Towards certifiable AI in medicine: illustrated for multi-label ECG classification performance metrics

Miriam Elia, Fabian Stieler, Fabian Ripke, Marius Nann, Sarah Dopfer, Bernhard Bauer

Angaben zur Veröffentlichung / Publication details:

Elia, Miriam, Fabian Stieler, Fabian Ripke, Marius Nann, Sarah Dopfer, and Bernhard Bauer. 2024. "Towards certifiable AI in medicine: illustrated for multi-label ECG classification performance metrics." In *2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS), 23-24 May 2024, Madrid, Spain*, edited by José Antonio Iglesias, Rashmi Dutta Baruah, Araceli Sanchis, Mario Muñoz Organero, Dimitry Kangin, and Paulo Vitor De Campos Souza, 1–8. Piscataway, NJ: IEEE. https://doi.org/10.1109/eais58494.2024.10570023.

Towards Certifiable AI in Medicine: Illustrated for Multi-label ECG Classification Performance Metrics

1st Miriam Elia
Faculty of Applied Computer Science
University of Augsburg
Augsburg, Germany
miriam.elia@uni-a.de

2nd Fabian Stieler
Faculty of Applied Computer Science
University of Augsburg
Augsburg, Germany
fabian.stieler@uni-a.de

3th Fabian Ripke
GS Elektromedizinische Geräte
G. Stemple GmbH
Kaufering, Germany
fabian.ripke@corpuls.com

4th Marius Nann
GS Elektromedizinische Geräte
G. Stemple GmbH
Kaufering, Germany
marius.nann@corpuls.com

5rd Sarah Dopfer
GS Elektromedizinische Geräte
G. Stemple GmbH
Kaufering, Germany
sarah.dopfer@corpuls.com

6rd Bernhard Bauer
Faculty of Applied Computer Science
University of Augsburg
Augsburg, Germany
bernhard.bauer@uni-a.de

***Abstract*—Cardiovascular diseases count among the most critical and propagated diseases worldwide. Artificial Intelligence [AI] generates promising results across various medical domains and can enhance cardiovascular diagnosis and treatment. However, contextual knowledge is crucial for multiple design decisions to avoid pitfalls, and the true success of the intelligent application can only be measured in relation to its intended clinical setting. Machine learning [ML] models are evaluated based on performance metrics whose interpretation is influenced by data distribution and tuning objective. The present paper introduces a domain-embedded approach aiming towards a reliable performance evaluation of ML models throughout the complete lifecycle. Our findings are illustrated for multi-label ECG classification in an emergency setting and calculated on open-source Physionet data. Finally, the resulting procedure is embedded as a contribution to Quality Gate Metrics within our generic and customizable methodology based on Quality Gates towards certifiable AI in medicine.**

***Index Terms*—Artificial Intelligence, ECG Multi-Label Classification, Machine Learning Performance Metrics, Interdisciplinary Approach, Domain Embedding, Stakeholder Inclusion**

I. Introduction

In high-risk domains such as medicine, certification is mandatory regarding the transition from research to medical product. Thus, standardized process steps and reasonable design decisions are crucial to ensure the legally required quality. AI has proven it's ability to enhance healthcare, but thanks to its novelty, standards are currently being developed. As of now, no standardized application of performance metrics for the multitude of medical use cases exists, which poses a risk due to possible oversimplification and possibly overlooking important contextual factors [1, 6]. Every medical AI project is based on a unique combination of data type(s), tuning objective(s) and data set distribution(s) that all impact the chosen metrics' interpretation [2], among others. Consequently, unifying performance evaluation approaches is not a trivial task, but necessary for reliable quality and risk management [QRM], which is legally required for intelligent medical products within the EU, as clarified by the AI Act and Medical Device Regulation [MDR]. The present paper is embedded within our research towards certifiable AI in medicine: A Generic and Customizable Methodology based on Quality Gates [QG] [3] that aims to include regulatory requirements towards trustworthy AI in the intelligent system "by design". As a concept to structure relevant information, QGs are located at various stages along the intelligent system's transition from idea to application, i.e. the system's lifecycle. QG-realization aims to transform interrelated, abstract ML concepts into directions of thought that can be concretized tailored to the respective use case and stakeholder. The present paper introduces an empirical approach contributing to QG Metrics definition through the proposition of an approach towards reliable performance metrics, which will be refined in the future. The method is envisioned to be extendable to other ML design decisions along the AI lifecycle, and to contribute to reliable long-term monitoring. Our findings are based on the ML part of a fictional use case centered around ECG multi-label classification in an emergency setting, but aims to be generalizable to other (multi-label) use cases as well. The proposed process is structured into pre-, intra- and post-selection steps to arrange important directions of thought. Specifying the abstract method, concrete guidelines for a reliable metrics application in the context of (ECG) multi-label classification and high data imbalance are extracted based on our experiments. Consequently, our focus lies on the extraction of guidelines and directions of thought, rather than the perfect model. A comprehensive overview of our results, as well as the proposed method's integration on a process level within the QG structure covering the AI lifecycle can be found on the open research repository Zenodo.[1] Our

This work was partially funded by the German Federal Ministry of Education and Research (BMBF) under reference number 031L9196B.

[1]https://zenodo.org/records/10931084

experiments are calculated with data from the Physionet-Challenge 2021, where participants competed to train a model that can identify clinical diagnosis from ECG-recordings. One sample can include multiple, differently correlated heart arrhythmias/rhythms and other heart diseases [HARHD], or *labels* from a technical viewpoint. Further, we focused on 12-lead ECGs only. To include the, from a medical viewpoint necessary information on label correlations, our domain embedded procedure is centered around a cost-sensitive post-processing thresholding method that also considers category imbalance [4], as introduced in section 4.3.

In the following, important points of consideration regarding a reliable metrics selection will be outlined and illustrated based on ECG multi-label classification, while our proposed method aims to transform abstract into arranged and more practical guidelines to simplify auditing and clinical workflow integration for all stakeholders - starting with the arrangement and identification of relevant directions of thought for performance metrics selection. The paper is structured as follows: First, the presented method is embedded within our methodology towards certifiable AI in medicine and our experiments are briefly introduced (II-III). Next, the proposed design decision selection method, structured into pre-, intra- and post-selection steps, is presented. Concretizing our proposed approach, guidelines for reliable multi-label (ECG) classification metrics are demonstrated based on our experiments (IV-VI).

II. Trustworthy AI Lifecycle Management

A reliable metrics selection is not a trivial process and is marked by complex, use case-specific interdependencies and stakeholder perspectives throughout the AI lifecycle, which can result in an incorrect application and poses a risk. Consequently, to realize a trustworthy integration of AI as envisioned by the EU AI Act, the development of standardized methods is crucial. Risk management is an ongoing process that needs to be considered during the complete AI lifecycle, as pointed out by ISO 5338 on AI system lifecycle processes. Different levels of risk need to be (constantly) identified and evaluated, as well as risk control measures implemented and monitored. As a result, the implementation of AI risk management is highly domain- and use case-dependent. Among others, this openness regarding continuous AI risk identification, evaluation, and control, poses a challenge for auditing in the long-term, which necessitates a concretization of comprehensive methods that are to a certain extent automatable. According to which standards intelligent (medical) system processes will be defined is currently being researched around the world.

The present paper aims to contribute a generalizable method for reliable design decision making along the AI lifecycle as part of on-going research and is illustrated for ECG multi-label classification metrics.

A. A Methodology based on Quality Gates

Our methodology, as introduced in [3], aims to provide a generic and customizable conceptual structure with QGs at its core to realize risk management and ensure trustworthy AI on a process-level "by design". We aim to illustrate how to comprise existing research and new experiments on the development of AI models into concrete guidelines that are adaptable for developers and auditing institutions, while paying special attention to end-user perspectives, the inclusion of domain knowledge, and ML method- and data-specific interdependencies. On an abstract level, the AI lifecycle consists of multiple (sub-)process steps regarding data, the model, its on-boarding, deployment and maintenance in the clinical setting, as well as the integration of new data, and its decommissioning after possibly multiple cycles of evolution. These stages describe the identified common ground across different AI use cases, or the flow of AI-specific processes, as identified by ISO/IEC FDIS 5338:2023(E) on AI system lifecycle processes.[2] Overall, AI lifecycle management is complex and based on a great number of interdependencies, which necessitates a detailed and well structured overview of the complete design process. Especially, with respect to substantial modifications and Predetermined Change Control Plans [5], [6] answering to the system's capability to continuously learn.

Quality Gates: The concept of QGs aims to provide a foundation to structure the collection and creation of concrete approaches for reliable AI lifecycle management with QRM at their core and aims to capture intelligent system's inherent dynamics. High-level QGs represent specific lifecycle stages that consist of multiple design decisions, which are represented by lower-level QGs and gradually become more use case-specific. Further, QGs are envisioned to offer means to evaluate the respective design decision against pre-defined criteria. In summary, they correspond to specific, hierarchically structured lifecycle process steps and envisioned to consider relevant interdependencies with other QGs on all levels and lifecycle iterations.

Guided Questions: QG-Creation is centered around the concept of Guided Questions [GQ] to first identify relevant information in a comprehensive manner aiming to promote the system's trustworthiness. The definition of GQ is based on the abstractable identification of important considerations that are applicable across use cases in form of questions, as illustrated at the end of section III.

Our philosophy is in alignment with the NIST AI Risk Management Framework [AI RMF], promoting that "[...] AI risk management depends on a sense of collective responsibility among AI actors"[10] [1].[3] Consequently, in addition to (methods for) the extraction of QG-relevant information, the content is presented adapted to different stakeholder-views. For instance, metrics representation on monitoring dashboards differs for physicians and data scientists.

[2]Other relevant processes are envisaged to be integrated as part of QG creation for the respective design decision depending on the specific use case.

[3]AI actors are "[...] those who play an active role in the AI system lifecycle, including organizations and individuals that deploy or operate AI" [1, 2], following OECD.

B. Embedding ML Performance Metrics into Quality Gates

In general, our empirical approach for QG-Creation envisages the deduction of directions of thought in relation to other QGs, as well as their arrangement for relevant content extraction tailored to specific use cases, stakeholder-views, process steps, and design decisions. Our later presented method (IV-VI) contributes information on performance evaluation metrics to the definition of the more abstract QG Metrics, which is heavily influenced by QG Data and impacts other QGs. Our proposed procedure is structured into pre-, intra-, and post- metrics selection process steps and envisioned to be extendable to other design decisions, including fairness metrics as further contribution to QG Metrics, for instance. Concretely, the presented process can be referenced as a template that comprises structured lines of thought, which are currently being researched for the multitude of medical AI use cases by the research community, and the industry.

As a start, we extracted concrete guidelines regarding multi-label (ECG) classification metrics, which we believe are extendable to multi-label use cases with label correlations that are characterized by a high data imbalance, while the absence of a label is less meaningful than its presence:[4]

- A compilation of ML performance metrics consisting of confusion-matrix and ranking metrics, that are selected with respect to their use case-specific interpretation
- A mindful consultation of the Receiver Operating Characteristic Area under the Curve metric [ROC AUC]
- A detailed analysis of class probabilities and additional material for a comprehensive overview, especially regarding the interplay of metrics and data distribution
- Preferably the application of macro- or weighted-averaged multi-label metrics for a more balanced view on the model's performance
- A metric optimization and monitoring strategy
- An analysis of use-case specific challenges and possible solutions regarding label correlations (especially, for the multi-label case)

In general, a profound understanding of the clinical setting including the system's objective as foundation, an interdisciplinary team, which is based on the identification of relevant disciplines, as well as the continuous identification of interdependencies along the lifecycle for a long-term monitoring strategy are crucial. These tendencies can be formulated as GQs that are focused on the trustworthy QG Metrics lifecycle integration from the start and aim for reliable long-term monitoring towards risk management. Tailored to the developer's and possibly auditor's perspective, they include:

- What clinical impact is caused / desired by the model's predictions?
- Which metrics are relevant for which stakeholders at what point of the ML lifecycle?
- What other stakeholder-views need to be considered? (E.g. patients)
- Who has a global overview and is responsible to ensure guidance? What are other responsibilities and who are their representatives for accountability management?
- How should metrics (including supplementary information) be visualized depending on their application?
- In what cycles during the intelligent application's evolution and depending on which parameters, the decisions need to be monitored and possibly adjusted?
- Which interdependencies regarding other design decisions or lifecycle stages (QGs) exist?

[4]In contrast to binary classification, if a label is not present on an ECG-sample, this does not automatically signify that the patient is healthy, since other labels might exist.

III. Empirical Approach towards Metrics Selection and ML Design Decisions

Our proposed method for a reliable metrics selection was extracted in an empirical manner based on our experiments and accompanying research. On a more abstract level, we aim to provide a generalizable approach for reliable AI lifecycle design decisions that are adapted to specific medical use cases, analoguously to our approach for multi-label ECG classification in an emergency setting. Regarding our experiments, we extracted the following tendencies for several design decisions centered around ML performance metrics: (1) a use case-adapted metrics selection, (2) an analysis of ROC AUC vs. Precision-Recall AUC [PR AUC] expressiveness, (3) multi-label averaging strategies, (4) different label structures based on label support, i.e. each label's count in the data or how strongly it is represented compared with other labels (5) the metric to be monitored during training, for early stopping and hyperparameter optimization, as well as (6) different post-processing thresholding methods: the most common fixed threshold of 0.5, CIST [4] as well as ROC and PR curve based thresholds for comparison and as an additional perspective on the two rank-based metrics' behavior. The following sections' structure is intended as a template to arrange the definition process of ML design decisions - in form of pre-, intra- and post-selection steps. First, our novel use case-specific label structure is introduced. Our experiments are trained on Physionet data and were mapped to our label structure via their Systemized Nomenclature of Medicine – Clinical Terms Code [SNOMED CT].

A. Medical Label Structure

Labels refer to all possible HARHD that can occur in an ECG recording and (co-)exist in a multi-label setting. As a starting point for our use case-adapted and domain-embedded label compilation, we consulted the comprehensive PTB-XL mapping [7]. We then excluded clinically meaningless artifacts. Further, not all labels automatically result in a diagnosis from a medical viewpoint in the context of emergency medicine. However, a clear course of action is required, which, in our fictional use case is based on the direct follow-up step after interpreting the model's prediction. Therefore, we introduced two novel *container-labels* [CL] that consist of a combination of more specific labels: *other* and *norm*. The latter

refers to a healthy patient, including all unproblematic labels, as sinus rhythm variants and physiological electrical cardiac axis. It cannot coexist with any other label, while *other* and abnormalities can coexist. All remaining pathological ECG labels that do not allow for a specific therapy without further examinations are summarized under *other*. If one of these pathologies is identified in an ECG, the patient will most likely be transferred to a hospital. The proposed label structure for application in an emergency setting consists of 13 labels in total. This includes the addition of a CL for ischemias (except ST depression and elevation), based on similarity of follow-up treatment.

B. Experimental Label Structure

For our first experiments, we decided to itemize the CLs, which resulted in 21 labels in total. The mapping of our label structure to the Physionet data via their SNOMED CT code led to some incompleteness since not all, for our use case reasonable labels exist sufficiently in the data, and Pericarditis is not represented. We first trained a model on 20 labels. However, some labels could not be evaluated properly due to a very low support and they were excluded. This decision is based on per-label metrics that are not calculated ($= 0$) when the label is underrepresented. Finally, we continued our experiments with a combination of 8 labels whose support did not fall below ≈ 1100 samples, which is still very low. The following process and results are based on 8 labels.

IV. Pre-Selection

A reliable application of ML performance metrics always includes a selection of multiple metrics for a comprehensive evaluation tailored to different stakeholder perspectives, since one single metric is not capable to reflect on all desired capabilities [8, 3]. Consequently, the first step to choose an accurate metrics compilation is to understand the interplay between all requirements of the use case, which advisably should happen before implementation. This includes data, tuning objective, the metrics' contextual meaning as well as possible use case-specific draw backs that might require creative workarounds. For the multi-label case, label-correlations, possibly differing per-label tuning objectives, and the averaged multi-label metrics' expressiveness are important considerations.

1) Use Case: The fictional real-world setting of the intelligent medical product at the core of our research is an alarming guard functionality anywhere patients at risk could appear and the person in charge is not well trained in ECG diagnostics, e.g., in an emergency setting or at the general practitioner's office. Consequently, the direct clinical follow-up step after the model's prediction is relevant to define other related design decisions.

A. Domain Embedding

Generally, domain knowledge is crucial for trustworthy design decision-making and a continuous part of the process. For our use case, depending on the respective label, a higher Precision or Recall is desired, which leaves the questions open, how to tune (metric-to-monitor) and evaluate (averaged multi-label metrics) the model overall. Should the system be more sensitive, and false alarms tend to be more acceptable than missing a label based on the follow-up step, or not? Our approach is evaluated in the following sections.

1) Data Distribution: Considering the (dynamic) data distribution throughout the complete lifecycle is crucial for model evaluation and comparison. A data set distribution monitoring strategy can mitigate risks posed by biased behavior,[5] while continuously (re-)assessing data suitability for the intended purpose of the intelligent system. For instance, the medical prevalence could provide insights what a "realistic" data distribution could mean. Physionet data is comparably diverse and consists of "[...] ten databases from [...] several countries across three continents" [11, 3], i.e. USA, Europe and China. This information provides valuable insights into the behavior of performance metrics.

2) Benefit-Matrix and Post-Processing Thresholding: Data imbalance can be addressed through post-processing thresholding, i.e. individually adjusted thresholds. Aiming to include medical knowledge on all label correlations in addition to information on support, Category Imbalance & Cost-Sensitive Thresholding (CIST), as proposed in [4], is a suitable approach. Correlations are summarized in form of a Benefit-Matrix (BM) that contains similarities in follow-up treatment triggered by the model's prediction compared to the groundtruth over all labels. The authors based their experiments on the Physionet BM[6] [4, 7]. We generated a novel BM in cooperation with a medical domain expert that was adapted to our label structure tailored to emergency medical care. The values are maximized the closer the follow-up steps for differing labels - aiming to minimize the negative impact on the patient, and ranging from 0.1 to 1. The latter equals complete similarity in treatment. This information is then translated into misclassification costs that are afterwards combined with information on label support, and transformed into label-specific thresholds as in [4, 6f.]. Their interplay is regulated by the hyperparameter $modulating\ alpha$ ranging from $0-1$: class imbalance has $alpha$ and the misclassification costs $1 - alpha$ influence on the calculated thresholds [4, 9]. An evaluation of CIST performance compared to ROC and PR curves-based thresholding is part of the post-selection process. These methods extract the threshold that best balances both axis represented by different metrics respectively.

B. Metrics Analysis

Depending on the concrete task at hand, different choices of metrics are reasonable. Our use case is centered around multi-label classification, since ECG samples can be labelled with multiple, not mutually exclusive HARHD. The following considerations further explain our domain-embedded metrics selection process. First, binary classification-based multi-label averaging is explained in more detail. Next, common binary

[5] See ISO 24027 on bias in AI systems and aided decision making [16].

[6] https://github.com/physionetchallenges/evaluation-2021

classification metrics are introduced and evaluated against our use case, see figure 3. We based our main selection on medical interpretability and added some additional metrics to illustrate different perspectives on the model's discriminatory power. This step is highly use case-dependent and should be conducted based on a comprehensive analysis.[7]

1) Multi-Label Classification: For performance evaluation, it is common to transform the multi-label case into a multitude of independent binary classification problems, where each label is evaluated as present vs. not present. Then, some sort of averaging method to calculate classification metrics for overall model performance is applied. However, this conceptually simple approach does not exploit label correlations [12, 1820].[8] The previously introduced CIST method is applied to include medical information on label correlations. To calculate overall model performance, four distinct averaging methods exist: *macro* (or *weighted*), *micro* and *samples*. The latter means returning the mean over the complete test or validation set after calculating the performance measure for each sample.[9] Macro- and weighted-averaging refers to metrics calculated label-wise, and then their (weighted) average. Micro-averaging considers the complete prediction matrix at once, which is why it "[...] favors bigger classes" [13, 430]. [9, 1822], [14, 3] Consequently, analyzing per-class performance in addition to the averaged metrics, as explored in the next section, is important. Further, some labels prefer Precision over Recall from a clinical perspective or are underrepresented in the data, which is not always well mirrored by the averaged metrics.

Guideline Averaging-Method Macro-averaging is the preferred choice for our translational approach.[10] It evaluates each label equally, and thus also reflects on rare classes [13, 430] which displays the model's performance as realistically as possible. ECG multi-label classification is characterized by high data imbalance, like many medical use cases, and some HARHD are less propagated than others, but still relevant. The following tendencies can be derived for class-imbalance: if macro-averaged metrics are lower than micro-, the model is comparably worse at identifying rare classes, but competent with the more common ones, and vice versa. The former is mostly the case for our experiments, only ROC AUC displays a high performance for both averaging methods.

2) Confusion Matrix based Metrics for Binary Classification: A multitude of the existing binary classification metrics are calculated based on the four values of the confusion matrix [2, 5]. The model's predictions are compared to a threshold and then categorized into either $true$ or $false$. Next, the results are compared with the groundtruth depending on whether they belong to the $negative$ or $positive$ label. We mainly focus on the interplay and performance of Specificity, Precision, Recall, F1- and FBeta-Score, since their medical interpretation is reasonable and especially Recall and Specificity are already applied for common ECG-classification [15]. For a more comprehensive overview, Jaccard Score (insight on the distribution of the intersections between predictions and groundtruth), False Positive Rate [FPR] (as an additional perspective on false alarms), and Matthew's Correlation Coefficient [MCC] (it considers all four values of the confusion matrix), were also included. Most of the selected metrics range from $0-1$, a higher value indicating better performance. MCC from -1 to 1 and FPR are minimized.

3) ROC and PR AUC: These ranking-perspective based metrics access the real valued function learned by the model. They are calculated as the AUC signifying model performance for all possible thresholds regarding Recall and Precision [PR AUC], or False-Positive Rate [ROC AUC]. In contrast to confusion matrix-based metrics, this approach is a threshold-independent performance measurement, and can be referenced for post-processing thresholding. Both metrics reaching from 0 to 1 are maximized, and a high value signifies model confidence towards its predictions. As already discussed in our latest research, there is an ongoing discussion whether ROC AUC is suitable for heavily imbalanced data sets or if PR AUC should be preferred [3, 2]. Our experiments demonstrate that this is indeed true for our multi-label use case, as demonstrated in figure 1: ROC AUC displays a very similar behavior across all labels in contrast to the PR curves, which is suspicious based on the highly imbalanced label distribution.

Guideline ROC vs. PR AUC ROC AUC displays very optimistic behavior across all labels compared to PR AUC, which might lead to an unreliable application in the medical setting, since the model's actual performance is worse.

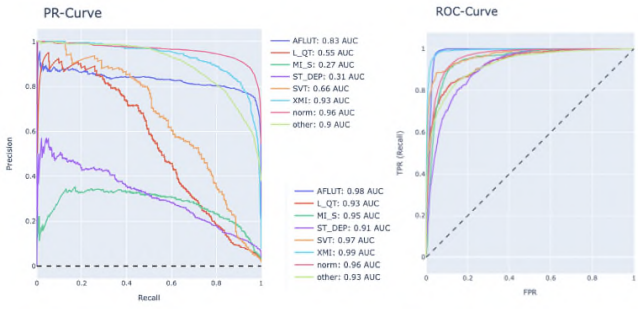

Fig. 1. PR vs. ROC AUC performance visualized for 8 labels.

V. Intra-Selection

After having researched several metrics (and approaches) that are suitable for the respective use case, their performance should be analyzed in more detail through experiments that might result in modifications regarding the previously selected metrics-structure. Also, from a long-term perspective, information on when, how and by whom the metrics will be consulted needs to be organized in a way that supports post-market

[7]Another interesting line of thought, which is not focus of our analysis, is to develop a custom evaluation metric in close cooperation with domain experts and potentially based on an existing metric. Refer to the Physionet Scoring Metric for an example: https://moody-challenge.physionet.org/2021/

[8]Other multi-label strategies include either pairwise or multiple label correlations with augmenting degree of computational complexity, but won't be the focus of this analysis [12, 1820].

[9]The multi-label metric Hammingscore is calculated as the samples averaged jaccard score.

[10]Weighted averaging depends on the selected weights and might need further tuning, which is why we did not analyze it in further detail for our experiments.

release monitoring. Consequently, this stage is intended to be considered alternately with post-selection to cover optimization and the intelligent system's dynamic nature. In addition, this includes that relevant process steps/design decisions need to be documented, refer to Zenodo for a proposed QG folder structure, and then translated for the respective stakeholder to achieve a smooth integration into the clinical workflow. For our multi-label use case and mainly focusing on the developer's perspective, this includes but is not limited to analyzing metrics per class in more detail under consideration of additional material, while considering clinical knowledge for reliable ML design decisions.

A. Additional Material

The inclusion of additional material, which highlights the data set, supports a more reliable comprehension and should be referenced during metrics planning and interpretation. For instance, this includes referring to the true label and probability distribution, as well as each label's support in the data. In figure 2, the groundtruth vs. probability distribution of our test data is displayed (red is 0, blue is 1) for $norm$ and general myocardial infarction signs. As can be deducted, a fixed threshold of 0.5 would lead to the latter almost never being classified as $true$ due to the model's low confidence. Regarding the majority label *norm*, the model displays desired tendencies. Here, explainable AI [XAI] methods, for instance could offer a deeper understanding.

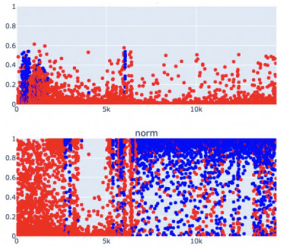

Fig. 2. Groundtruth and probability distribution for $norm$ and myocaridal infarction signs.

B. Class Probabilities

On top of additional material, contrasting per label scores with the averaged multi-label metrics highlights the model's performance for each label, as well as reflects on the averaging method's expressiveness. For our experiments, we did not consider class probabilities from the start. However, the differences between macro- and micro-averaged metrics led to a retrospective analysis. Without further consideration, we calculated four metrics class-wise as default: Precision, Recall, F1- and Jaccard-Score, but based on a more profound retrospective domain analysis, Specificity and FPR would have been interesting as well. Our experiments demonstrate that for multi-label classification the analysis of per-label scores leads to a more reliable view on the model's performance, which is strongly impacted by the label's support in the data. Further, our use case is charachterized by different per-label tuning objectives.

C. Domain Knowledge

Domain embedding is crucial for reliable ML, it enhances trustworthiness, provides a more realistic categorization of the intelligent system and contributes to monitoring the AI lifecycle. In alignment with our results, the following considerations can be analyzed further for a more comprehensive overview of multi-label (ECG) classification and are included as part of GQs for QG Metrics creation.

- Clinically different tuning objectives per class: Based on the interplay of Precision and Recall for instance, the trade-off between false alarms or missed diseases could be addressed more label-specific.
- Alternatively, a combination of ensemble models, i.e. one model per label.
- New BM for comparison and a possibly closer real-world adaptation: This could be based on a re-evaluation of the first approach including a clinical prioritization of class labels, for instance.
- Stakeholder & usability: For long-term monitoring, the baseline metrics-compilation can be adjusted or extended with respect to different stakeholder views and lifecycle stages, which should be assessed during clinical trials.
- Benchmarking: a performance comparison with a current clinical state-of-the-art is necessary for optimization and to substantiate the use of AI.

VI. Post-Selection

A reliable metric selection builds the foundation for further tuning, as well as monitoring or using the intelligent system, and is correlated with other design decisions. Based on the previously defined metrics compilation and all documented design decisions, the final step of the first cycle for metrics design is optimization. Therefore, multiple experiments whose structure is documented are conducted to optimize hyperparameters under consideration of use case-specific challenges. This includes planning during which cycles a re-evaluation in alignment with intra-selection is necessary. For our experiments, we focused on two design decisions embedded within our ECG use case that impact performance metrics to extract generalizable guidelines: the metric-to-monitor [MM] and post-processing threshold optimization, while other hyperparameters are left unchanged.

A. Metric to Monitor

We analyzed the MM during training for validation, monitoring early stopping and learning rate adjustment through comparison of F1- vs. FBeta-Score. These considerations are based on the use case-adapted tuning objective displayed as the interplay between Precision and Recall.

Guideline Metric-to-Monitor Following our domain-embedded approach, the chosen MM best summarizes the intended real-world objective. Our experiments show that this design decision impacts performance metrics in a very fine-grained way. For a fixed threshold of 0.5 micro- and macro-averaged results were almost identical, since the model's confidence tends to be rather low. However, regarding lower CIST ($alpha = 0.3$) thresholds, for both averaging methods, FBeta-Score favors Recall while F1-Score yields a more balanced performance improvement of a few percent regarding

Metric	Formula	Definition	Context	Evaluation
Specificity	TN / (TN + FP)	How good is the model at predicting the negative class? (True Negative Rate)	How good is the model at correctly predicting the absence of a label?	Recall for negative label
Precision	TP / (TP + FP)	How many as true predicted labels are present? (Positive Predictive Value)	How accurate is the model at predicting the presence of a label?	Some labels could lead to fatal follow-up procedures for the patient and false alarms should be avoided
Recall/Sensitivity	TP / (TP + FN)	How good is the model at predicting the positive class? (True Positive Rate)	How many of all true labels did the model predict correctly?	Some labels could lead to fatal consequences for the patient if missed
Fbeta/F1-Score	((1+beta**2)*TP) / ((1+beta**2)*TP + (beta**2*FN) + FP)	Fusion of Precision and Recall	Hyperparameter beta regulates allocation	In our experiments beta=2, i.e. Recall is twice as important as Precision or beta=1, which equals the F1-Score and is defined as the harmonic mean
Jaccard Score	A n B / A v B	Measures distance between two sets	The two sets are the predicted labels and groundtruths	How close are the predictions to the test or validation set?
MCC	(TP*TN – FP*FN) / √(TP+FP)(TP+FN)(TN+FP)(TN+FN)	Correlation coefficient value is between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction and -1 an inverse prediction	How correlated are the model's predictions with the test set distribution?	high score only if the prediction obtained good results in all of the four confusion matrix categories
False Positive Rate	FP / (FP + TN)	Probability of a false alarm, it should be minimized	How often does the model wrongly predict a positive label?	Inverse of Recall
ROC AUC	Area under the Receiver Operating Characteristic Curve	Performance of model measured for all possible thresholds by comparing TPR and FPR	How good is the model's performance regarding all possible thresholds?	See PR AUC, ROC AUC tends to be very optimistic for rare classes and should be used with caution
PR AUC	Area under the Precision-Recall Curve	Performance of model measured for all possible thresholds by comparing Recall and Precision	How good is the model's performance regarding all possible thresholds?	Measures the model's prediction confidence, a high AUC indicating high distinctive qualities

Fig. 3. Metrics interpretation as binary calculation before multi-label averaging: $true$ means the label is present and $false$ equals the label's absence.

the other confusion matrix-based metrics. ROC AUC and PR AUC performance is almost identical for both MM. Generally, the performance values are similar for both MM, with macro-averaging displaying a higher variance of $> 10\%$ than micro-averaged metrics with $\approx 10\%$. With respect to the calculated thresholds, the following behavior was observed as a consequence of the MM:

- CIST: Monitoring the F1-Score results in a slightly higher threshold for rare classes than with FBeta. It is slightly smaller for majority classes, which appears reasonable since the F1-Score values Precision more.
- PR: Displays a high variance regarding direction and value of the threshold deviations caused by the MM, which might be caused by equally balancing Recall and Precision.
- ROC: Monitoring the FBeta-Score always results in a higher threshold, probably because of Recall, but in general the difference is infinitesimal.

B. Threshold Optimization

The threshold applied to the model's prediction after training can be fine-tuned for evaluation. This post-processing method is a reasonable step to address data imbalance, and performance should be monitored based on the dynamics of the relevant data distribution to possibly adjust the calculated thresholds in a dynamic manner. In addition to a fixed threshold of 0.5 for all labels, we compared three label-specific methods: ROC and PR AUC-based thresholding, as well as CIST with $alpha = 0.3$ for both MM. CIST displays the most balanced behavior across all metrics.

Guideline Threshold Optimization Post-processing threshold optimization addresses the effects of high data imbalance and offers a means to include label correlations for multi-label classification. Multiple methods exist and it is strongly correlated with the data distribution and should be monitored accordingly. Each thresholding method displays a significant difference in performance values, especially regarding Recall and Precision, with augmenting behavior the scarcer the data. ROC-thresholds have the highest range of values, with close to 0 for rare classes and 0.66 for *norm*. It seems to perform well only if the label is prevalent and can be applied, when a high precision is desired. PR and CIST seemingly adjust better to class-imbalance, however the PR-thresholds' high variance should be analyzed further. CIST appears to follow the strategy the more represented the label in the data, the higher the threshold. Comparably, it generates rather low thresholds and pushes Recall. PR-thresholds generally tend to be higher than CIST-thresholds, which is why, similar to 0.5, it favors Precision, FPR and Specificity. Regarding macro-averaged metric performance, CIST achieves the most balanced and satisfying scores across all metrics, while it scores highest for F1-Score, Recall, MCC, Jaccard-Score. However, it is outperformed by 0.5 and PR regarding Precision, FPR and Specificity, which is based on its lower thresholds. A more comprehensive analysis can be found on Zenodo.

VII. Quality Gate Evaluation

We propose a practical approach for a reliable metrics selection embedded within our methodology for AI lifecycle QRM as part of QG Metrics. The presented interdisciplinary and domain-adapted approach, as well as the derivable guidelines for multi-label (ECG) performance metrics contribute towards trustworthy AI lifecycle design. For the present paper, our model is not optimized outside the described scope centered around performance metrics, but rather serves as a basis to derive empirical tendencies. Reliable metrics are highlighted as part of general AI risk management [1, 6], and we believe it is advisable to start with a solid foundation early on.

Pre-, Intra-, Post-Selection: Depending on the level of abstraction, our proposed method is extendable to other (multi-label) use cases and design decisions such as QG model architecture or QG loss function. Further, after the first cycle of metrics-selection, depending on the dynamic real-world context and other design decisions (QG) that impact metrics, intra-, and post-selection stages are envisaged to be revisited.

Finally, our presented approach towards multi-label ECG classification might not be the best answer, for instance, depending on the label's individual tuning objectives, a different solution such as ensemble learning might be more suitable. These considerations belong to a comprehensive pre-selection, and possibly multiple methods are first evaluated in parallel.

Global Embedding: Generally, QRM frameworks are based on norms and other recommendations, while incorporating important horizontal [17] and, depending on the intended use, sector-specific legislation such as the *International Medical Device Regulators Forum* as well as country-specific information provided by notified bodies [18]. Standards that aim at regulating intelligent systems and their novel challenges are currently being developed and refined. They are integrated as process steps along the AI lifecycle, which is also desired from an auditing perspective. The IG-NB, the German association of notified bodies, for instance, base their guidelines on "[...] the idea that the safety of AI-based medical devices can only be achieved through a process-oriented approach, whereby all relevant processes and phases of the lifecycle must be considered" [18, 1]. Within the broader context of AI QRM, our presented method aligns with the propagated process view. A practical approach for general AI QRM is presented by the NIST AI RMF which promotes risk adjustment to the real-world setting, transparent documentation [1, 6], the inclusion of a " [...] broad set of perspectives and actors across the AI lifecycle" [1, 9], as well as continuous risk management [1, 20]. Our methodology can be integrated with the AI RMF as a use case for the medical sector that incorporates GOVERN, i.e. the legal culture of risk management, through application of MAP, i.e. domain embedding to healthcare in form of MANAGE, i.e. a comprehensive QRM strategy, to develop MEASURE, i.e. concrete methods for a reliable lifecycle process [1, 20f.], such as the present research. The development of standards to guide developers along the AI lifecycle and its evolution to resolve design decisions in alignment with regulation is crucial. This investment levels the path for a trustworthy integration of AI into society and requires all stakeholders to be involved. Other approaches can be found in [19] across sectors, and in [20] for medicine.

VIII. Outlook

The present paper describes a generalizable method for reliable design decision-making illustrated for performance metrics selection, and structured into pre-, intra- and post-process steps. It is embedded within our methodology towards certifiable AI in medicine as a contribution to QG Metrics. Extracted guidelines are envisioned to be applicable across use cases. Each stage is characterized by important directions of thought that are accompanied by domain-embedded considerations for ECG multi-label classification. Our experiments exhibit exemplary characteristics and do not assert completeness. Finally, a reliable performance metrics selection process is not a trivial task, but crucial for a multitude of reasons. Especially, since performance metrics are connected with many other design decisions throughout the AI lifecycle. Further research regarding QG Metrics is necessary, among others:

- Interdependencies with other QGs need to be identified
- Explainable AI (XAI) could lead to a more profound interpretation of the model's output, the metric's performance and additional material
- Metrics that measure other characteristics such as fairness, for instance should be considered, as in ISO 24027
- Stakeholder-adaptation of relevant results along the lifecycle for post-market surveillance and usability, which includes an on-boarding strategy to educate recipients on the metrics' interpretation, for instance in form of interactive monitoring dashboards
- Representation for auditors could include a textual template following the presented approach, filled with use case-specific information including a reference to supplementary material with detailed results
- A metric that monitors metrics could be designed for surveillance-monitoring support

References

[1] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.

[2] Jussi et al., Evaluation of machine learning algorithms for health and wellness applications: A tutorial, Computers in Biology and Medicine, 2021.

[3] Elia et al., A methodology based on quality gates for certifiable AI in medicine: towards a reliable application of metrics in machine learning, ICSOFT, 2023.

[4] Liu et al., Automatic Multi-Label ECG Classification with Category Imbalance and Cost-Sensitive Thresholding, Biosensors, 2021.

[5] U.S. Food and Drug Agency, Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD), Discussion Paper and Request for Feedback. 2021.

[6] Prinz et al, Market access of continuous learning AI systems in medicine, VDE DGBMT, 2023.

[7] Wagner et al., PTB-XL, a large publicly available electrocardiography dataset, nature, 2020.

[8] Kelly et al., Key challenges for delivering clinical impact with artificial intelligence, BMC Medicine, 2019.

[9] Rahmani et al., Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction, International Journal of Medical Informatics, 2023.

[10] Schwartz et al., Prevalence of the congenital long-QT syndrome, Circulation vol. 120,18, 2009.

[11] Reyna et al., Issues in the automated classification of multilead ecgs using heterogeneous labels and populations, Physiological measurement vol. 43, 2022.

[12] Zhang et al., A review on multilabel learning algorithms, IEEE Transactions on Knowledge and Data Engineering, 2014.

[13] Sokolova et al., A systematic analysis of performance measures for classification tasks, Information Processing Management, 2009.

[14] Wu et al., A unified view of multi-label performance measures, CoRR abs/1609.00288, 2016.

[15] Macfarlane et al., Automated ECG Interpretation—A Brief History from High Expectations to Deepest Networks. Hearts, 2, 433-448, 2021.

[16] International Standards Organization, ISO/EC–TR 24027:2021(E) on Bias in AI Systems and Aided Decision Making, 2021.

[17] Floridi et al, capAI: A procedure for conducting conformity assessment of AI systems in line with the EU Artifcial Intelligence Act, 2022.

[18] Rad et al., Questionnaire "Artificial Intelligence (AI) in medical devices", IG-NB, 2023.

[19] Brajovic, et al., Model Reporting for Certifiable AI: A Proposal from Merging EU Regulation into AI Development, 2023.

[20] Reddy et al., Evaluation framework to guide implementation of AI systems into healthcare settings, BMJ health & care informatics, 2021.