learn local relationships among neighboring descriptors. The first layer uses 128 filters with a kernel size of three, and the second uses 256 filters with the same kernel size. Both Conv1D layers employ ReLU activation and are followed by MaxPooling1D with a pool size of two. The convolution operation is expressed in Equation (5).

$$z_k(t) = \varphi(\Sigma c\ \Sigma_{\tau=0}^{K-1}\ w_k, c, \tau\ x\_c(t - \tau) + b_k) \quad (5)$$

The convolutional output is flattened and passed to a 512-unit Dense layer with ReLU activation and Batch Normalization. The representation is then reshaped into three dimensions and processed by three LSTM layers, each containing 128 units. The first two LSTM layers use return_sequences=True, whereas the final LSTM produces a single representation vector. Each LSTM layer is followed by Dropout at a rate of 0.5. The hidden-state and cell-state update mechanism is summarized in Equation (6). In this architecture, the LSTM block performs a gated nonlinear transformation of the latent representation rather than processing a sequence of raw ECG waveforms.

$$(h_t, c_t) = \mathrm{LSTM}(z_t, h_{t-1}, c_{t-1}) \quad (6)$$

The final representation is passed to a 512-unit Dense layer, Batch Normalization, and a ten-neuron output layer with Softmax activation. The probability of each category is calculated using Equation (7). The model is optimized using Adam with a learning rate of 0.001 and sparse categorical cross-entropy as the loss function.

$$\hat{p}_k = \exp(o_k) / \Sigma_{j=1}^{c} \exp(o_j) \quad (7)$$

Table 2. Configuration of the interval-informed hybrid CNN-LSTM

Stage	Layer	Configuration
Input	Feature representation	24 statistical descriptors from RR, PR, QT, and QRS
1	Conv1D	128 filters; kernel size 3; ReLU activation
2	MaxPooling1D	Pool size 2
3	Conv1D	256 filters; kernel size 3; ReLU activation
4	MaxPooling1D	Pool size 2
5	Flatten	Flattens the convolutional output
6	Dense	512 units; ReLU
7	BatchNormalization	Normalizes latent activations
8	Reshape	Forms a three-dimensional input for LSTM
9	LSTM	128 units; return_sequences=True
10	Dropout	0.5
11	LSTM	128 units; return_sequences=True
12	Dropout	0.5
13	LSTM	128 units
14	Dropout	0.5
15	Flatten	Flattens the LSTM output
16	Dense + BatchNormalization	512 units; ReLU
Output	Dense-Softmax	10 classes
Optimization	Adam	Learning rate 0.001; sparse categorical cross-entropy

As shown in Table 2, the architecture first extracts local relationships among the 24 interval-statistical descriptors through two Conv1D blocks, then transforms the learned representation through a Dense stage and three LSTM layers before ten-class Softmax classification.

2.6 Model Development and Hyperparameter Setting

The 3,302-sample development set was evaluated using StratifiedKFold with five folds, shuffle=True, and random_state=42. For every fold, preprocessing parameters and SMOTE were fitted only on the four training partitions, and the remaining partition retained its original class distribution for validation. A new model was initialized for each fold. Hyperparameters were selected from the mean validation performance across folds rather than from a single best fold.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 859−880
DOI 10.47065/bits.v8i2.10916
https://ejurnal.seminar-id.com/index.php/bits

The search evaluated batch sizes of 32, 64, and 128 in combination with 25, 50, 75, and 100 epochs. The configuration with the highest mean validation accuracy was selected. A final CNN-LSTM was then reinitialized and trained on the complete development set using the selected configuration, with scaling and SMOTE restricted to that development data, before a single evaluation on the untouched test set. Predictions were obtained from Softmax probabilities and the final class was selected using the argmax operation in Equation (12).

2.7 Evaluation Metrics

Performance was assessed using accuracy, precision, sensitivity or recall, and F1-score. Accuracy in Equation (8) represents the proportion of all samples classified correctly.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{8}$$

Precision in Equation (9) is the proportion of correct predictions for a category among all samples predicted as that category.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{9}$$

Sensitivity, or recall, is calculated using Equation (10) and represents the proportion of samples from a category that are correctly identified by the model.

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \tag{10}$$

F1-score in Equation (11) is the harmonic mean of precision and recall and therefore summarizes performance while accounting for both types of error.

$$\text{F1} = 2(\text{Precision} \times \text{Sensitivity})/(\text{Precision} + \text{Sensitivity}) \tag{11}$$

For multiclass classification, performance is reported using class-specific precision, recall, and F1-score together with macro and weighted averages. Macro-F1 is emphasized because it assigns equal importance to each category and is therefore less dominated by the large Normal and PVC classes. Weighted averages are retained as secondary summaries of sample-level performance. The confusion matrix is used to show the complete distribution of errors between true and predicted categories.

$$\hat{y}_i = \arg\max_k \hat{p}_{i,k}, \; k \in \{1, \dots, 10\} \tag{12}$$

3. RESULT AND DISCUSSION

3.1 Preprocessing and Class Distribution

The filtering stage yielded a Mean Squared Error (MSE) of 0.12 and a Root Mean Squared Error (RMSE) of 0.34, as reported in Table 3. These values quantify the difference between the signals compared during filtering. Because the study did not use a noise-free reference signal or compare multiple denoising methods, these measures are treated as descriptive preprocessing results and are not used to claim that Butterworth filtering is superior.

Table 3. Signal measurements after denoising

Metric	Value
Mean Squared Error	0.12
Root Mean Squared Error	0.34

After interval construction, six statistical descriptors were extracted from each of the four interval families. Table 4 reports the original class distribution and the leakage-controlled 80/20 development/test partition. No global "after SMOTE" sample count is reported because synthetic observations are generated only within each training partition and are never included in the validation or test distributions.

Table 4. Original class distribution and leakage-controlled data partition

Code	Original total	Development (80%)	Held-out test (20%)
/	342	274	68
A	302	242	60
F	58	46	12
L	339	271	68
N	1,603	1,282	321
R	315	252	63
V	1,094	875	219
j	27	22	5
x	22	18	4
—	26	20	6

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 859−880
DOI 10.47065/bits.v8i2.10916
https://ejurnal.seminar-id.com/index.php/bits

Code	Original total	Development (80%)	Held-out test (20%)
Total	4,128	3,302	826

Table 4 shows the substantial natural class imbalance: Normal and PVC contribute most original samples, whereas Fusion Beat, Junctional Escape Beat, Blocked APB, and Isolated QRS-like Artifact are rare. The held-out test set preserves this original distribution. SMOTE is used only within training data to improve exposure to minority categories; consequently, performance on rare test categories is interpreted with caution because only a small number of original observations are available.

3.2 Hyperparameter Selection

Leakage-controlled hyperparameter-tuning results are presented in Table 5. Validation accuracy generally increased with training duration across the three batch sizes. For a batch size of 32, validation accuracy increased from 88.45% at 25 epochs to 92.15% at 100 epochs. The corresponding highest accuracies for batch sizes of 64 and 128 were 91.68% and 91.05%, respectively.

Table 5. Batch-size and epoch tuning results

Batch size	Epoch	Validation accuracy (%)
32	25	88.45
32	50	90.72
32	75	91.85
32	100	92.15
64	25	87.90
64	50	89.95
64	75	91.10
64	100	91.68
128	25	86.50
128	50	89.15
128	75	90.45
128	100	91.05

As shown in Table 5, a batch size of 32 and 100 epochs produced the highest validation accuracy of 92.15% within the evaluated search space and was therefore selected for subsequent evaluation. At this batch size, the improvement from 75 to 100 epochs was only 0.30 percentage points, indicating that validation performance was approaching a plateau.

3.3 Five-Fold Validation Performance

The selected configuration was evaluated using leakage-controlled stratified five-fold cross-validation on the development set. In each fold, feature scaling and SMOTE were fitted exclusively to the training partition, while the validation partition retained its original class distribution. As reported in Table 6, validation accuracy ranged from 91.70% to 92.80%, while macro-F1 ranged from 85.95% to 87.70%.

Table 6. Five-fold validation performance of the tuned CNN-LSTM

Fold	Training accuracy (%)	Accuracy (%)	Macro precision (%)	Macro recall (%)	Macro F1-score (%)
1	96.10	91.70	86.10	85.80	85.95
2	96.85	92.35	87.45	86.90	87.17
3	96.40	91.95	86.80	86.50	86.65
4	97.10	92.80	87.90	87.50	87.70
5	96.55	91.95	86.65	86.20	86.42
Mean ± SD	96.60 ± 0.37	92.15 ± 0.43	86.98 ± 0.69	86.58 ± 0.65	86.78 ± 0.68

Table 6 shows a mean validation accuracy of 92.15% ± 0.43%, macro precision of 86.98% ± 0.69%, macro recall of 86.58% ± 0.65%, and macro-F1 of 86.78% ± 0.68%. The relatively narrow variation across folds indicates that performance remained consistent across different development-set partitions. The gap between the mean training accuracy of 96.60% and validation accuracy of 92.15% also indicates a measurable generalization gap.

3.4 Training Behavior

The training and validation curves for the selected CNN-LSTM configuration are shown in Figure 2. Training accuracy progressively increased toward approximately 96–97%, whereas validation accuracy stabilized near 92%. A similar separation was observed in the loss curves: training loss continued to decrease after validation loss had largely plateaued.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 859−880
DOI 10.47065/bits.v8i2.10916
https://ejurnal.seminar-id.com/index.php/bits

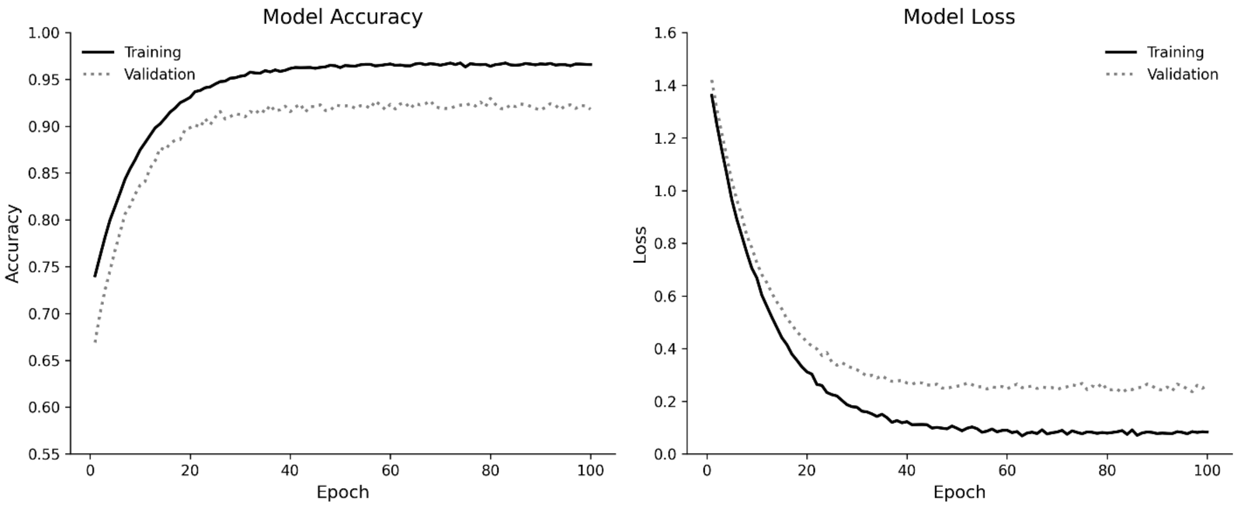

Figure 2. Training and validation curves of the tuned CNN-LSTM

As shown in Figure 2, the model converged but exhibited a persistent training–validation gap. Validation accuracy remained broadly stable after approximately the middle stage of training, while additional epochs continued to reduce training loss without a corresponding improvement in validation performance. This behavior indicates moderate overfitting and is consistent with the limited gain observed between 75 and 100 epochs in Table 5.

3.5 Unseen Test Performance and Class-Wise Error Analysis

After hyperparameter selection, a final CNN-LSTM model was reinitialized and trained using the complete development set, with feature scaling and SMOTE restricted to the training data, and was evaluated once on the untouched 826-sample test set. The model correctly classified 765 samples, corresponding to an accuracy of 92.62%. Macro precision, recall, and F1-score were 87.82%, 88.12%, and 87.49%, respectively, while the weighted F1-score was 92.68%. Figure 3 presents the corresponding confusion matrix under the original test-set class distribution.

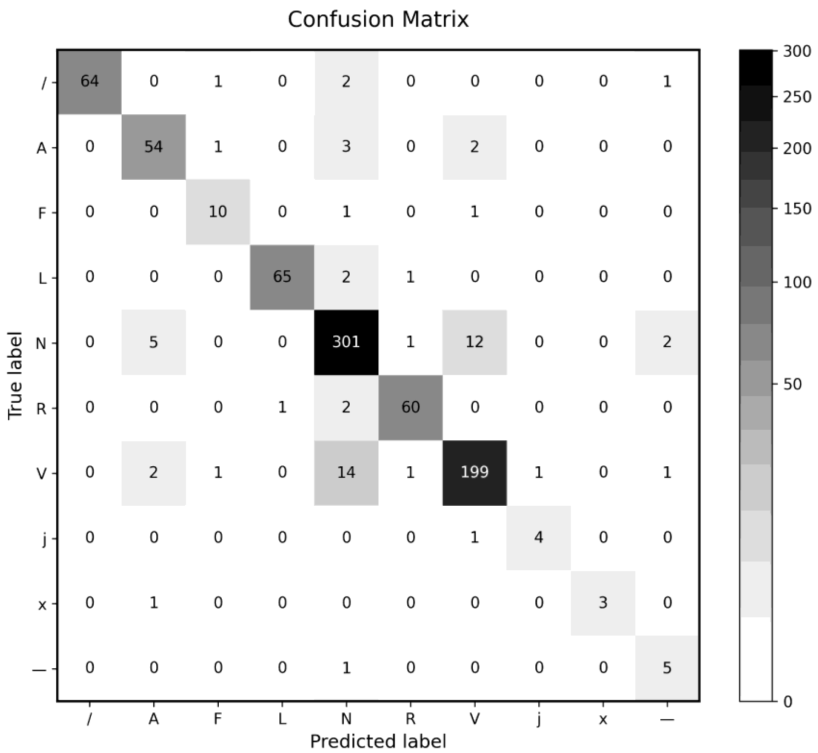

Figure 3. Confusion matrix of the hybrid CNN-LSTM on unseen test data

As shown in Figure 3, predictions were concentrated along the main diagonal, although errors were not distributed uniformly across categories. The most prominent bidirectional confusion occurred between Normal and PVC, with 12 Normal samples classified as PVC and 14 PVC samples classified as Normal. Because the held-out test set retained its original distribution, the confusion matrix also reveals the uncertainty associated with categories

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 859−880
DOI 10.47065/bits.v8i2.10916
https://ejurnal.seminar-id.com/index.php/bits

represented by only a small number of original samples. Class-specific, macro-average, and weighted-average metrics are reported in Table 7.

Table 7. Class-wise performance on unseen test data

Category	Precision (%)	Recall (%)	F1-score (%)
Paced Beat /	100.00	94.12	96.97
Atrial Premature Contraction A	87.10	90.00	88.52
Fusion Beat F	76.92	83.33	80.00
Left Bundle Branch Block L	98.48	95.59	97.01
Normal N	92.33	93.77	93.04
Right Bundle Branch Block R	95.24	95.24	95.24
Premature Ventricular Contraction V	92.56	90.87	91.71
Junctional Escape Beat j	80.00	80.00	80.00
Blocked APB x	100.00	75.00	85.71
Isolated QRS-like Artifact	55.56	83.33	66.67
Macro-average	87.82	88.12	87.49
Weighted-average	92.84	92.62	92.68

As shown in Table 7, class-specific performance was heterogeneous. Normal and PVC achieved F1-scores of 93.04% and 91.71%, respectively, while Paced Beat, LBBB, and RBBB each achieved F1-scores above 95%. In contrast, the rare Fusion Beat, Junctional Escape Beat, Blocked APB, and Isolated QRS-like Artifact categories achieved F1-scores of 80.00%, 80.00%, 85.71%, and 66.67%, respectively. Consequently, the macro-F1 of 87.49% was lower than the weighted F1 of 92.68%, demonstrating the importance of reporting class-balanced metrics rather than relying only on aggregate or weighted performance.

The Normal–PVC pair remained the dominant source of bidirectional error. The interval-statistical representation summarizes timing, conduction, and within-segment interval variability but does not retain the complete amplitude morphology available in the original ECG waveform. Therefore, although Figure 3 clearly identifies where these errors occur, the present experiment does not establish a definitive mechanism for each misclassification.

For the rare categories, the held-out sample counts were small: 12 Fusion Beats, 5 Junctional Escape Beats, 4 Blocked APB events, and 6 Isolated QRS-like Artifacts. Their class-specific scores should therefore be interpreted as preliminary estimates rather than evidence of robust generalization across broader clinical populations.

3.6 Comparison with Existing Hybrid Deep Learning Studies

Table 8 compares the proposed approach with previous hybrid deep-learning studies in terms of dataset, input representation, classification scope, validation protocol, and principal result. The comparison is contextual rather than a direct performance ranking because the studies use different class definitions, input forms, imbalance strategies, and partitioning protocols.

Table 8. Comparison with hybrid deep-learning studies for ECG classification

Study	Dataset / categories	Input	Model / validation	Main result	Position relative to this study
Xu et al. [2]	MIT-BIH; 5 classes	Raw ECG beat	CNN-BiLSTM; transfer evaluation	Acc. 95.90%; F1 NR	Raw input; fewer classes
Obeidat and Alqudah [6]	MIT-BIH; 6 classes	Segmented ECG beat	Lightweight 1D CNN-LSTM; train/test	Acc. 98.22%; Sens. 98.23%	Does not use statistics from four interval families
Rahul and Sharma [9]	MIT-BIH; 5 AAMI classes	ECG + preprocessing	1D CNN-BiLSTM; 10-fold CV	Acc. 99.41%; F1 NR	Fewer classes; waveform input
HARDC [12]	MIT-BIH; multiclass	Raw ECG + CGAN	Dilated CNN + BiGRU/BiLSTM + attention	Acc. 99.60%; F1 98.21%	More complex architecture and augmentation
Alamatsaz et al. [13]	MIT-BIH + LTAFDB; 8 arrhythmias + Normal	Raw ECG	Lightweight CNN-LSTM + SHAP	Acc. 98.24%; F1 NR	Waveform-level XAI; no PR/QT/QRS statistics
Sun et al. [15]	MIT-BIH; multiclass	ECG after EEMD	CNN-LSTM-SE	Acc. 98.50%; F1 >98%	Uses channel attention

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 859−880
DOI 10.47065/bits.v8i2.10916
https://ejurnal.seminar-id.com/index.php/bits

Study	Dataset / categories	Input	Model / validation	Main result	Position relative to this study
Ardiyansyah and Mandala [25]	MIT-BIH; multiclass	ECG features	Attention-CNN-GRU + LRP	Not reported in a directly comparable format	Previous work focused on attention and XAI
Ba Mahel et al. [19]	Public ECG; 5 classes	Raw ECG	Attention CNN-GRU + XAI	Acc. 99.16%; F1 99.16%	Fewer classes; attention-based
Present study	MIT-BIH; 10 categories	24 RR, PR, QT, and QRS statistics	2 Conv1D + 3 LSTM; leakage-controlled 5-fold development + untouched test	Test Acc. 92.62%; macro-F1 87.49%	Compact interval-statistical representation; ten-category scope; no claim of superiority

As shown in Table 8, several previous studies reported accuracies above 98%. However, these results were obtained using different numbers of categories, input representations, imbalance strategies, and validation protocols and therefore should not be interpreted as direct performance rankings. The present study achieved 92.62% test accuracy and 87.49% macro-F1 using only 24 statistical descriptors derived from RR, PR, QT, and QRS intervals. Its contribution is therefore the evaluation of a compact and explicit interval-statistical representation for a ten-category task rather than superior accuracy over waveform-based approaches.

3.7 Discussion and Limitations

The leakage-controlled results indicate that statistical descriptors derived from RR, PR, QT, and QRS intervals retain useful discriminative information for ten-category ECG beat and event classification. Five-fold validation produced an accuracy of 92.15% ± 0.43% and macro-F1 of 86.78% ± 0.68%, while evaluation on the untouched test set yielded 92.62% accuracy and 87.49% macro-F1. The close agreement between cross-validation and held-out test performance suggests that the observed classification capability is reasonably consistent across independent partitions.

The six descriptors summarize the center, range, asymmetry, and tail behavior of each interval distribution within a seven-second segment. This gives the model structured information about interval variability without exposing the full amplitude sequence. The representation is compact and physiologically interpretable, but its dimensional efficiency necessarily removes morphology that may be useful for difficult pairs such as Normal and PVC. The CNN-LSTM therefore learns relationships among interval-derived descriptors rather than directly learning waveform morphology. This distinction also clarifies the role of the LSTM block: it performs gated transformations of a latent feature representation and should not be interpreted as processing the original ECG temporal sequence.

The class-wise results show that aggregate accuracy alone does not fully characterize model performance. Although the weighted F1-score reached 92.68%, macro-F1 was 87.49% because performance was substantially lower for several rare categories. In particular, Isolated QRS-like Artifact achieved an F1-score of 66.67%, while Fusion Beat and Junctional Escape Beat each achieved 80.00%. These categories contained only 4–12 observations in the held-out test set, and their class-specific metrics should therefore be interpreted cautiously.

Normal and PVC represented the most prominent bidirectional confusion, with 12 Normal samples classified as PVC and 14 PVC samples classified as Normal. Both classes nevertheless maintained F1-scores above 90%. Because the proposed model operates on interval-statistical features rather than full waveform morphology, the present results identify the location of these errors but do not establish a definitive morphological mechanism for individual misclassifications.

A practical strength of the approach is its compact 24-feature input representation. However, input compactness does not make the entire classifier computationally trivial. Based on the stated architecture and a 24×1 input, the CNN-LSTM contains approximately 1.09 million parameters, corresponding to about 4.2 MB of FP32 weights. This scale is manageable on general-purpose computing platforms but remains non-trivial for low-power embedded deployment, and device-specific latency, memory, and energy consumption have not yet been measured.

Practical performance also depends on reliable ECG fiducial-point delineation. FastNVG/FastWHVG supplies the R-peak anchors, while PR, QT, and QRS-width estimation additionally depends on the P-, Q-, S-, and T-related landmarks used by the feature-extraction routine. Any localization error propagates directly into the interval measurements and the six statistical descriptors presented to the classifier. Because fiducial accuracy was not benchmarked as a separate endpoint, broader application requires validated delineation and quality-control procedures for segments in which landmark extraction is unreliable.

The study has several remaining limitations. First, the training curves indicate moderate overfitting. Second, the dataset contains very few original observations for several minority categories, making class-specific estimates unstable even when SMOTE is restricted to training data. Third, evaluation is limited to the MIT-BIH Arrhythmia Database without external validation. Fourth, a seven-second segment may contain heterogeneous beat patterns, while the compact representation does not retain complete waveform morphology. Fifth, classification remains dependent on accurate fiducial-point localization, which was not evaluated independently from the downstream model. Finally,

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 859−880
DOI 10.47065/bits.v8i2.10916
https://ejurnal.seminar-id.com/index.php/bits

computational feasibility on portable medical hardware has not yet been established. These limitations define the next steps: external validation, validated fiducial delineation, and deployment-oriented optimization and benchmarking.

4. CONCLUSION

This study investigated an interval-informed hybrid CNN-LSTM for ten-category ECG beat and event classification using 24 statistical features derived from RR, PR, QT, and QRS intervals. To prevent information leakage, the original samples were partitioned before oversampling, with feature scaling and SMOTE restricted to training data while validation and held-out test sets retained their original distributions. Leakage-controlled five-fold cross-validation achieved 92.15% ± 0.43% accuracy and 86.78% ± 0.68% macro-F1. On the untouched 826-sample test set, the model achieved 92.62% accuracy, 87.49% macro-F1, and 92.68% weighted F1. Normal and PVC remained the most frequent bidirectional confusion, while performance for several rare categories was less stable because of their small original sample counts. These findings demonstrate that a compact interval-statistical representation can retain useful discriminative information for multicategory ECG classification, although it does not preserve complete waveform morphology. The observed training–validation gap, dependence on reliable fiducial-point delineation, limited minority-class support, absence of external validation, and computational requirements of the approximately 1.09-million-parameter model remain important limitations. Future work should focus on external validation, validated ECG delineation, and deployment-oriented optimization and evaluation.

5. DECLARATIONS

5.1 Author Contributions

Conceptualization: S.M.D., I.P.W. and A.A.;
Methodology: S.M.D., and I.P.W.;
Software: S.M.D., and A.A.;
Validation: S.M.D., and I.P.W.;
Formal Analysis: S.M.D., I.P.W., and AA;
Investigation: S.M.D., I.P.W., and A.A.;
Resources: I.P.W., and A.A.;
Data Curation: S.M.D., I.P.W., and A.A.;
Writing Original Draft Preparation: S.M.D., and I.P.W.;
Writing Review and Editing: S.M.D., and I.P.W.;
Visualization: I.P.W., and A.A.;
All authors have read and agreed to the published version of the manuscript.

5.2 Data Availability

The data presented in this study are available on request from the corresponding author.

5.3 Funding

This work was financially supported by the Directorate of Research, Technology, and Community Service, Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia, under Grant No. 352/PNLT2/PPM/2023 and Agreement No. 003/SP2H/RT-MONO/LL4/2023. The authors gratefully acknowledge Telkom University for the academic support and research facilities provided. The authors also acknowledge PhysioNet and the maintainers of the MIT-BIH Arrhythmia Database for making the ECG recordings and annotations available for research purposes.

5.4 Institutional Review Board Statement

Not applicable.

5.5 Informed Consent Statement

Not applicable.

5.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

5.7 Generative AI and AI-assisted Technologies in The Writing Process

The authors used Generative AI to improve the writing clarity of this paper. They reviewed and edited the AI-assisted content and took full responsibility for the final publication.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 859−880
DOI 10.47065/bits.v8i2.10916
https://ejurnal.seminar-id.com/index.php/bits

REFERENCES

[1] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH Arrhythmia Database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001, doi: 10.1109/51.932724.
[2] X. Xu, S. Jeong, and J. Li, "Interpretation of Electrocardiogram (ECG) rhythm by combined CNN and BiLSTM," *IEEE Access*, vol. 8, pp. 125380–125388, 2020, doi: 10.1109/ACCESS.2020.3006707.
[3] J. Zhang, A. Liu, M. Gao, X. Chen, X. Zhang, and X. Chen, "ECG-based multi-class arrhythmia detection using spatio-temporal attention-based convolutional recurrent neural network," *Artif. Intell. Med.*, vol. 106, p. 101856, 2020, doi: 10.1016/j.artmed.2020.101856.
[4] G. Petmezas and others, "Automated atrial fibrillation detection using a hybrid CNN-LSTM network on imbalanced ECG datasets," *Biomed. Signal Process. Control*, vol. 63, p. 102194, 2021, doi: 10.1016/j.bspc.2020.102194.
[5] M. Jiang *et al.*, "HADLN: Hybrid attention-based deep learning network for automated arrhythmia classification," *Front. Physiol.*, vol. 12, p. 683025, 2021, doi: 10.3389/fphys.2021.683025.
[6] Y. M. Obeidat and A. M. Alqudah, "A hybrid lightweight 1D CNN-LSTM architecture for automated ECG beat-wise classification," *Traitement du Signal*, vol. 38, no. 5, pp. 1281–1291, 2021, doi: 10.18280/ts.380503.
[7] E. Essa and X. Xie, "An ensemble of deep learning-based multi-model for ECG heartbeats arrhythmia classification," *IEEE Access*, vol. 9, pp. 103452–103464, 2021, doi: 10.1109/ACCESS.2021.3098986.
[8] P. Madan, V. Singh, D. P. Singh, M. Diwakar, B. Pant, and A. Kishor, "A hybrid deep learning approach for ECG-based arrhythmia classification," *Bioengineering*, vol. 9, no. 4, p. 152, 2022, doi: 10.3390/bioengineering9040152.
[9] J. Rahul and L. D. Sharma, "Automatic cardiac arrhythmia classification based on hybrid 1-D CNN and Bi-LSTM model," *Biocybern. Biomed. Eng.*, vol. 42, no. 1, pp. 312–324, 2022, doi: 10.1016/j.bbe.2022.02.006.
[10] S. Sowmya and D. Jose, "Contemplate on ECG signals and classification of arrhythmia signals using CNN-LSTM deep learning model," *Measurement: Sensors*, vol. 24, p. 100558, 2022, doi: 10.1016/j.measen.2022.100558.
[11] T. H. Rafi and Y. W. Ko, "HeartNet: Self multihead attention mechanism via convolutional network with adversarial data synthesis for ECG-based arrhythmia classification," *IEEE Access*, vol. 10, pp. 100501–100512, 2022, doi: 10.1109/ACCESS.2022.3206431.
[12] M. S. Islam and others, "HARDC: A novel ECG-based heartbeat classification method to detect arrhythmia using hierarchical attention based dual structured RNN with dilated CNN," *Neural Networks*, vol. 162, pp. 271–287, 2023, doi: 10.1016/j.neunet.2023.03.004.
[13] N. Alamatsaz, L. Tabatabaei, M. Yazdchi, H. Payan, N. Alamatsaz, and F. Nasimi, "A lightweight hybrid CNN-LSTM explainable model for ECG-based arrhythmia detection," *Biomed. Signal Process. Control*, vol. 90, p. 105884, 2024, doi: 10.1016/j.bspc.2023.105884.
[14] A. Eleyan and E. Alboghbaish, "Electrocardiogram signals classification using deep-learning-based incorporated convolutional neural network and long short-term memory framework," *Computers*, vol. 13, no. 2, p. 55, 2024, doi: 10.3390/computers13020055.
[15] A. Sun, W. Hong, J. Li, and J. Mao, "An arrhythmia classification model based on a CNN-LSTM-SE algorithm," *Sensors*, vol. 24, no. 19, p. 6306, 2024, doi: 10.3390/s24196306.
[16] N. Sharma, R. K. Sunkaria, and A. Kaur, "Electrocardiogram heartbeat classification using machine learning and ensemble convolutional neural network-bidirectional long short-term memory technique," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 6, pp. 2816–2827, 2024, doi: 10.1109/TAI.2023.3324627.
[17] H. Anu, S. Rathnakara, and S. Mallikarjunaswamy, "Enhanced ECG signal classification with CNN-LSTM networks using Aquila optimization," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23461–23466, 2025, doi: 10.48084/etasr.10492.
[18] Y. Ye, K. Chipusu, M. A. Ashraf, B. Ding, Y. Huang, and J. Huang, "Hybrid CNN-BLSTM architecture for classification and detection of arrhythmia in ECG signals," *Sci. Rep.*, vol. 15, p. 34510, 2025, doi: 10.1038/s41598-025-17671-1.
[19] A. S. Ba Mahel, F. M. G. Alotaibi, Z. M. Lonseko, and N.-N. Rao, "XABH-CNN-GRU: Explainable attention-based hybrid CNN-GRU model for accurate identification of common arrhythmias," *Journal of Electronic Science and Technology*, vol. 23, no. 3, p. 100322, 2025, doi: 10.1016/j.jnlest.2025.100322.
[20] K. Saranya, U. Karthikeyan, A. S. Kumar, A. O. Salau, and T. T. Tin, "DenseNet-ABiLSTM: Revolutionizing multiclass arrhythmia detection and classification using hybrid deep learning approach leveraging PPG signals," *International Journal of Computational Intelligence Systems*, vol. 18, p. 33, 2025, doi: 10.1007/s44196-025-00765-z.
[21] Y. Wang and S. Rani, "Hybrid CNN-LSTM model for ECG-based arrhythmia detection in Internet of Medical Things," *Discover Internet of Things*, vol. 6, no. 1, p. 7, 2026, doi: 10.1007/s43926-025-00131-7.
[22] G. S. Manivannan and others, "An efficient deep CNN based BiLSTM framework with RanA optimization for accurate cardiac arrhythmia detection," *Sci. Rep.*, vol. 16, p. 7156, 2026, doi: 10.1038/s41598-026-38227-x.
[23] E. A. Dinsel and H. Kodaz, "A lightweight and robust parallel CNN-LSTM network with random crop augmentation for single-lead raw ECG arrhythmia classification," *Engineering Science and Technology, an International Journal*, vol. 79, p. 102417, 2026, doi: 10.1016/j.jestch.2026.102417.
[24] M. T. Al-Bairmani, M. Yazdchi, and F. Nasimi, "A joint CNN-Bi-LSTM-Transformer architecture with SHAP explanations for multi-label arrhythmia detection from 12-lead ECGs," *Sci. Rep.*, vol. 16, p. 11123, 2026, doi: 10.1038/s41598-026-40925-5.
[25] A. Ardiyansyah and S. Mandala, "Accurate model for arrhythmia classification using hybrid attention-based convolutional neural network and gated recurrent units with explainable artificial intelligence," in *Proc. 12th International Conference on Information and Communication Technology (ICoICT)*, 2024, pp. 57–64. doi: 10.1109/ICoICT61617.2024.10698622.
[26] A. Singh and S. Krishnan, "ECG signal feature extraction trends in methods and applications," *Biomed. Eng. Online*, vol. 22, p. 22, 2023, doi: 10.1186/s12938-023-01075-1.
[27] N. Rastogi and R. Mehra, "Analysis of Butterworth and Chebyshev filters for ECG denoising using wavelets," *IOSR Journal of Electronics and Communication Engineering*, vol. 6, no. 6, pp. 37–44, 2013, doi: 10.9790/2834-0663744.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 859−880
DOI 10.47065/bits.v8i2.10916
https://ejurnal.seminar-id.com/index.php/bits

[28] J. Emrich, T. Koka, S. Wirth, and M. Muma, "Accelerated sample-accurate R-peak detectors based on visibility graphs," in *Proc. 31st European Signal Processing Conference (EUSIPCO)*, Helsinki, Finland, 2023, pp. 1090–1094. doi: 10.23919/EUSIPCO58844.2023.10290007.

[29] T. N. Alotaiby, S. R. Alrshoud, S. A. Alshebeili, and L. M. Aljafar, "ECG-based subject identification using statistical features and random forest," *J. Sens.*, vol. 2019, p. 6751932, 2019, doi: 10.1155/2019/6751932.

[30] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

