A Comparative Study of Clinical Rule-Based and Deep Learning-Based Diagnosis of Myocardial Infarction Using Electrocardiograms

Silvia Ibrahimi[1], Massimo Walter Rivolta[1], Roberto Sassi[1]

[1] Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy

Abstract

Myocardial infarction (MI) diagnosis relies on established clinical criteria, primarily involving the identification of electrocardiographic (ECG) abnormalities, such as ST-segment elevation in anatomically contiguous leads. Although rule-based algorithms grounded in these guidelines remain prevalent in clinical practice, recent advances in deep learning (DL) have demonstrated promising performance. However, there is a lack of studies comparing these two methods. To this aim, we developed a multitask DL model and evaluated its performance in comparison with a rule-based algorithm for MI detection, anatomical localization (anterior, lateral, inferior, and septal), and stadium classification (acute, chronic, and normal). The model was trained using 12-lead median beats from the PTB-XL+ dataset. The DL model achieved a sensitivity (Se) of 0.89 and a specificity (Sp) of 0.96 for MI detection, outperforming the rule-based algorithm, which achieved a Se of 0.69 and a Sp of 0.94. For MI localization, the DL model achieved an average F1 score across regions of 0.72, while the rule-based algorithm 0.55. In MI stadium classification, the DL model attained an average F1 score of 0.68, compared to 0.58 for the rule-based method. Overall, the DL model outperformed the rule-based algorithm across all tasks.

1. Introduction

Myocardial infarction (MI), commonly known as a heart attack, is a life-threatening condition that occurs due to the interruption of blood flow to a part of the heart, leading to myocardial tissue damage or necrosis. Early and accurate diagnosis is critical to improve clinical outcomes and reducing mortality. One of the primary tools for MI diagnosis is the 12-lead electrocardiogram (ECG), which captures the heart's electrical activity from multiple angles and provides valuable information about cardiac function.

Typically, MI is diagnosed by analyzing specific alterations in the 12-lead ECG. For example, according to the clinical guidelines [1], one of the primary criteria for acute MI detection is the presence of ST-segment elevation in at least two anatomically contiguous leads. Although rule-based clinical guidelines are still the standard in clinical practice and are integrated into various diagnostic software systems [2], recent advancements in artificial intelligence particularly in deep learning (DL) methodologies have drawn significant attention for their promising performance in ECG analysis and cardiac abnormality detection [3, 4]. In contrast to rule-based methods that depend on handcrafted feature engineering, DL models are capable of automatically learning hierarchical and abstract representations directly from raw ECG signals.

To illustrate the potential of DL, Guo *et al.* [3] employed a transformer-based DL model for MI localization and reported an accuracy of 0.80. In another study, Prabhakararao *et al.* [5] utilized a multi-scale convolutional neural network for MI detection, achieving an average F1 score of 0.84 on the PTB-XL dataset. However, these studies focused on individual tasks and did not simultaneously address MI detection, localization, and stadium classification. Moreover, they did not include a comparison with the clinicalical rule-based algorithm. To the best of our knowledge, no prior work has comprehensively tackled all three MI-related tasks (*i.e.*, detection, localization, and stadium classification) while also comparing their performance with rule-based algorithms.

The main contributions of this study are twofold: i) the development of a multitask DL model for performing MI detection, localization, and stadium classification using 12-lead ECGs, and ii) a comparison of its performance with a rule-based algorithm across these tasks.

2. Materials and methods

2.1. Dataset

We used the PTB-XL+ dataset [6], an extension of the PTB-XL dataset [7] publicly available on PhysioNet. Specifically, PTB-XL+ includes 21,799 median beats sampled at 500 Hz extracted, using three software tools, with comprehensive ECG features, such as amplitudes and intervals of various ECG segments and waves.

Table 1: Clinical guidelines implemented in the rule-based algorithm.

MI ECG changes	Description
ST-elevation	New ST-elevation at the J-point in two contiguous leads with the cut-point: $\geq$1 mm in all leads other than leads V2-V3 where the following cut-points apply: $\geq$2 mm in men $\geq$40 years; $\geq$2.5 mm in men $<$40 years, or $\geq$1.5 mm in women regardless of age.
ST-depression and T wave changes	New horizontal or downsloping ST-depression 0.5 mm in two contiguous leads and/or T inversion $>$ 1 mm in two contiguous leads with prominent R wave or R/S ratio $>$ 1.
Pathological Q wave	Any Q wave in leads V2–V3$>$0.02 s or QS complex in leads V2–V3. Q wave $\geq$0.03 s and $\geq$1 mm deep or QS complex in leads I, II, aVL, aVF or V4–V6 in any two leads of a contiguous lead grouping (I, aVL; V1–V6; II, III, aVF).

In this work, we specifically relied on the features provided by Glasgow (Uni-G) ECG Analysis Program [2]. This software applies clinically standard ECG preprocessing, including filtering and construction of a representative median beat, to ensure consistency with clinical interpretation. From the resulting median beat, the features of interest included the amplitude of the J-point, Q, R, S, and T waves, as well as the duration of the Q, R, and S waves.

From the PTB-XL+, we selected all healthy (normal + sinus rhythm) and MI median beats available. In the MI group, ECG recordings presenting bundle branch block and ventricular hypertrophy were excluded, as the diagnostic rules for MI detection outlined in [1] did not apply to these conditions. Posterior MI cases were also excluded due to their low prevalence in the dataset. After selection, the final dataset was composed of 7,054 healthy and 3,631 MI median beats belonging to 9,731 patients.

The dataset included metadata about the MI stadium and localization. For stadium classification, we binarized the stadium labels. ECGs labels as "Stadium I" and "Stadium I-II" were grouped under the class "acute", while "Stadium II", "Stadium II-III" and "Stadium III" were labeled as "chronic". These stadium classifications delineate the evolution of the infarction. Acute phases are typically characterized by ST-T segment abnormalities, whereas chronic MI is associated with pathological Q waves.

For MI localization, the dataset included codified labels, referred as "SCP code" [8] which specified the affected anatomical regions. Based on these labels, we codified the information to a binary vector of four elements. Each element indicated the presence or absence of MI in one of the following anatomical regions, as defined by their associated ECG leads: septal (V1, V2), anterior (V3, V4), inferior (II, III, aVF) and lateral (I, aVL, V5, V6). This binary encoding allowed us to capture both isolated and overlapping infarct patterns across different regions of the heart.

Approximately 60% of the selected MI ECGs lacked stadium annotations. For both the rule-based algorithm and the DL model, these samples were excluded from the performance evaluation. In contrast, the DL model was trained using pseudo-labels generated by the rule-based algorithm.

2.2. Rule-based algorithm

The rule-based algorithm was designed to simulate clinical criteria for MI detection based on 12-lead ECG interpretation reported in [1]. According to these guidelines, MI in its acute or prior forms is primarily characterized by abnormalities in the ST segment, T and Q wave morphology, and these changes should occur in at least two or more anatomically contiguous leads. Since patients with posterior MI were excluded from the dataset, we also excluded the clinical rules to detect it. In addition to ECG features, the algorithm incorporated patient demographics, including age and sex, to adjust the thresholds of the rules accordingly. To determine the affected region and the MI stadium, the algorithm evaluated the presence of each condition (*e.g.*, ST elevation, pathological Q wave) in the appropriate group of leads. The final localization was determined by aggregating the identified regions. When none of the guideline-based conditions were satisfied, the ECG was classified as normal.

The MI stadium was inferred based on the types of abnormalities quantified: i) ST segment elevation/depression and/or T wave changes indicated acute MI; ii) the presence of pathological Q waves suggested prior/chronic MI; and iii) the presence of both ST changes and Q waves classified MI as acute. The rules implemented are shown in Table 1.

2.3. DL model and experiments

The dataset was divided into training (80%), validation (10%), and test (10%) sets, according to the split proposed in [6]. All ECG recordings belonging to the same patient

were assigned exclusively to the same subset. In this study, a multitask DL framework was used, comprising a shared backbone architecture followed by three task-specific output heads, each targeting one of the following objectives: i) MI detection, ii) MI localization, and iii) MI stadium classification. The shared backbone was based on a modified one-dimensional ResNet architecture, designed to process 12-lead median beats, each beat with a length of 600 samples. The backbone architecture started with a one-dimensional convolutional layer, followed by four residual blocks. Each residual block consisted of two convolutional layers, batch normalization, ReLU activation, and a dropout layer with a dropout rate of 0.3. The three task-specific heads were implemented as fully connected layers: the first head contained a single output neuron for binary MI detection, the second head included four output neurons corresponding to the anatomical localization categories (septal, anterior, lateral, and inferior), and the third head comprised three output neurons for stadium classification (acute, chronic, and normal). Sigmoid activation functions were used for the first and second heads to support multilabel classification, while a softmax activation function was used in the third head for multiclass stadium classification.

The total loss function was defined as:

$$\mathcal{L} = \mathcal{L}_{\text{MI}} + \mathcal{L}_{\text{stadium}} + \mathcal{L}_{\text{localization}} \quad (1)$$

Each loss term was formulated using either cross-entropy (CE) or binary cross-entropy (BCE), depending on the task. Specifically, for the MI detection and MI localization tasks, we adopted weighted BCE loss. The localization task was framed as a multilabel classification problem using four sigmoid outputs from the deep learning model. For the stadium identification task, we applied weighted CE loss to handle class imbalance. Moreover, for the stadium identification task, the loss function, $\mathcal{L}_{\text{stadium}}$, was further weighted based on the reliability of the ground truth labels. Specifically, when the ground truth labels were based on estimated (pseudo-labeled) stadium classifications, a weight of 0.5 was applied to the loss associated with those cases. For instances where the ground truth labels were reliable, a full weight of 1 was assigned. This weighting strategy was employed to reduce the influence of potentially erroneous labels and to encourage the model to prioritize learning from the more reliable labeled data.

The DL model was trained for 20 epochs with an initial learning rate of 10^{-3} and early stopping was applied to prevent overfitting (patience$=7$).

3. Results and discussions

We evaluated the performance of both the DL model and the rule-based algorithm across the three tasks. The performance of the methods was assessed using two metrics: recall and F1-score, based on a testing set of 759 patients. The results of the methods are presented in Table 2.

Overall, the DL model outperformed the rule-based algorithm in MI detection task. Specifically, the DL model achieved a recall of 0.89 (sensitivity; Se) and an F1-score of 0.87 for the MI class, while the rule-based algorithm reached a recall of 0.69 and an F1-score of 0.71. For the normal class, the DL model also showed superior performance, with a recall of 0.96 (specificity; Sp) and an F1-score of 0.97, compared to the rule-based algorithm's recall of 0.69 and F1-score of 0.71. These results indicate that the DL model is more effective in distinguishing both MI and normal cases.

In the MI localization task, the DL model consistently outperformed the rule-based algorithm across all anatomical regions. Recall was computed in a multilabel setting on a per-class basis, with a prediction considered correct if the predicted label was present in the ground truth. For septal and anterior MI, the DL model achieved the highest recall (0.96 for both) with F1-scores of 0.84 and 0.83, exceeding those of the rule-based method (0.73 and 0.70). The performance gap was even more pronounced for lateral and inferior MI, where the DL model achieved recalls of 0.76 and 0.81 and F1-scores of 0.42 and 0.76, compared to 0.24 and 0.39 for recall and 0.24 and 0.53 for F1-score with the rule-based algorithm. In addition to recall and F1-score, we also computed accuracy for both methods. In terms of exact match accuracy (a match was defined when all predicted regions matched the ground truth; healthy cases included), the DL model achieved a higher value (0.84) compared to the rule-based method (0.80).

For MI stadium classification, the DL model showed better results compared to the rule-based algorithm. In the acute stadium, it achieved a higher F1-score (0.33 vs. 0.26) and same recall (0.75), indicating greater precision with same sensitivity. In the chronic stadium, the DL model outperformed the rule-based method with higher recall (0.66 vs. 0.46) and F1-score (0.73 vs. 0.56). In the normal class, the DL model achieved a strong recall (0.97) and F1-score (0.97), surpassing the rule-based algorithm (recall: 0.94, F1-score: 0.93).

Overall, the DL model demonstrated superior performance compared to the rule-based algorithm across the three tasks. While DL models showed superior performance, their "black-box" nature can pose challenges in clinical practice. Rule-based algorithms, in contrast, are more interpretable, providing rationales. However, this interpretability comes at the cost of performance in complex tasks. Moreover, annotation (label) errors in the PTB-XL dataset [6, 9] likely reduce the effectiveness of the training of the DL model. However, DL models can leverage their ability to learn complex patterns and adapt to noisy

Table 2: Comparison of performance between the DL model and rule-based algorithm on the test set.

MI Task	Class	DL Model		Rule-based Algorithm	
		Recall	F1-score	Recall	F1-score
Detection	MI	0.89	0.87	0.69	0.71
	Normal	0.96	0.97	0.94	0.93
Localization	Septal	0.96	0.84	0.86	0.73
	Anterior	0.96	0.83	0.68	0.70
	Lateral	0.76	0.42	0.24	0.24
	Inferior	0.81	0.76	0.39	0.53
Stadium	Acute	0.75	0.33	0.75	0.26
	Chronic	0.66	0.73	0.46	0.56
	Normal	0.97	0.97	0.94	0.93

annotations, offering a more robust solution for addressing these challenges. Despite the promising results of the proposed DL model, this study has several limitations. First, the evaluation was conducted on a single dataset (PTB-XL+), which may limit the generalizability of the findings to other populations or clinical settings. External validation using diverse, multi-center datasets is necessary to confirm the robustness and applicability of the DL model. Additionally, this study was limited to the analysis of 12-lead median beats, without incorporating temporal dynamics or patient history and relied exclusively on ECG data, whereas clinical practice typically integrates additional information such as biomarkers and imaging to improve diagnostic accuracy.

4. Conclusion

In this study, we presented a multitask DL model for MI detection, localization, and stadium classification using 12-lead ECG signals, and compared its performance with a clinical rule-based algorithm. The DL model demonstrated superior performance across all tasks. These findings highlight the DL model's potential as a reliable and flexible alternative to traditional rule-based algorithms, offering enhanced accuracy and generalizability for automated MI diagnosis from ECGs.

Acknowledgments

SI acknowledges support for her PhD fellowship by Novartis.

References

[1] Joint ESC / ACC / AHA / WHF Task Force for the Universal Definition of Myocardial Infarction. Fourth universal definition of myocardial infarction (2018). Circulation 2018; 138(20):e618–e651.

[2] Macfarlane P, Devine B, Clark E. The university of Glasgow (Uni-G) ECG analysis program. In Comput Cardiol. 2005; 451–454.

[3] Guo L, Zhan Q, Yang J, An Y, et al. Lead-grouped multi-stage learning for myocardial infarction localization. Methods 2025;.

[4] Xiong P, Lee SMY, Chan G. Deep learning for detecting and locating myocardial infarction by electrocardiogram: A literature review. Front Cardiovasc Med 2022;9:860032.

[5] Prabhakararao E, Dandapat S. Multi-scale convolutional neural network ensemble for multi-class arrhythmia classification. IEEE J Biomed Health Inform 2021;26(8):3802–3812.

[6] Strodthoff N, Mehari T, Nagel C, et al. PTB-XL+, a comprehensive electrocardiographic feature dataset. Sci Data 2023; 10(1):279.

[7] Wagner P, Strodthoff N, Bousseljot RD, et al. PTB-XL, a large publicly available electrocardiography dataset. Scientific data 2020;7(1):1–15.

[8] Rubel P, Fayn J, Macfarlane PW, Pani D, et al. The history and challenges of SCP-ECG: The standard communication protocol for computer-assisted electrocardiography. Hearts 2021;2(3):384–409.

[9] Doggart P, Kennedy A, Foreman E, Finlay D, Bond R. Automated identification of label errors in large electrocardiogram datasets. In Comput Cardiol, volume 498. 2022; 1–4.

Address for correspondence:

Silvia Ibrahimi
Dipartimento di Informatica, Università degli Studi di Milano,
Via Celoria 18, Milan 20133, Italy
silvia.ibrahimi@unimi.it