Radboud University Nijmegen

Faculty of Science

Regressing Creatinine from Single-Lead Electrocardiograms: Assessing Self-Supervised Architectures

in cooperation with TCC GmbH

Thesis MSc Data Science

Author:
Laura Stritzel
s1130411

Supervisor Radboud:
dr. Inge Wortel

Supervisor TCC:
Dominik Thiele

Second reader:
dr. Johannes Textor

May 2026

Contents

1 Abbreviations

Medical Terms	
ECG	Electrocardiogram
ICU	Intensive care unit
HR	Heart Rate
AKI	Acute Kidney Injury
CKD	Chronic Kidney Disease
GFR	Glomerular Filtration Rate
hyperkalemia	high potassium
normokalemia	normal potassium
EHR	Electronic Health Record
resprate	Respiratory Rate
o2sat	Oxygen Saturation
sbp	Systolic Blood Pressure
dbp	Diastolic Blood Pressure
NTproBNP	N-terminal Pro-B-type Natriuretic Peptide
pO2	Partial Pressure of Oxygen
WBC	White Blood Cell
LVEF	Left Ventricular Ejection Fraction
Datasets	
TCC	Telehealth Competence Center
MIMIC-IV	Medical Information Mart for Intensive Care - IV
Machine Learning	
ML	Machine Learning
DL	Deep Learning
AI	Artificial Intelligence
DNNs	Deep Neural Networks
CNNs	Convolutional Neural Networks
XGBoost	Extreme Gradient Boosting
ResNet	Residual Network
FM	Foundation Model
SSL	Self-Supervised Learning
CLEF	Clinically-Guided Contrastive Learning for ECG Foundation Models
CLEF-S, CLEF-M, CLEF-L	CLEF small, medium, large variants
UMAP	Uniform Manifold Approximation and Projection
PCA	Principal Component Analysis
PCs	Principal Components
GradCAM	Gradient-weighted Class Activation Mapping
Evaluation Metrics	
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
R^2	R-squared (Coefficient of Determination)
SD	Standard Deviation
Metrics	
Hz	Hertz
bpm	beats per minute
μmol	micromole
mg/dL	milligrams per deciliter
μV	microvolts
mV	millivolts

2 Abstract

Serum creatinine is a vital biomarker, yet its measurement typically relies on invasive and time-consuming blood sampling. While recent advances in Deep Learning (DL) have shown promise in identifying biomarker-related changes from electrocardiograms (ECGs), continuous regression of creatinine has not been proven feasible in the literature yet. This thesis investigates the efficacy of various machine learning (ML) paradigms, particularly the potential role of self-supervised foundation models (FM), for the non-invasive regression of biomarker levels from single-lead ECG signals. These FMs may offer significant advantages by utilizing self-supervised pre-training on vast unlabeled datasets to mitigate the scarcity of paired ECG-biomarker labels. This approach can enable the extraction of complex, non-linear representations for downstream tasks, which traditional supervised architectures often fail to capture from scratch, especially when target biomarker signatures are subtle. Leveraging two diverse datasets, consisting of the large-scale MIMIC-IV database and a smaller clinical dataset from the Telehealth Competence Center (TCC), this research benchmarks traditional Gradient Boosting (XGBoost) and standard Residual Network (ResNet) architectures against the emerging Clinically-Guided Contrastive Learning for ECG Foundation Models (CLEF). The results indicate that continuous creatinine values cannot be reliably predicted from ten-second single-lead ECGs, likely because these brief recordings do not contain the physiological signatures necessary for accurate creatinine regression. This observation is supported by the high accuracy achieved when applying the identical pipeline to heart rate (HR) prediction, which serves as a validation for the employed methodology. Although the use of the CLEF FM did not yield a performance breakthrough, its backbone embeddings capture meaningful ECG characteristics and provide benefits such as reduced computational costs compared to training DL models from scratch. Consequently, this work establishes a framework for future research into non-invasive and continuous biomarker monitoring, particularly for applications where labeled data is scarce and Self-Supervised Learning (SSL) techniques can be leveraged effectively.

3 Introduction

In medical practice, biomarkers serve as crucial indicators of a patient's health status [1]. Once extracted, the laboratory values can give important quantitative insights about a patients physiological condition and possibly life threatening risks, supporting clinicians to make informed decisions [1, 2]. Traditionally, these markers are measured through the analysis of bodily fluids such as blood or urine [1, 3]. While highly accurate, this process is inherently invasive, resource-intensive, and captures only a single point in time rather than enabling continuous monitoring [1–3]. Creatinine, a primary biomarker used to assess kidney function, is essential in detecting Acute Kidney Injury (AKI) or Chronic Kidney Disease (CKD) [4]. The transition toward digital biomarkers, extracting physiological insights through non-invasive and faster methods, offers a paradigm shift [3]. Specifically, the Electrocardiogram (ECG) has emerged as a rich source of data, as abnormal laboratory values often manifest as subtle alterations in cardiac electrical activity [1, 2, 5–8].

Deep Neural Networks (DNNs) have been proven successful in detecting complex patterns within ECG signals [1,3,9,10], and have suggested that laboratory values may be partially predictable from ECGs. A benchmark study by von Bachmann et al. (2024) [3] demonstrated that, while Machine Learning (ML) can approximate potassium levels from ECG signals, the predictive value to date is not sufficient to replace blood sampling ($R^2 \approx 0.34$[1]). The prediction of creatinine levels was even more challenging ($R^2 \approx 0.12$[2]). This benchmark study had access to $\sim 300,000$ ECG recordings with 8 leads paired with potassium and creatinine recordings. In the context of clinical diagnostics, datasets in which ECG signals are precisely timestamped with synchronized laboratory values are rare [11]. Moreover, this research was restricted to single-lead ECG analysis, as only three channels were recorded at the participating hospital, of which only one provided signals of sufficient quality for regression analysis. This raises the question whether creatinine prediction would be feasible in this context.

Consequently, this thesis aims to evaluate the feasibility of continuous prediction of creatinine and other biomarkers from single-lead ECG data. Performance is assessed across two distinct datasets: a small local clinical database from the Telehealth Competence Center (TCC) in Germany, and the large-scale MIMIC-IV research database from the USA. These sources provide a valuable contrast, as the TCC dataset reflects the practical challenges of limited data availability, while the MIMIC-IV dataset serves as a robust and widely utilized benchmark in the field [1, 2, 8, 12, 13]. To mitigate the challenges associated with predicting subtle patterns from limited data, the potential use of SSL and the development of ECG Foundation Models (FMs) [7, 12, 14–18] is assessed. These models do not rely on labeled data for training. Instead, they are pre-trained on self-supervised tasks. This enables models trained on millions of unlabeled waveforms to learn robust latent representations of cardiac physiology. Such a model should, in theory, transform raw waveforms into a feature space where clinical targets can be predicted through simple linear aggregation or fine-tuning, even with limited labeled data. By comparing the Clinically-Guided Contrastive Learning for ECG Foundation Models (CLEF) approach to methods based on traditional Gradient Boosting (XGBoost), and standard Deep Learning (DL) (ResNet), this work addresses the following research question: *To what extent can self-supervised DL methods improve performance of Electrocardiogram based biomarker regression when dealing with small amounts of data and which methodological choices, if any, contribute most to this improvement?*

[1] Note that von Bachmann et al. reported a Pearson correlation coefficient of $R \approx 0.58$, which translates to $R^2 \approx 0.34$; this may differ slightly from the direct R^2 calculation used in this study.

[2] As noted in the preceding footnote regarding the study by von Bachmann et al., $R \approx 0.35$ translates to $R^2 \approx 0.12$.

Given the inherent complexity of creatinine regression even in the idealized case of a large dataset with 8 leads [3], it remains uncertain to what extent this task is solvable at all. Therefore, a multi-stage validation strategy is employed and the benefit of using CLEF is additionally evaluated on the task of heart rate (HR) prediction, a more tractable regression task with known ECG correlates [19]. This establishes whether the difficulty in regressing creatinine is inherent to the biomarker's representation or a result of the methodological approach itself. Furthermore, the study incorporates a rigorous assessment of data quality and an examination of the latent representations produced by the CLEF approach, determining the extent to which self-supervised features capture the essential physiological markers required for clinical monitoring.

The remainder of this thesis is structured as follows. First, a review of relevant literature establishes the current state of ECG-based biomarker estimation and the evolution of FMs in cardiology. This is followed by a detailed description of the methodological framework, including the data pre-processing pipeline and the architectures of the XGBoost, ResNet, and CLEF models. In the results section the models performances are evaluated, initially pinpointing the physiological and technical challenges of creatinine regression before establishing a performance ceiling by applying the same methodology to the more tractable task of HR prediction. This section further validates the latent representations of the self-supervised models to ensure clinical relevance. Finally, in the discussion the findings are interpreted in the context of existing literature, addresses the study's limitations, and proposes avenues for future work before providing concluding remarks.

4 Related Work

This section establishes the study's theoretical framework by presenting background on serum creatinine, electrocardiograms (ECGs), the evolution of ML methodologies for ECG analysis, and a comprehensive review of relevant literature regarding non-invasive ECG based biomarker estimations, with a special focus on the use of self-supervised FMs.

4.1 Serum Creatinine: Clinical Relevance

Serum creatinine is an important biomarker in clinical medicine, used as an indicator for muscle mass, as well as a critical estimator for kidney function [4, 20]. As the metabolic end product of creatine phosphate metabolism in skeletal muscle, it is among the most frequently measured biomarkers in clinical chemistry laboratories around the world [4, 20, 21]. In a healthy state creatinine is produced in dependence on the muscle mass, as well as age, and gender at a constant rate [20, 21]. A rise in serum creatinine levels often signals AKI or CKD [22, 23].

The typical reference interval for creatinine is 0.63–1.16 mg/dL for men and 0.48–0.93 mg/dL for women, with a bias in measurement of around 0.06–0.31 mg/dL [4, 21]. However, in order to detect kidney injury or dysfunction, this reference range is not meaningful enough. Instead, a patients baseline creatinine value is more commonly used as a measure of comparison to detect possible changes in creatinine levels [21, 23]. This baseline value can be defined as the creatinine amount at admission to the doctor or hospital, the lowest inpatient creatinine measured, or other surrogates [23]. An AKI is diagnosed when the serum creatinine value is 1.5 times as large as the baseline value, or if there is an increase of 0.3 mg/dL within 48 hours [22]. CKD is detected by measuring serum creatinine, or more specifically the Glomerular Filtration Rate (GFR), often in

combination with cystatin C [24]. A generally very low amount of serum creatinine (e.g. < 0.6) was also linked to increased mortality and longer stay in intensive care units [4].

Traditionally, serum creatinine measurements are derived through blood sampling which is a reliable and cheap measure [1, 3, 20, 21]. However, this method is also invasive, resource-intensive, and time-consuming [1–3]. Moreover, the need for specialized equipment, trained personnel, and laboratory processing introduces delays that can limit the utility of these tests in fast-paced or resource-limited settings such as emergency care, rural clinics, or hospitals [1, 3, 25]. These challenges point out the need for alternative non-invasive measurement approaches.

4.2 Electrocardiogram (ECG)

In this context, the Electrocardiogram (ECG) emerges as a promising solution, as it is non-invasive, inexpensive, and easily obtainable [3, 18]. In clinical practice it serves as a primary diagnostic tool to evaluate a patients cardiac condition, as well as underlying physiological states. It is therefore commonly used in emergency care where patients undergo regular medical assessments [1, 7, 26].

A conventional 12-lead ECG is obtained using ten electrodes positioned on the patients limbs and chest, generating multiple voltage signals that reflect cardiac dynamics over time [12, 27]. These time-series signals contain characteristic waveforms which provide clinically meaningful information concerning the heart's state (see fig. 1). The limb leads (I, II, III, aVR, aVL, aVF) view the heart in the frontal plane, with Lead II typically offering the highest signal quality for rhythm analysis and R-peak detection due to its alignment with the heart's main electrical axis [28, 29]. Due to their proximity to the heart, the precordial leads (V1–V6) record ECG waveforms with higher amplitudes, making peak extraction simpler [28]. However, precordial leads also typically include more noise, resulting from respiration or torso movement. Wearable devices such as smartwatches, an ECG patch, or chest straps are generally limited to a single-channel recording, typically Lead I, II, or a localized chest lead. While this offers high patient compliance and convenience, they often exhibit a lower signal-to-noise ratio compared to the clinical standard of Lead II and lack the spatial diversity of a clinical 12-lead setup [30].

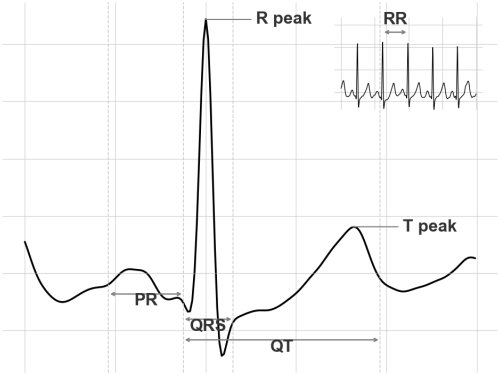

Figure 1: Example ECG strip from the TCC dataset labeled with PR, QRS, QT, and RR intervals, as well as R and T waveform peaks.

ECG recordings are susceptible to various types of physiological and environmental noise that can obscure cardiac features and lead to inaccurate clinical interpretations. The four primary noise sources are baseline wander, powerline interference, muscle artifacts, and electrode motion artifacts [31–33]. Baseline wander, typically occurring

below 1 Hz, stems from respiration, or poor electrode contact and can be mitigated using high-pass filters such as Butterworth, to stabilize the signal [31–33]. Powerline interference appears as a 50 or 60 Hz sinusoidal wave that is commonly addressed via high-pass filters, while high-frequency muscle artifacts and other rapid artifacts caused my body movement are reduced using low-pass filters [31, 33]. By implementing these specific filtering techniques, the clinical integrity of the waveform is preserved, ensuring that the physiological analysis remains accurate and free from misleading interference.

ECG waveforms can provide crucial insights into health outcomes, such as morbidity and mortality [27]. Since ECG acquisition is non-invasive and simple, it is well suited for continuous monitoring, and early identification of cardiovascular abnormalities such as arrhythmias, myocardial infarctions and HR variability [8]. Beyond its established role in cardiology, the ECG is increasingly viewed as a valuable signal for remote healthcare applications, such as telemonitoring with Holter devices, and as a potential source of information for estimating non-cardiac physiological markers such as laboratory values [18, 34].

Nevertheless, the interpretation of ECG waveforms remains a challenging task, even for experienced cardiologists [34]. This difficulty is partly explained by "individual variations, complex waveforms, and susceptibility to noises" that can distort the recordings [8]. Moreover, the dependence on expert clinicians for ECG interpretation is resource intensive and limits scalability, which can result in higher diagnostic error rates [8]. Therefore, automated analysis of ECGs through ML has gained popularity [8, 27].

4.3 ECGs and Machine Learning

The automated analysis of ECGs has been explored by researchers since the the early 1960s [27]. Due to its long research history, various ML approaches have been explored [27].

Early automated systems such as XGBoost algorithms or logistic regression relied on handcrafted ECG features as inputs [3, 9]. These features consist of ECG characteristics such as the time interval between successive waves, their slopes, or the amplitude of peaks [3]. While interpretable, handcrafted features often fail to capture the high-dimensional, more subtle non-linear relationships inherent in complex waveforms.

Most recent studies focus on DL techniques using DNNs, which have been shown to consistently outperform traditional ML methods in various ECG analysis tasks [3, 7, 27]. This is due to their ability to automatically extract meaningful features and therewith, detect patterns which are not visible to the human eye [3, 8]. DL models can for instance identify myocardial infarction from ECGs without typical ST-segment elevation and predict risks such as mortality, atrial fibrillation, and left ventricular dysfunction [3]. Mathematically, DL enables the modeling of high-dimensional, hierarchical, and non-linear relationships that are not representable by conventional methods [9].

Hence, DL based approaches have been reported to achieve diagnostic performance exceeding that of human experts, which has stimulated interest in artificial intelligence (AI) enhanced ECG interpretation and reshaped diagnostic perspectives [8, 26]. DNNs include various architectural designs, of which Convolutional Neural Networks (CNNs), particularly ResNet- and Inception-based architectures have demonstrated the strongest performance for ECG prediction tasks [34]. These AI driven prediction models have shown strong potential to support clinical decision-making in acute and emergency care, where fast and accurate assessment is critical [13]. Such models may facilitate early diagnosis, guide diagnostic workups, predict hospital admissions, estimate survival, and enable faster, low-cost clinical assessments [13].

Besides the detection of heart related disorders such as cardiac arrhythmia's, DL models were also able to identify non-cardiac factors from ECGs [26, 27]. Those include

patient characteristics such as age and gender, as well as electrolyte imbalances [3, 8, 34]. Various articles prove that DL systems can recognize ECG patterns associated with a wide range of other pathological conditions, thereby extending the clinical relevance of the ECG to early detection of patient deterioration, as well as predicting a diverse range of conditions [1, 26, 27]. Holmstrom et al. for instance used a CNN to detect CKD from ECG waveforms, achieving an AUC of 0.767 [35].

Nevertheless, it is important to mention, that DL may also introduce challenges and risks compared to simpler ML models. This includes the problem that DL models offer limited insight into the reasoning behind their predictions which makes it difficult to explain why certain results occur [27]. Hence, clinicians often prefer simpler models, which have a greater interpretability, although their performance might be worse. Transparency, trust, and interpretability are especially essential in clinical settings with real-world adoption. To address this, techniques such as GradCAM have been developed to visualize and identify which parts of the input a DL model is using to base their prediction on. Hicks et al. (2021) have even developed a specialized version (ECGradCAM) for ECGs which can be applied not only to understand outcomes, but also to detect new ECG features [27]. Furthermore, DL models typically require large and diverse labeled datasets to learn effectively [1, 3, 9, 10]. When trained on limited sample sizes, they frequently struggle to generalize to unseen populations, making overfitting a persistent challenge [9,36]. This issue is especially pronounced in the prediction of laboratory values from ECGs, as precisely paired ECG–laboratory data are often scarce [11]. This study aims to address this limitation by exploring the use of FMs to leverage pre-learned ECG features for laboratory value prediction, potentially improving model performance and robustness.

4.4 ECG Based Prediction of Laboratory Values

Extracting laboratory values from ECG signals is an inherently challenging task, as these values do not always manifest as significant changes in cardiac rhythm [5,18]. Nevertheless, there is a known fundamental connection between ECG signals and laboratory values, which is rooted in cardiac electrophysiology [37]. Cardiac electrical activity is driven by the movement of ions across cell membranes which creates the action potentials recorded by the ECG [37,38]. Because electrolytes and metabolites such as potassium and creatinine directly influence these cellular processes, the resulting ECG waveforms serve as a non-invasive reflection of a patient's systemic biochemical state [3,38]. This relationship is most evident in the study of electrolytes, especially potassium, with hypokalemia representing the most frequently observed clinical electrolyte abnormality [5, 6, 10, 37, 39, 40].

Early studies provide evidence that the classification of biomarker levels based on ECGs is possible in principle, yet it remains a difficult task. These studies relied on classifying abnormal hypo (low) and hyper (high) concentration levels from handcrafted ECG features, using simple ML models, or discretized concentration levels with employed methods comparable to ordinal regression [3]. Frohnert et al. (1970) first manually derived a link between such features and potassium concentrations [41]. Since then, various articles reported associations between ECG morphologies and high potassium values (hyperkalemia), such as large peaking T waves, prolonged QRS and PR intervals, smaller P wave amplitudes, and reduced QT intervals [1, 5, 6, 42]. Nevertheless, these more recent works also highlight that these abnormalities are only found in around 22% of hyperkalemic patients, therefore underscoring the limited diagnostic specificity of these handcrafted markers [3, 5].

As for the field of diagnostics using ECG, recent research on biomarker prediction has also shifted towards DL methods [3]. Particularly CNNs such as ResNets have been found suitable for ECG based prediction tasks [3, 9, 11, 18]. While the majority

of research has focused on potassium, recent work has expanded into other biomarkers and more sophisticated architectures [1, 3]. Alcaraz et al. (2025) for instance made use of Structured State Space Sequence (S4) models to predict abnormalities of various biomarkers [1]. S4 models have shown superiority over standard CNNs and Transformers in capturing the long-range temporal dependencies inherent in ECG signals. Overall, both ResNet style CNNs and S4 models remain prevalent in current literature on biomarker prediction [1, 12, 17, 36].

While the vast majority of research in automated laboratory value estimation utilizes the standard clinical 12-lead ECG, there is a growing interest in the potential of reduced lead systems [1, 12]. The 12-lead ECG remains the gold standard because it provides a comprehensive, three-dimensional view of the heart's electrical axis, offering diverse perspectives that are crucial for detecting subtle systemic changes [12]. For instance, Ribeiro et al. (2020) demonstrated the high diagnostic performance (F1 $>$ 80%) of DNNs in predicting ECG abnormalities when trained on these rich, multi-dimensional signals [43]. However, the requirement for a 12-lead setup, involving ten electrodes and specialized clinical equipment, restricts laboratory monitoring to hospitals or clinics, making continuous, long-term tracking in everyday life impractical [1, 17].

Therefore, single-lead ECG recordings have become prevalent due to the rise of consumer wearables and smartwatches [1, 10, 18]. These devices offer a more frequent and accessible means of data collection, yet they present a significant technical challenge: a single-lead provides only a restricted perspective of the heart's electrical activity, potentially missing biomarkers that appear only in specific chest or limb leads [12]. Despite this information gap, recent studies have started to prove the feasibility of single-lead prediction. Chiu et al. (2024) introduced an AI-enabled smartwatch model capable of monitoring serum potassium levels in patients with end-stage renal disease [1, 10]. Furthermore, FMs such as CLEF and AnyECG-Lab have shown that representations learned from 12-lead ECG data can be transferred to single-lead scenarios, with the latter reaching an AUC of 0.709 in creatinine classification [17, 18]. Exploring creatinine prediction using a single-lead is thus a highly important direction. It marks a key step toward shifting kidney function monitoring from occasional clinical blood draws to continuous, wearable-driven health monitoring.

While many studies rely solely on ECG data as input, other researchers integrate multimodal data to provide a more comprehensive context for the model. Kwon et al. (2021) moved beyond single-lead analysis to multimodal inputs, combining single-lead ECGs with age and sex to classify potassium, calcium, and sodium abnormalities, achieving an AUC of 0.903 for hyperkalemia detection [9]. More recently, Alcaraz et al. (2025) utilized extensive multimodal data including demographics, biometrics, vitals, and 12-lead ECG recordings to predict abnormalities in laboratory values, achieving AUROCs of 0.788 and 0.746 for creatinine abnormality classification ($\geq$ 1.2 mg/dL and $\leq$ 0.5 mg/dL, respectively) [1]. In another paper Alcaraz et al. (2025) demonstrate that multimodal models incorporating 12-lead ECG waveforms and other clinical data significantly outperform unimodal 12-lead ECG waveform models on diagnostic tasks, achieving a macro AUROC of approximately 0.83 compared to 0.77 [13]. They also highlight the added benefit of utilizing raw ECG signals in combination with DL models, rather than handcrafted ECG features with simple tree based models for clinical prediction tasks. Although the integration of multimodal inputs may enhance predictive performance, the clinical data required for such models may not always be available in practice, underscoring the need for robust signal-only models that ensure broad applicability across diverse real-world settings.

Despite the progress in binary classification, treating laboratory value estimation as a continuous regression problem remains rather underexplored [2, 3, 25]. Deep direct regression, where a model minimizes the Mean squared Error (MSE) to predict precise

values, offers more granular monitoring compared to binary classification [3]. Furthermore, because the clinical classification of creatinine levels is fundamentally dependent on an individual's established baseline [21, 23], performing direct regression becomes a more significant and clinically relevant objective.

Von Bachmann et al. (2024) provided the first benchmark for the continuous regression of potassium, calcium, sodium, and creatinine from ECG signals using DL techniques [3]. Their employed dataset comprised a substantial cohort of 166,908 patients and 295,606 ECG recordings, from six different emergency departments in Sweden. They used 10 second, 8-lead ECG samples at 400Hz and matched them with laboratory values laying within ± 60 minutes. Their study population reflects a balanced gender distribution, had a mean age of around 61 years, and a mean creatinine concentration of approximately 1.02 mg/dL ($\approx 90.55\ \mu mol/l^3$), with a standard deviation (SD) of 0.80 mg/dL ($\approx 71\ \mu mol/l^3$). Consequently, the dataset by von Bachmann et al. is characterized by its large sample magnitude, high-resolution comprehensive spatial representation, and multicenter diversity. While their results confirmed that DNNs consistently outperform traditional ML baselines such as XGBoost and Random Forest, they also revealed a significant performance gap. While potassium regression showed potential (Pearson correlation of 0.58), predictive accuracies for continuous concentrations of creatinine (Pearson correlation of 0.35) remained rather poor with MSE of 0.48 mg/dL ($\approx 3719\ \mu mol/l^3$), and Mean Absolute Error (MAE), of 0.30 mg/dL ($\approx 26.69\ \mu mol/l^3$). This suggests that creatinine prediction from ECG waveforms is a more complex task than electrolyte prediction, potentially requiring more advanced architectures or larger labeled datasets. Alternatively, the inherent difficulty of the task could indicate that the upper limit of achievable accuracy has already been reached.

4.5 Self-Supervised Learning - Foundation Models

A primary bottleneck in training high performing DL models for laboratory regression is the limited availability of datasets where ECGs are paired with precise laboratory labels. To address this, recent literature has proposed the use of Self-Supervised Learning (SSL) and the development of FMs [44].

The objective of this approach is to leverage large volumes of unlabeled ECG data to learn generalized representations of cardiac health before fine-tuning the model to a downstream task on a smaller, labeled dataset [12, 36]. This two-stage process of pre-training and then transfer learning enables models to acquire domain knowledge, e.g. understanding complex waveforms and individual variations, without relying solely on scarce human annotated samples [8, 11]. An assumption of this method is that the knowledge transfer from the pre-trained model backbone consistently enhances the downstream performance, in contrast to training from random initialization [11].

Standard practice for evaluating these FMs involves either linear probing or fine-tuning [12,44]. In linear probing, the pre-trained backbone is frozen and only a single linear head is trained on top, corresponding to the outputs of the downstream task [17,38,44,45]. This method is designed to measure the quality of learned representations by examining how well the learned representations can be linearly separated [17, 44]. It ensures that any performance gains are a direct result of the rich representations learned during pre-training rather than the complexity of the downstream architecture. Meanwhile, for fine-tuning, the pre-trained backbone weights are unfrozen and updated together with the task specific model head [44]. While fine-tuning typically yields the best performance on small datasets, its advantages such as accelerated convergence, quicker training, and

[3]While von Bachmann et al. (2024) [3] reported values in $mmol/l$, the data distribution suggests $\mu mol/l$ was the intended unit. For comparability, these values were converted to mg/dL using the standard conversion factor of 0.0113.

reduced computational costs, may diminish as downstream dataset size increases, and it does not necessarily improve performance in all cases [11, 34].

While the field has long been dominated by CNNs, specifically 1D-ResNet architectures, which are highly effective at spatial and temporal feature extraction, new architectures are emerging to challenge this standard [7,36]. Models such as ECGFounder and ECG-FM rely on ResNet backbones to process raw waveform data, preserving critical information that traditional feature engineering might lose [12, 16]. More recent research by Mehari and Strodthoff (2023) has introduced S4 models as a more expressive alternative [7]. In comparative benchmarks, S4 architectures combined with SSL (such as Contrastive Predictive Coding) have outperformed both ResNets and transformers, achieving state-of-the-art results in 12-lead ECG analysis (macro AUC ≈ 0.94 for 71 diagnostic, form-related, and rhythm-related labels) [7]. Transformer based architectures have been explored for their ability to focus on specific segments of the ECG, though they can struggle with the computational load of very long sequences compared to S4 models [7, 16].

A prominent and novel ECG FM is CLEF (Clinically-Guided Contrastive Learning for Electrocardiogram Foundation Models) [17]. Most traditional contrastive learning frameworks (e.g., SimCLR) learn by comparing two augmented views of the same signal, pulling similar samples together and pushing dissimilar samples apart using an InfoNCE loss. However, this can be problematic in clinical settings where two different patients might have identical cardiac pathologies. A standard model would erroneously penalize the system for placing these clinically similar signals close together in the latent space [17]. To resolve this, CLEF introduces a clinically-guided contrastive loss coupled with a dissimilarity alignment loss in the pre-training phase. By incorporating clinical risk scores from Electronic Health Records (EHR) into a metadata similarity framework, the model organizes its latent space based on underlying systemic health states rather than just waveform geometry. Instead of treating all other patients as strict negatives, the Clinically-Guided Contrastive Loss uses the similarity matrix to apply soft weights [17]. If two patients share similar clinical risk profiles, the repelling force between their embeddings is lowered [17]. By training on a ResNet backbone and evaluating via linear probing, CLEF demonstrates that incorporating clinically-guided signals allows the model to learn more robust features for downstream laboratory prediction tasks, such as estimating biomarker levels from single-lead or 12-lead ECGs [17,38]. Furthermore, while pre-trained on 12-lead data, the model supports fine-tuning on specific leads, ensuring broad applicability across various clinical monitoring setups.

4.6 This Thesis

Building upon past research, this thesis seeks to bridge the gap in continuous creatinine prediction. By employing traditional ML (XGBoost) and standard DL (ResNet) as baselines, and comparing them against the representations learned by the CLEF FM, this study aims to determine if SSL-derived features can overcome the inherent complexity and data limitations associated with non-invasive creatinine monitoring. The performance of the smaller TCC dataset, is compared to MIMIC-IV, a publicly available dataset, to determine the usefulness of pre-learned representations at different dataset sizes and population cohorts.

5 Methods

The following sections describe the methods used for all pipeline steps in more detail. Figure 2 visualizes these steps in a simplified flowchart format. The pipeline begins with data loading and moves through data preparation and feature extraction, before splitting the data for model training and evaluation. The training phase evaluates three distinct architectures, namely XGBoost, ResNet, and CLEF, after which the process concludes with a performance evaluation and final validation steps including a task validation, data validation, pipeline validation, and CLEF latent representation validation. All steps were implemented in Python 3.12.10 [46].

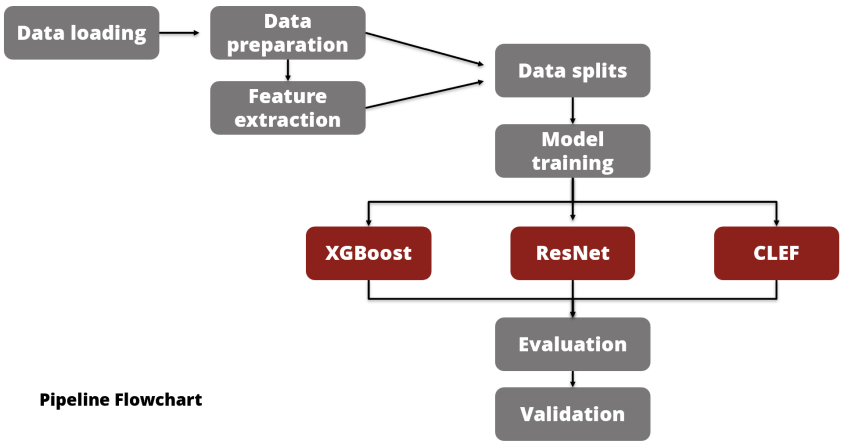

Figure 2: Flowchart of employed pipeline.

5.1 Data Loading

All models were trained on two datasets: TCC, a dataset containing data from German hospitals, and MIMIC-IV, a popular research dataset from the US. In this section the collection of the datasets is described in detail. All relevant steps were performed using a combination of SQL and Python commands.

5.1.1 TCC

The TCC stores data of their customers in a central, anonymized PostgreSQL database. The relevant data for the purpose of this study was extracted using SQL queries. Data samples were filtered for the intensive care unit (ICU) of a German hospital. For the purpose of this study, only the second lead (ECG_II) which measures ECGs from the patient's extremities was used, as only one of the three channels recorded at the clinical site maintained the signal-to-noise ratio necessary for robust regression analysis. The monitors which record the ECG waveforms are Dräger monitors [47]. The study population was restricted to adult patients aged 18 to 120, with data samples spanning measurements recorded between May 2, 2025, and December 13, 2025. All ECG recordings in the TCC dataset have a standardized sampling frequency of 200 Hz.

Laboratory serum creatinine results were extracted from the TCC database and to ensure a robust and clinically relevant target distribution, creatinine values were constrained to a range of $0 < x \leq 15$ mg/dL. For each laboratory creatinine measurement, a corresponding ECG window was defined by extracting waveform data spanning 5 minutes before and 5 minutes after the recorded laboratory timestamp, resulting in a

total interval of 10 minutes per laboratory identifier. This approach ensured temporal alignment between physiological signals and laboratory values while capturing relevant pre- and post-measurement cardiac activity. This led to a final dataset consisting of 61,651 datarows of 1,133 unique patients with 2,837 unique creatinine measurements. A comprehensive summary of the TCC creatinine dataset characteristics is available in appendix 8.

Moreover, HR measurements were retrieved from the TCC database and were aligned with the ECG recordings which had already been extracted for the creatinine dataset, using a nearest-neighbor, time-based merge, such that each ECG recording was paired with the closest HR measurement within a ±15 minute window. A plausibility filter was not necessary, as values lied within the plausible range of $25 < x < 300$. Therefore, the HR dataset incorporated a total of 59,515 datarows of 1,123 unique patients with 2,755 unique HR measurements. More details on the TCC HR dataset are presented in appendix 10.

For a sociodemographic description of these datasets, age and gender were added for each patient. This description can be found in appendix A.

5.1.2 MIMIC-IV

For comparison, the publicly available MIMIC-IV dataset from Physionet was used, including ECG data from MIMIC-IV-ECG [48] and creatinine and HR recordings from MIMIC-IV-v3.1 [49]. The MIMIC-IV-ECG database consists of standard 12-lead diagnostic ECG recordings, each 10 seconds long and sampled at a frequency of 500 Hz. Meanwhile, the MIMIC-IV-v3.1 database contains various laboratory data. The employed MIMIC-IV databases are anonymized and contain medical records of patients admitted to the ICU or emergency department of the Beth Israel Deaconess Medical Center, in the USA, between the years 2008 and 2022. The ECG recording monitors are from manufacturers including Burdick/Spacelabs, Philips, and General Electric [50–52].

Three source tables were used from MIMIC-IVs database to combine ECG measurements with matching serum creatinine recordings using SQL queries. ECG timestamps, along with patient and study identifiers, were extracted from the machine measurements table in MIMIC-IV-ECG. Creatinine measurements were obtained from the labevents within the hospital events recorded in MIMIC-IV-v3.1.

The datasets were merged using the common patient identifier, and only laboratory values within a ±60 minute time interval around each ECG were kept. Similarly to the TCC data, a clinically plausible range of $0 < x \leq 15$ mg/dL was employed, leading to a total of 202,749 datarows of 103,062 unique patients with 201,901 unique creatinine recordings. A more detailed description of the data is presented in appendix 9.

To obtain a matched dataset of ECG recordings with corresponding HR values, the previously constructed MIMIC ECG–creatinine dataset was expanded with HR measurements. These measurements were extracted from the vitalsign table of the MIMIC-IV-ED 2.2 database [53], which contains time-stamped physiological measurements for each patient. The two datasets were then linked using a nearest-neighbor join on the patient identifier, matching each ECG recording to the temporally closest HR measurement within a tolerance window of ±15 minutes. Rows without a corresponding match within this interval were excluded, resulting in a dataset of ECG recordings paired with the nearest available HR measurements in close temporal proximity. Consistent with the filtering criteria used for the TCC dataset, HR values were filtered with a plausibility range of $25 < x < 300$. This led to a total of 36,117 datarows from 28,402 unique patients, with 36,117 unique HR measurements. A summary of the data can be found in appendix 12.

Similarly to the TCC dataset, age and gender variables were gathered from MIMIC-

IV-v3.1's patients table to get an insight into the sociodemographic distributions of the datasets which are described in detail in appendix A.

5.2 Data Preparation

This section details the comprehensive pre-processing pipeline conducted prior to model training, which encompassed waveform standardization, signal cleaning, filtering, and normalization. Additionally, it describes the quality assessment of ECGs, a capping method designed to mitigate patient-level bias, feature extraction techniques, and the data splitting method.

5.2.1 ECG Waveforms: Standardization

For the TCC datasets, ECG recordings had to be segmented into fixed-length chunks to ensure suitability for downstream tasks. A segment duration of approximately 10 seconds was targeted, as this length is sufficient to capture multiple cardiac cycles and maintaining manageable input sizes for modeling. At a sampling frequency of 200 Hz, this would correspond to 2000 samples. However, the segment length was adjusted to 2048 samples (i.e., 2^{11}) before training to align with computationally efficient input sizes. This results in an effective segment duration of 10.24 seconds, which remains close to the intended temporal resolution while offering practical advantages for signal processing and model implementation. This led to approximately 58 discrete waveform segments per 10 minute interval. To preserve data integrity and avoid information leakage during model training and evaluation, it was ensured that ECG segments are non-overlapping in time.

Moreover, to ensure a consistent comparison between the TCC and MIMIC-IV datasets, MIMIC-IV ECG samples were resampled to 200 Hz. Given that the MIMIC-IV recordings were pre-segmented into 10 second intervals, this resampling initially yielded a sequence length of 2000 samples. To align these sequences with the TCC data format, the segments were zero-padded to a fixed length of 2048 samples prior to model training. Furthermore, a secondary version of the MIMIC-IV HR dataset was preserved at its original 500 Hz sampling rate. This higher resolution version was kept to evaluate performance when used as input for the CLEF approach. At this sampling frequency, the 10 second temporal windows correspond to a sequence length of 5000 samples.

5.2.2 ECG Waveforms: Cleaning, Filtering, Normalizing

To ensure high quality ECG signals, various cleaning and filtering steps were performed. Empty signals or signals containing no valid finite numbers, as well as signals with implausibly high amplitudes were removed.

Moreover, various functions of the popularly known NeuroKit2 library [54] were used for ECG pre-processing [31, 32, 55]. ECGs were cleaned using NeuroKit2's ecg_clean function with the neurokit method. This function removes artifacts by employing a high-pass Butterworth filter with a cutoff frequency of 0.5 Hz and order 5 to eliminate baseline wander. Furthermore, the signal is passed through a Powerline noise filter to suppress 50 Hz noise within the signal. The result is a cleaner ECG waveform with reduced low-frequency drift and high-frequency electrical noise, making it more suitable for accurate downstream analysis. Furthermore, inverted ECGs were detected and inverted back to the regular format using NeuroKit2's ecg_invert function.

To ensure physiological consistency and numerical stability, signal normalization was performed on TCC ECG data by converting raw amplitude values from microvolts (μV) to millivolts (mV). A scale factor of 10^{-3} was applied across all samples, standardizing the waveforms amplitudes to a range consistent with standard clinical ECG representation.

This step prevents large integer inputs from causing gradient instability during the backpropagation process when training DNNs.

5.2.3 Data Quality

Data Quality was assessed before and after all pre-processing steps were performed by calculating the quality score using NeuroKit2's ecg_quality function with the templatematch setting. This score ranges between -1 and 1, where higher values indicate better signal integrity. Boxplots visualizing the data quality can be found in appendix D.

The initial unprocessed data shows a high median quality score for both TCC and MIMIC-IV datasets, positioned near one. The TCC creatinine dataset has a notably tighter interquartile range, suggesting that the middle 50% of the data is very consistent and concentrated at the top of the scale. Meanwhile the MIMIC-IV creatinine dataset shows a slightly larger spread in its interquartile range, indicating more variability in signal quality compared to the TCC data. Both datasets show a significant number of negative outliers. MIMIC-IV appears to have a higher density and a lower reach of poor quality samples, with some scores dropping toward -0.4. TCC has fewer extreme outliers, though one single point reaches a low of -0.5.

After pre-processing the data, a quality threshold of ≥ 0.75 was set, ensuring that both datasets contain ECGs of high quality. The introduction of the quality threshold reduced the TCC dataframe by 11,685 samples, and the MIMIC-IV dataframe by 14,159 samples.

5.2.4 Mitigating Patient-Level Bias Through Capping

Two capping strategies were applied to normalize the contribution of individual patients and prevent those with numerous laboratory entries from disproportionately biasing the model.

First, the number of laboratory entries per case was capped at six. This was achieved by binning ECG-creatinine samples into three categories (low, mid, and high) based on creatinine values using the bounds $[0.5, 1.2]$. Within each bin, reservoir sampling was used in a round-robin manner to select representative entries. This binning procedure ensures diversity across the creatinine range and maintains a balanced distribution that prevents the model from favoring a specific concentration level. Thus, the limit of six was chosen to allow for a maximum of two entries per bin in the highest-laboratory scenario, ensuring that each of the three categories is sufficiently represented.

Second, the same approach was used to cap the number of ECG segments per patient at 58. This specific limit was derived from the temporal constraints of the available data: for each laboratory entry, a 10-minute (600-second) window of ECG data was considered. Dividing this total window by the length of a single ECG snippet (10.24 seconds) results in approximately 58 possible segments ($600/10.24 \approx 58.59$), which was also the most common number of ECGs available per patient. This capping step ensures that patients with a greater number of cases, and consequently more associated ECG data, do not disproportionately influence the model. By limiting the number of segments per patient, the model maintains a more balanced representation of the population cohort across the entire dataset. Limiting the amount of ECGs per patient is a strategy that was also taken by other authors such as for instance Kwon et al. who were even stricter and allowed one ECG sample per patient only [9].

The complete pre-processing pipeline led to an overall reduction of 36,945 samples within the MIMIC-IV creatinine dataset, with the original dataset containing 239,694 data rows and the cleaned dataset 202,749 data rows. Due to this removal of data rows,

1,759 patients were excluded from the original dataset. For the TCC creatinine dataset significantly more data rows were removed, namely 232,402 rows, reducing the total number of rows to 61,651. Thus, data of 20 unique patients was removed from the TCC creatinine dataset.

5.2.5 Feature Extraction

To compare against learned features obtained through DL and to train the XGBoost model, a total of 23 handcrafted features were extracted, characterizing the morphology and rhythm of the cleaned ECG signals. The ECG signals were processed using NeuroKit2's ecg_process library which provided standard ECG features. These include general statistical descriptors such as the *mean*, *median*, *SD*, *range*, *skewness*, and *kurtosis* of the waveform, as well as the 'ECG rate mean' feature which represents the mean HR. Furthermore, domain-specific morphological features were included to capture the characteristics of individual cardiac components (P, Q, R, S, and T wave indices) of which the amplitudes mean and SD were calculated. Finally, temporal features including the mean and SD of the RR interval, the means of PR, QRS, and QT intervals, as well as the mean of the HR-corrected QTc interval were derived to quantify the cardiac rhythm.

5.2.6 Dataset Splits

To ensure model generalizability and prevent data leakage, a robust data splitting method was implemented. The dataset was partitioned into train, validation, and test sets (70/15/15) using a Greedy Stratified Group Shuffle Split algorithm. This approach addresses two critical requirements in clinical ML: maintaining independence between patient cohorts and preserving the statistical distribution of the target variable across all subsets.

Prior to splitting, the continuous target variables were discretized into three distinct physiological categories representing low, middle, and high values. These categories were assigned based on the 25^{th} and 75^{th} percentiles of the overall target data distribution of the relevant dataset. Because patients often contribute a varying number of records, the algorithm employed a greedy optimization heuristic to assign entire patient groups rather than individual rows. By sorting patients by their total record count and iteratively assigning them to the partition that minimized the MSE between current split proportions and the global targets, the pipeline effectively balanced clinical prevalence across the subsets while ensuring that no individual patient's data spanned multiple splits.

5.3 Model Training

Three models were trained in this study: XGBoost, ResNet, and CLEF. This section describes each model in detail. XGBoost and ResNet served as baseline models for comparison with CLEF, and additional reference models were trained to support performance evaluation, as described below. All models were trained with fixed seeds, to ensure reproducibility.

5.3.1 XGBoost - Simple ML Baseline

The first baseline model was employed using an Extreme Gradient Boosting (XGBoost) regression architecture implemented via the XGBRegressor from the XGBoost library. The extracted 23 handcrafted features were used as inputs for training.

The model was configured to optimize the squared error objective function. The standard hyperparameters were set to 125 estimators, a maximum tree depth of 6,

a learning rate of 0.01, and a subsampling rate of 0.8 to introduce stochasticity and mitigate overfitting. Besides these standard settings, a hyperparameter tuning pipeline was created using grid search with 3-fold cross-validation from the sklearn library. The search space systematically evaluated combinations of the number of estimators (100, 125, 150), maximum tree depths (4, 6, 8), learning rates (0.01, 0.05, 0.1), and subsample ratios (0.8, 1.0), selecting the optimal configuration based on the lowest MSE. Hyperparameter tuning was employed when predicting biomarkers (creatinine, HR).

To assess the predictive performance of each experiment, the experimental design incorporated two baseline paradigms alongside the standard model initialization. First, a shuffled baseline is used to empirically assess the extent to which the model exploits spurious correlations or overfits the training distribution. Therefore, the target values in the training dataset are randomly shuffled. This breaks any real connection between the input features and the target, while keeping the overall distribution of values the same. The performance of this model shows the error rate achievable purely by chance, confirming that the main model's success is based on actual patterns in the data. Second, the mean baseline establishes a naive, zero-skill predictive threshold. For this evaluation, all target observations in the training dataset are replaced with their overall arithmetic mean. Consequently, the model ignores all input features and only learns to predict this constant average value. This sets an absolute minimum performance score, allowing us to accurately measure how much value the XGBoost model and the engineered features add compared to simply repeatedly guessing the average.

5.3.2 ResNet - DNN Baseline

As a baseline DNN model, the ResNet architecture used by von Bachmann et al. was utilized and slightly adapted [3]. The ResNet architecture is a custom one-dimensional Residual Network, implemented using PyTorch, specifically designed to process sequential, unidimensional signals such as ECGs. The network architecture starts with an initial sequence of a 1D convolutional layer, a batch normalization layer, and a ReLU activation function. This is followed by a series of interconnected residual blocks. Each residual block consists of two stacked 1D convolutional layers, both followed by batch normalization, ReLU activation, and a dropout layer with a rate of 10% to mitigate overfitting.

To prevent the vanishing gradient problem commonly found in deep networks, each block utilizes a skip connection that adds the block's input directly to its output. When a block requires downsampling to compress the signal's temporal dimension, the skip connection applies a 1D max pooling layer. If the number of feature filters increases between blocks, a 1x1 convolution is used within the skip connection to match the dimensions. The model progressively extracts higher-level features using filter sizes of 64, 128, 196, 256, and 320, while consistently applying a kernel size of 15. Simultaneously, the temporal sequence is downsampled through the blocks from its original length to consecutive lengths of 2048, 512, 128, 32, and 8. Finally, the output of the last residual block is flattened and passed through a fully connected linear layer to yield the final continuous regression prediction.

The model training was executed in a cloud environment using Google Cloud Vertex AI, leveraging compute clusters equipped with four GPUs to handle the high computational demands. The network was trained to optimize continuous target predictions using the MSE loss function over a maximum of 50 epochs with a batch size of 256. Early stopping was implemented with a patience of 5, stopping training if the validation loss did not decrease after five consecutive epochs. To ensure a stable and effective training process, the initial learning rate was set to 0.0005 alongside a weight decay of 0.001 for regularization. To optimize convergence and further combat overfitting, a dynamic learning rate scheduler was used, reducing the learning rate by a factor of 0.5 if the

validation loss did not improve for 3 consecutive epochs. Furthermore, the gradient norm was clipped at a maximum of 1.0 to prevent exploding gradients. A random signal shift limit of 10% was employed as on-the-fly data augmentation to improve model robustness. Consistent with the experimental framework established for the XGBoost models, the ResNet training also incorporated both shuffled and mean baseline paradigms. Finally, the model weights from the epoch that achieved the lowest validation loss were restored for the final evaluation on the test set.

5.3.3 CLEF - Foundation Model

For the specific tasks of predicting creatinine and HR, the pre-trained CLEF backbone was chosen as a feature extraction engine, processing the raw ECG segments through the frozen ResNet backbone encoder to generate high-dimensional global representation vectors [17].

CLEF was selected as the primary FM backbone for several reasons. First, its architecture is optimized for single-lead configurations, making it compatible with the data available from the TCCs database. Second, the model is pre-trained in a self-supervised setting, therewith giving the TCC the opportunity to pre-train the model themselves in the future, as no labeled ground-truth data is needed. Third, the backbone network architecture of CLEF is based on a ResNet architecture, making it comparable to the ResNet baseline model. Finally, CLEF was pre-trained on the MIMIC-IV-ECG dataset, encompassing over 161,000 patients. This ensures that the model's latent space is already tuned to the specific signal characteristics, noise profiles, and clinical population density of the MIMIC-IV database used in this research.

To balance computational efficiency with representational capacity, the CLEF backbone is available in three configurable scales, which differ primarily in their network depth and the resulting dimensionality of their feature embeddings [17]. The CLEF-Small (CLEF-S) variant serves as a compact model with approximately 448,000 parameters and produces a 256-dimensional vector. For more complex tasks, the CLEF-Medium (CLEF-M) configuration consists of 30.7 million parameters that yield a 1,024-dimensional vector. Finally, the CLEF-Large (CLEF-L) variant utilizes 296 million parameters, outputting a high-capacity 2,048-dimensional embedding. For the purpose of this study, all three backbone CLEF sizes were explored.

A significant challenge in utilizing the CLEF FM was the dimensional mismatch between the architecture's expected input scale and the varying sample lengths of the experimental datasets. While the architecture of the CLEF backbone can technically accommodate varying input lengths through its pooling layers, it was specifically optimized and pre-trained on a standardized sequence length of 5,000 samples. This raised the question of how to best present our inputs to the model, given that the TCC and MIMIC data provide 2,048 and 2,000 samples respectively. To address this issue, three distinct input strategies were implemented: single, double, and interpolated. In the single approach, the ECG waveform was maintained in its native form, passing the original 2,000 or 2,048 samples directly to the model to preserve the raw temporal resolution. The double strategy aimed to align the input with the 5,000 sample benchmark by concatenating the waveform with an exact duplicate of itself and zero-padding the remaining sequence, providing the model with repetitive morphological context within the standardized window. The interpolated methodology employed a continuous signal transformation to align the waveforms with the model's required input dimensions. Utilizing a Fourier-based resampling algorithm via the scipy.signal.resample function, the raw signals were uniformly interpolated across the temporal axis to reach 5,000 samples. Furthermore, to compare these input options with the CLEF model performance for inputs of 5,000 samples, the MIMIC-IV HR dataset with sample rate 500Hz was used.

Thus, to evaluate the predictive utility of the learned representations, CLEF embeddings were extracted across the three model scales (CLEF-S, CLEF-M, and CLEF-L) using single, double, and interpolated input methods. These same input formats were also applied to the raw waveforms, providing a direct baseline for comparing the predictive performance of abstract CLEF representations against that of the original signal data.

To map the complex embeddings produced by the CLEF backbone to the continuous clinical targets (creatinine, and HR), a supervised prediction head was appended. For this prediction head, two distinct architectures were implemented and evaluated. The first model head architecture is a single fully connected linear layer without activation function. The input dimension d is tied directly to the backbone size (e.g., 2048 for CLEF-L). This approach is used to assess the linear separability and quality of the self-supervised representations. As a more complex alternative, a Simple 1D-CNN architecture was tested to evaluate whether further refinement of the latent representations through convolutional filtering improves prediction accuracy. This head treats the embedding vector as a one-dimensional feature map, processing it through a series of three convolutional blocks, which are each followed by batch normalization and ReLU activation. Following these operations, the resulting feature activations are compressed through a global average pooling operation and passed into a final linear projection layer, which maps the aggregated feature maps to a single scalar regression output. While the linear head tests the inherent expressivity of the latent space, the simple CNN head might allow the model to extract and align localized morphological markers specifically tuned on the target biomarker, potentially capturing non-linear relationships that a simple linear mapping might overlook.

The methodological setup yielded a total of 3 x (3 + 1) x 2 = 24 (Input methods x CLEF model scales + raw waveforms x model heads) potential configurations per target (see e.g., appendix F table 14). The networks were trained using a standard supervised learning procedure optimizing for MSE loss over the regression outputs. This process was designed to systematically assess how different combinations of input data, FM scales, and head architectures influence the accuracy of creatinine and HR estimation. The network heads were optimized using the Adam optimizer with a fixed learning rate of 10^{-3} and a batch size of 256. To ensure computational efficiency and prevent overfitting, early stopping was implemented with a patience of 5 epochs, monitoring the validation loss over a maximum of 50 epochs. After the early stopping criterion was met, the model state from the epoch with the lowest validation loss was restored as the final version for testing.

5.4 Evaluation Metrics

After models were trained, evaluations were performed across the original train, validation, and test splits computing Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) scores. All metrics were calculated using the sklearn.metrics library.

MSE and RMSE were prioritized to assess the models' sensitivity to outliers; by squaring the error terms, these metrics disproportionately penalize large deviations. This is particularly relevant in clinical contexts where significant prediction errors could lead to incorrect medical interventions. While MSE provides a strict penalty for variance, RMSE offers a more intuitive view by returning the error to the original scale of the data. In contrast, MAE provides a robust and direct interpretation of the average error magnitude in the target's physical units (e.g., mg/dL for creatinine or bpm for HR).

To assess the overall explanatory power of the models, the R^2 coefficient was calculated. Within the sklearn implementation, R^2 is defined as:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y})^2}$$

where y represents the ground-truth target labels, $\hat{y}$ the predicted labels, n the total number of samples, and $\bar{y} = \frac{1}{n}\sum_{i=0}^{n-1} y_i$ the arithmetic mean of the observed data. An R^2 score of one indicates a perfect fit, whereas a score of zero suggests the model performs no better than predicting the mean. Notably, this specific formula allows for negative values, which serves as a critical diagnostic signal that the model's predictions are worse than predicting the dataset's mean.

To quantify the uncertainty of these metrics, 1,000 bootstrap resampling iterations were performed on the model's predictions for each respective dataset split (train, validation, and test). By sampling the paired ground-truth and predicted labels with replacement, 95% confidence intervals were calculated for all target metrics, ensuring the statistical robustness of the comparative analysis. Following evaluation, all derived model artifacts, configuration parameters, execution times, and evaluation subsets were serialized, versioned and saved for subsequent comparative analysis.

5.5 Data and Task Validation

To ensure the integrity of the data processing pipeline and the physiological relevance of the model's representations, various validation steps were undertaken which are described in detail in this section. These steps include a physician-led pilot study to establish a clinical baseline for waveform analysis, and an analysis of the correlations between ECG morphology and laboratory biomarkers. Finally, the latent spaces generated by the CLEF approach are evaluated through manifold visualizations, correlations with handcrafted features, and predictive modeling to determine how effectively the backbone captures essential ECG waveform characteristics.

5.5.1 Task Validation: Clinical ECG Analysis

To validate the the feasibility of biomarker prediction from single-lead ECG data, as well as the quality of our ECG signals for this task, a clinical validation pilot was conducted in collaboration with a doctor at the TCC. Potassium was selected as the target for this pilot because it is a biomarker known to have relatively strong correlations with specific ECG morphological changes, such as T-wave peaking, QRS widening, and QT shortening [5, 6]. A balanced subset of $n = 50$ ECG samples, previously matched with confirmed laboratory potassium values was selected for manual annotation. This subset comprised 25 cases of hyperkalemia and 25 cases of normokalemia (normal potassium). The physician, blinded to the ground-truth laboratory results, performed a qualitative assessment of the ECG waveforms to identify characteristic morphological changes associated with potassium imbalances and classify them into hyperkalemia or normokalemia cases. This qualitative review aimed to establish a clinical baseline for the quality of ECG signals evaluated by a professional clinician, compared to the automated quality assessment performed using NeuroKit2.

5.5.2 Biomarkers and ECG Feature Correlations

As a first indicator of the feasibility of predicting creatinine from ECG recordings, the handcrafted ECG features were correlated with creatinine values using Spearman correlation. Besides the correlation to creatinine, other biomarkers and vital signs were extracted from the MIMIC-IV database as well, with the goal of finding a variable with more significant correlations to be used for general pipeline validation. HR, Temperature,

Respiratory Rate (resprate), Oxygen Saturation (o2sat), Systolic Blood Pressure (sbp), and Diastolic Blood Pressure (dbp) were all extracted from the vitalsign table of the MIMIC-IV-ED 2.2 database [53]. Moreover, Potassium, White Blood Cells (WBC), N-terminal Pro-B-type Natriuretic Peptide (NTproBNP), Lactate, D-Dimer, Partial Pressure of Oxygen (pO2), and Cortisol were all extracted from the MIMIC-IV-v3.1 database [49], similar to how creatinine values were extracted, using a ± 60 minute time window for ECG and biomarker matches. Data for the Left Ventricular Ejection Fraction (LVEF) was used from the LVEF table which was provided on the GitHub repository of the ECGFounder model [56]. This file already contained LVEF values matched with patient identifiers and ECG waveform file paths. This led to a total of 15 unique physiological and laboratory biomarkers used for the correlation analysis.

5.5.3 CLEF Representations

To assess the quality of the latent spaces generated by the CLEF approach, a series of validation experiments were conducted. The primary objective was to assess whether handcrafted ECG features are reflected in the high dimensional representations learned by the model's backbone through SSL, thereby indicating that the FM captures meaningful signal within the data.

Representations were visualized using Uniform Manifold Approximation and Projection (UMAP) to project the CLEF embeddings into a two-dimensional space. To further quantify the relationship between learned representations and established clinical metrics, the latent vectors were reduced to 10 principal components (PCs) via Principal Component Analysis (PCA). A Spearman correlation matrix was then constructed between these PCs and the 23 handcrafted features to identify the degree of alignment between the model's abstract features and known ECG features.

Furthermore, instead of utilizing the handcrafted features, CLEF embeddings were used as inputs to the XGBoost model to predict HR. For this prediction no hyperparameter tuning was used. This vanilla implementation was intentional. By keeping the downstream classifier simple and unoptimized, the performance of the model becomes a direct reflection of the quality of the embeddings rather than the power of the regressor. Successful HR prediction validates that during the self-supervised contrastive learning phase, CLEF successfully learns to represent the periodicity and temporal density of the ECG waveform within its latent space. It confirms that the learned features are not merely statistical noise, but are physiologically grounded representations that can be mapped to real-world clinical metrics such as HR. This approach serves as a more rigorous validation of the representations compared to a simple linear prediction head, as it evaluates whether the embeddings incorporate non-linear structures which a gradient-boosting framework can exploit.

Finally, to further assess the quality of the CLEF representations, the XGBoost framework (without hyperparameter tuning) was used to predict five main handcrafted features using CLEF representations extracted from the ECGs of the 200Hz MIMIC-IV dataset. Therefore, the multi-dimensional CLEF embeddings were stacked into dense feature matrices. The features to predict were 'ECG rate mean', 'RR interval', 'QT interval','PR interval', and 'T peak amplitude', as these are among the most common ECG features [1, 5, 6, 42]. The analysis accounted for variability in both model scale (CLEF-S, CLEF-M, CLEF-L) and input pre-processing (single, double, interpolated). This experimental setting gives an estimate of the extent to which the CLEF representations capture crucial ECG features, relevant for downstream clinical tasks. Furthermore, it can be used as an indicator for which input pre-processing method is most beneficial in extracting high quality CLEF representations.

6 Results

This section gives a detailed summary of experimental results, starting with the prediction of serum creatinine from ECG signals, task validation via a clinical pilot study, correlation between biomarkers and handcrafted ECG features, followed by a methodological validation of the pipeline though HR prediction, as well as various quality evaluations of CLEF representations.

6.1 Predicting Creatinine from ECGs

The following section details the performance of the three tested model architectures XGBoost, ResNet, and CLEF-based models, across the TCC and MIMIC-IV datasets for predicting creatinine from ECG recordings. A summary of the best results is presented in table 1. All detailed results can be found in appendix F. Across both datasets and model types, the primary finding is that predicting creatinine levels directly from ECG waveforms yielded poor results, with R^2 values consistently laying near or below zero. This suggests that none of the tested models were able to use a 10 second, single-lead ECG waveform to predict serum creatinine any better than the dataset mean.

Table 1: Summary of best test results for training different models (XGBoost, ResNet, CLEF) on creatinine prediction using either TCC data or MIMIC-IV data. For XGBoost and ResNet the best performing run and the mean baseline runs are listed. For CLEF, different input methods (single, double, interpolated), backbone sizes (S,M,L) and model heads (CNN, linear) were explored. The best performing CLEF representation ("Best") is paired with the corresponding result of using the raw ECG representation ("Reference") for the same architecture and input type. Results are reported as point estimates with 95% confidence intervals indicated in brackets.

Results: Creatinine Prediction						
Model	**Datasource**	**Configuration**	**MSE**	**RMSE**	**MAE**	**R2**
XGBoost	TCC	Regular (200Hz)	2.71 [2.43, 3.01]	1.65 [1.56, 1.74]	0.84 [0.81, 0.87]	0.00 [-0.01, 0.01]
		Baseline: Mean	2.72 [2.43, 3.02]	1.65 [1.56, 1.74]	0.85 [0.82, 0.88]	-0.00 [-0.00, -0.00]
	MIMIC-IV	Regular (200Hz)	1.35 [1.26, 1.44]	1.16 [1.12, 1.20]	0.56 [0.55, 0.57]	0.05 [0.05, 0.06]
		Baseline: Mean	1.43 [1.33, 1.52]	1.19 [1.15, 1.23]	0.61 [0.60, 0.62]	-0.00 [-0.00, -0.00]
ResNet	TCC	Regular (200Hz)	3.52 [3.22, 3.85]	1.87 [1.79, 1.96]	0.97 [0.94, 1.00]	-0.30 [-0.34, -0.26]
		Baseline: Mean	2.71 [2.43, 3.02]	1.65 [1.56, 1.74]	0.84 [0.82, 0.88]	-0.00 [-0.00, 0.00]
	MIMIC-IV	Regular (200Hz)	1.45 [1.36, 1.54]	1.20 [1.17, 1.24]	0.61 [0.60, 0.62]	-0.02 [-0.04, 0.00]
		Baseline: Mean	1.43 [1.32, 1.53]	1.19 [1.15, 1.24]	0.61 [0.60, 0.62]	-0.00 [-0.00, -0.00]
CLEF	TCC	Interpolated, Linear, CLEF-M (Best)	2.67 [2.39, 2.96]	1.63 [1.55, 1.72]	0.81 [0.79, 0.85]	0.02 [0.01, 0.02]
		Interpolated, CNN, Raw (Best)	2.66 [2.39, 2.94]	1.63 [1.54, 1.71]	0.91 [0.88, 0.94]	0.02 [0.01, 0.03]
		Interpolated, Linear, Raw (Reference)	2.92 [2.63, 3.23	1.71 [1.62, 1.80]	0.93 [0.90, 0.96]	-0.08 [-0.09, -0.07]
	MIMIC-IV	Interpolated, Linear, CLEF-M (Best)	1.37 [1.27, 1.48]	1.17 [1.13, 1.22]	0.54 [0.53, 0.55]	0.04 [0.03, 0.04]
		Interpolated, Linear, Raw (Reference)	1.45 [1.35, 1.56]	1.20 [1.16, 1.25]	0.62 [0.61, 0.63]	-0.02 [-0.02, -0.01]

Initial experiments used XGBoost and ResNet to establish performance benchmarks on the TCC and MIMIC-IV datasets. Analysis of the TCC dataset revealed that traditional gradient boosting architectures struggled to surpass baseline mean predictors, with R^2 values remaining near zero values remaining near zero. Most ResNet models had negative R^2 values, suggesting the model's predictions being even less accurate than the mean and shuffled baselines. This test result is visualized in the predicted vs. true creatinine levels plot (see figure 3) which clearly depicts this weak predictive performance. Contrary to the initial hypothesis, the use of CLEF representations did not yield a significant breakthrough in predictive accuracy. The low performance of the models across the board was not due to some specific feature of the smaller TCC dataset, as the total explained variance of these models remained very low even when applied to the much larger MIMIC-IV dataset.

Even though there were no major performance differences between the models, an analysis of the training and evaluation metrics reveals varying degrees of overfitting across

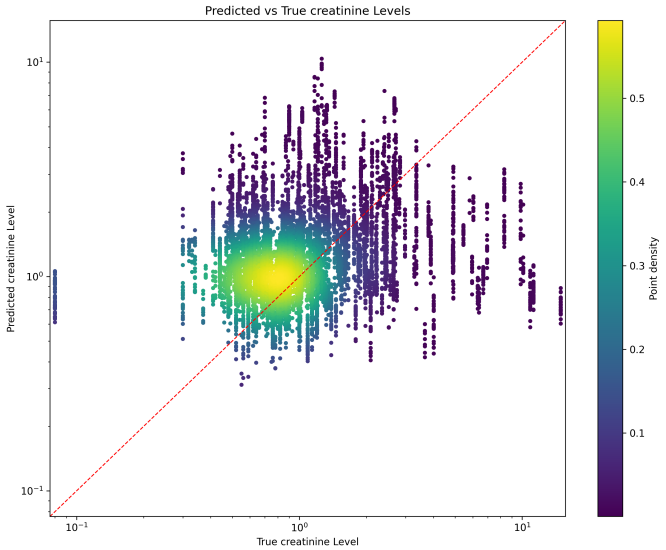

Figure 3: Predicted (y-axis) vs. true (x-axis) creatinine levels on the test set, after training with ResNet, using the TCC dataset (log scaled). The red dashed line denotes perfect agreement. Most predicted values are clustered around creatinine values of 1, regardless of their true value. The model tends to overestimate low creatinine and underestimate high creatinine, indicating regression toward the mean and limited sensitivity to extreme values. The vertical clustering of data points observed in the scatterplot is a direct consequence of the data's structure. Due to the capping methods employed, each patient can be represented by a maximum of 58 ECG recordings and a maximum of six laboratory entries within the dataset. Consequently, the dataset contains multiple row entries for each patient where a single, static laboratory value is mapped to several distinct ECG segments. This many-to-one relationship manifests as the discrete groupings seen in the visualization.

the different model architectures, particularly when trained on the smaller TCC dataset. The ResNet models exhibited the most significant evidence of overfitting. For instance, on the TCC data, the ResNet achieved a training R^2 of 0.20, yet the test is drastically lower with an R^2 of -0.30, with MSE increasing from 1.62 in training to 3.52 on the test set. Even on the larger MIMIC-IV dataset, the ResNet showed a performance drop from a positive training R^2 (0.12) to a negative test R^2 (-0.02), suggesting that the high-capacity convolutional layers struggled to generalize the subtle characteristics of creatinine within the ECG. In contrast, XGBoost demonstrated better stability, particularly on the TCC dataset where training and validation MAE remained nearly identical (0.75 vs. 0.76), though a rise in MSE on the TCC test set (2.71) compared to training (1.73) still indicates a struggle with out-of-distribution samples. CLEF representations showed the most robust resistance to overfitting with no large variations between train, validation, and test results.

All in all, these results highlight the potential of using CLEF representations to prevent overfitting in low-data regimes, but also demonstrate that predicting creatinine might be a too complex task to be feasible based on single-lead ECG recordings.

6.2 Task Validation: Manual Clinical ECG Analysis

To contextualize the findings, a clinical pilot study was conducted centered on potassium, as this biomarker has a much more established relationship with ECGs [1, 3, 5, 6, 41, 42]. The objective was to establish a human-expert benchmark for how much biochemical

information can realistically be extracted from TCC's ECG signals.

The clinical validation pilot yielded an overall accuracy of 54.0%, with the physician correctly identifying the potassium category in 27 out of 50 cases, which is statistically not significantly better than a coin flip. A detailed error analysis revealed a significant disparity in classification performance: while normal kalemia was accurately identified in 38.0% of all cases (with only a 10.0% false-positive rate), hyperkalemia was correctly detected in only 16.0% of samples. Strikingly, 34.0% of the high-potassium cases were missed entirely, which indicates that many pathological signatures are likely too subtle for manual review, were obscured by signal noise, or are not present in the single-lead ECG data.

These findings demonstrate that even when simplified into a classification task for potassium, a biomarker with a well-established relation to cardiac activity, performance remained limited for an experienced clinician. This outcome establishes a baseline for human-expert performance and further underscores the inherent difficulty of predicting laboratory values from the employed single-lead ECG data.

6.3 Biomarkers and ECG Feature Correlations

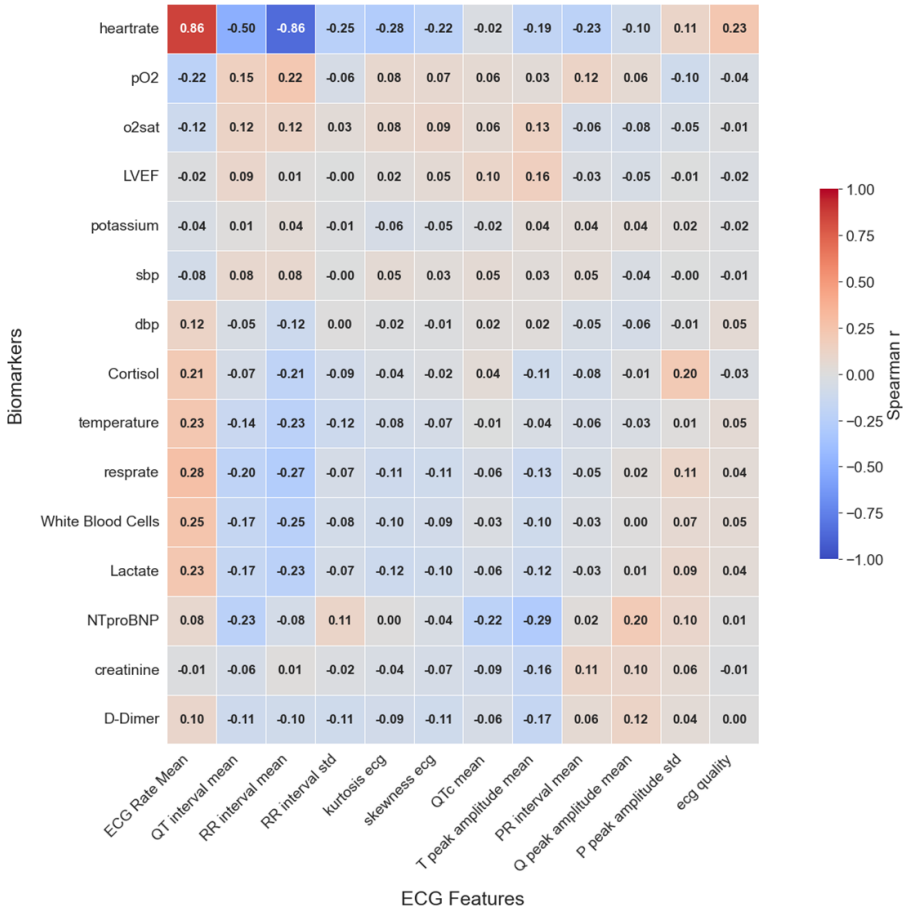

Figure 4: Spearman correlation of biomarkers and vital signs (with correlations ≥ 0.2) with handcrafted ECG features. Darker red tones show higher positive correlation, while darker blue tones show higher negative correlation.

Since earlier results suggested that creatinine regression from single-lead ECG might be infeasible, the correlations of other biomarkers with ECGs were explored. The goal was to find a biomarker for which the prediction task should be feasible, so that the potential value of CLEF embeddings can be evaluated in a task that is known to be solvable.

Thus, a Spearman rank correlation analysis was performed between 15 unique biomarkers, vital signs, and handcrafted ECG features. An extensive correlation heatmap, also indicating the sample sizes for each biomarker and correlations with all handcrafted features can be found in appendix E. A more compressed version including only handcrafted features with correlations ≥ 0.2 is depicted in figure 4.

In line with the previous findings that predicting creatinine from single-lead ECGs is hard to impossible, correlations between creatinine and ECG features were extremely low (see figure 4). The same applied to most other biomarkers, including potassium, for which features traditionally associated with hyperkalemia in medical literature, such as T-peak amplitudes, showed almost no correlation ($r \approx 0.04$). Furthermore, the correlation analysis shows statistically identifiable relationships of several biomarkers with key handcrafted ECG features, including the ECG rate mean, QT interval mean, RR interval mean, and T peak amplitude mean. However, these associations remain generally weak, with absolute correlation coefficients consistently falling below 0.3.

The strongest and clearest correlations with ECG features are visible for HR. Therefore, the task of HR prediction should be solvable and provides an opportunity to test the ability of CLEF-based models to learn the ECG patterns associated with HR.

6.4 Pipeline Validation: Predicting Heart Rate from ECGs

Because none of the biomarkers except HR were strongly correlated with ECG features (see figure 4), HR regression was used as a way to validate the pipeline, testing if the poor performance on creatinine prediction was due to architectural failures or the inherent difficulty of the task. Unlike creatinine, HR regression should be feasible, due to its direct physiological link to specific ECG features, most notably the RR interval. As the mathematical reciprocal of HR ($RR \approx 60/HR$ [19]), the RR interval should provide a clear signal that allows for a robust validation of the model's predictive capacity.

Table 2 gives a summarized overview of the best experimental runs, as well as their reference runs (mean for XGBoost and ResNet, and raw for CLEF) among all three model types. A more extensive overview of the performances of all configurations of models, across the TCC and MIMIC-IV datasets can be found in appendix G.

As would be expected given the direct relationship between HR and ECG, both XGBoost and ResNet models were able to predict HR well with R^2 values above 0.8 (see table 2, and figure 5) and no major performance difference on TCC or MIMIC data. Performance on MIMIC for this task was also generally robust to resampling the data at a lower frequency (200 Hz).

Overall performance of CLEF was slightly below XGBoost and ResNet baselines, and the model achieved a peak R^2 of 0.73 on MIMIC-IV at 500Hz, likely because this matched the backbone's pre-training conditions. On the TCC dataset, the best CLEF configuration achieved an R^2 of 0.58 and results remained consistent across single, double, and interpolated input types. Notably, the CNN head performed best on raw inputs, whereas the linear head reached its peak when utilizing larger CLEF-M or CLEF-L representations. On MIMIC-IV, the interpolated input configuration yielded the best results. Furthermore, the CLEF-L backbone consistently outperformed smaller variants, and the linear head generally proved superior to the CNN head.

The evaluation metrics indicate varying robustness to overfitting across the three models, relative to the scale of the dataset. For the XGBoost model applied to the TCC

Table 2: Summary of best test results for training different models (XGBoost, ResNet, CLEF) on HR prediction using either TCC data or MIMIC-IV data. For XGBoost and ResNet the best performing run and the mean baseline runs are listed. For CLEF, different input methods (single, double, interpolated), backbone sizes (S,M,L) and model heads (CNN, linear) were explored. The best performing CLEF representation ("Best") is paired with the corresponding result of using the raw ECG representation ("Reference") for the same architecture and input type. Results are reported as point estimates with 95% confidence intervals indicated in brackets. Rows highlighted in grey represent the best performance (lowest MSE) for each model type. The globally best result is highlighted in bold.

Results: Heart Rate Prediction						
Model	**Datasource**	**Configuration**	**MSE**	**RMSE**	**MAE**	**R2**
XGBoost	TCC	Regular (200Hz)	73.51 [68.14, 78.91]	8.57 [8.25, 8.88]	5.03 [4.89, 5.17]	0.81 [0.79, 0.82]
		Baseline: Mean	384.56 [371.16, 397.85]	19.61 [19.27, 19.95]	15.34 [15.09, 15.60]	-0.00 [-0.00, -0.00]
	MIMIC-IV	Regular (200Hz)	105.85 [93.05, 119.66]	10.28 [9.65, 10.94]	6.40 [6.20, 6.61]	0.80 [0.78, 0.83]
		Baseline: Mean	543.04 [519.60, 566.97]	23.30 [22.79, 23.81]	18.42 [18.03, 18.82]	-0.00 [-0.00, -0.00]
	MIMIC-IV	Regular (500Hz)	111.37 [101.35, 123.30]	10.55 [10.07, 11.10]	6.99 [6.79, 7.20]	0.80 [0.78, 0.82]
		Baseline: Mean	554.00 [529.08, 578.16]	23.54 [23.00, 24.05]	18.63 [18.23, 19.01]	-0.00 [-0.00, -0.00]
ResNet	**TCC**	**Regular (200Hz)**	**66.39 [62.86, 70.27]**	**8.15 [7.93, 8.38]**	**5.50 [5.38, 5.62]**	**0.83 [0.82, 0.84]**
		Baseline: Mean	384.76 [371.33, 398.05]	19.62 [19.27, 19.95]	15.34 [15.09, 15.60]	-0.00 [-0.00, -0.00]
	MIMIC-IV	Regular (200Hz)	103.15 [90.60, 117.60]	10.16 [9.52, 10.84]	6.61 [6.42, 6.82]	0.81 [0.79, 0.83]
		Baseline: Mean	542.88 [520.05, 565.44]	23.30 [22.80, 23.78]	18.40 [18.03, 18.74]	-0.00 [-0.00, 0.00]
	MIMIC-IV	Regular (500Hz)	104.53 [93.81, 116.17]	10.22 [9.69, 10.78]	6.78 [6.59, 6.99]	0.81 [0.79, 0.83]
		Baseline: Mean	550.70 [527.86, 575.02]	23.47 [22.98, 23.98]	18.57 [18.21, 18.93]	-0.00 [-0.00, 0.00]
CLEF	TCC	Double, Linear, CLEF-M (Best)	161.49 [154.95, 167.81]	12.71 [12.45, 12.95]	9.43 [9.26, 9.60]	0.58 [0.57, 0.59]
		Double, CNN, Raw (Best)	157.20 [150.71, 163.56]	12.54 [12.28, 12.79]	9.57 [9.40, 9.75]	0.59 [0.58, 0.61]
		Double, Linear, Raw (Reference)	5566.38 [5499.78, 5628.23]	74.61 [74.16, 75.02]	71.57 [71.14, 71.97]	-13.48 [-13.93, -13.06]
	MIMIC-IV	Interp., Linear, CLEF-L (Best, 200Hz)	156.12 [142.09, 170.57]	12.49 [11.92, 13.06]	8.63 [8.40, 8.87]	0.71 [0.69, 0.73]
		Interp, Linear, Raw (Reference, 200Hz)	7356.71 [7235.92, 7476.83]	85.77 [85.06, 86.47]	81.76 [81.11, 82.42]	-12.57 [-13.07, -12.09]
	MIMIC-IV	Single, Linear, CLEF-L(Best, 500Hz)	147.25 [136.55, 158.82]	12.13 [11.69, 12.60]	8.64 [8.41, 8.88]	0.73 [0.71, 0.75]
		Single, Linear, Raw (Reference, 500Hz)	7361.91 [7243.42, 7485.81]	85.80 [85.11, 86.52]	81.72 [81.05, 82.39]	-12.38 [-12.86, -11.90]

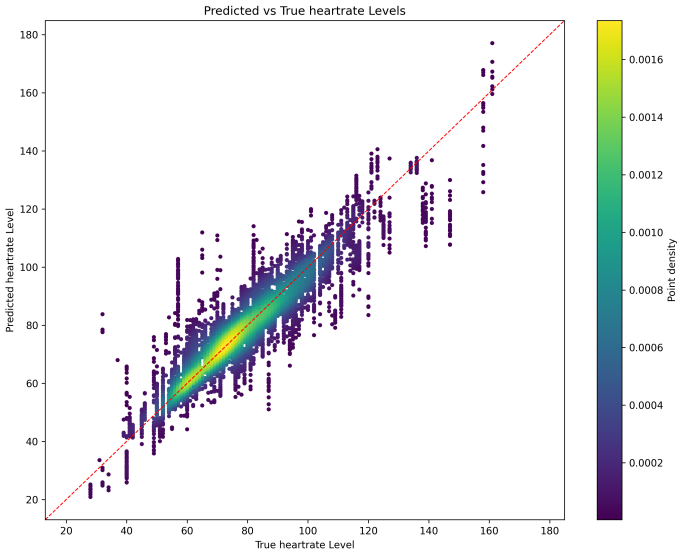

Figure 5: Predicted (y-axis) vs. true (x-axis) HR levels on the test set, after training with ResNet, using the TCC dataset. The red dashed line denotes perfect agreement. Overall, the data points show good alignment. While data points in the mid-range are located close to the identity line, there is slightly more variance at the extreme ends of the spectrum, specifically below 40 bpm and above 140 bpm.

dataset, the R^2 decreased from 0.91 during training to 0.81 on the test set, while the MIMIC-IV (200 Hz) dataset showed a smaller margin of 0.86 to 0.8. ResNet demonstrated similar patterns between train and test results with a decrease in R^2 from 0.9 to 0.83

on the TCC dataset, compared to a slight shift in R^2 from 0.84 to 0.81 on MIMIC-IV (200Hz). For CLEF, performance discrepancies between training and testing remained much lower on MIMIC-IV than on TCC, with no or only very small changes in R^2 values.

In summary, while the successful regression of HR across all model architectures validates the integrity of the data pipeline, the performance gap between CLEF and the supervised benchmarks (XGBoost and ResNet) is notable. These results suggest that although the task is inherently solvable, the current CLEF embeddings may not prioritize the primary physiological features necessary for accurate HR estimation. This discrepancy necessitates a deeper investigation into the specific information encoded within the CLEF latent space.

6.5 Testing CLEF Representations

This section describes the various steps taken to explore and validate the quality of the representations produced by the CLEF backbones. This includes the visualization of the CLEF embeddings, their correlation with handcrafted features, and the prediction of HR, as well as handcrafted features with XGBoost using CLEF representations as inputs.

6.5.1 Visualization of CLEF Representations

To evaluate the physiological information captured within the latent space of CLEF representations, UMAP projections were generated for both the raw input vectors and the CLEF representations using the various input configurations. Coloring the points by HR provides a first qualitative assessment of how HR relates to variation in the latent space. Detailed UMAP visualizations for the single, double, and interpolated input methods across CLEF-S, CLEF-M, and CLEF-L representations are provided in appendix J.

The visualizations show that HR is visible to some extent within all the latent spaces. This observation provides a first indication that the self-supervised embeddings are capturing a latent structure that is at least partially organized along clinically relevant variables. A representative comparison between the UMAP projections of the raw interpolated data and the learned CLEF-L interpolated embeddings, presented in figure 6, gives a visual indicator of latent structures.

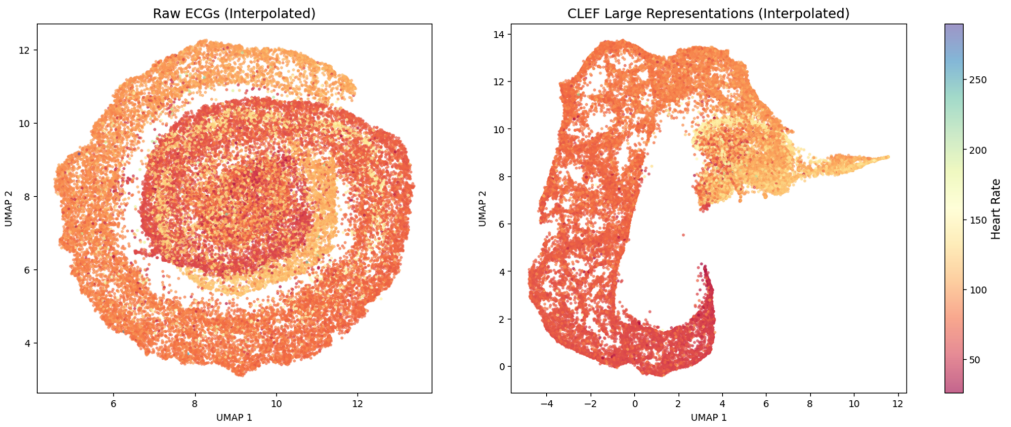

Figure 6: Comparison of UMAP representations between raw ECG inputs and CLEF-L representations, using the interpolated input option. Points are colour coded by HR value, ranging from 26 bpm (lowest) to 290 bpm (highest), utilizing the MIMIC-IV (200Hz) HR dataset.

6.5.2 Correlation of CLEF Embeddings with Handcrafted Features

To further test the information encoded in the CLEF representations, it was evaluated which of the handcrafted ECG features, as well as HR, were clearly represented in the embeddings by assessing their correlation with the first 10 PCs of the CLEF embeddings. Correlation heatmaps for the single, double, and interpolated input methods across CLEF-S, CLEF-M, and CLEF-L representations can be found in appendix J.

The raw ECG inputs showed high correlations only with the mean ECG feature (e.g., figure 7), while the CLEF representations correlated with more ECG features (e.g., figure 8). Across all input types (single, double, interpolated) the CLEF-S and CLEF-M representations show very strong correlations ($|r| > 0.85$) in the first PC with the ECG SD, ECG range, as well as the R and S peak amplitudes. Meanwhile, the CLEF-L representations showed the strongest correlation with HR related features. These findings confirm that the major sources of variation in the CLEF embeddings reflect handcrafted features.

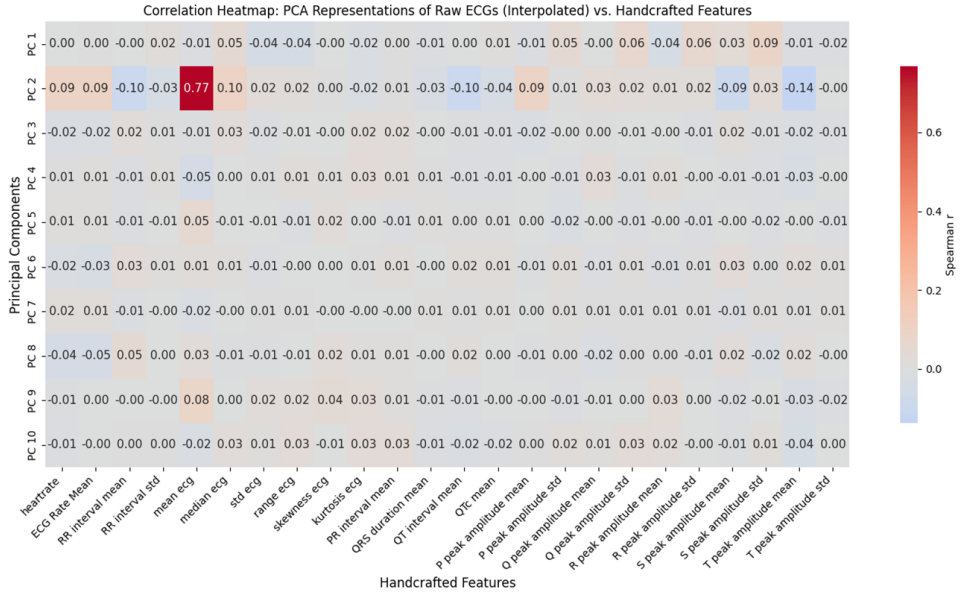

Figure 7: Spearman correlation heatmap between ten PCs extracted from raw ECGs from the MIMIC-IV HR dataset and handcrafted features, as well as HR.

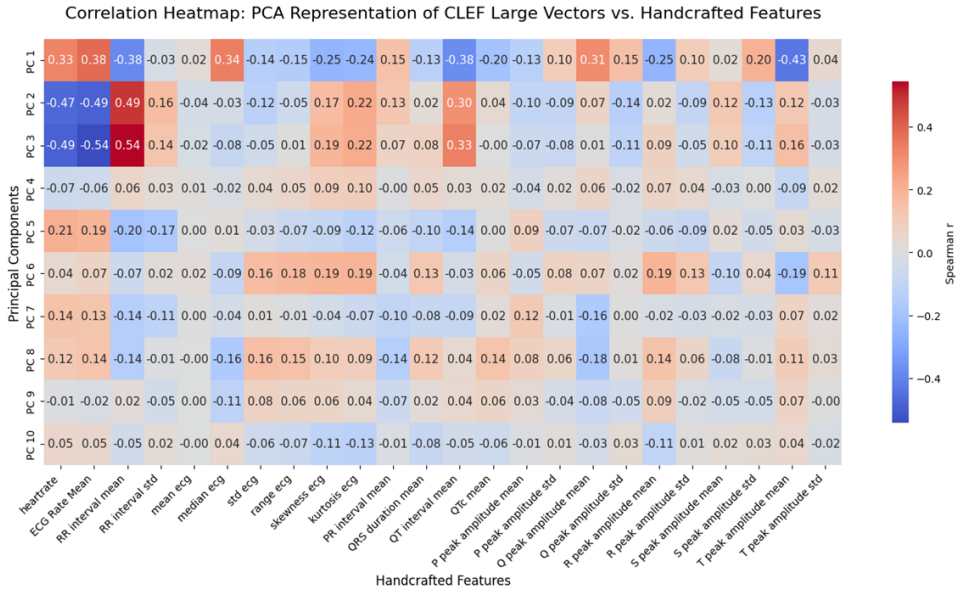

Figure 8: Spearman correlation heatmap between ten PCs extracted from CLEF-L representation with interpolated inputs from the MIMIC-IV HR dataset and handcrafted features, as well as HR.

6.5.3 Heart Rate Prediction from CLEF Embeddings using XGBoost

To further validate that the CLEF latent space captures physiologically relevant information, CLEF representations, extracted from MIMIC-IV, were used as input for the XGBoost model to predict HR. This analysis serves as a validation for the utility of the learned features.

All results are summarized in appendix H. Performance varied slightly between the combinations of input format and CLEF representation size. The double input configuration combined with the CLEF-S representation achieved the highest overall performance on the test set, reaching an R^2 of 0.64 [0.63, 0.66] and the lowest MSE of 194.09 [181.01, 207.79]. The interpolated setting maintained high stability, with both CLEF-M and CLEF-L sizes yielding a test R^2 of 0.63 [0.61, 0.65] and MSEs near 200. A notable divergence occurred regarding representation size. The CLEF-S representation generally performed best in the single and double input settings. Conversely, for the interpolated input, the CLEF-M and CLEF-L representations significantly outperformed the CLEF-S version. Overall, all regular runs substantially outperformed the baseline mean and shuffled runs, however, they substantially underperform in comparison to the XGBoost predictions made on handcrafted features (see table 2). While the self-supervised model captures several key physiological features to some extent, including handcrafted features, the resulting embeddings do not yet seem to match the richness or granularity provided by traditional handcrafted feature sets.

6.5.4 Feature Prediction from CLEF Embeddings using XGBoost

To further evaluate the performance differences between the different input (single, double, interpolated) and representational (CLEF-S, CLEF-M, CLEF-L) settings, the learned CLEF embeddings were passed as inputs to the XGBoost model to predict some of the handcrafted ECG features. This served as a complementary approach to HR prediction, validating to what extent the handcrafted features have been captured in the

self-supervised setting and are preserved in the learned embeddings. Table 3 provides a condensed summary of the findings and more detailed performance metrics for all experimental configurations are available in appendix I.

Table 3: Comprehensive summary of predicting several handcrafted features from CLEF representations, using the XGBoost model. Results are reported as R^2 [95% CI] for regular test runs across the five predicted handcrafted features, categorized by input format and CLEF representation. The highest R^2 value per input format is highlighted in grey for each feature, and the globally highest R^2 per feature is marked in bold.

Input Format	Repr.	ECG rate mean	RR Interval	QT Interval	PR Interval	T peak amplitude
single	CLEF-S	0.72 [0.71, 0.74]	0.66 [0.63, 0.68]	0.37 [0.35, 0.38]	0.42 [0.41, 0.44]	**0.47 [0.41, 0.52]**
	CLEF-M	0.54 [0.52, 0.55]	0.52 [0.49, 0.54]	0.28 [0.27, 0.30]	0.31 [0.29, 0.32]	0.42 [0.37, 0.46]
	CLEF-L	0.50 [0.48, 0.51]	0.50 [0.47, 0.52]	0.24 [0.22, 0.25]	0.15 [0.14, 0.16]	0.31 [0.27, 0.34]
double	CLEF-S	**0.74 [0.72, 0.75]**	0.67 [0.65, 0.70]	**0.38 [0.37, 0.40]**	0.44 [0.42, 0.46]	**0.47 [0.42, 0.52]**
	CLEF-M	0.46 [0.44, 0.48]	0.45 [0.43, 0.47]	0.24 [0.22, 0.25]	0.25 [0.24, 0.27]	0.35 [0.30, 0.38]
	CLEF-L	0.50 [0.49, 0.52]	0.51 [0.49, 0.53]	0.25 [0.23, 0.26]	0.15 [0.14, 0.16]	0.31 [0.27, 0.34]
interp.	CLEF-S	0.57 [0.55, 0.59]	0.53 [0.51, 0.55]	0.35 [0.33, 0.36]	**0.48 [0.46, 0.50]**	0.46 [0.40, 0.50]
	CLEF-M	0.72 [0.70, 0.73]	**0.69 [0.66, 0.71]**	0.36 [0.34, 0.37]	0.32 [0.30, 0.34]	0.44 [0.38, 0.48]
	CLEF-L	0.72 [0.70, 0.73]	0.68 [0.65, 0.71]	0.36 [0.35, 0.38]	0.24 [0.23, 0.26]	0.43 [0.38, 0.48]

Predictive performance was highest for rhythm-based features (ECG rate mean and RR Interval), which achieved R^2 scores between 0.45 and 0.74. In contrast, QT Interval, PR Interval, and T peak amplitude showed moderate recovery, suggesting that while handcrafted features are reconstructible to some extent from CLEF embeddings, accuracy varies significantly by feature type.

Regarding input and CLEF size configurations, CLEF-S emerged as the most effective representation for single and double input formats, while for interpolated input formats performance varied between CLEF sizes. A performance decay was observed as CLEF size increased across all features for single inputs, as well as for the PR interval and T peak amplitude across all input types. Although the double and interpolated configurations yielded the global maxima for nearly all features, the minimal variance between input types suggests that no specific input format is universally superior.

Ultimately, CLEF embeddings demonstrate recoverability especially for rhythmic features, yet the moderate performance across other features highlights limitations in total feature recoverability.

7 Discussion

This section contextualizes the experimental results by assessing the extent to which biomarkers are represented within ECG signals, while systematically analyzing and comparing the methodological factors influencing model performance across the TCC and MIMIC-IV datasets. Therefore, the inherent challenges of continuous creatinine regression is outlined in comparison to literature, and both the data and methodological approach are validated. In addition, self-supervised representations are examined and CLEF configurations compared, before limitations and future directions are acknowledged.

7.1 Regressing Creatinine from ECGs

When tasked with predicting continuous creatinine values from ECG recordings, no model configuration achieved an R^2 significantly higher than 0.05 and MSEs no lower

than 1.35. This suggests that creatinine levels may not be sufficiently encoded in the ten-second ECG waveforms to allow for reliable regression.

A primary objective of this study was to determine if the poor performance observed on the smaller TCC dataset could be mitigated by the significantly larger scale of the MIMIC-IV dataset. However, while the MIMIC-IV experiments resulted in lower error metrics, this improvement is largely due to other factors such as the distribution of the data rather than an increase in predictive accuracy, as R^2 remained near zero. This suggests that the reliance on large amounts of data in DL was not the main limitation. Instead, the task itself may be intrinsically too difficult, as predicting creatinine levels from ECG signals lacks a well-established physiological relationship. The comparative analysis across the two datasets also highlighted varying degrees of model stability. On the smaller TCC dataset, the ResNet architecture showed strong overfitting, with R^2 dropping from 0.20 in training to -0.30 on the test set. While the larger MIMIC-IV dataset helped stabilize these metrics, it did not enable the model to capture meaningful physiological patterns. Furthermore, while the CLEF FM demonstrated the most robust resistance to overfitting, the employment of this self-supervised backbone did not provide the expected performance breakthrough.

Ultimately, these findings suggest that regardless of dataset scale or architectural complexity, the models struggle to move beyond predicting the distribution mean. Whether using handcrafted features, end-to-end ResNets, or pre-trained CLEF embeddings, the consistent failure to achieve clinical utility across both TCC and MIMIC-IV suggests that the creatinine-ECG relationship lacks the necessary signal strength for reliable continuous monitoring.

Compared to the benchmark established in prior literature by von Bachmann et al. (2024) [3], a noticeable performance gap is visible in the prediction of creatinine using the ResNet architecture. Although the benchmark study also highlighted the difficulty of correctly regressing creatinine from ECGs, the results of this study yielded a MSE of 1.45 and a MAE of 0.61, which are approximately three times and twice as high as the benchmark values of 0.48 and 0.30, respectively. These discrepancies may be partially attributed to a more homogeneous target space of the benchmark study. While von Bachmann et al. reported a mean concentration of 1.02 with a SD of 0.80, the populations used in this study both exhibit higher means and greater variance (MIMIC-IV: 1.28 ± 1.24; TCC: 1.34 ± 1.45). Thus, in a less dispersed target distribution, a model that defaults to predicting the mean value will naturally achieve lower error metrics than it would on the more dispersed and complex distributions of the TCC and MIMIC-IV datasets.

Moreover, the R^2 value achieved in this study was -0.02, falling significantly short of the benchmark R^2 by von Bachmann et al. (2024) [3] of approximately 0.12^2. Since the general architecture of the ResNet model remained similar, this discrepancy likely stems from the benchmark study benefiting from a more robust signal, driven by both input resolution and dataset scale. While this study utilized single-lead 200Hz data, von Bachmann et al. employed 8-lead ECG samples at 400Hz. This configuration provides a more comprehensive spatial representation of cardiac electrical activity and a more detailed temporal resolution, possibly offering a richer feature set for the model to exploit. Although the TCC ECG-creatinine samples might be more accurately aligned due to the stricter $\pm$ 5 minute time window, compared to the benchmarks $\pm$ 60 minute interval, this does not seem to compensate for the loss of information inherent in single-lead, lower-resolution inputs. This technical advantage was further amplified by the scale and diversity of the benchmarks cohort. With nearly 300,000 ECG recordings from six different emergency departments, the benchmarks model was exposed to a much greater number and variety of unique physiological examples than the single-hospital populations

employed in this study (MIMIC-IV: 200,000; TCC: 60,000). This reduction in dataset scale, particularly for the TCC cohort, significantly limits the ability of high-capacity models such as ResNet to generalize, as there are fewer unique physiological examples from which to learn any subtle, non-linear relationships between cardiac activity and creatinine. Therefore, the combination of higher resolution multi-lead ECG recordings and a more diverse and larger dataset likely enabled the benchmark model to learn slightly more robust and generalizable feature representations.

Finally, while the TCC population is demographically distinct from the benchmark cohort, demographic variation does not appear to be an explanation of poor performance. Although age and sex shifts could theoretically increase task difficulty on TCC data, creatinine regression was also not possible on MIMIC data, which shares a demographic profile highly similar to the benchmarks population. This suggests that the barriers to successful regression are rooted in the available signal and dataset distribution rather than demographic characteristics.

Ultimately, the consistent inability of both traditional and DL models to surpass the mean baseline indicates that the ECG-based prediction of creatinine is an extraordinarily complex task, a finding that aligns with the performance challenges previously documented in the literature. The fact that even a large scale dataset such as MIMIC-IV and a sophisticated FM such as CLEF failed to yield a significant breakthrough suggests that the bottleneck is not merely a matter of data volume or architectural depth. This leads to a fundamental question: Is regressing creatinine from a single-lead ECG simply a mission-impossible task due to a lack of a causal link between serum creatinine levels and the cardiac electrical activity detectable by an ECG, is it a result of limitations in the signal quality of the underlying input data, or is the methodology itself flawed?

7.2 Data Validation

Given that the data quality had been validated through NeuroKit2's quality assessment library, and the extracted RR interval feature aligned with established clinical benchmarks, confirming the feature quality, it remained unclear whether the ECG waveforms provided sufficient signal clarity for a complex regression task such as creatinine prediction. To further investigate this, the following sections provide a systematic assessment of dataset integrity for the task at hand, and the direct correlation between specific ECG characteristics and creatinine concentrations. To establish a broader context, these correlations are also calculated for other biomarkers.

7.2.1 Task Validation: Manual Clinical ECG Analysis

A clinical pilot study was conducted to determine whether the human eye, even with specialized training, could detect subtle abnormalities and variations within the used ECG waveforms to make informed decisions about changes in biomarker values. To establish a baseline for human performance, potassium was utilized as a reference biomarker, as its impact on ECG shape is well documented and biologically established [1, 5, 6, 41, 42].

The results revealed a high rate of false negatives, suggesting that the structural ECG patterns, even for a known signal like hyperkalemia, are often too subtle for human practitioners to isolate from background noise. With an experienced specialist identifying only 16.0% of hyperkalemia cases, it is evident that this diagnostic task lies at the very edge of what can be reliably observed in the signal. This difficulty is further compounded by the use of a single-lead configuration, which provides a significantly more limited window into cardiac electrical activity than a full 12-lead setup.

These findings provide critical context for the poor performance observed in the creatinine regression task. If a strong and predictable biomarker such as potassium is

nearly invisible to a trained clinician in this specific data format, it suggests that the changes associated with creatinine may be even more obscured. While DL models are theoretically capable of identifying microscopic fluctuations invisible to humans, the consistently low performance of the ResNet and CLEF architectures indicates that they are encountering the same visual obscurity and signal-to-noise limitations.

Ultimately, this pilot study raises the question of whether necessary information is truly present within the 10 second single-lead ECG snippets or if the signal is simply too weak to be captured. Consequently, rather than assuming a failure of model architecture, the following section shifts the focus toward a rigorous investigation of the direct correlation between waveform characteristics and laboratory values.

7.2.2 Biomarkers and ECG Feature Correlations

The correlation analysis between creatinine and handcrafted ECG features suggests that creatinine levels do not significantly manifest within the measured electrical signal. This finding aligns with current medical literature. To our knowledge, there are no established reports of a strong, direct link between ECG morphology and creatinine. Even in larger benchmark studies, such as von Bachmann et al. (2024), only modest Pearson correlations of approximately 0.35 were reported [3]. Furthermore, no meaningful correlation was found even for features traditionally associated with ECG changes, such as potassium abnormalities showing as changes in T-wave amplitude, QRS duration, PR interval, and P-wave morphology [1, 5, 6, 42]. This reinforces the conclusion that identifying creatinine concentrations from single-lead, 10 second ECG snapshots might not be feasible.

In contrast to the weak correlation observed for creatinine, the correlation analysis showed that HR was still clearly detectable from the single-lead ECG samples. Therefore, this biomarker was utilized as a validation target to confirm that the employed models and pre-processing steps are capable of capturing and regressing meaningful information when it is actually present in the ECG waveform.

7.3 Approach Validation: Heart Rate Prediction

Success in the prediction of HR confirms that the architectural framework is capable of performing high accuracy regression on ECG data. While HR estimation is a significantly less complex task than creatinine prediction, these results suggest that the primary barrier to the latter is likely the biomarker's limited representation within the signal, rather than severe deficiencies in the modeling approach or data quality.

Across all three models the ResNet architecture achieved the best performance for HR prediction, closely followed by XGBoost, reaching R^2 values of around 0.8, therewith significantly outperforming mean and shuffled baselines. This confirms that both handcrafted features and DL representations contain robust physiological information, with the marginal superiority of the DL approach aligning with existing literature [13]. The CLEF model achieved lower overall accuracy compared to the two baseline models. Peak performance with CLEF was reached on the MIMIC-IV dataset at 500Hz ($R^2 = 0.73$), likely due to optimal alignment with CLEF's 500Hz MIMIC-based pre-training. Performance dropped on the 200Hz MIMIC-IV dataset ($R^2 = 0.52$ on single-input, see table 18 in appendix G) and the 200Hz TCC dataset ($R^2 = 0.56$ on single-input, see table 17 in appendix G). These results suggest that the CLEF backbone is sensitive to the sampling frequency and dataset characteristics of its pre-training source, limiting its immediate generalizability to different data sources or frequencies.

To determine if sample frequency mismatches of the pre-trained backbone and the downstream data can be mitigated, various input formatting configurations were evaluated. On the MIMIC-IV dataset, utilizing the interpolation method to upsample 200Hz data

to 500Hz significantly increased performance, yielding an R^2 of 0.71. This suggests that interpolation can possibly restore the richness of the input which is required by the backbone. However, because interpolation provided no measurable benefit on the TCC cohort, where performance remained consistent across all input formats, no universal conclusion can be drawn regarding optimal input handling. Future research should consider multiple input configurations as the impact appears to be dataset dependent.

The effectiveness of the CLEF FM also depended on its scale and model head architecture. Performance generally scaled with model size. While the trend was less distinct on TCC data, the CLEF-L variant consistently outperformed smaller versions across all MIMIC-IV configurations. Furthermore, a notable interaction between head architectures and representations was observed. On TCC data, CNN heads performed best on raw inputs, while linear heads were better suited to leverage the high-dimensional feature space of CLEF-L embeddings. This suggests that for prominent signals such as HR, additional architectural complexity in the model head may obscure information already encoded by the backbone. It is important to note that these findings are based solely on HR prediction and may not universally apply to more complex or subtle biomarkers.

The comparative analysis suggests that dataset volume and domain alignment play critical roles in model performance. The larger performance gaps observed on the TCC dataset for all models between train, validation and test sets likely reflects the challenges of learning complex patterns from limited samples, leading to higher sensitivity and moderate overfitting. High stability was observed between splits when using CLEF on the MIMIC-IV dataset, highlighting the advantages of in-distribution feature extraction, as the CLEF backbone was pre-trained on MIMIC-IV data, and therefore developed a latent space optimized for these specific waveforms.

Initially, it was hypothesized that the CLEF backbone would provide a performance breakthrough, especially on the smaller TCC dataset, by leveraging general ECG features learned during self-supervised pre-training. The observed gap between CLEF and the baselines can be attributed to several factors. While ResNet and XGBoost are trained specifically for the regression task, the CLEF results reported here rely primarily on a single linear head to learn the regression task within a few epochs of training, from quality ECG representations. Since only the model head was pre-trained, and the backbone was kept frozen, this limits the model's ability to retrain more weights specific for the regression task. Alternatively, it is possible that the self-supervised representations do not yet capture sufficient information relevant to the specific biomarkers, suggesting that the pre-training objective may not fully encode the subtle physiological features necessary for high-accuracy regression.

The successful prediction of HR serves as a critical proof of concept. Besides showing that the deployed data processing and model pipelines work for regressing biomarkers with known correlations to ECG waveforms, it also proves that the TCC dataset, despite its smaller size, is of sufficient quality to support complex DL tasks. Although the CLEF model underperformed relative to the baselines, its ability to achieve an R^2 of 0.73 on the MIMIC-IV dataset, and an R^2 of 0.58 on TCC data using only a simple linear head proves that the self-supervised representations contain physiological relevant information to some extent. Moreover, the CLEF approach demonstrated unique strengths that justify its utility. Notably, CLEF exhibited the highest degree of stability across the MIMIC-IV data splits. While this consistency is partially attributable to the alignment between the backbone's pre-training distribution and the downstream task, it highlights the model's capacity for domain-specific representational alignment. Furthermore, because only the model head required training on labeled data, the CLEF configurations maintained a significantly lower computational load compared to training a ResNet from scratch. This still makes CLEF a promising model, even if its performance could not yet compete. To

better understand what information the FM is capturing within its backbone embeddings, a thorough analysis of the self-supervised representations was performed.

7.4 Self-Supervised Representations

The quality of the CLEF backbone embeddings, as well as which specific information they capture was explored across various experiments, namely latent space structure (UMAP), feature alignment (PCA correlations), and downstream predictive performance (HR and broader ECG biomarker prediction). The findings of these representation evaluations are described in detail in this section.

7.4.1 Correlation of CLEF Embeddings with Handcrafted Features

Comparison of the correlation heatmaps between raw ECG embeddings and CLEF embeddings showed that raw ECG embeddings had almost no correlation with handcrafted features, except for mean ECG, suggesting that linear decomposition mainly captures basic amplitude shifts rather than complex clinical markers. In contrast, CLEF embeddings showed broader correlations with handcrafted features, indicating that the model has effectively disentangled various physiological properties into distinct dimensions of its latent space.

Among all three input types, it seems that the CLEF-S and CLEF-M latent dimensions force the model to prioritize and compress the most prominent physiological features into its primary components. The CLEF-L representations yielded the strongest correlations with HR metrics. Thus, while the specific orientation of the latent space shifts depending on the model capacity, the fundamental ability to extract and disentangle clinical features remains a core characteristic of the CLEF representations.

7.4.2 Heart Rate Prediction from CLEF Embeddings using XGBoost

The results of predicting HR with XGBoost, using CLEF representations as inputs, demonstrate that CLEF embeddings contain predictive features for HR. Among the tested configurations, results did not provide a very clear "one-size-fits-all" winner in terms of input choice or CLEF model variant. While the findings confirm that CLEF representations capture essential physiological information without the need for manual feature engineering, they still underperform relative to handcrafted features. Specifically, the best CLEF-based XGBoost configuration ($R^2 = 0.64$) remains notably inferior to the handcrafted feature baseline ($R^2 = 0.80$) on the same dataset. Therefore, although CLEF representations contain characteristic ECG features to some extent, they do not serve as a full replacement for traditional handcrafted features.

7.4.3 Feature Prediction from CLEF Embeddings using XGBoost

To assess in a task independent way how well the learned CLEF representations capture clinically relevant ECG information and to assess the differences in CLEF inputs and sizes, they were used as inputs to XGBoost, predicting a range of handcrafted ECG features.

The results show that CLEF embeddings consistently encode meaningful ECG features across different input formats and representation sizes. For simpler, rhythm-related features such as ECG rate mean and RR intervals, performance was high ($R^2 = 0.74$ and $R^2 = 0.69$ respectively) across most configurations, indicating that global temporal structure is robustly captured. In contrast, more complex features such as QT and PR intervals were predicted with lower accuracy ($R^2 = 0.38$ and $R^2 = 0.48$ respectively), suggesting that finer waveform characteristics are only partially represented. Thus,

although features seem to be recoverable to some extent, there are still limitations, which might give an explanation to why the CLEF-based models underperformed compared to XGBoost and ResNet on the HR prediction task.

Results for all three input types remained similar for all predicted features. A key observation is that CLEF-S representations often outperformed larger ones, particularly for PR interval prediction, where performance decreased notably with increasing embedding size. This likely indicates that the subtle atrial signatures required to predict this interval are best preserved in a more distilled latent space, as the increased complexity of CLEF-M or CLEF-L models did not translate to better predictive accuracy for this specific feature.

7.5 Computational Load and Hyperparameters

Notable differences were observed regarding execution speed and therewith computational load among the different models (see appendix K). While XGBoost remains the fastest overall, CLEF provides a similarly rapid training profile with the added benefit of deep automatic feature representation, whereas ResNet remains the most resource intensive architecture. Dependent on the amount of input samples, total training and evaluation times for ResNet could last from around 16 minutes to one hour and 47 minutes. In contrast, XGBoost and CLEF demonstrated great efficiency with total training and evaluation times lasting around 30 seconds to 3 minutes and 30 seconds only.

All HR prediction experiments utilizing the CLEF representations with a linear head reached the maximum limit of 50 epochs. This suggests that the models had not yet reached convergence, and more training cycles could potentially yield further performance gains. Additionally, more extensive hyperparameter tuning likely offers a path to improved accuracy. For instance, reducing the batch size from 256 to 16 in the TCC cohort (using the linear head, CLEF-M representation, and doubled input) already improved the R^2 from 0.58 to 0.66 and lowered the MSE from 161.49 to 129.22 (see appendix 29).

7.6 CLEF Configurations: Input Formats, CLEF Sizes, Model Heads

Overall, the results of the HR prediction using CLEF, as well as the thorough assessments of the self-supervised representations provide insights on how input formats, CLEF backbone sizes, and model heads influence the quality and utility of the learned embeddings.

Across all evaluations, the choice of input format and therewith sampling frequency was only relevant when the downstream data matched the characteristics of the pre-training dataset, as it seems to tune its latent space to the specific data employed. Consequently, future applications of CLEF should either utilize downstream data that mirrors the pre-training set or involve fine-tuning the model's backbone on data more representative of the target task.

Throughout the various experiments, no single CLEF size variant emerged as a definitive winner, and the optimal selection seems to depend on the specific prediction task, input characteristics, and model head employed. Larger CLEF variants proved superior for HR prediction when utilizing simple linear heads. This highlights the utility of FM: they transform ECG waveforms into a linearly separable feature space, allowing even a simple linear aggregator to achieve reasonable performance despite its lower complexity compared to a deep convolutional. Nevertheless, when training CLEF for HR prediction on the smaller TCC dataset, the competitive performance of the raw input CNN head suggests, that for simpler datasets a high capacity convolutional head can

extract sufficient features without the need for sophisticated pre-trained embeddings. Furthermore, the CLEF-S model demonstrated better performance for both HR and handcrafted feature prediction tasks when paired with an XGBoost regressor. This suggests that larger embeddings are more linearly separable, whereas smaller embeddings may provide more compact representations better suited for nonlinear regressors.

These findings partially align with those of the CLEF authors, who report that CLEF-S and CLEF-M perform best on regression tasks with backbone fine-tuning, while CLEF-L is superior on classification tasks with linear probing [17]. Although this study only evaluated regression, the stronger performance of larger variants with linear heads is consistent with Shu et al.'s (2025) [17] explanation that larger models appear to produce feature representations that are more easily separable when training simple linear heads.

7.7 Limitations and Future Work

While this study provides a comprehensive evaluation of ML methods for biomarker regression using ECG waveforms as inputs, as well as a detailed analysis of FM representations, several limitations must be acknowledged. These constraints highlight the complexity of the task and provide a roadmap for future research.

A significant challenge lies in the limited spatial resolution of the data, as this research was restricted to a single-lead ECG configuration. While ECG_II is a standard extremity lead, most successful studies in digital biomarker detection utilize a full 12-lead setup, suggesting that a single-lead might lack the spatial resolution required to capture the subtle structural variations caused by systemic biomarker changes. Therefore, future research should scale up the number of used ECG leads and possibly incorporate additional data sources such as vital signs and biometrics to improve model performance. Incorporating age and sex as model inputs can be essential for refining predictive accuracy and ensuring robust predictions, since serum creatinine levels are heavily influenced by muscle mass, which correlates with sex and age. Nevertheless, incorporation of more inputs also reduces the applicability of the model to real-world settings in which such data might be scarce.

In addition to expanding the input variables, the existing framework could be tested on the prediction of various other digital biomarkers beyond creatinine and HR.

Furthermore, a mean bias is present in both datasets, where extreme high and low laboratory values are underrepresented, making it difficult for models to learn the full clinical spectrum of the biomarker. This issue could be targeted in future work by reweighting training data through up-sampling minority entries by applying data augmentation and down-sampling majority entries, or by utilizing a weighted loss function.

Additionally, the validity of the creatinine labels remains a point of uncertainty. Creatinine levels are dynamic, and the optimal time interval between a blood draw and an ECG recording is not yet universally defined. Thus, matching creatinine measurements with the right ECG recordings highly depends on the employed tolerance window. Investigating the temporal dynamics and the impact of time delays between lab results and ECG recordings could help identify the specific window size within which laboratory measurements can be coupled with ECG recordings, potentially improving the alignment between the physiological signal and the chemical ground truth.

Methodological constraints post another limitation, as the CLEF FM was not subjected to extensive hyperparameter tuning. Thus, future work should explore variations in performance under hyperparameter tuning, including for instance training CLEF for more epochs, and utilizing a smaller batch size. Moroever, alternative signal cleaning methods beyond the neurokit method could be explored, as well as the different available options for ECG quality assessment other than the employed template matching option.

Moreover, future studies should focus on incorporating larger, multi-site datasets to improve model robustness and validate the pipeline on fully independent test sets from diverse hospital systems.

Beyond expanding to multi-site data, future work should systematically investigate the relationship between dataset scale and model performance. Specifically, it remains to be determined at what sample size self-supervised representations provide a definitive advantage over supervised baselines. By evaluating the pipeline on varying sub-samples of size N, researchers could identify the threshold at which predictive accuracy begins to degrade, thereby defining the minimum data requirements for effective FM deployment in clinical settings.

Another promising avenue for future work is the full fine-tuning of the CLEF backbone on domain specific data, compared to performing linear probing only. To go even further, the CLEF backbone could be pre-trained on TCC data only to adapt to their characteristics of 200Hz single-lead TCC data with input dimension of 2048 for the 10 second ECG snippets. For this pre-training a large enough amount of ECG samples would be necessary. Given that the original CLEF model was pre-trained over 160,000 unique patients with one ECG per patient, future efforts should aim to compile a similar scale of local ECG data to effectively pre-train the model backbone and extract qualitative self-supervised representations for this specific clinical environment.

Lastly, future work could explore other methods to further evaluate the quality of FM representations and enhance model explainability. Specifically, GradCAM could be applied on the last convolutional layer within the networks backbone, or Integrated Gradients could be used to visualize feature importance and provide insights into how specific ECG segments influence model predictions.

8 Conclusion

This research investigated the feasibility of using DL and self-supervised FMs to predict continuous serum creatinine levels from single-lead ECG signals. Specifically, it addressed to what extent self-supervised methods improve performance on limited data and which methodological choices drive these gains.

The results highlight that continuous creatinine regression remains an exceptionally challenging task. Despite the implementation of sophisticated models, predictive performance for creatinine remained poor compared to more tractable tasks such as HR prediction. The CLEF approach did not provide a breakthrough in accuracy, with results consistently falling below baseline models.

A key finding was that creatinine levels do not appear to manifest significantly within a 10 second single-lead ECG recording. Correlation analysis between creatinine and handcrafted ECG features revealed no meaningful direct relationships, supporting the conclusion that the required biological signals may be too weak or entirely absent in the employed ECG recordings. This conclusion is reinforced by the high accuracy achieved when the same pipeline was applied to HR prediction, confirming that the methodology is sound.

Ultimately, while CLEF did not reach the same performance levels of the XGBoost and ResNet baselines, it demonstrated a strong representational alignment with its source domain. Its superior performance on the MIMIC-IV dataset compared to the TCC cohort suggests that the learned latent space is highly tuned to the morphological signatures of the pre-training population. While this limits immediate transferability to external cohorts, the FM remains a compelling framework due to its potential for optimization across more diverse populations. Furthermore, the self-supervised embeddings showed correlations with established clinical features that were absent in raw ECG signals, indicating that the backbone was able to learn relevant ECG features. Moreover, CLEF

has inherent benefits in model efficiency. By utilizing a frozen backbone and training only a linear head, the approach significantly reduces computational load compared to training deep architectures from scratch, establishing a scalable feature space for clinical monitoring.

In conclusion, while self-supervised FMs such as CLEF can improve representation quality, they are restricted by the inherent signal strength of the target biomarker. Moving toward precise, continuous creatinine monitoring via single-lead signals is an important clinical goal. However, these results suggest that 10-second single-lead ECG samples lack the necessary resolution to replace traditional blood sampling for creatinine. Future research should explore multi-lead input configurations as well as fine-tuning or even pre-training the FM model backbone on the task specific data.

References

[1] J. M. L. Alcaraz and N. Strodthoff, "Abnormality prediction and forecasting of laboratory values from electrocardiogram signals using multimodal deep learning," *Scientific Reports*, vol. 15, p. 39362, Nov. 2025.

[2] J. M. L. Alcaraz and N. Strodthoff, "CardioLab: Laboratory Values Estimation from Electrocardiogram Features - An Exploratory Study," Nov. 2025. arXiv:2407.18629 [eess].

[3] P. von Bachmann, D. Gedon, F. K. Gustafsson, A. H. Ribeiro, E. Lampa, S. Gustafsson, J. Sundström, and T. B. Schön, "Evaluating regression and probabilistic methods for ECG-based electrolyte prediction," *Scientific Reports*, vol. 14, p. 15273, July 2024.

[4] K. Kashani, M. H. Rosner, and M. Ostermann, "Creatinine: From physiology to clinical application," *European Journal of Internal Medicine*, vol. 72, pp. 9–14, Feb. 2020.

[5] A. A. Hirve, A. B. Repp, and L. K. Burgess, "Things We Do for No Reason™: Obtaining an electrocardiogram for managing mild hyperkalemia in hospitalized adults," *Journal of Hospital Medicine*, vol. 19, pp. 841–844, Sept. 2024.

[6] A. Webster, W. Brady, and F. Morris, "Recognising signs of danger: ECG changes resulting from an abnormal serum potassium concentration," *Emergency Medicine Journal : EMJ*, vol. 19, pp. 74–77, Jan. 2002.

[7] T. Mehari and N. Strodthoff, "Towards Quantitative Precision for ECG Analysis: Leveraging State Space Models, Self-Supervision and Patient Metadata," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, pp. 5326–5334, Nov. 2023.

[8] Y. Xu, J. Lu, S. Ding, D. Cao, X. Hu, and C. Yang, "An Electrocardiogram Multi-task Benchmark with Comprehensive Evaluations and Insightful Findings," Nov. 2025. arXiv:2512.08954 [cs].

[9] J. Kwon, M. Jung, K. Kim, Y. Jo, J. Shin, Y. Cho, Y. Lee, J. Ban, K. Jeon, S. Y. Lee, J. Park, and B. Oh, "Artificial intelligence for detecting electrolyte imbalance using electrocardiography," *Annals of Noninvasive Electrocardiology*, vol. 26, p. e12839, May 2021.

[10] I.-M. Chiu, P.-J. Wu, H. Zhang, J. W. Hughes, A. J. Rogers, L. Jalilian, M. Perez, C.-H. R. Lin, C.-T. Lee, J. Zou, and D. Ouyang, "Serum Potassium Monitoring Using AI-Enabled Smartwatch Electrocardiograms," *JACC: Clinical Electrophysiology*, vol. 10, pp. 2644–2654, Dec. 2024.

[11] C. V. Nguyen and C. D. Do, "Transfer learning in ECG diagnosis: Is it effective?," *PLOS ONE*, vol. 20, pp. 1–13, May 2025. Publisher: Public Library of Science.

[12] J. Li, A. Aguirre, J. Moura, C. Liu, L. Zhong, C. Sun, G. Clifford, B. Westover, and S. Hong, "An Electrocardiogram Foundation Model Built on over 10 Million Recordings with External Evaluation across Multiple Domains," Aug. 2025. arXiv:2410.04133 [cs].

[13] J. M. L. Alcaraz, H. Bouma, and N. Strodthoff, "Enhancing clinical decision support with physiological waveforms — A multimodal benchmark in emergency care," *Computers in Biology and Medicine*, vol. 192, p. 110196, June 2025.

[14] Y. He, F. Huang, X. Jiang, Y. Nie, M. Wang, J. Wang, and H. Chen, "Foundation Model for Advancing Healthcare: Challenges, Opportunities, and Future Directions," Apr. 2024. arXiv:2404.03264 [cs].

[15] A. Vaid, J. Jiang, A. Sawant, S. Lerakis, E. Argulian, Y. Ahuja, J. Lampert, A. Charney, H. Greenspan, J. Narula, B. Glicksberg, and G. N. Nadkarni, "A foundational vision transformer improves diagnostic performance for electrocardiograms," *npj Digital Medicine*, vol. 6, pp. 1–8, June 2023.

[16] K. McKeen, S. Masood, A. Toma, B. Rubin, and B. Wang, "ECG-FM: an open electrocardiogram foundation model," *JAMIA Open*, vol. 8, p. ooaf122, Sept. 2025.

[17] Y. Shu, P. H. Charlton, F. Kawsar, J. Hernesniemi, and M. Malekzadeh, "CLEF: Clinically-Guided Contrastive Learning for Electrocardiogram Foundation Models," Dec. 2025. arXiv:2512.02180 [cs].

[18] Y. Xiao, G. Tang, W. Liu, J. Li, G. Nie, Z. Kan, D. Zhang, Q. Zhao, and S. Hong, "AnyECG-Lab: An Exploration Study of Fine-tuning an ECG Foundation Model to Estimate Laboratory Values from Single-Lead ECG Signals," Oct. 2025. arXiv:2510.22301 [cs].

[19] E. S. Prakash and Madanmohan, "How to Tell Heart Rate From an ECG? (Learning Objects #769 and #878)," *Advances in Physiology Education*, vol. 29, pp. 57–57, June 2005.

[20] S. S. Patel, M. Z. Molnar, J. A. Tayek, J. H. Ix, N. Noori, D. Benner, S. Heymsfield, J. D. Kopple, C. P. Kovesdy, and K. Kalantar-Zadeh, "Serum creatinine as a marker of muscle mass in chronic kidney disease: results of a cross-sectional study and review of literature," *Journal of Cachexia, Sarcopenia and Muscle*, vol. 4, pp. 19–29, Mar. 2013.

[21] P. Delanaye, E. Cavalier, and H. Pottel, "Serum Creatinine: Not So Simple!," *Nephron*, vol. 136, no. 4, pp. 302–308, 2017.

[22] C. Ronco, R. Bellomo, and J. A. Kellum, "Acute kidney injury," *The Lancet*, vol. 394, pp. 1949–1964, Nov. 2019.

[23] E. D. Siew and M. E. Matheny, "Choice of Reference Serum Creatinine in Defining AKI," *Nephron*, vol. 131, no. 2, pp. 107–112, 2015.

[24] L. A. Inker, C. H. Schmid, H. Tighiouart, J. H. Eckfeldt, H. I. Feldman, T. Greene, J. W. Kusek, J. Manzi, F. Van Lente, Y. L. Zhang, J. Coresh, and A. S. Levey, "Estimating Glomerular Filtration Rate from Serum Creatinine and Cystatin C," *New England Journal of Medicine*, vol. 367, pp. 20–29, July 2012.

[25] P. von Bachmann, D. Gedon, F. K. Gustafsson, and S. Gustafsson, "ECG-Based Electrolyte Prediction: Evaluating Regression and Probabilistic Methods," 2022.

[26] N. Strodthoff, J. M. Lopez Alcaraz, and W. Haverkamp, "Prospects for artificial intelligence-enhanced electrocardiogram as a unified screening tool for cardiac and non-cardiac conditions: an explorative study in emergency care," *European Heart Journal - Digital Health*, vol. 5, pp. 454–460, July 2024.

[27] S. A. Hicks, J. L. Isaksen, V. Thambawita, J. Ghouse, G. Ahlberg, A. Linneberg, N. Grarup, I. Strümke, C. Ellervik, M. S. Olesen, T. Hansen, C. Graff, N.-H. Holstein-Rathlou, P. Halvorsen, M. M. Maleckar, M. A. Riegler, and J. K. Kanters, "Explaining deep neural networks for knowledge discovery in electrocardiogram analysis," *Scientific Reports*, vol. 11, p. 10949, May 2021.

[28] A. Abdou and S. Krishnan, "Horizons in Single-Lead ECG Analysis From Devices to Data," *Frontiers in Signal Processing*, vol. 2, Apr. 2022.

[29] C. Lai, S. Zhou, and N. A. Trayanova, "Optimal ECG-lead selection increases generalizability of deep learning on ECG abnormality classification,"

[30] X. Guan, Y. Lai, J. Jin, J. Li, H. Wang, Q. Zhao, D. Zhang, S. Geng, and S. Hong, "Reconstructing 12-Lead ECG from 3-Lead ECG using Variational Autoencoder to Improve Cardiac Disease Detection of Wearable ECG Devices," Oct. 2025. arXiv:2510.11442 [cs].

[31] L. H. Nguyen, T. H. Diep, A. Q. Phan, A. Q. Le, Q. H. Le, and A. Turnip, "Heart Rate Feature Extraction based on Neurokit2 with Python," *Internetworking Indonesia Journal*, vol. 13, pp. 39–43, June 2021.

[32] S. Kalpande, N. K. Sahu, and H. R. Lone, "Investigating the Generalizability of ECG Noise Detection Across Diverse Data Sources," in *Proceedings of the 2025 ACM International Symposium on Wearable Computers*, ISWC '25, (New York, NY, USA), pp. 113–119, Association for Computing Machinery, 2025.

[33] T. Bouthillet, "Guide to Understanding ECG Artifact — ACLS Training Blog." `https://www.aclsmedicaltraining.com/blog/guide-to-understanding-ecg-artifact`, 2024. [Accessed 02-04-2026].

[34] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 1519–1528, May 2021.

[35] L. Holmstrom, M. Christensen, N. Yuan, J. Weston Hughes, J. Theurer, M. Jujjavarapu, P. Fatehi, A. Kwan, R. K. Sandhu, J. Ebinger, S. Cheng, J. Zou, S. S. Chugh, and D. Ouyang, "Deep learning-based electrocardiographic screening for chronic kidney disease," *Communications Medicine*, vol. 3, p. 73, May 2023.

[36] Z.-Y. Liu, C.-H. Lin, Y.-C. Hsu, J.-S. Chen, P.-C. Chang, M.-S. Wen, and C.-F. Kuo, "Universal representations in cardiovascular ECG assessment: A self-supervised learning approach," *International Journal of Medical Informatics*, vol. 195, p. 105742, Mar. 2025.

[37] N. El-Sherif and G. Turitto, "Electrolyte disorders and arrhythmogenesis," *Cardiology Journal*, vol. 18, no. 3, pp. 233–245, 2011.

[38] M. A. Al-Masud, J. M. L. Alcaraz, and N. Strodthoff, "Benchmarking ECG Foundational Models: A Reality Check Across Clinical Tasks," Sept. 2025. arXiv:2509.25095 [eess].

[39] B. Surawicz, "Relationship between electrocardiogram and electrolytes," *American Heart Journal*, vol. 73, pp. 814–834, June 1967.

[40] R. Noordam, W. J. Young, R. Salman, J. K. Kanters, M. E. van den Berg, D. van Heemst, H. J. Lin, S. M. Barreto, M. L. Biggs, G. Biino, E. Catamo, M. P. Concas, J. Ding, D. S. Evans, L. Foco, N. Grarup, L.-P. Lyytikäinen, M. Mangino, H. Mei, P. J. van der Most, M. Müller-Nurasyid, C. P. Nelson, Y. Qian, L. Repetto, M. A. Said, N. Shah, K. Schramm, P. G. Vidigal, S. Weiss, J. Yao, N. R. Zilhao, J. A. Brody, P. S. Braund, M. Brumat, E. Campana, P. Christofidou, M. J. Caulfield, A. De Grandi, A. F. Dominiczak, A. S. Doney, G. Eiriksdottir, C. Ellervik, L. Giatti, M. Gögele, C. Graff, X. Guo, P. van der Harst, P. K. Joshi, M. Kähönen, B. Kestenbaum, M. F.

Lima-Costa, A. Linneberg, A. C. Maan, T. Meitinger, S. Padmanabhan, C. Pattaro, A. Peters, A. Petersmann, P. Sever, M. F. Sinner, X. Shen, A. Stanton, K. Strauch, E. Z. Soliman, K. V. Tarasov, K. D. Taylor, C. H. Thio, A. G. Uitterlinden, S. Vaccargiu, M. Waldenberger, A. Robino, A. Correa, F. Cucca, S. R. Cummings, M. Dörr, G. Girotto, V. Gudnason, T. Hansen, S. R. Heckbert, C. R. Juhl, S. Kääb, T. Lehtimäki, Y. Liu, P. A. Lotufo, C. N. Palmer, M. Pirastu, P. P. Pramstaller, A. L. P. Ribeiro, J. I. Rotter, N. J. Samani, H. Snieder, T. D. Spector, B. H. Stricker, N. Verweij, J. F. Wilson, J. G. Wilson, J. W. Jukema, A. Tinker, C. H. Newton-Cheh, N. Sotoodehnia, D. O. Mook-Kanamori, P. B. Munroe, and H. R. Warren, "Effects of Calcium, Magnesium, and Potassium Concentrations on Ventricular Repolarization in Unselected Individuals," *JACC*, vol. 73, pp. 3118–3131, June 2019. Publisher: American College of Cardiology Foundation.

[41] P. P. Frohnert, E. R. Gluliani, M. Friedberg, W. J. Johnson, and W. N. Tauxe, "Statistical Investigation of Correlations Between Serum Potassium Levels and Electrocardiographic Findings in Patients on Intermittent Hemodialysis Therapy," *Circulation*, vol. 41, pp. 667–676, Apr. 1970.

[42] C. D. Galloway, A. V. Valys, J. B. Shreibati, D. L. Treiman, F. L. Petterson, V. P. Gundotra, D. E. Albert, Z. I. Attia, R. E. Carter, S. J. Asirvatham, M. J. Ackerman, P. A. Noseworthy, J. J. Dillon, and P. A. Friedman, "Development and Validation of a Deep-Learning Model to Screen for Hyperkalemia From the Electrocardiogram," *JAMA Cardiology*, vol. 4, pp. 428–436, May 2019.

[43] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. S. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira, T. B. Schön, and A. L. P. Ribeiro, "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nature Communications*, vol. 11, p. 1760, Apr. 2020.

[44] T. Mehari and N. Strodthoff, "Self-supervised representation learning from 12-lead ECG data," *Computers in Biology and Medicine*, vol. 141, p. 105114, Feb. 2022.

[45] H. Liu, Z. Zhao, and Q. She, "Self-supervised ECG pre-training," *Biomedical Signal Processing and Control*, vol. 70, p. 103010, Sept. 2021.

[46] "Python 3.12.10." `https://www.python.org/downloads/release/python-31210/`, 2025. [Accessed 13-04-2026].

[47] "Infinity-delta-series monitor." `https://www.draeger.com/de_de/Products/Infinity-Delta-Series`. [Accessed 17-04-2026].

[48] B. Gow, T. Pollard, L. A. Nathanson, A. Johnson, B. Moody, C. Fernandes, N. Greenbaum, J. W. Waks, P. Eslami, T. Carbonati, A. Chaudhari, E. Herbst, D. Moukheiber, S. Berkowitz, R. Mark, and S. Horng, "MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset," *PhysioNet*, Sept. 2023. Version 1.0.

[49] A. Johnson, L. Bulgarelli, T. Pollard, B. Gow, B. Moody, S. Horng, L. A. Celi, and R. Mark, "MIMIC-IV," *PhysioNet*, Oct. 2024. Version 3.1.

[50] "Space labs healthcare." `https://spacelabshealthcare.com/`. [Accessed 17-04-2026].

[51] "Philips: IntelliSpace, EKG-Management-System." `https://www.philips.de/healthcare/product/{H}{C}860426/ekgmanagementsysteme-intellispaceecg-ekgmanagementsystem`. [Accessed 17-04-2026].

[52] "Diagnostische Kardiologie: Geräte & Software." https://www.gehealthcare.com/de-de/products/diagnostic-cardiology. [Accessed 17-04-2026].

[53] A. Johnson, L. Bulgarelli, T. Pollard, L. A. Celi, R. Mark, and S. Horng, "MIMIC-IV-ED," *PhysioNet*, Jan. 2023. Version 2.2.

[54] "NeuroKit2 documentation." https://neuropsychology.github.io/{N}euro{K}it/functions/ecg.html. [Accessed 04-04-2026].

[55] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, pp. 1689–1696, Aug. 2021.

[56] "GitHub - PKUDigitalHealth/ECGFounder." https://github.com/{P}{K}{U}{D}igital{H}ealth/{E}{C}{G}{F}ounder, 2025. [Accessed 05-04-2026].

Appendices

A Dataset Demographics

For the TCC and MIMIC-IV datasets, information on patients' sex and age was available. Detailed distributions of these two variables among the creatinine and HR datasets are visualized below.

Across both TCC subsets, the patient population is skewed toward males and older age groups. Within the complete datasets, male patients comprise approximately 60% of the populations. While this distribution is generally consistent across splits, there is a slight fluctuation in the HR dataset, where the validation set reaches an even higher male peak of 65.72%. Regarding age, the 60–79 year age group constitutes the clear majority, representing roughly 55% of the total population in both datasets which have a mean age of about 67 years. In contrast, younger demographics are significantly underrepresented. Specifically, the 18–39 age group accounts for only about 6% of the total population in both datasets. Some internal variations across the data subsets are worth noting. This includes for the TCC creatinine dataset the higher concentration of the 18–39 age group in the validation set (10.52%) compared to the training set (4.80%). Moreover, for the HR dataset the test set features a larger proportion of patients aged 80+ (23.91%) compared to the training set (18.71%), and the test set also includes a lower amount of people aged 40-59 (14.32%) compared to the training and validation sets (around 20%).

In contrast, the MIMIC-IV datasets exhibit a much more balanced distribution regarding sex. In the MIMIC-IV creatinine dataset, females represent 49.74% of the total population, and in the HR dataset, they account for 49.26%. This near-equal split is maintained consistently across the training, validation, and test sets. Similar to the TCC dataset, the age distribution in MIMIC-IV remains skewed toward older populations, though less severely than in TCC with a mean age of around 61 years for both datasets. Therefore, the 60–79 age group is still the largest demographic, representing 40.25% of the creatinine dataset and 38.68% of the HR dataset. The 18–39 demographic is more prominent in MIMIC-IV than in TCC, making up roughly 13–15% of the total populations. Interestingly, the MIMIC-IV validation and test sets show a higher percentage of patients in the 20–39 range (up to 16.95%) compared to their respective training sets.

Table 4: Demographic distributions across TCC creatinine dataset.

Category	**Group**	**Train**	**Validation**	**Test**	**Total**
Sex	Female	39.79%	38.35%	42.42%	39.97%
	Male	60.21%	61.65%	57.58%	60.03%
Age	Mean	67.29	64.45	67.20	66.85
	18–39	4.80%	10.52%	7.55%	6.07%
	40–59	19.71%	20.14%	16.64%	19.31%
	60–79	55.91%	50.11%	55.09%	54.92%
	80+	19.57%	19.24%	20.73%	19.70%

Table 5: Demographic distributions across TCC HR dataset.

Category	Group	Train	Validation	Test	Total
Sex	Female	41.46%	34.28%	39.13%	40.04%
	Male	58.54%	65.72%	60.87%	59.96%
Age	Mean	66.59	66.39	68.28	66.81
	18–39	6.18%	5.62%	6.10%	6.08%
	40–59	20.15%	19.78%	14.32%	19.22%
	60–79	54.96%	56.35%	55.67%	55.28%
	80+	18.71%	18.25%	23.91%	19.42%

Table 6: Demographic distributions across MIMIC-IV creatinine dataset.

Category	Group	Train	Validation	Test	Total
Sex	Female	50.49%	47.79%	48.17%	49.74%
	Male	49.51%	52.21%	51.83%	50.26%
Age	Mean	62.70	59.45	59.26	61.70
	18–39	10.48%	17.69%	17.94%	12.68%
	40–59	29.07%	29.34%	29.29%	29.14%
	60–79	42.25%	35.66%	35.52%	40.25%
	80+	18.20%	17.32%	17.25%	17.93%

Table 7: Demographic distributions across MIMIC-IV HR dataset.

Category	Group	Train	Validation	Test	Total
Sex	Female	48.96%	49.39%	50.54%	49.26%
	Male	51.04%	50.61%	49.46%	50.74%
Age	Mean	61.30	58.97	59.04	60.61
	18–39	13.75%	17.96%	18.24%	15.06%
	40–59	28.74%	29.75%	29.44%	29.00%
	60–79	39.71%	36.40%	36.13%	38.68%
	80+	17.80%	15.89%	16.19%	17.27%

B Dataset Distributions and Statistics

Dataset distributions of TCC and MIMIC-IV datasets for creatinine and HR overall show a strong degree of alignment suggesting that the physiological profiles of both populations are similar. The creatinine distributions in both cohorts exhibit a severe right skew, with a long tail of extreme values extending beyond 10.0 mg/dL. While the TCC dataset shows a slightly higher density of patients in the lower range, the box plots confirm that both datasets are characterized by a high volume of outliers. Similarly, the HR distributions follow a roughly normal, though right-skewed, profile with significant overlap between the two datasources.

The statistical characteristics of the MIMIC-IV and TCC datasets reveal significant structural differences, particularly in patient-to-observation ratios and target value distributions (see appendix tables below). The most prominent distinction lies in the density of unique measurements. While the MIMIC-IV creatinine dataset is nearly entirely unique with $201,901$ unique creatinine measurements and $202,749$ ECGs, the TCC dataset is highly longitudinal, with over $61,000$ ECG entries derived from only

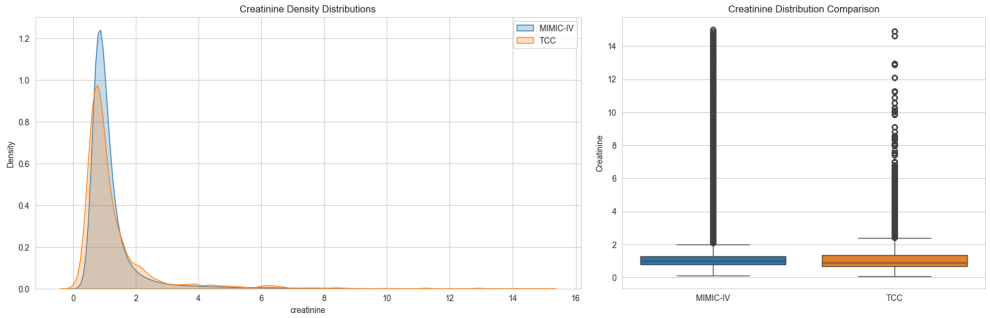

Figure 9: Histogram (left) and boxplot (right) of creatinine distributions in the TCC (blue) and MIMIC-IV (orange) dataset.

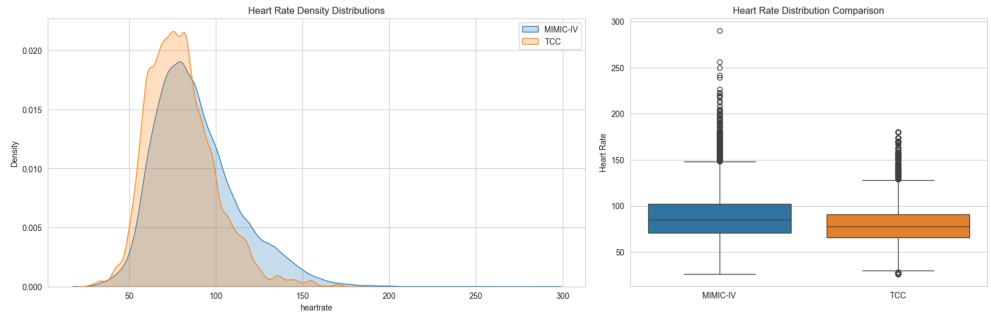

Figure 10: Histogram (left) and boxplot (right) of HR distributions in the TCC (blue) and MIMIC-IV (orange) dataset.

1,133 patients. Regarding creatinine, both datasets show a low median (0.91 mg/dL for TCC and 1.00 mg/dL for MIMIC-IV) compared to the mean (1.34 mg/dL and 1.28 mg/dL, respectively). Furthermore, the TCC creatinine dataset exhibits slightly higher variance (SD = 1.45) compared to the MIMIC-IV creatinine dataset (SD = 1.24). The HR statistics highlight that the MIMIC-IV cohort exhibits a higher prevalence of elevated HRs compared to the TCC population. With a median HR of 85 bpm and a SD of approximately 24 bpm, the MIMIC-IV data reflects a more clinically acute population, frequently reaching tachycardic levels that exceed the upper physiological limits observed in the TCC dataset.

Table 8: TCC Dataset characteristics and creatinine statistics for training, validation, test, and total dataset. Creatinine values are reported in mg/dL.

Statistic	**Train**	**Validation**	**Test**	**Total**
Number of rows	43,150	9,252	9,249	61,651
Number of patients	768	184	181	1,133
Number of cases	815	190	187	1,192
Unique creatinine measurements	2,131	358	348	2,837
Creatinine mean	1.34	1.30	1.40	1.34
Creatinine median	0.91	0.91	0.91	0.91
Creatinine SD	1.42	1.39	1.65	1.45
Creatinine minimum	0.07	0.12	0.08	0.07
Creatinine maximum	14.61	12.11	14.90	14.90

Table 9: MIMIC dataset characteristics and creatinine statistics for training, validation, test, and total dataset. Creatinine values are reported in mg/dL.

Statistic	**Train**	**Validation**	**Test**	**Total**
Number of rows	141,923	30,413	30,413	202,749
Number of patients	53,615	24,724	24,723	103,062
Unique creatinine measurements	141,199	30,349	30,353	201,901
Creatinine mean	1.29	1.27	1.26	1.28
Creatinine median	1.00	1.00	1.00	1.00
Creatinine SD	1.27	1.19	1.19	1.24
Creatinine minimum	0.10	0.10	0.10	0.10
Creatinine maximum	15.00	15.00	15.00	15.00

Table 10: TCC HR dataset characteristics and statistics for training, validation, test, and total dataset. HR values are reported in 1/min.

Statistic	**Train**	**Validation**	**Test**	**Total**
Number of rows	41,663	8,932	8,920	59,515
Number of patients	750	185	188	1,123
Number of cases	805	188	182	1,175
Unique HR measurements	1,996	376	383	2,755
HR mean	79.83	79.87	79.71	79.81
HR median	78	78	78	78
HR SD	19.59	20.17	19.61	19.70
HR minimum	26	33	28	26
HR maximum	180	163	161	180

Table 11: MIMIC HR (200Hz) dataset characteristics and statistics for training, validation, test, and total dataset. HR values are reported in 1/min.

Statistic	**Train**	**Validation**	**Test**	**Total**
Number of rows	25,282	5,418	5,417	36,117
Number of patients	17,567	5,418	5,417	28,402
Unique HR measurements	25,282	5,418	5,417	36,117
HR mean	88.96	88.18	87.95	88.40
HR median	85	85	85	85
HR SD	25.05	23.26	23.30	24.17
HR minimum	28	30	26	26
HR maximum	290	219	205	290

Table 12: MIMIC HR 500Hz dataset characteristics and statistics for training, validation, test, and total dataset. HR values are reported in 1/min.

Statistic	**Train**	**Validation**	**Test**	**Total**
Number of rows	25,301	5,423	5,420	36,144
Number of patients	17,549	5,423	5,420	28,392
Number of cases	–	–	–	–
Unique HR measurements	25,301	5,423	5,420	36,144
HR mean	88.86	88.04	88.20	88.58
HR median	85	85	85	85
HR SD	25.01	23.39	23.47	24.25
HR minimum	26	28	26	26
HR maximum	290	205	241	290

C Feature Validation: RR Interval Correlation

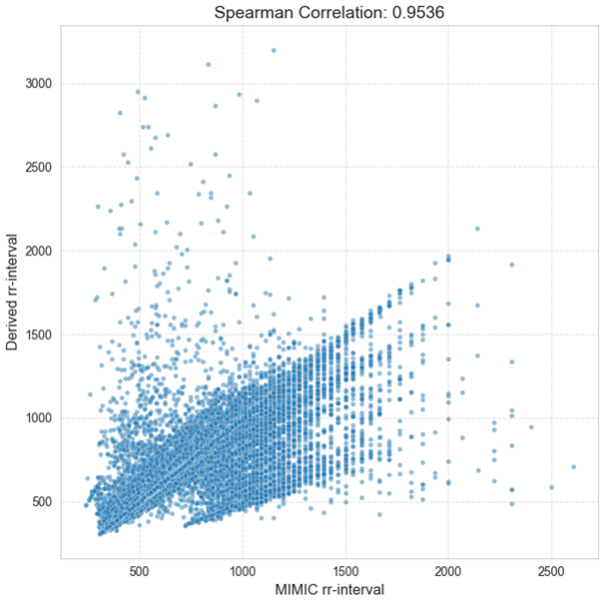

Figure 11: To validate the reliability and quality of the feature extraction process, the RR interval was utilized as a benchmark to compare the clinically derived measurements in the MIMIC-IV-ECG dataset (x-axis) against those algorithmically generated by the NeuroKit2 library (y-axis). Specifically, a Spearman rank correlation was performed between the derived and the recorded RR intervals. The resulting correlations are depicted in this scatterplot, indicating that the extracted NeuroKit2 RR intervals are useful with an overall correlation score of ≈ 0.95. The scatterplot demonstrates a strong linear alignment between the two data sources across a wide physiological range (approximately 300 ms to 2500 ms). Minor discrepancies, such as the vertical outliers are observed, which may be the result of missed R-peaks by the automated algorithm in exceptionally noisy segments.

D Data Quality

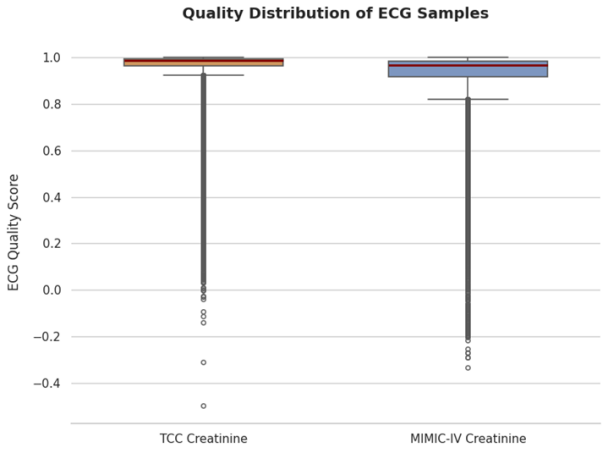

Figure 12: Quality comparison between TCC and MIMIC-IV data **before** data preparation. Quality scores were calculated using the 'ecg_quality' function and the 'templatematch' setting from the NeuroKit2 library.

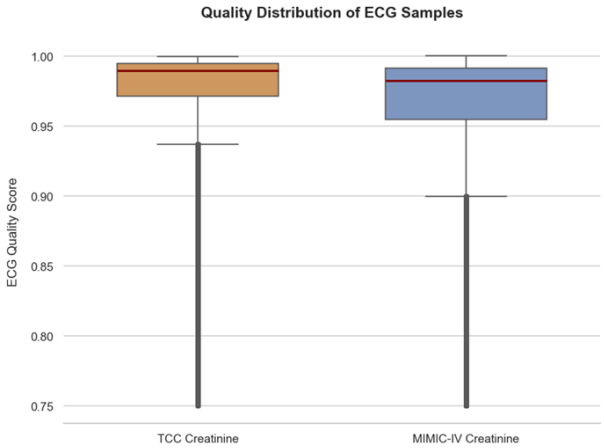

Figure 13: Quality comparison between TCC and MIMIC-IV data **after** data preparation. Quality scores were calculated using the 'ecg_quality' function and the 'templatematch' setting from the NeuroKit2 library. The quality threshold was set to ≥ 0.75.

E Biomarkers and ECG Feature Correlations - Extensive Results

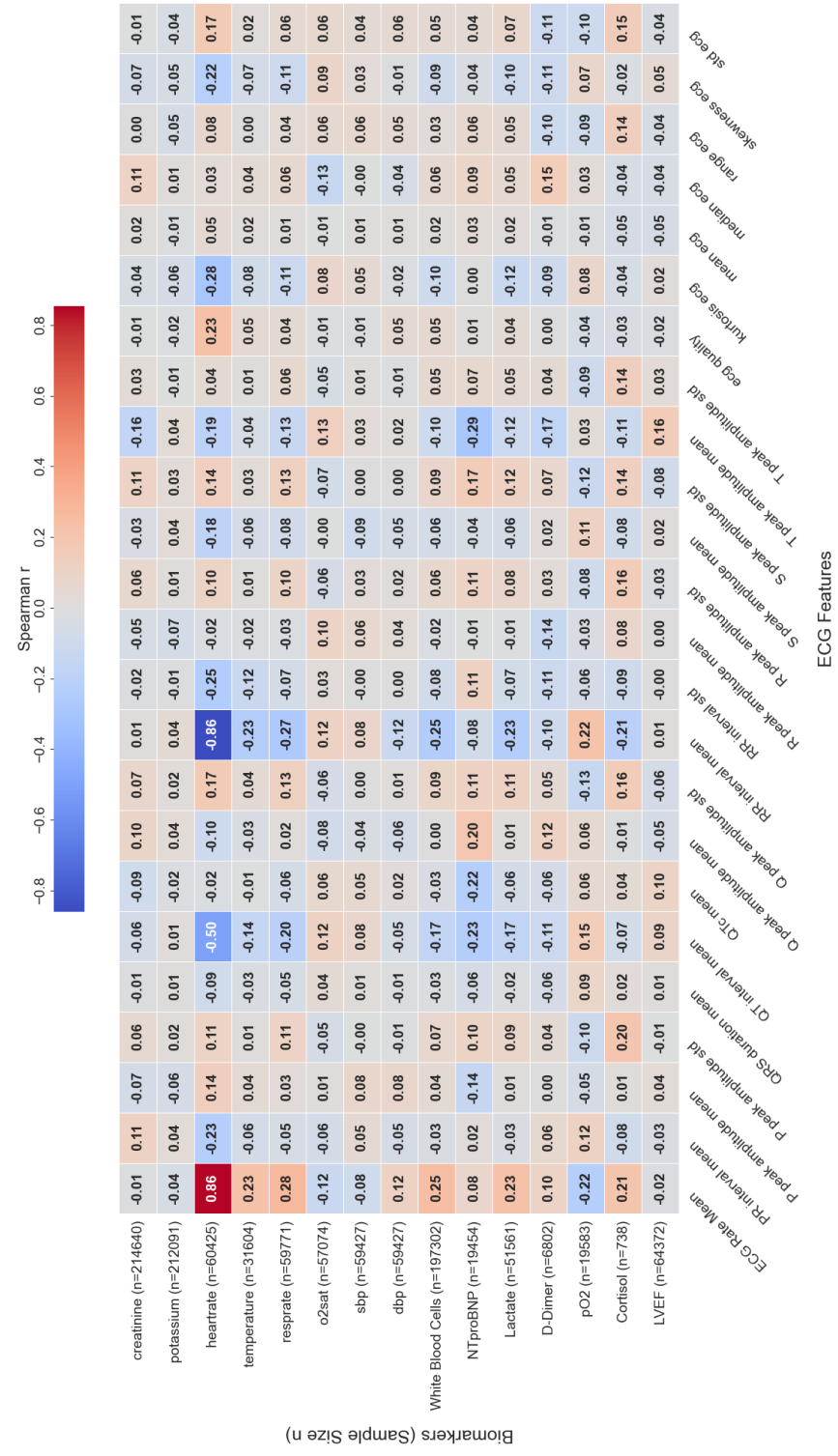

Figure 14: Spearman correlation of various biomarkers and vital signs with handcrafted ECG features. Darker red tones show higher positive correlation, while darker blue tones show higher negative correlation. The number of data samples used for the correlation analysis are defined in brackets for each biomarker.

F Creatinine Prediction - Extensive Results

Table 13: Evaluation results of predicting creatinine using XGBoost and ResNet. Results are reported as point estimates with 95% confidence intervals indicated in brackets. Test results of regular runs are highlighted in grey, with the lowest MSE per model type shown in bold.

Results: Creatinine Prediction								
Model	Datasource	ECG Hz	Run	Split	MSE	RMSE	MAE	R2
XGBoost	TCC	200	Regular	train	1.73 [1.65, 1.81]	1.32 [1.28, 1.34]	0.75 [0.74, 0.76]	0.14 [0.13, 0.16]
				validation	1.95 [1.76, 2.13]	1.40 [1.33, 1.46]	0.76 [0.74, 0.78]	-0.01 [-0.02, 0.00]
				test	2.71 [2.43, 3.01]	1.65 [1.56, 1.74]	0.84 [0.81, 0.87]	0.00 [-0.01, 0.01]
			Baseline: Mean	train	2.02 [1.93, 2.12]	1.42 [1.39, 1.45]	0.80 [0.79, 0.81]	-0.00 [-0.00, -0.00]
				validation	1.94 [1.75, 2.13]	1.39 [1.32, 1.46]	0.77 [0.74, 0.79]	-0.00 [-0.00, -0.00]
				test	2.72 [2.43, 3.02]	1.65 [1.56, 1.74]	0.85 [0.82, 0.88]	-0.00 [-0.00, -0.00]
			Baseline: Shuffled	train	2.02 [1.93, 2.12]	1.42 [1.39, 1.45]	0.80 [0.79, 0.81]	0.00 [0.00, 0.00]
				validation	1.94 [1.74, 2.12]	1.39 [1.32, 1.46]	0.77 [0.75, 0.79]	0.00 [-0.00, 0.00]
				test	2.72 [2.43, 3.02]	1.65 [1.56, 1.74]	0.85 [0.82, 0.88]	-0.00 [-0.00, -0.00]
	MIMIC-IV	200	Regular	train	1.50 [1.45, 1.54]	1.22 [1.20, 1.24]	0.61 [0.60, 0.61]	0.06 [0.06, 0.07]
				validation	1.34 [1.24, 1.44]	1.16 [1.11, 1.20]	0.56 [0.55, 0.57]	0.05 [0.04, 0.05]
				test	**1.35 [1.26, 1.44]**	**1.16 [1.12, 1.20]**	**0.56 [0.55, 0.57]**	**0.05 [0.05, 0.06]**
			Baseline: Mean	train	1.60 [1.55, 1.65]	1.27 [1.25, 1.28]	0.64 [0.63, 0.64]	-0.00 [-0.00, -0.00]
				validation	1.41 [1.31, 1.51]	1.19 [1.14, 1.23]	0.61 [0.60, 0.63]	-0.00 [-0.00, -0.00]
				test	1.43 [1.33, 1.52]	1.19 [1.15, 1.23]	0.61 [0.60, 0.62]	-0.00 [-0.00, -0.00]
			Baseline: Shuffled	train	1.60 [1.55, 1.65]	1.27 [1.25, 1.28]	0.64 [0.63, 0.65]	-0.00 [-0.00, -0.00]
				validation	1.41 [1.31, 1.51]	1.19 [1.14, 1.23]	0.61 [0.60, 0.63]	-0.00 [-0.00, -0.00]
				test	1.43 [1.33, 1.52]	1.19 [1.15, 1.23]	0.61 [0.60, 0.62]	-0.00 [-0.00, -0.00]
ResNet	TCC	200	Regular	train	1.62 [1.54, 1.68]	1.27 [1.24, 1.30]	0.74 [0.73, 0.75]	0.20 [0.19, 0.21]
				validation	2.20 [2.00, 2.41]	1.48 [1.41, 1.55]	0.81 [0.78, 0.83]	-0.13 [-0.15, -0.12]
				test	3.52 [3.22, 3.85]	1.87 [1.79, 1.96]	0.97 [0.94, 1.00]	-0.30 [-0.34, -0.26]
			Baseline: Mean	train	0.00 [0.00, 0.00]	0.03 [0.03, 0.03]	0.02 [0.02, 0.02]	0.00 [0.00, 0.00]
				validation	1.94 [1.75, 2.14]	1.39 [1.32, 1.46]	0.76 [0.74, 0.78]	0.00 [-0.00, 0.00]
				test	2.71 [2.43, 3.02]	1.65 [1.56, 1.74]	0.84 [0.82, 0.87]	-0.00 [-0.00, 0.00]
			Baseline: Shuffled	train	2.02 [1.92, 2.12]	1.42 [1.39, 1.46]	0.76 [0.75, 0.77]	-0.00 [-0.00, 0.00]
				validation	1.93 [1.75, 2.13]	1.39 [1.32, 1.46]	0.73 [0.71, 0.76]	0.00 [-0.01, 0.01]
				test	2.87 [2.58, 3.19]	1.69 [1.61, 1.79]	0.87 [0.84, 0.90]	-0.06 [-0.07, -0.05]
	MIMIC-IV	200	Regular	train	1.40 [1.36, 1.44]	1.18 [1.17, 1.20]	0.65 [0.64, 0.66]	0.12 [0.12, 0.13]
				validation	1.31 [1.23, 1.39]	1.15 [1.11, 1.18]	0.60 [0.59, 0.61]	0.07 [0.05, 0.08]
				test	**1.45 [1.36, 1.54]**	**1.20 [1.17, 1.24]**	**0.61 [0.60, 0.62]**	**-0.02 [-0.04, 0.00]**
			Baseline: Mean	train	0.00 [0.00, 0.00]	0.03 [0.03, 0.03]	0.03 [0.03, 0.03]	-8.28e10
				validation	1.40 [1.31, 1.50]	1.19 [1.15, 1.22]	0.60 [0.59, 0.61]	0.00 [0.00, 0.00]
				test	1.43 [1.32, 1.53]	1.19 [1.15, 1.24]	0.61 [0.60, 0.62]	-0.00 [-0.00, -0.00]
			Baseline: Shuffled	train	1.61 [1.56, 1.66]	1.27 [1.25, 1.29]	0.65 [0.64, 0.65]	-0.00 [-0.00, -0.00]
				validation	1.41 [1.32, 1.50]	1.19 [1.15, 1.23]	0.62 [0.61, 0.63]	-0.00 [-0.01, -0.00]
				test	1.44 [1.34, 1.54]	1.20 [1.16, 1.24]	0.65 [0.64, 0.66]	-0.01 [-0.01, -0.01]

Table 14: Evaluation results of predicting creatinine with CLEF across input types, model heads, and CLEF representations, using TCC data. Results are reported as point estimates with 95% confidence intervals indicated in brackets. Grey rows indicate the representation with the lowest MSE on the test dataset for each input type and model head. The globally lowest test MSE is highlighted in bold.

Input	Model head	Representation	Dataset	MSE	RMSE	MAE	R2
single	CNN	raw	train	1.89 [1.81, 1.99]	1.38 [1.35, 1.41]	0.72 [0.71, 0.73]	0.06 [0.06, 0.07]
			validation	1.97 [1.78, 2.18]	1.40 [1.34, 1.48]	0.72 [0.69, 0.74]	-0.02 [-0.02, -0.01]
			test	2.73 [2.44, 3.03]	1.65 [1.56, 1.74]	0.77 [0.74, 0.80]	-0.00 [-0.01, 0.00]
		CLEF-S	train	2.02 [1.96, 2.10]	1.42 [1.40, 1.45]	1.04 [1.03, 1.05]	0.00 [-0.01, 0.01]
			validation	2.30 [2.15, 2.47]	1.52 [1.46, 1.57]	1.09 [1.07, 1.11]	-0.19 [-0.23, -0.15]
			test	2.98 [2.72, 3.27]	1.73 [1.65, 1.81]	1.09 [1.06, 1.12]	-0.10 [-0.12, -0.08]
		CLEF-M	train	1.96 [1.87, 2.04]	1.40 [1.37, 1.43]	0.83 [0.82, 0.84]	0.03 [0.03, 0.03]
			validation	2.02 [1.84, 2.22]	1.42 [1.36, 1.49]	0.82 [0.80, 0.84]	-0.04 [-0.05, -0.03]
			test	2.70 [2.42, 2.99]	1.64 [1.56, 1.73]	0.88 [0.85, 0.91]	0.00 [-0.00, 0.01]
		CLEF-L	train	2.00 [1.91, 2.10]	1.41 [1.38, 1.45]	0.72 [0.71, 0.74]	0.01 [0.01, 0.01]
			validation	1.96 [1.77, 2.16]	1.40 [1.33, 1.47]	0.70 [0.67, 0.72]	-0.01 [-0.02, -0.00]
			test	2.74 [2.45, 3.05]	1.65 [1.56, 1.75]	0.77 [0.74, 0.80]	-0.01 [-0.02, -0.00]
	linear	raw	train	2.05 [1.96, 2.15]	1.43 [1.40, 1.46]	0.82 [0.81, 0.83]	-0.01 [-0.02, -0.01]
			validation	2.04 [1.85, 2.25]	1.43 [1.36, 1.50]	0.81 [0.78, 0.83]	-0.05 [-0.06, -0.04]
			test	2.79 [2.51, 3.10]	1.67 [1.58, 1.76]	0.88 [0.85, 0.91]	-0.03 [-0.04, -0.02]
		CLEF-S	train	1.92 [1.83, 2.00]	1.38 [1.35, 1.42]	0.79 [0.78, 0.80]	0.05 [0.05, 0.06]
			validation	2.03 [1.85, 2.22]	1.42 [1.36, 1.49]	0.83 [0.80, 0.85]	-0.05 [-0.06, -0.03]
			test	2.75 [2.47, 3.05]	1.66 [1.57, 1.75]	0.87 [0.84, 0.90]	-0.01 [-0.02, -0.01]
		CLEF-M	train	1.84 [1.75, 1.93]	1.35 [1.32, 1.39]	0.75 [0.74, 0.77]	0.09 [0.09, 0.10]
			validation	2.20 [2.02, 2.41]	1.48 [1.42, 1.55]	0.85 [0.83, 0.88]	-0.14 [-0.18, -0.10]
			test	2.78 [2.51, 3.07]	1.67 [1.58, 1.75]	0.85 [0.82, 0.88]	-0.02 [-0.04, -0.01]
		CLEF-L	train	1.88 [1.80, 1.97]	1.37 [1.34, 1.40]	0.75 [0.74, 0.76]	0.07 [0.07, 0.08]
			validation	2.05 [1.86, 2.24]	1.43 [1.36, 1.50]	0.78 [0.75, 0.80]	-0.06 [-0.07, -0.04]
			test	2.75 [2.48, 3.06]	1.66 [1.57, 1.75]	0.82 [0.79, 0.85]	-0.01 [-0.03, -0.00]
double	CNN	raw	train	1.87 [1.79, 1.96]	1.37 [1.34, 1.40]	0.74 [0.73, 0.75]	0.08 [0.07, 0.08]
			validation	1.94 [1.76, 2.12]	1.39 [1.33, 1.46]	0.74 [0.72, 0.76]	0.00 [-0.01, 0.01]
			test	2.73 [2.44, 3.04]	1.65 [1.56, 1.74]	0.80 [0.77, 0.83]	-0.01 [-0.02, 0.00]
		CLEF-S	train	1.88 [1.79, 1.98]	1.37 [1.34, 1.41]	0.85 [0.84, 0.86]	0.07 [0.06, 0.08]
			validation	2.09 [1.91, 2.27]	1.44 [1.38, 1.51]	0.90 [0.88, 0.92]	-0.08 [-0.10, -0.06]
			test	2.77 [2.49, 3.07]	1.66 [1.58, 1.75]	0.93 [0.91, 0.96]	-0.02 [-0.03, -0.01]
		CLEF-M	train	1.99 [1.91, 2.08]	1.41 [1.38, 1.44]	0.90 [0.89, 0.91]	0.01 [0.01, 0.02]
			validation	2.05 [1.88, 2.23]	1.43 [1.37, 1.49]	0.90 [0.88, 0.93]	-0.06 [-0.07, -0.04]
			test	2.71 [2.45, 2.99]	1.65 [1.56, 1.73]	0.96 [0.93, 0.99]	0.00 [-0.01, 0.01]
		CLEF-L	train	2.01 [1.92, 2.11]	1.42 [1.38, 1.45]	0.77 [0.76, 0.78]	0.01 [0.00, 0.01]
			validation	1.94 [1.75, 2.13]	1.39 [1.32, 1.46]	0.74 [0.72, 0.76]	0.00 [0.00, 0.00]
			test	2.72 [2.43, 3.03]	1.65 [1.56, 1.74]	0.82 [0.79, 0.85]	-0.00 [-0.00, 0.00]
	linear	raw	train	2.12 [2.03, 2.22]	1.46 [1.42, 1.49]	0.83 [0.82, 0.84]	-0.05 [-0.06, -0.04]
			validation	2.13 [1.94, 2.32]	1.46 [1.39, 1.52]	0.83 [0.80, 0.85]	-0.10 [-0.11, -0.08]
			test	2.90 [2.61, 3.21]	1.70 [1.61, 1.79]	0.89 [0.86, 0.92]	-0.07 [-0.08, -0.06]
		CLEF-S	train	1.92 [1.83, 2.02]	1.39 [1.35, 1.42]	0.78 [0.77, 0.79]	0.05 [0.04, 0.05]
			validation	1.99 [1.81, 2.17]	1.41 [1.34, 1.47]	0.80 [0.78, 0.83]	-0.02 [-0.04, -0.01]
			test	2.76 [2.48, 3.06]	1.66 [1.57, 1.75]	0.86 [0.83, 0.89]	-0.02 [-0.02, -0.01]
		CLEF-M	train	1.82 [1.73, 1.90]	1.35 [1.31, 1.38]	0.76 [0.75, 0.77]	0.10 [0.09, 0.11]
			validation	2.08 [1.90, 2.29]	1.44 [1.38, 1.51]	0.82 [0.79, 0.84]	-0.07 [-0.11, -0.05]
			test	2.75 [2.49, 3.05]	1.66 [1.58, 1.75]	0.85 [0.82, 0.88]	-0.02 [-0.03, -0.00]
		CLEF-L	train	1.88 [1.80, 1.97]	1.37 [1.34, 1.40]	0.78 [0.77, 0.79]	0.07 [0.06, 0.07]
			validation	2.07 [1.89, 2.26]	1.44 [1.37, 1.50]	0.81 [0.78, 0.83]	-0.07 [-0.09, -0.05]
			test	2.75 [2.48, 3.05]	1.66 [1.57, 1.75]	0.86 [0.83, 0.89]	-0.01 [-0.03, -0.00]
interpolated	CNN	raw	train	1.93 [1.85, 2.02]	1.39 [1.36, 1.42]	0.86 [0.85, 0.87]	0.04 [0.04, 0.05]
			validation	1.99 [1.81, 2.17]	1.41 [1.35, 1.47]	0.84 [0.81, 0.86]	-0.03 [-0.04, -0.01]
			test	**2.66 [2.39, 2.94]**	**1.63 [1.54, 1.71]**	**0.91 [0.88, 0.94]**	**0.02 [0.01, 0.03]**
		CLEF-S	train	1.96 [1.86, 2.06]	1.40 [1.36, 1.44]	0.73 [0.72, 0.74]	0.03 [0.03, 0.03]
			validation	1.94 [1.75, 2.14]	1.39 [1.32, 1.46]	0.71 [0.69, 0.74]	-0.00 [-0.01, 0.00]
			test	2.70 [2.41, 3.01]	1.64 [1.55, 1.73]	0.78 [0.75, 0.82]	0.00 [-0.00, 0.01]
		CLEF-M	train	2.05 [1.96, 2.15]	1.43 [1.40, 1.47]	0.72 [0.71, 0.73]	-0.01 [-0.02, -0.01]
			validation	2.17 [1.95, 2.39]	1.47 [1.40, 1.54]	0.73 [0.71, 0.76]	-0.12 [-0.13, -0.11]
			test	2.88 [2.57, 3.20]	1.69 [1.60, 1.79]	0.77 [0.74, 0.81]	-0.06 [-0.07, -0.05]
		CLEF-L	train	1.98 [1.89, 2.08]	1.41 [1.37, 1.44]	0.75 [0.74, 0.76]	0.02 [0.02, 0.02]
			validation	1.94 [1.75, 2.14]	1.39 [1.32, 1.46]	0.72 [0.69, 0.74]	-0.00 [-0.01, 0.00]
			test	2.68 [2.39, 2.98]	1.64 [1.55, 1.73]	0.79 [0.76, 0.82]	0.01 [0.01, 0.02]
	linear	raw	train	2.16 [2.07, 2.25]	1.47 [1.44, 1.50]	0.86 [0.85, 0.87]	-0.07 [-0.08, -0.06]
			validation	2.17 [1.98, 2.38]	1.47 [1.41, 1.54]	0.86 [0.83, 0.88]	-0.12 [-0.14, -0.10]
			test	2.92 [2.63, 3.23]	1.71 [1.62, 1.80]	0.93 [0.90, 0.96]	-0.08 [-0.09, -0.07]
		CLEF-S	train	1.94 [1.85, 2.03]	1.39 [1.36, 1.42]	0.79 [0.78, 0.81]	0.04 [0.04, 0.05]
			validation	1.93 [1.75, 2.11]	1.39 [1.32, 1.45]	0.79 [0.77, 0.82]	0.01 [-0.00, 0.02]
			test	2.74 [2.46, 3.04]	1.65 [1.57, 1.74]	0.87 [0.84, 0.90]	-0.01 [-0.01, -0.00]
		CLEF-M	train	1.78 [1.70, 1.87]	1.33 [1.30, 1.37]	0.73 [0.72, 0.74]	0.12 [0.11, 0.13]
			validation	2.07 [1.88, 2.27]	1.44 [1.37, 1.51]	0.78 [0.76, 0.81]	-0.07 [-0.08, -0.05]
			test	2.67 [2.39, 2.96]	1.63 [1.55, 1.72]	0.81 [0.79, 0.85]	0.02 [0.01, 0.02]
		CLEF-L	train	1.77 [1.70, 1.84]	1.33 [1.30, 1.36]	0.81 [0.80, 0.82]	0.12 [0.12, 0.13]
			validation	2.13 [1.94, 2.34]	1.46 [1.39, 1.53]	0.86 [0.83, 0.88]	-0.10 [-0.12, -0.08]
			test	2.71 [2.43, 2.98]	1.64 [1.56, 1.73]	0.91 [0.88, 0.94]	0.00 [-0.01, 0.01]

Table 15: Evaluation results of predicting creatinine with CLEF across input types, model heads, and CLEF representations, using MIMIC-IV data. Results are reported as point estimates with 95% confidence intervals indicated in brackets. Grey rows indicate the representation with the lowest MSE on the test dataset for each input and model head. The globally lowest test MSE is highlighted in bold.

Input	Model head	Representation	Dataset	MSE	RMSE	MAE	R2
single	CNN	raw	train	1.56 [1.51, 1.61]	1.25 [1.23, 1.27]	0.57 [0.57, 0.58]	0.03 [0.02, 0.03]
			validation	1.36 [1.27, 1.46]	1.17 [1.13, 1.21]	0.53 [0.52, 0.54]	0.03 [0.03, 0.04]
			test	1.38 [1.28, 1.48]	1.17 [1.13, 1.22]	0.53 [0.52, 0.54]	0.03 [0.03, 0.04]
		CLEF-S	train	2.31 [2.25, 2.36]	1.52 [1.50, 1.54]	0.87 [0.86, 0.87]	-0.44 [-0.45, -0.43]
			validation	2.11 [1.99, 2.22]	1.45 [1.41, 1.49]	0.86 [0.85, 0.88]	-0.50 [-0.52, -0.47]
			test	2.12 [2.00, 2.24]	1.46 [1.41, 1.50]	0.86 [0.85, 0.88]	-0.49 [-0.51, -0.47]
		CLEF-M	train	1.77 [1.73, 1.81]	1.33 [1.31, 1.35]	0.91 [0.91, 0.92]	-0.10 [-0.11, -0.10]
			validation	1.57 [1.49, 1.65]	1.25 [1.22, 1.29]	0.87 [0.86, 0.88]	-0.12 [-0.13, -0.10]
			test	1.59 [1.50, 1.68]	1.26 [1.22, 1.30]	0.87 [0.86, 0.88]	-0.11 [-0.13, -0.10]
		CLEF-L	train	1.71 [1.66, 1.76]	1.31 [1.29, 1.33]	0.54 [0.53, 0.54]	-0.07 [-0.07, -0.06]
			validation	1.50 [1.40, 1.60]	1.22 [1.18, 1.27]	0.51 [0.50, 0.52]	-0.07 [-0.07, -0.06]
			test	1.52 [1.41, 1.63]	1.23 [1.19, 1.28]	0.51 [0.50, 0.52]	-0.06 [-0.07, -0.06]
	linear	raw	train	1.60 [1.55, 1.65]	1.26 [1.25, 1.28]	0.65 [0.64, 0.65]	0.00 [0.00, 0.00]
			validation	1.42 [1.33, 1.51]	1.19 [1.15, 1.23]	0.62 [0.61, 0.63]	-0.01 [-0.01, -0.00]
			test	1.43 [1.33, 1.54]	1.20 [1.15, 1.24]	0.62 [0.61, 0.63]	-0.01 [-0.01, -0.00]
		CLEF-S	train	1.57 [1.52, 1.62]	1.25 [1.23, 1.27]	0.65 [0.64, 0.66]	0.02 [0.02, 0.02]
			validation	1.37 [1.28, 1.46]	1.17 [1.13, 1.21]	0.60 [0.59, 0.62]	0.03 [0.02, 0.03]
			test	1.39 [1.29, 1.49]	1.18 [1.13, 1.22]	0.60 [0.59, 0.61]	0.03 [0.02, 0.03]
		CLEF-M	train	1.58 [1.53, 1.62]	1.26 [1.24, 1.27]	0.68 [0.68, 0.69]	0.02 [0.01, 0.02]
			validation	1.37 [1.29, 1.46]	1.17 [1.13, 1.21]	0.63 [0.62, 0.64]	0.02 [0.02, 0.03]
			test	1.39 [1.29, 1.49]	1.18 [1.14, 1.22]	0.63 [0.62, 0.64]	0.03 [0.02, 0.03]
		CLEF-L	train	1.62 [1.57, 1.67]	1.27 [1.25, 1.29]	0.75 [0.75, 0.76]	-0.01 [-0.02, -0.01]
			validation	1.45 [1.36, 1.53]	1.20 [1.17, 1.24]	0.71 [0.69, 0.72]	-0.03 [-0.03, -0.02]
			test	1.46 [1.36, 1.56]	1.21 [1.17, 1.25]	0.70 [0.69, 0.71]	-0.03 [-0.03, -0.02]
double	CNN	raw	train	1.59 [1.54, 1.64]	1.26 [1.24, 1.28]	0.55 [0.55, 0.56]	0.01 [0.00, 0.01]
			validation	1.40 [1.30, 1.49]	1.18 [1.14, 1.22]	0.51 [0.50, 0.52]	0.01 [0.00, 0.01]
			test	1.41 [1.30, 1.51]	1.19 [1.14, 1.23]	0.51 [0.50, 0.52]	0.01 [0.01, 0.02]
		CLEF-S	train	1.61 [1.57, 1.66]	1.27 [1.25, 1.29]	0.54 [0.53, 0.55]	-0.01 [-0.01, -0.00]
			validation	1.41 [1.32, 1.51]	1.19 [1.15, 1.23]	0.50 [0.49, 0.51]	-0.00 [-0.01, 0.00]
			test	1.43 [1.32, 1.54]	1.19 [1.15, 1.24]	0.50 [0.49, 0.51]	-0.00 [-0.01, 0.00]
		CLEF-M	train	1.60 [1.56, 1.65]	1.27 [1.25, 1.28]	0.69 [0.68, 0.69]	-0.00 [-0.00, 0.00]
			validation	1.41 [1.32, 1.50]	1.19 [1.15, 1.22]	0.65 [0.64, 0.66]	0.00 [-0.00, 0.00]
			test	1.42 [1.32, 1.52]	1.19 [1.15, 1.23]	0.65 [0.64, 0.66]	0.00 [-0.00, 0.01]
		CLEF-L	train	2.22 [2.18, 2.25]	1.49 [1.48, 1.50]	1.19 [1.19, 1.20]	-0.38 [-0.40, -0.37]
			validation	2.04 [1.96, 2.12]	1.43 [1.40, 1.46]	1.16 [1.15, 1.17]	-0.45 [-0.50, -0.41]
			test	2.06 [1.97, 2.15]	1.44 [1.41, 1.46]	1.16 [1.15, 1.17]	-0.45 [-0.49, -0.40]
	linear	raw	train	1.61 [1.56, 1.65]	1.27 [1.25, 1.29]	0.64 [0.63, 0.65]	-0.00 [-0.00, -0.00]
			validation	1.43 [1.33, 1.52]	1.19 [1.15, 1.23]	0.61 [0.60, 0.62]	-0.01 [-0.02, -0.01]
			test	1.44 [1.34, 1.55]	1.20 [1.16, 1.24]	0.61 [0.60, 0.62]	-0.01 [-0.01, -0.01]
		CLEF-S	train	1.57 [1.53, 1.62]	1.25 [1.24, 1.27]	0.63 [0.62, 0.64]	0.02 [0.02, 0.02]
			validation	1.37 [1.28, 1.46]	1.17 [1.13, 1.21]	0.59 [0.57, 0.60]	0.03 [0.02, 0.03]
			test	1.38 [1.28, 1.49]	1.18 [1.13, 1.22]	0.58 [0.57, 0.59]	0.03 [0.03, 0.03]
		CLEF-M	train	1.57 [1.53, 1.62]	1.25 [1.24, 1.27]	0.66 [0.66, 0.67]	0.02 [0.02, 0.02]
			validation	1.38 [1.29, 1.47]	1.17 [1.13, 1.21]	0.62 [0.61, 0.63]	0.02 [0.02, 0.03]
			test	1.39 [1.29, 1.49]	1.18 [1.14, 1.22]	0.61 [0.60, 0.63]	0.02 [0.02, 0.03]
		CLEF-L	train	1.59 [1.54, 1.64]	1.26 [1.24, 1.28]	0.59 [0.58, 0.59]	0.01 [0.01, 0.01]
			validation	1.42 [1.32, 1.51]	1.19 [1.15, 1.23]	0.55 [0.54, 0.57]	-0.01 [-0.01, -0.00]
			test	1.43 [1.33, 1.54]	1.20 [1.15, 1.24]	0.55 [0.54, 0.56]	-0.00 [-0.01, -0.00]
interpolated	CNN	raw	train	1.62 [1.57, 1.67]	1.27 [1.25, 1.29]	0.71 [0.71, 0.72]	-0.01 [-0.01, -0.01]
			validation	1.42 [1.32, 1.50]	1.19 [1.15, 1.23]	0.66 [0.65, 0.67]	-0.00 [-0.01, 0.00]
			test	1.43 [1.33, 1.53]	1.19 [1.15, 1.24]	0.66 [0.65, 0.67]	-0.00 [-0.01, 0.01]
		CLEF-S	train	1.63 [1.58, 1.67]	1.28 [1.26, 1.29]	0.78 [0.77, 0.78]	-0.02 [-0.02, -0.01]
			validation	1.42 [1.34, 1.51]	1.19 [1.16, 1.23]	0.73 [0.72, 0.74]	-0.01 [-0.02, -0.00]
			test	1.44 [1.34, 1.53]	1.20 [1.16, 1.24]	0.72 [0.71, 0.74]	-0.01 [-0.02, -0.00]
		CLEF-M	train	1.62 [1.57, 1.67]	1.27 [1.25, 1.29]	0.73 [0.72, 0.74]	-0.01 [-0.01, -0.01]
			validation	1.42 [1.33, 1.51]	1.19 [1.15, 1.23]	0.68 [0.67, 0.69]	-0.01 [-0.01, -0.00]
			test	1.43 [1.33, 1.53]	1.20 [1.15, 1.24]	0.68 [0.67, 0.69]	-0.00 [-0.01, 0.00]
		CLEF-L	train	1.88 [1.83, 1.94]	1.37 [1.35, 1.39]	0.61 [0.61, 0.62]	-0.17 [-0.18, -0.17]
			validation	1.68 [1.58, 1.79]	1.30 [1.26, 1.34]	0.60 [0.58, 0.61]	-0.20 [-0.20, -0.19]
			test	1.69 [1.58, 1.81]	1.30 [1.26, 1.34]	0.60 [0.58, 0.61]	-0.19 [-0.19, -0.18]
	linear	raw	train	1.61 [1.57, 1.66]	1.27 [1.25, 1.29]	0.65 [0.64, 0.66]	-0.01 [-0.01, -0.01]
			validation	1.44 [1.34, 1.53]	1.20 [1.16, 1.24]	0.63 [0.61, 0.64]	-0.02 [-0.02, -0.02]
			test	1.45 [1.35, 1.56]	1.20 [1.16, 1.25]	0.62 [0.61, 0.63]	-0.02 [-0.02, -0.01]
		CLEF-S	train	1.58 [1.54, 1.63]	1.26 [1.24, 1.28]	0.66 [0.65, 0.66]	0.01 [0.01, 0.01]
			validation	1.39 [1.29, 1.47]	1.18 [1.14, 1.21]	0.62 [0.61, 0.63]	0.02 [0.01, 0.02]
			test	1.40 [1.30, 1.50]	1.18 [1.14, 1.23]	0.62 [0.60, 0.63]	0.02 [0.02, 0.02]
		CLEF-M	train	1.55 [1.50, 1.60]	1.25 [1.23, 1.26]	0.59 [0.58, 0.59]	0.03 [0.03, 0.03]
			validation	1.36 [1.27, 1.45]	1.17 [1.13, 1.21]	0.54 [0.53, 0.55]	0.03 [0.03, 0.04]
			test	**1.37 [1.27, 1.48]**	**1.17 [1.13, 1.22]**	**0.54 [0.53, 0.55]**	**0.04 [0.03, 0.04]**
		CLEF-L	train	1.56 [1.52, 1.61]	1.25 [1.23, 1.27]	0.65 [0.65, 0.66]	0.03 [0.02, 0.03]
			validation	1.39 [1.30, 1.48]	1.18 [1.14, 1.22]	0.61 [0.60, 0.62]	0.01 [0.01, 0.02]
			test	1.40 [1.30, 1.51]	1.18 [1.14, 1.23]	0.61 [0.60, 0.62]	0.02 [0.01, 0.02]

G Heart Rate Prediction - Extensive Results

Table 16: Comprehensive evaluation results of predicting HR using XGBoost and ResNet. All values are rounded to two decimal places. Results are reported as point estimates with 95% confidence intervals indicated in brackets. Test results of regular runs are highlighted in grey, with the lowest MSE per model type shown in bold.

Results: Heart Rate Prediction								
Model	Datasource	ECG Hz	Run	Split	MSE	RMSE	MAE	R2
XGBoost	TCC	200	Regular	train	34.30 [33.07, 35.57]	5.86 [5.75, 5.96]	3.76 [3.72, 3.80]	0.91 [0.91, 0.91]
				validation	65.08 [60.34, 70.21]	8.07 [7.77, 8.38]	4.97 [4.84, 5.10]	0.84 [0.83, 0.85]
				test	**73.51 [68.14, 78.91]**	**8.57 [8.25, 8.88]**	**5.03 [4.89, 5.17]**	**0.81 [0.79, 0.82]**
			Baseline: Mean	train	383.71 [377.10, 390.70]	19.59 [19.42, 19.77]	15.15 [15.04, 15.27]	-0.00 [-0.00, -0.00]
				validation	405.78 [391.70, 421.31]	20.14 [19.79, 20.53]	15.42 [15.15, 15.70]	-0.00 [-0.00, -0.00]
				test	384.56 [371.16, 397.85]	19.61 [19.27, 19.95]	15.34 [15.09, 15.60]	-0.00 [-0.00, -0.00]
			Baseline: Shuffled	train	384.91 [378.37, 392.01]	19.62 [19.45, 19.80]	15.16 [15.05, 15.28]	-0.00 [-0.00, -0.00]
				validation	405.87 [391.68, 421.56]	20.15 [19.79, 20.53]	15.41 [15.14, 15.69]	-0.00 [-0.00, 0.00]
				test	386.33 [372.80, 399.70]	19.65 [19.31, 19.99]	15.36 [15.11, 15.62]	-0.00 [-0.01, -0.00]
	MIMIC-IV	200	Regular	train	89.03 [84.62, 93.63]	9.44 [9.20, 9.68]	5.97 [5.89, 6.07]	0.86 [0.85, 0.86]
				validation	107.49 [95.42, 121.27]	10.36 [9.77, 11.01]	6.52 [6.31, 6.73]	0.80 [0.78, 0.82]
				test	105.85 [93.05, 119.66]	10.28 [9.65, 10.94]	6.40 [6.20, 6.61]	0.80 [0.78, 0.83]
			Baseline: Mean	train	625.03 [611.54, 639.19]	25.00 [24.73, 25.28]	19.31 [19.13, 19.50]	-0.00 [-0.00, -0.00]
				validation	541.82 [520.31, 565.88]	23.28 [22.81, 23.79]	18.51 [18.16, 18.89]	-0.00 [-0.00, -0.00]
				test	543.04 [519.60, 566.97]	23.30 [22.79, 23.81]	18.42 [18.03, 18.82]	-0.00 [-0.00, -0.00]
			Baseline: Shuffled	train	619.74 [606.30, 633.96]	24.89 [24.62, 25.18]	19.19 [19.01, 19.37]	0.01 [0.01, 0.01]
				validation	536.70 [515.52, 560.41]	23.17 [22.71, 23.67]	18.38 [18.04, 18.76]	0.01 [0.01, 0.01]
				test	537.91 [514.60, 561.75]	23.19 [22.68, 23.70]	18.29 [17.90, 18.69]	0.01 [0.00, 0.01]
	MIMIC-IV	500	Regular	train	96.16 [92.08, 100.51]	9.81 [9.60, 10.03]	6.43 [6.34, 6.53]	0.85 [0.84, 0.85]
				validation	111.64 [99.97, 126.08]	10.56 [10.00, 11.23]	6.89 [6.68, 7.12]	0.80 [0.77, 0.82]
				test	111.37 [101.35, 123.30]	10.55 [10.07, 11.10]	6.99 [6.79, 7.20]	0.80 [0.78, 0.82]
			Baseline: Mean	train	625.11 [609.76, 640.38]	25.00 [24.69, 25.31]	19.31 [19.11, 19.51]	-0.00 [-0.00, -0.00]
				validation	549.27 [524.97, 572.69]	23.43 [22.91, 23.93]	18.59 [18.18, 18.97]	-0.00 [-0.00, -0.00]
				test	554.00 [529.08, 578.16]	23.54 [23.00, 24.05]	18.63 [18.23, 19.01]	-0.00 [-0.00, -0.00]
			Baseline: Shuffled	train	625.98 [610.79, 641.28]	25.02 [24.71, 25.32]	19.32 [19.13, 19.52]	-0.00 [-0.00, -0.00]
				validation	550.13 [525.74, 573.51]	23.45 [22.93, 23.95]	18.60 [18.19, 18.99]	-0.00 [-0.01, -0.00]
				test	554.91 [529.94, 579.04]	23.56 [23.02, 24.06]	18.64 [18.25, 19.02]	-0.00 [-0.00, -0.00]
ResNet	TCC	200	Regular	train	38.62 [37.34, 40.02]	6.21 [6.11, 6.33]	4.35 [4.31, 4.40]	0.90 [0.90, 0.90]
				validation	57.70 [54.17, 61.25]	7.60 [7.36, 7.83]	5.22 [5.11, 5.33]	0.86 [0.85, 0.87]
				test	**66.39 [62.86, 70.27]**	**8.15 [7.93, 8.38]**	**5.50 [5.38, 5.62]**	**0.83 [0.82, 0.84]**
			Baseline: Mean	train	0.94 [0.94, 0.95]	0.97 [0.97, 0.97]	0.96 [0.95, 0.96]	-1.62e10
				validation	406.42 [391.88, 422.98]	20.16 [19.80, 20.57]	15.32 [15.05, 15.60]	0.00 [-0.00, 0.00]
				test	384.76 [371.33, 398.05]	19.62 [19.27, 19.95]	15.34 [15.09, 15.60]	-0.00 [-0.00, -0.00]
			Baseline: Shuffled	train	388.18 [381.32, 395.18]	19.70 [19.53, 19.88]	15.32 [15.20, 15.43]	-0.01 [-0.01, -0.01]
				validation	384.94 [371.47, 400.33]	19.62 [19.27, 20.01]	15.04 [14.78, 15.31]	0.05 [0.05, 0.06]
				test	395.05 [381.51, 408.38]	19.88 [19.53, 20.21]	15.60 [15.35, 15.86]	-0.03 [-0.03, -0.02]
	MIMIC-IV	200	Regular	train	101.06 [96.45, 105.99]	10.05 [9.82, 10.30]	6.38 [6.29, 6.47]	0.84 [0.83, 0.85]
				validation	97.41 [86.66, 110.32]	9.87 [9.31, 10.50]	6.40 [6.20, 6.60]	0.82 [0.80, 0.84]
				test	103.15 [90.60, 117.60]	10.16 [9.52, 10.84]	6.61 [6.42, 6.82]	0.81 [0.79, 0.83]
			Baseline: Mean	train	0.87 [0.86, 0.88]	0.93 [0.93, 0.94]	0.82 [0.81, 0.82]	-1.50e10
				validation	534.49 [512.05, 557.35]	23.12 [22.63, 23.61]	18.49 [18.13, 18.88]	0.01 [0.01, 0.01]
				test	542.88 [520.05, 565.44]	23.30 [22.80, 23.78]	18.40 [18.03, 18.74]	-0.00 [-0.00, 0.00]
			Baseline: Shuffled	train	622.74 [608.20, 637.62]	24.95 [24.66, 25.25]	18.99 [18.80, 19.19]	0.01 [0.00, 0.01]
				validation	524.80 [501.37, 549.24]	22.91 [22.39, 23.44]	17.98 [17.61, 18.37]	0.03 [0.02, 0.03]
				test	668.92 [638.51, 700.49]	25.86 [25.27, 26.47]	20.29 [19.84, 20.71]	-0.23 [-0.26, -0.20]
	MIMIC-IV	500	Regular	train	108.11 [102.36, 114.10]	10.40 [10.12, 10.68]	6.51 [6.42, 6.61]	0.83 [0.82, 0.84]
				validation	96.62 [85.53, 109.59]	9.83 [9.25, 10.47]	6.36 [6.17, 6.57]	0.82 [0.80, 0.84]
				test	104.53 [93.81, 116.17]	10.22 [9.69, 10.78]	6.78 [6.59, 6.99]	0.81 [0.79, 0.83]
			Baseline: Mean	train	0.07 [0.07, 0.07]	0.26 [0.26, 0.27]	0.23 [0.22, 0.23]	0.00 [0.00, 0.00]
				validation	544.06 [521.03, 567.76]	23.33 [22.83, 23.83]	18.50 [18.11, 18.88]	0.01 [0.00, 0.01]
				test	550.70 [527.86, 575.02]	23.47 [22.98, 23.98]	18.57 [18.21, 18.93]	-0.00 [-0.00, 0.00]
			Baseline: Shuffled	train	610.33 [594.76, 623.38]	24.70 [24.39, 24.97]	18.94 [18.74, 19.13]	0.02 [0.02, 0.03]
				validation	510.33 [488.46, 532.54]	22.59 [22.10, 23.08]	17.85 [17.48, 18.22]	0.07 [0.06, 0.07]
				test	747.82 [720.78, 775.23]	27.35 [26.85, 27.84]	22.09 [21.67, 22.50]	-0.36 [-0.40, -0.32]

Table 17: Model performance for predicting HR across input types, model heads, and CLEF representations, using TCC data. Results are reported as point estimates with 95% confidence intervals indicated in brackets. Grey rows indicate the representation with the lowest MSE on the test dataset for each input and model head. The globally lowest test MSE is highlighted in bold.

Input	Model head	Representation	Dataset	MSE	RMSE	MAE	R2
single	CNN	raw	train	157.64 [153.71, 162.12]	12.56 [12.40, 12.73]	8.80 [8.72, 8.89]	0.59 [0.58, 0.60]
			validation	171.84 [163.76, 180.38]	13.11 [12.80, 13.43]	9.29 [9.11, 9.49]	0.58 [0.56, 0.59]
			test	189.07 [180.06, 197.83]	13.75 [13.42, 14.07]	10.09 [9.90, 10.27]	0.51 [0.49, 0.53]
		CLEF-S	train	380.15 [375.49, 384.67]	19.50 [19.38, 19.61]	16.77 [16.67, 16.87]	0.01 [-0.01, 0.03]
			validation	357.23 [348.24, 366.00]	18.90 [18.66, 19.13]	16.11 [15.90, 16.31]	0.12 [0.09, 0.16]
			test	376.96 [367.12, 386.91]	19.41 [19.16, 19.67]	16.37 [16.16, 16.59]	0.02 [-0.02, 0.05]
		CLEF-M	train	161.63 [157.38, 166.21]	12.71 [12.55, 12.89]	8.69 [8.60, 8.78]	0.58 [0.57, 0.59]
			validation	204.20 [194.34, 214.83]	14.29 [13.94, 14.66]	9.91 [9.69, 10.13]	0.50 [0.48, 0.51]
			test	190.53 [182.33, 198.86]	13.80 [13.50, 14.10]	9.87 [9.67, 10.06]	0.50 [0.49, 0.52]
		CLEF-L	train	230.16 [224.77, 235.88]	15.17 [14.99, 15.36]	10.84 [10.74, 10.94]	0.40 [0.39, 0.41]
			validation	285.54 [273.10, 299.86]	16.90 [16.53, 17.32]	12.16 [11.91, 12.40]	0.30 [0.28, 0.31]
			test	248.86 [238.03, 259.58]	15.77 [15.43, 16.11]	11.39 [11.18, 11.63]	0.35 [0.34, 0.37]
	Linear	raw	train	5477.65 [5446.36, 5508.61]	74.01 [73.80, 74.22]	71.13 [70.95, 71.32]	-13.26 [-13.48, -13.05]
			validation	5629.41 [5558.67, 5703.91]	75.03 [74.56, 75.52]	71.93 [71.49, 72.37]	-12.84 [-13.27, -12.41]
			test	5560.85 [5494.45, 5621.41]	74.57 [74.12, 74.98]	71.60 [71.18, 71.98]	-13.47 [-13.91, -13.05]
		CLEF-S	train	368.31 [360.29, 376.42]	19.19 [18.98, 19.40]	14.10 [13.98, 14.22]	0.04 [0.03, 0.05]
			validation	395.79 [380.83, 413.06]	19.89 [19.51, 20.32]	14.56 [14.28, 14.86]	0.03 [-0.00, 0.06]
			test	474.80 [455.12, 494.39]	21.79 [21.33, 22.23]	16.02 [15.72, 16.33]	-0.23 [-0.28, -0.19]
		CLEF-M	train	142.11 [138.17, 146.07]	11.92 [11.75, 12.09]	8.40 [8.32, 8.48]	0.63 [0.62, 0.64]
			validation	176.41 [163.67, 192.13]	13.28 [12.79, 13.86]	9.17 [8.96, 9.38]	0.57 [0.53, 0.59]
			test	167.35 [161.06, 174.10]	12.94 [12.69, 13.19]	9.57 [9.39, 9.75]	0.56 [0.55, 0.58]
		CLEF-L	train	135.58 [132.10, 139.27]	11.64 [11.49, 11.80]	8.35 [8.27, 8.43]	0.65 [0.64, 0.65]
			validation	173.97 [160.41, 191.51]	13.19 [12.67, 13.84]	9.32 [9.14, 9.51]	0.57 [0.53, 0.60]
			test	172.89 [166.39, 180.03]	13.15 [12.90, 13.42]	9.82 [9.63, 9.99]	0.55 [0.54, 0.56]
double	CNN	raw	train	134.45 [131.09, 138.02]	11.59 [11.45, 11.75]	8.49 [8.41, 8.57]	0.65 [0.64, 0.66]
			validation	145.11 [139.19, 151.45]	12.05 [11.80, 12.31]	9.03 [8.85, 9.19]	0.64 [0.63, 0.66]
			test	**157.20 [150.71, 163.56]**	**12.54 [12.28, 12.79]**	**9.57 [9.40, 9.75]**	**0.59 [0.58, 0.61]**
		CLEF-S	train	213.28 [208.47, 218.24]	14.60 [14.44, 14.77]	10.83 [10.74, 10.91]	0.44 [0.44, 0.45]
			validation	251.42 [240.68, 262.88]	15.86 [15.51, 16.21]	11.82 [11.60, 12.06]	0.38 [0.36, 0.40]
			test	222.88 [214.26, 230.74]	14.93 [14.64, 15.19]	11.63 [11.44, 11.82]	0.42 [0.40, 0.44]
		CLEF-M	train	435.52 [429.36, 442.12]	20.87 [20.72, 21.03]	16.76 [16.63, 16.87]	-0.13 [-0.16, -0.11]
			validation	479.43 [464.38, 494.36]	21.90 [21.55, 22.23]	17.65 [17.39, 17.91]	-0.18 [-0.23, -0.13]
			test	531.88 [517.11, 546.56]	23.06 [22.74, 23.38]	18.76 [18.49, 19.03]	-0.38 [-0.43, -0.33]
		CLEF-L	train	271.83 [266.03, 277.58]	16.49 [16.31, 16.66]	11.99 [11.88, 12.09]	0.29 [0.28, 0.30]
			validation	332.54 [318.01, 348.62]	18.23 [17.83, 18.67]	13.06 [12.80, 13.33]	0.18 [0.16, 0.20]
			test	312.66 [300.09, 325.93]	17.68 [17.32, 18.05]	13.03 [12.79, 13.27]	0.19 [0.17, 0.21]
	Linear	raw	train	5467.82 [5434.17, 5501.25]	73.94 [73.72, 74.17]	71.01 [70.81, 71.22]	-13.24 [-13.48, -13.02]
			validation	5638.09 [5565.73, 5713.23]	75.09 [74.60, 75.59]	71.91 [71.47, 72.38]	-12.87 [-13.29, -12.43]
			test	5566.38 [5499.78, 5628.23]	74.61 [74.16, 75.02]	71.57 [71.14, 71.97]	-13.48 [-13.93, -13.06]
		CLEF-S	train	378.15 [370.17, 385.76]	19.45 [19.24, 19.64]	14.46 [14.33, 14.58]	0.01 [0.00, 0.03]
			validation	401.07 [385.42, 418.97]	20.03 [19.63, 20.47]	14.68 [14.40, 14.97]	0.01 [-0.01, 0.04]
			test	458.45 [440.97, 475.90]	21.41 [21.00, 21.82]	16.08 [15.78, 16.37]	-0.19 [-0.23, -0.15]
		CLEF-M	train	142.30 [138.28, 146.24]	11.93 [11.76, 12.09]	8.40 [8.32, 8.48]	0.63 [0.62, 0.64]
			validation	176.28 [163.54, 192.63]	13.27 [12.79, 13.88]	9.19 [8.98, 9.40]	0.57 [0.53, 0.59]
			test	161.49 [154.95, 167.81]	12.71 [12.45, 12.95]	9.43 [9.26, 9.60]	0.58 [0.57, 0.59]
		CLEF-L	train	138.19 [134.62, 141.86]	11.76 [11.60, 11.91]	8.33 [8.25, 8.41]	0.64 [0.63, 0.65]
			validation	186.68 [164.11, 224.54]	13.65 [12.81, 14.98]	9.34 [9.14, 9.55]	0.54 [0.45, 0.59]
			test	172.78 [165.92, 180.06]	13.14 [12.88, 13.42]	9.73 [9.55, 9.91]	0.55 [0.53, 0.56]
interpolated	CNN	raw	train	181.11 [176.68, 185.43]	13.46 [13.29, 13.62]	10.03 [9.95, 10.11]	0.53 [0.52, 0.54]
			validation	197.23 [189.02, 206.26]	14.04 [13.75, 14.36]	10.54 [10.35, 10.75]	0.52 [0.50, 0.53]
			test	192.34 [184.45, 200.10]	13.87 [13.58, 14.15]	10.60 [10.43, 10.79]	0.50 [0.48, 0.52]
		CLEF-S	train	1010.83 [1003.92, 1017.58]	31.79 [31.68, 31.90]	29.40 [29.29, 29.52]	-1.63 [-1.68, -1.58]
			validation	984.40 [968.05, 1000.44]	31.37 [31.11, 31.63]	28.49 [28.19, 28.76]	-1.42 [-1.51, -1.32]
			test	1017.35 [999.25, 1034.12]	31.90 [31.61, 32.16]	28.72 [28.44, 29.01]	-1.65 [-1.75, -1.55]
		CLEF-M	train	354.98 [350.47, 359.26]	18.84 [18.72, 18.95]	15.81 [15.72, 15.91]	0.07 [0.06, 0.09]
			validation	386.12 [375.78, 397.07]	19.65 [19.38, 19.93]	16.48 [16.26, 16.69]	0.05 [0.02, 0.08]
			test	427.79 [415.49, 440.03]	20.68 [20.38, 20.98]	17.24 [17.00, 17.48]	-0.11 [-0.15, -0.08]
		CLEF-L	train	232.80 [228.56, 237.56]	15.26 [15.12, 15.41]	11.53 [11.44, 11.63]	0.39 [0.39, 0.40]
			validation	269.07 [258.19, 281.70]	16.40 [16.07, 16.78]	12.09 [11.87, 12.32]	0.34 [0.32, 0.35]
			test	257.35 [247.61, 267.92]	16.04 [15.74, 16.37]	12.09 [11.87, 12.32]	0.33 [0.32, 0.35]
	Linear	raw	train	5465.25 [5433.49, 5496.80]	73.93 [73.71, 74.14]	70.98 [70.78, 71.19]	-13.24 [-13.46, -13.02]
			validation	5641.17 [5569.55, 5716.85]	75.11 [74.63, 75.61]	71.91 [71.48, 72.37]	-12.87 [-13.30, -12.43]
			test	5567.30 [5500.53, 5629.49]	74.61 [74.17, 75.03]	71.55 [71.13, 71.96]	-13.48 [-13.93, -13.07]
		CLEF-S	train	597.51 [587.88, 606.93]	24.44 [24.25, 24.64]	18.74 [18.58, 18.88]	-0.56 [-0.58, -0.53]
			validation	602.89 [582.38, 626.95]	24.55 [24.13, 25.04]	18.65 [18.32, 18.98]	-0.48 [-0.53, -0.44]
			test	608.19 [588.34, 627.73]	24.66 [24.26, 25.05]	19.38 [19.07, 19.68]	-0.58 [-0.63, -0.54]
		CLEF-M	train	141.91 [137.22, 147.38]	11.91 [11.71, 12.14]	8.21 [8.13, 8.29]	0.63 [0.62, 0.64]
			validation	168.77 [158.87, 180.03]	12.99 [12.60, 13.42]	8.98 [8.80, 9.19]	0.59 [0.56, 0.60]
			test	176.98 [168.00, 186.53]	13.30 [12.96, 13.66]	9.50 [9.31, 9.69]	0.54 [0.52, 0.56]
		CLEF-L	train	123.03 [119.71, 126.39]	11.09 [10.94, 11.24]	7.71 [7.63, 7.79]	0.68 [0.67, 0.69]
			validation	160.86 [152.90, 169.74]	12.68 [12.37, 13.03]	8.81 [8.63, 9.00]	0.60 [0.59, 0.62]
			test	165.39 [156.85, 174.19]	12.86 [12.52, 13.20]	8.95 [8.77, 9.13]	0.57 [0.55, 0.59]

Table 18: Model performance for predicting HR across input types, model heads, and CLEF representations, using MIMIC-IV data at 200Hz sample rate. Results are reported as point estimates with 95% confidence intervals indicated in brackets. Grey rows indicate the representation with the lowest MSE on the test dataset for each input and model head. The globally lowest test MSE is highlighted in bold.

Input	Model head	Representation	Dataset	MSE	RMSE	MAE	R2
single	CNN	raw	train	2027.59 [2005.62, 2051.12]	45.03 [44.78, 45.29]	39.86 [39.61, 40.13]	-2.23 [-2.32, -2.15]
			validation	2064.72 [2017.64, 2115.15]	45.44 [44.92, 45.99]	40.33 [39.79, 40.89]	-2.82 [-3.01, -2.63]
			test	2079.35 [2030.71, 2127.87]	45.60 [45.06, 46.13]	40.49 [39.95, 41.01]	-2.84 [-3.01, -2.66]
		CLEF-S	train	2570.74 [2548.73, 2595.90]	50.70 [50.48, 50.95]	48.05 [47.87, 48.28]	-3.09 [-3.18, -3.02]
			validation	2509.05 [2466.70, 2553.81]	50.09 [49.67, 50.54]	47.79 [47.41, 48.19]	-3.64 [-3.83, -3.46]
			test	2517.71 [2472.09, 2563.96]	50.18 [49.72, 50.64]	47.83 [47.44, 48.23]	-3.64 [-3.83, -3.48]
		CLEF-M	train	729.50 [712.16, 747.33]	27.01 [26.69, 27.34]	19.49 [19.26, 19.72]	-0.16 [-0.18, -0.15]
			validation	635.66 [604.35, 668.52]	25.21 [24.58, 25.86]	18.60 [18.16, 19.07]	-0.17 [-0.21, -0.14]
			test	631.30 [601.70, 661.14]	25.12 [24.53, 25.71]	18.55 [18.11, 19.00]	-0.16 [-0.20, -0.13]
		CLEF-L	train	385.32 [375.39, 395.66]	19.63 [19.38, 19.89]	14.76 [14.61, 14.92]	0.39 [0.38, 0.39]
			validation	338.66 [321.95, 355.63]	18.40 [17.94, 18.86]	14.16 [13.84, 14.46]	0.37 [0.35, 0.39]
			test	337.79 [322.08, 355.41]	18.38 [17.95, 18.85]	14.10 [13.82, 14.39]	0.38 [0.36, 0.40]
	Linear	raw	train	7512.96 [7450.95, 7575.02]	86.68 [86.32, 87.03]	82.59 [82.26, 82.89]	-10.97 [-11.21, -10.75]
			validation	7359.31 [7235.40, 7488.80]	85.79 [85.06, 86.54]	82.16 [81.50, 82.86]	-12.60 [-13.12, -12.13]
			test	7348.38 [7237.10, 7464.09]	85.72 [85.07, 86.39]	82.10 [81.49, 82.73]	-12.55 [-13.05, -12.07]
		CLEF-S	train	721.43 [703.92, 739.92]	26.86 [26.53, 27.20]	20.26 [20.05, 20.47]	-0.15 [-0.16, -0.14]
			validation	622.30 [593.78, 648.96]	24.94 [24.37, 25.47]	19.42 [19.00, 19.84]	-0.15 [-0.16, -0.14]
			test	622.63 [594.88, 650.27]	24.95 [24.39, 25.50]	19.25 [18.86, 19.65]	-0.15 [-0.16, -0.13]
		CLEF-M	train	454.99 [444.92, 466.22]	21.33 [21.09, 21.59]	16.53 [16.37, 16.70]	0.27 [0.27, 0.28]
			validation	400.20 [381.70, 420.22]	20.00 [19.54, 20.50]	15.89 [15.56, 16.22]	0.26 [0.24, 0.28]
			test	401.23 [383.75, 419.78]	20.03 [19.59, 20.49]	16.00 [15.69, 16.34]	0.26 [0.24, 0.28]
		CLEF-L	train	309.81 [300.37, 318.92]	17.60 [17.33, 17.86]	12.62 [12.46, 12.76]	0.51 [0.50, 0.51]
			validation	263.87 [247.94, 280.15]	16.24 [15.75, 16.74]	11.93 [11.64, 12.23]	0.51 [0.49, 0.53]
			test	258.74 [243.75, 275.39]	16.08 [15.61, 16.60]	11.75 [11.49, 12.03]	0.52 [0.50, 0.54]
double	CNN	raw	train	541.29 [530.22, 551.89]	23.27 [23.03, 23.49]	18.47 [18.30, 18.64]	0.14 [0.12, 0.16]
			validation	498.78 [477.99, 519.75]	22.33 [21.86, 22.80]	17.94 [17.58, 18.31]	0.08 [0.04, 0.12]
			test	498.01 [475.46, 520.29]	22.31 [21.81, 22.81]	17.83 [17.48, 18.17]	0.08 [0.04, 0.12]
		CLEF-S	train	5576.94 [5539.54, 5614.82]	74.68 [74.43, 74.93]	72.22 [71.99, 72.44]	-7.89 [-8.06, -7.72]
			validation	5495.36 [5425.16, 5571.84]	74.13 [73.66, 74.64]	71.95 [71.51, 72.42]	-9.16 [-9.56, -8.77]
			test	5483.40 [5410.97, 5557.60]	74.05 [73.56, 74.55]	71.88 [71.43, 72.35]	-9.11 [-9.51, -8.74]
		CLEF-M	train	915.62 [904.46, 927.28]	30.26 [30.07, 30.45]	26.21 [26.02, 26.39]	-0.46 [-0.49, -0.43]
			validation	880.11 [857.32, 903.24]	29.67 [29.28, 30.05]	25.68 [25.28, 26.05]	-0.63 [-0.69, -0.56]
			test	902.89 [878.88, 926.05]	30.05 [29.65, 30.43]	26.05 [25.65, 26.47]	-0.67 [-0.73, -0.60]
		CLEF-L	train	503.60 [490.24, 517.37]	22.44 [22.14, 22.75]	16.11 [15.90, 16.30]	0.20 [0.19, 0.21]
			validation	437.13 [412.01, 461.14]	20.91 [20.30, 21.47]	15.48 [15.11, 15.87]	0.19 [0.17, 0.22]
			test	427.48 [404.28, 451.36]	20.67 [20.11, 21.25]	15.21 [14.85, 15.56]	0.21 [0.19, 0.23]
	Linear	raw	train	7475.05 [7409.31, 7540.87]	86.46 [86.08, 86.84]	82.03 [81.70, 82.38]	-10.92 [-11.17, -10.70]
			validation	7350.10 [7221.06, 7479.56]	85.73 [84.98, 86.48]	81.78 [81.08, 82.49]	-12.59 [-13.11, -12.11]
			test	7350.72 [7232.19, 7468.67]	85.74 [85.04, 86.42]	81.81 [81.16, 82.46]	-12.56 [-13.06, -12.08]
		CLEF-S	train	713.83 [698.05, 730.32]	26.72 [26.42, 27.02]	20.36 [20.16, 20.56]	-0.14 [-0.14, -0.13]
			validation	616.18 [588.39, 642.14]	24.82 [24.26, 25.34]	19.52 [19.10, 19.94]	-0.14 [-0.15, -0.13]
			test	619.50 [592.30, 646.45]	24.89 [24.34, 25.43]	19.37 [18.98, 19.76]	-0.14 [-0.16, -0.13]
		CLEF-M	train	474.18 [462.99, 485.64]	21.78 [21.52, 22.04]	16.81 [16.63, 16.98]	0.24 [0.24, 0.25]
			validation	413.65 [393.84, 433.75]	20.34 [19.85, 20.83]	16.09 [15.74, 16.42]	0.24 [0.22, 0.25]
			test	415.24 [396.47, 433.73]	20.38 [19.91, 20.83]	16.22 [15.91, 16.55]	0.23 [0.21, 0.25]
		CLEF-L	train	313.07 [303.68, 321.57]	17.69 [17.43, 17.93]	12.58 [12.42, 12.73]	0.50 [0.49, 0.51]
			validation	274.01 [257.68, 291.22]	16.55 [16.05, 17.07]	12.06 [11.75, 12.36]	0.49 [0.47, 0.51]
			test	261.66 [244.82, 280.23]	16.17 [15.65, 16.74]	11.72 [11.45, 12.01]	0.52 [0.49, 0.54]
interpolated	CNN	raw	train	357.97 [349.68, 366.79]	18.92 [18.70, 19.15]	14.12 [13.98, 14.28]	0.43 [0.42, 0.44]
			validation	329.41 [311.56, 347.01]	18.15 [17.65, 18.63]	13.63 [13.30, 13.96]	0.39 [0.36, 0.43]
			test	323.64 [304.88, 341.92]	17.99 [17.46, 18.49]	13.48 [13.15, 13.78]	0.40 [0.37, 0.43]
		CLEF-S	train	2266.28 [2239.22, 2291.50]	47.61 [47.32, 47.87]	43.63 [43.39, 43.87]	-2.61 [-2.68, -2.55]
			validation	2194.43 [2147.19, 2246.13]	46.84 [46.34, 47.39]	43.29 [42.83, 43.76]	-3.06 [-3.20, -2.92]
			test	2184.52 [2137.00, 2237.47]	46.74 [46.23, 47.30]	43.15 [42.72, 43.60]	-3.03 [-3.18, -2.89]
		CLEF-M	train	261.31 [253.63, 268.63]	16.16 [15.93, 16.39]	11.94 [11.80, 12.08]	0.58 [0.57, 0.59]
			validation	233.65 [221.16, 247.60]	15.28 [14.87, 15.74]	11.47 [11.22, 11.74]	0.57 [0.55, 0.59]
			test	232.72 [217.67, 248.38]	15.25 [14.75, 15.76]	11.29 [11.03, 11.56]	0.57 [0.55, 0.59]
		CLEF-L	train	456.80 [443.49, 470.20]	21.37 [21.06, 21.68]	15.47 [15.29, 15.64]	0.27 [0.26, 0.28]
			validation	382.44 [362.90, 402.36]	19.55 [19.05, 20.06]	14.74 [14.40, 15.10]	0.29 [0.27, 0.31]
			test	388.87 [367.16, 411.30]	19.72 [19.16, 20.28]	14.81 [14.46, 15.15]	0.28 [0.26, 0.31]
	Linear	raw	train	7468.38 [7407.28, 7533.13]	86.42 [86.07, 86.79]	81.91 [81.59, 82.25]	-10.91 [-11.13, -10.67]
			validation	7352.67 [7221.91, 7480.57]	85.75 [84.98, 86.49]	81.70 [80.99, 82.40]	-12.59 [-13.11, -12.11]
			test	7356.71 [7235.92, 7476.83]	85.77 [85.06, 86.47]	81.76 [81.11, 82.42]	-12.57 [-13.07, -12.09]
		CLEF-S	train	685.00 [669.02, 701.97]	26.17 [25.87, 26.49]	19.98 [19.79, 20.17]	-0.09 [-0.10, -0.09]
			validation	583.36 [556.60, 608.12]	24.15 [23.59, 24.66]	19.07 [18.67, 19.47]	-0.08 [-0.09, -0.07]
			test	587.27 [561.99, 612.48]	24.23 [23.71, 24.75]	18.97 [18.60, 19.36]	-0.08 [-0.09, -0.07]
		CLEF-M	train	301.27 [291.09, 310.65]	17.36 [17.06, 17.63]	11.85 [11.70, 12.00]	0.52 [0.51, 0.53]
			validation	246.41 [230.95, 263.08]	15.70 [15.20, 16.22]	11.12 [10.83, 11.42]	0.54 [0.53, 0.56]
			test	247.03 [228.95, 265.83]	15.71 [15.13, 16.30]	10.92 [10.63, 11.22]	0.54 [0.52, 0.57]
		CLEF-L	train	184.27 [177.09, 191.56]	13.57 [13.31, 13.84]	9.16 [9.05, 9.29]	0.71 [0.70, 0.71]
			validation	155.61 [144.81, 167.89]	12.47 [12.03, 12.96]	8.74 [8.49, 8.97]	0.71 [0.69, 0.73]
			test	**156.12 [142.09, 170.57]**	**12.49 [11.92, 13.06]**	**8.63 [8.40, 8.87]**	**0.71 [0.69, 0.73]**

Table 19: Model performance for predicting HR across model heads, and CLEF representations, using MIMIC-IV data at 500Hz sample rate. Results are reported as point estimates with 95% confidence intervals indicated in brackets. Grey rows indicate the representation with the lowest MSE on the test dataset for each input and model head. The globally lowest test MSE is highlighted in bold.

Input	Model head	Representation	Dataset	MSE	RMSE	MAE	R2
single (500Hz)	CNN	raw	train	465.50 [456.55, 474.79]	21.58 [21.37, 21.79]	17.17 [17.01, 17.34]	0.26 [0.24, 0.27]
			validation	412.89 [394.97, 431.15]	20.32 [19.87, 20.76]	16.18 [15.85, 16.50]	0.24 [0.21, 0.29]
			test	418.76 [401.80, 437.21]	20.46 [20.04, 20.91]	16.34 [16.01, 16.68]	0.24 [0.20, 0.28]
		CLEF-S	train	4722.98 [4684.79, 4762.82]	68.72 [68.45, 69.01]	65.56 [65.31, 65.82]	-6.56 [-6.70, -6.41]
			validation	4628.62 [4551.06, 4709.16]	68.03 [67.46, 68.62]	65.22 [64.70, 65.76]	-7.46 [-7.79, -7.16]
			test	4661.49 [4589.98, 4733.53]	68.27 [67.75, 68.80]	65.44 [64.95, 65.94]	-7.47 [-7.77, -7.17]
		CLEF-M	train	3919.21 [3876.93, 3958.07]	62.60 [62.26, 62.91]	57.40 [57.07, 57.70]	-5.27 [-5.43, -5.11]
			validation	3912.73 [3830.32, 3995.36]	62.55 [61.89, 63.21]	57.56 [56.89, 58.23]	-6.16 [-6.52, -5.81]
			test	3862.13 [3787.09, 3936.41]	62.15 [61.54, 62.74]	57.30 [56.66, 57.96]	-6.02 [-6.37, -5.71]
		CLEF-L	train	788.75 [773.11, 805.35]	28.08 [27.80, 28.38]	23.41 [23.22, 23.60]	-0.26 [-0.28, -0.25]
			validation	702.86 [676.55, 732.49]	26.51 [26.01, 27.06]	22.69 [22.31, 23.07]	-0.29 [-0.32, -0.25]
			test	723.59 [699.35, 750.37]	26.90 [26.45, 27.39]	23.02 [22.68, 23.39]	-0.32 [-0.35, -0.28]
	Linear	raw	train	7441.42 [7375.40, 7509.01]	86.26 [85.88, 86.65]	81.75 [81.39, 82.09]	-10.90 [-11.13, -10.66]
			validation	7355.86 [7226.43, 7493.57]	85.77 [85.01, 86.57]	81.68 [80.97, 82.38]	-12.45 [-12.95, -11.97]
			test	7361.91 [7243.42, 7485.81]	85.80 [85.11, 86.52]	81.72 [81.05, 82.39]	-12.38 [-12.86, -11.90]
		CLEF-S	train	684.69 [667.98, 702.76]	26.17 [25.85, 26.51]	19.98 [19.77, 20.19]	-0.09 [-0.10, -0.09]
			validation	587.74 [562.82, 614.19]	24.24 [23.72, 24.78]	19.07 [18.65, 19.46]	-0.07 [-0.08, -0.07]
			test	592.11 [568.07, 618.31]	24.33 [23.83, 24.87]	19.10 [18.73, 19.48]	-0.08 [-0.08, -0.07]
		CLEF-M	train	305.09 [295.32, 315.97]	17.47 [17.18, 17.78]	11.90 [11.76, 12.06]	0.51 [0.50, 0.52]
			validation	236.61 [220.44, 253.81]	15.38 [14.85, 15.93]	10.90 [10.62, 11.18]	0.57 [0.55, 0.59]
			test	245.24 [229.50, 261.28]	15.66 [15.15, 16.16]	11.15 [10.88, 11.44]	0.55 [0.53, 0.57]
		CLEF-L	train	185.85 [178.32, 194.11]	13.63 [13.35, 13.93]	9.19 [9.07, 9.31]	0.70 [0.69, 0.71]
			validation	154.14 [141.23, 167.22]	12.41 [11.88, 12.93]	8.67 [8.43, 8.89]	0.72 [0.70, 0.74]
			test	**147.25 [136.55, 158.82]**	**12.13 [11.69, 12.60]**	**8.64 [8.41, 8.88]**	**0.73 [0.71, 0.75]**

H Heart Rate Prediction from CLEF Embeddings using XGBoost - Extensive Results

Table 20: HR prediction with CLEF representations as inputs using XGBoost and the MIMIC-IV dataset. Results are reported as point estimates with 95% confidence intervals indicated in brackets. Grey rows mark the lowest test MSE per input and representation across settings. The globally best test MSE is highlighted in bold.

Input	Repr.	Run	Dataset	MSE	RMSE	MAE	R2
single	CLEF-S	regular	train	218.24 [211.70, 225.10]	14.77 [14.55, 15.00]	10.66 [10.54, 10.78]	0.65 [0.64, 0.66]
			validation	196.65 [183.68, 211.33]	14.02 [13.55, 14.54]	10.30 [10.05, 10.56]	0.64 [0.62, 0.65]
			test	203.53 [189.91, 217.44]	14.26 [13.78, 14.75]	10.35 [10.10, 10.60]	0.62 [0.61, 0.64]
		mean	train	627.17 [613.76, 642.43]	25.04 [24.77, 25.35]	19.34 [19.16, 19.53]	-0.00 [-0.00, -0.00]
			validation	541.97 [517.84, 564.79]	23.28 [22.76, 23.77]	18.52 [18.14, 18.90]	-0.00 [-0.00, -0.00]
			test	543.58 [521.00, 566.21]	23.31 [22.83, 23.80]	18.43 [18.07, 18.77]	-0.00 [-0.00, -0.00]
		shuffled	train	626.00 [612.61, 640.96]	25.02 [24.75, 25.32]	19.29 [19.11, 19.48]	0.00 [0.00, 0.00]
			validation	540.06 [516.12, 562.90]	23.24 [22.72, 23.73]	18.46 [18.08, 18.84]	0.00 [-0.00, 0.00]
			test	540.61 [517.96, 562.93]	23.25 [22.76, 23.73]	18.35 [17.98, 18.69]	0.00 [0.00, 0.01]
	CLEF-M	regular	train	303.35 [295.33, 312.25]	17.42 [17.19, 17.67]	12.64 [12.50, 12.79]	0.52 [0.51, 0.52]
			validation	279.67 [263.68, 296.93]	16.72 [16.24, 17.23]	12.37 [12.08, 12.67]	0.48 [0.47, 0.50]
			test	280.74 [264.14, 297.79]	16.75 [16.25, 17.26]	12.37 [12.08, 12.65]	0.48 [0.46, 0.50]
		mean	train	627.17 [613.76, 642.43]	25.04 [24.77, 25.35]	19.34 [19.16, 19.53]	-0.00 [-0.00, -0.00]
			validation	541.97 [517.84, 564.79]	23.28 [22.76, 23.77]	18.52 [18.14, 18.90]	-0.00 [-0.00, -0.00]
			test	543.58 [521.00, 566.21]	23.31 [22.83, 23.80]	18.43 [18.07, 18.77]	-0.00 [-0.00, -0.00]
		shuffled	train	626.32 [612.73, 641.44]	25.03 [24.75, 25.33]	19.31 [19.14, 19.51]	0.00 [-0.00, 0.00]
			validation	540.77 [516.81, 563.39]	23.25 [22.73, 23.74]	18.49 [18.11, 18.87]	0.00 [-0.00, 0.00]
			test	541.47 [518.95, 563.89]	23.27 [22.78, 23.75]	18.38 [18.02, 18.73]	0.00 [-0.00, 0.00]
	CLEF-L	regular	train	323.79 [315.87, 332.58]	17.99 [17.77, 18.24]	13.11 [12.98, 13.26]	0.48 [0.48, 0.49]
			validation	304.33 [285.87, 323.50]	17.44 [16.91, 17.99]	12.97 [12.66, 13.28]	0.44 [0.42, 0.45]
			test	300.18 [283.19, 318.01]	17.32 [16.83, 17.83]	12.75 [12.45, 13.05]	0.45 [0.43, 0.46]
		mean	train	627.17 [613.76, 642.43]	25.04 [24.77, 25.35]	19.34 [19.16, 19.53]	-0.00 [-0.00, -0.00]
			validation	541.97 [517.84, 564.79]	23.28 [22.76, 23.77]	18.52 [18.14, 18.90]	-0.00 [-0.00, -0.00]
			test	543.58 [521.00, 566.21]	23.31 [22.83, 23.80]	18.43 [18.07, 18.77]	-0.00 [-0.00, -0.00]
		shuffled	train	625.05 [611.89, 640.33]	25.00 [24.74, 25.30]	19.28 [19.10, 19.48]	0.00 [0.00, 0.00]
			validation	539.67 [516.44, 562.08]	23.23 [22.73, 23.71]	18.46 [18.09, 18.84]	0.00 [-0.00, 0.01]
			test	540.63 [518.28, 562.87]	23.25 [22.77, 23.72]	18.36 [17.99, 18.70]	0.00 [-0.00, 0.01]
double	CLEF-S	regular	train	211.20 [205.14, 218.03]	14.53 [14.32, 14.77]	10.52 [10.41, 10.65]	0.66 [0.66, 0.67]
			validation	192.02 [179.48, 206.16]	13.86 [13.40, 14.36]	10.18 [9.95, 10.43]	0.65 [0.63, 0.66]
			test	**194.09 [181.01, 207.79]**	**13.93 [13.45, 14.41]**	**10.17 [9.93, 10.41]**	**0.64** [0.63, 0.66]
		mean	train	627.17 [613.76, 642.43]	25.04 [24.77, 25.35]	19.34 [19.16, 19.53]	-0.00 [-0.00, -0.00]
			validation	541.97 [517.84, 564.79]	23.28 [22.76, 23.77]	18.52 [18.14, 18.90]	-0.00 [-0.00, -0.00]
			test	543.58 [521.00, 566.21]	23.31 [22.83, 23.80]	18.43 [18.07, 18.77]	-0.00 [-0.00, -0.00]
		shuffled	train	625.57 [612.22, 640.66]	25.01 [24.74, 25.31]	19.27 [19.09, 19.46]	0.00 [0.00, 0.00]
			validation	539.26 [515.65, 562.08]	23.22 [22.71, 23.71]	18.43 [18.06, 18.81]	0.00 [0.00, 0.01]
			test	540.45 [517.86, 562.90]	23.25 [22.76, 23.73]	18.35 [17.98, 18.69]	0.00 [0.00, 0.01]
	CLEF-M	regular	train	337.65 [329.27, 346.53]	18.37 [18.15, 18.62]	13.58 [13.43, 13.74]	0.46 [0.45, 0.47]
			validation	310.49 [294.28, 328.30]	17.62 [17.15, 18.12]	13.27 [12.99, 13.58]	0.43 [0.41, 0.44]
			test	313.08 [296.13, 330.08]	17.69 [17.21, 18.17]	13.36 [13.06, 13.66]	0.42 [0.40, 0.44]
		mean	train	627.17 [613.76, 642.43]	25.04 [24.77, 25.35]	19.34 [19.16, 19.53]	-0.00 [-0.00, -0.00]
			validation	541.97 [517.84, 564.79]	23.28 [22.76, 23.77]	18.52 [18.14, 18.90]	-0.00 [-0.00, -0.00]
			test	543.58 [521.00, 566.21]	23.31 [22.83, 23.80]	18.43 [18.07, 18.77]	-0.00 [-0.00, -0.00]
		shuffled	train	626.24 [612.74, 641.50]	25.02 [24.75, 25.33]	19.32 [19.14, 19.51]	0.00 [-0.00, 0.00]
			validation	540.88 [517.30, 563.49]	23.26 [22.74, 23.74]	18.49 [18.12, 18.86]	0.00 [-0.00, 0.00]
			test	541.93 [519.78, 564.92]	23.28 [22.80, 23.77]	18.40 [18.03, 18.74]	0.00 [-0.00, 0.00]
	CLEF-L	regular	train	319.43 [310.95, 328.30]	17.87 [17.63, 18.12]	13.02 [12.87, 13.16]	0.49 [0.48, 0.50]
			validation	304.48 [287.42, 322.15]	17.45 [16.95, 17.95]	13.00 [12.70, 13.31]	0.44 [0.42, 0.45]
			test	294.37 [277.55, 311.12]	17.16 [16.66, 17.64]	12.67 [12.37, 12.96]	0.46 [0.44, 0.47]
		mean	train	627.17 [613.76, 642.43]	25.04 [24.77, 25.35]	19.34 [19.16, 19.53]	-0.00 [-0.00, -0.00]
			validation	541.97 [517.84, 564.79]	23.28 [22.76, 23.77]	18.52 [18.14, 18.90]	-0.00 [-0.00, -0.00]
			test	543.58 [521.00, 566.21]	23.31 [22.83, 23.80]	18.43 [18.07, 18.77]	-0.00 [-0.00, -0.00]
		shuffled	train	625.79 [612.22, 641.15]	25.02 [24.74, 25.32]	19.28 [19.10, 19.48]	0.00 [0.00, 0.00]
			validation	540.09 [516.44, 563.07]	23.24 [22.73, 23.73]	18.46 [18.08, 18.84]	0.00 [-0.00, 0.00]
			test	541.29 [519.07, 563.91]	23.26 [22.78, 23.75]	18.39 [18.03, 18.73]	0.00 [-0.00, 0.00]
interpolated	CLEF-S	regular	train	283.68 [276.25, 291.65]	16.84 [16.62, 17.08]	12.41 [12.28, 12.55]	0.55 [0.54, 0.55]
			validation	262.49 [247.40, 278.82]	16.20 [15.73, 16.70]	12.07 [11.77, 12.36]	0.52 [0.50, 0.53]
			test	264.47 [248.99, 279.85]	16.26 [15.78, 16.73]	12.00 [11.73, 12.29]	0.51 [0.49, 0.53]
		mean	train	627.17 [613.76, 642.43]	25.04 [24.77, 25.35]	19.34 [19.16, 19.53]	-0.00 [-0.00, -0.00]
			validation	541.97 [517.84, 564.79]	23.28 [22.76, 23.77]	18.52 [18.14, 18.90]	-0.00 [-0.00, -0.00]
			test	543.58 [521.00, 566.21]	23.31 [22.83, 23.80]	18.43 [18.07, 18.77]	-0.00 [-0.00, -0.00]
		shuffled	train	626.13 [612.94, 641.01]	25.02 [24.76, 25.32]	19.32 [19.14, 19.51]	0.00 [0.00, 0.00]
			validation	541.22 [516.99, 564.01]	23.26 [22.74, 23.75]	18.50 [18.12, 18.88]	0.00 [-0.00, 0.00]
			test	541.97 [519.26, 564.36]	23.28 [22.79, 23.76]	18.41 [18.04, 18.75]	0.00 [-0.00, 0.00]
	CLEF-M	regular	train	217.83 [211.20, 225.05]	14.76 [14.53, 15.00]	10.32 [10.20, 10.45]	0.65 [0.65, 0.66]
			validation	199.71 [186.58, 213.34]	14.13 [13.66, 14.61]	10.09 [9.85, 10.37]	0.63 [0.61, 0.65]
			test	200.63 [186.28, 216.30]	14.16 [13.65, 14.71]	9.95 [9.69, 10.19]	0.63 [0.61, 0.65]
		mean	train	627.17 [613.76, 642.43]	25.04 [24.77, 25.35]	19.34 [19.16, 19.53]	-0.00 [-0.00, -0.00]
			validation	541.97 [517.84, 564.79]	23.28 [22.76, 23.77]	18.52 [18.14, 18.90]	-0.00 [-0.00, -0.00]
			test	543.58 [521.00, 566.21]	23.31 [22.83, 23.80]	18.43 [18.07, 18.77]	-0.00 [-0.00, -0.00]
		shuffled	train	626.86 [613.27, 642.11]	25.04 [24.76, 25.34]	19.28 [19.10, 19.48]	0.00 [-0.00, 0.00]
			validation	540.71 [516.59, 563.31]	23.25 [22.73, 23.73]	18.45 [18.07, 18.83]	0.00 [-0.00, 0.00]
			test	541.82 [519.13, 564.79]	23.28 [22.78, 23.77]	18.36 [18.00, 18.71]	0.00 [-0.00, 0.00]
	CLEF-L	regular	train	217.61 [210.59, 225.28]	14.75 [14.51, 15.01]	10.37 [10.25, 10.50]	0.65 [0.65, 0.66]
			validation	198.00 [185.22, 211.19]	14.07 [13.61, 14.53]	10.14 [9.90, 10.39]	0.63 [0.62, 0.65]
			test	201.72 [186.51, 217.52]	14.20 [13.66, 14.75]	10.00 [9.75, 10.26]	0.63 [0.61, 0.65]
		mean	train	627.17 [613.76, 642.43]	25.04 [24.77, 25.35]	19.34 [19.16, 19.53]	-0.00 [-0.00, -0.00]
			validation	541.97 [517.84, 564.79]	23.28 [22.76, 23.77]	18.52 [18.14, 18.90]	-0.00 [-0.00, -0.00]
			test	543.58 [521.00, 566.21]	23.31 [22.83, 23.80]	18.43 [18.07, 18.77]	-0.00 [-0.00, -0.00]
		shuffled	train	627.73 [613.91, 643.28]	25.05 [24.78, 25.36]	19.32 [19.14, 19.52]	-0.00 [-0.00, 0.00]
			validation	542.04 [518.42, 564.92]	23.28 [22.77, 23.77]	18.51 [18.12, 18.89]	-0.00 [-0.00, 0.00]
			test	543.80 [520.88, 566.64]	23.32 [22.82, 23.80]	18.41 [18.05, 18.76]	-0.00 [-0.01, 0.00]

I Feature Prediction from CLEF Embeddings using XGBoost - Extensive Results

Table 21: Results for predicting mean ECG rate from different CLEF representations and input formats. Results are reported as point estimates with 95% confidence intervals indicated in brackets. The row with the lowest test MSE per CLEF input format is highlighted in grey, and the global best result is emphasized in bold.

Results: ECG rate prediction from various CLEF representations, using XGBoost

Input	Repr.	Run	Dataset	MSE	RMSE	MAE	R2
single	CLEF-S	regular	train	149.16 [144.94, 153.18]	12.21 [12.04, 12.38]	8.84 [8.74, 8.94]	0.74 [0.74, 0.75]
			validation	143.71 [134.39, 153.59]	11.99 [11.59, 12.39]	8.63 [8.41, 8.85]	0.72 [0.70, 0.73]
			test	142.21 [132.96, 151.32]	11.92 [11.53, 12.30]	8.61 [8.39, 8.83]	0.72 [0.71, 0.74]
		mean	train	576.99 [565.74, 588.00]	24.02 [23.79, 24.25]	18.92 [18.74, 19.10]	-0.00 [-0.00, -0.00]
			validation	512.85 [492.31, 533.53]	22.65 [22.19, 23.10]	18.21 [17.86, 18.55]	-0.00 [-0.01, -0.00]
			test	517.14 [495.47, 537.96]	22.74 [22.26, 23.19]	18.21 [17.86, 18.57]	-0.00 [-0.01, -0.00]
		shuffled	train	576.11 [564.56, 587.07]	24.00 [23.76, 24.23]	18.84 [18.66, 19.01]	0.00 [0.00, 0.00]
			validation	510.82 [490.21, 531.43]	22.60 [22.14, 23.05]	18.11 [17.74, 18.46]	0.00 [-0.00, 0.00]
			test	515.16 [493.89, 536.07]	22.70 [22.22, 23.15]	18.11 [17.76, 18.47]	0.00 [-0.00, 0.00]
	CLEF-M	regular	train	248.76 [242.06, 255.75]	15.77 [15.56, 15.99]	11.35 [11.22, 11.50]	0.57 [0.56, 0.58]
			validation	235.78 [221.36, 250.42]	15.35 [14.88, 15.82]	11.08 [10.79, 11.37]	0.54 [0.52, 0.56]
			test	238.02 [224.43, 251.98]	15.43 [14.98, 15.87]	11.20 [10.91, 11.49]	0.54 [0.52, 0.55]
		mean	train	576.99 [565.74, 588.00]	24.02 [23.79, 24.25]	18.92 [18.74, 19.10]	-0.00 [-0.00, -0.00]
			validation	512.85 [492.31, 533.53]	22.65 [22.19, 23.10]	18.21 [17.86, 18.55]	-0.00 [-0.01, -0.00]
			test	517.14 [495.47, 537.96]	22.74 [22.26, 23.19]	18.21 [17.86, 18.57]	-0.00 [-0.01, -0.00]
		shuffled	train	576.40 [564.87, 587.37]	24.01 [23.77, 24.24]	18.85 [18.66, 19.02]	0.00 [-0.00, 0.00]
			validation	510.42 [489.64, 531.12]	22.59 [22.13, 23.05]	18.10 [17.74, 18.45]	0.00 [-0.00, 0.01]
			test	515.26 [493.84, 536.06]	22.70 [22.22, 23.15]	18.13 [17.77, 18.49]	-0.00 [-0.00, 0.00]
	CLEF-L	regular	train	272.13 [265.41, 279.04]	16.50 [16.29, 16.70]	12.07 [11.93, 12.21]	0.53 [0.52, 0.53]
			validation	265.10 [250.99, 281.45]	16.28 [15.84, 16.78]	12.02 [11.73, 12.32]	0.48 [0.47, 0.50]
			test	258.26 [244.25, 273.18]	16.07 [15.63, 16.53]	11.91 [11.62, 12.23]	0.50 [0.48, 0.51]
		mean	train	576.99 [565.74, 588.00]	24.02 [23.79, 24.25]	18.92 [18.74, 19.10]	-0.00 [-0.00, -0.00]
			validation	512.85 [492.31, 533.53]	22.65 [22.19, 23.10]	18.21 [17.86, 18.55]	-0.00 [-0.01, -0.00]
			test	517.14 [495.47, 537.96]	22.74 [22.26, 23.19]	18.21 [17.86, 18.57]	-0.00 [-0.01, -0.00]
		shuffled	train	577.93 [566.38, 588.74]	24.04 [23.80, 24.26]	18.91 [18.73, 19.08]	-0.00 [-0.00, -0.00]
			validation	512.62 [492.22, 533.56]	22.64 [22.19, 23.10]	18.18 [17.82, 18.53]	-0.00 [-0.01, 0.00]
			test	517.46 [496.10, 538.21]	22.75 [22.27, 23.20]	18.20 [17.85, 18.56]	-0.00 [-0.01, -0.00]
double	CLEF-S	regular	train	142.90 [138.73, 146.80]	11.95 [11.78, 12.12]	8.70 [8.60, 8.81]	0.75 [0.75, 0.76]
			validation	139.33 [130.54, 148.88]	11.80 [11.43, 12.20]	8.57 [8.36, 8.79]	0.73 [0.71, 0.74]
			test	**136.15 [127.64, 145.19]**	**11.67 [11.30, 12.05]**	**8.47 [8.26, 8.69]**	**0.74 [0.72, 0.75]**
		mean	train	576.99 [565.74, 588.00]	24.02 [23.79, 24.25]	18.92 [18.74, 19.10]	-0.00 [-0.00, -0.00]
			validation	512.85 [492.31, 533.53]	22.65 [22.19, 23.10]	18.21 [17.86, 18.55]	-0.00 [-0.01, -0.00]
			test	517.14 [495.47, 537.96]	22.74 [22.26, 23.19]	18.21 [17.86, 18.57]	-0.00 [-0.01, -0.00]
		shuffled	train	575.59 [564.17, 586.40]	23.99 [23.75, 24.22]	18.85 [18.67, 19.03]	0.00 [0.00, 0.00]
			validation	510.96 [490.56, 531.62]	22.60 [22.15, 23.06]	18.12 [17.76, 18.47]	0.00 [-0.00, 0.00]
			test	515.44 [494.57, 536.45]	22.70 [22.24, 23.16]	18.14 [17.78, 18.50]	-0.00 [-0.00, 0.00]
	CLEF-M	regular	train	287.63 [280.57, 294.81]	16.96 [16.75, 17.17]	12.52 [12.37, 12.66]	0.50 [0.49, 0.51]
			validation	272.46 [257.52, 287.92]	16.50 [16.05, 16.97]	12.29 [11.99, 12.58]	0.47 [0.45, 0.49]
			test	276.13 [262.34, 290.99]	16.62 [16.20, 17.06]	12.40 [12.10, 12.70]	0.46 [0.44, 0.48]
		mean	train	576.99 [565.74, 588.00]	24.02 [23.79, 24.25]	18.92 [18.74, 19.10]	-0.00 [-0.00, -0.00]
			validation	512.85 [492.31, 533.53]	22.65 [22.19, 23.10]	18.21 [17.86, 18.55]	-0.00 [-0.01, -0.00]
			test	517.14 [495.47, 537.96]	22.74 [22.26, 23.19]	18.21 [17.86, 18.57]	-0.00 [-0.01, -0.00]
		shuffled	train	576.64 [565.08, 587.66]	24.01 [23.77, 24.24]	18.89 [18.70, 19.06]	0.00 [-0.00, 0.00]
			validation	511.28 [490.71, 531.78]	22.61 [22.15, 23.06]	18.15 [17.78, 18.49]	0.00 [-0.00, 0.00]
			test	516.62 [495.08, 537.32]	22.73 [22.25, 23.18]	18.17 [17.82, 18.53]	-0.00 [-0.01, 0.00]
	CLEF-L	regular	train	267.68 [261.05, 274.27]	16.36 [16.16, 16.56]	12.02 [11.87, 12.16]	0.54 [0.53, 0.54]
			validation	262.86 [248.93, 277.85]	16.21 [15.78, 16.67]	11.96 [11.68, 12.26]	0.49 [0.47, 0.50]
			test	255.59 [242.73, 269.97]	15.99 [15.58, 16.43]	11.90 [11.63, 12.21]	0.50 [0.49, 0.52]
		mean	train	576.99 [565.74, 588.00]	24.02 [23.79, 24.25]	18.92 [18.74, 19.10]	-0.00 [-0.00, -0.00]
			validation	512.85 [492.31, 533.53]	22.65 [22.19, 23.10]	18.21 [17.86, 18.55]	-0.00 [-0.01, -0.00]
			test	517.14 [495.47, 537.96]	22.74 [22.26, 23.19]	18.21 [17.86, 18.57]	-0.00 [-0.01, -0.00]
		shuffled	train	578.92 [567.27, 590.00]	24.06 [23.82, 24.29]	18.94 [18.75, 19.11]	-0.00 [-0.00, -0.00]
			validation	513.11 [492.48, 533.89]	22.65 [22.19, 23.11]	18.21 [17.84, 18.55]	-0.00 [-0.01, -0.00]
			test	517.39 [495.91, 538.36]	22.74 [22.27, 23.20]	18.21 [17.86, 18.57]	-0.00 [-0.01, -0.00]
interp.	CLEF-S	regular	train	230.36 [224.41, 235.95]	15.18 [14.98, 15.36]	11.25 [11.12, 11.37]	0.60 [0.59, 0.61]
			validation	223.20 [210.30, 236.33]	14.94 [14.50, 15.37]	11.03 [10.77, 11.30]	0.56 [0.54, 0.58]
			test	221.73 [209.68, 234.51]	14.89 [14.48, 15.31]	11.05 [10.79, 11.33]	0.57 [0.55, 0.59]
		mean	train	576.99 [565.74, 588.00]	24.02 [23.79, 24.25]	18.92 [18.74, 19.10]	-0.00 [-0.00, -0.00]
			validation	512.85 [492.31, 533.53]	22.65 [22.19, 23.10]	18.21 [17.86, 18.55]	-0.00 [-0.01, -0.00]
			test	517.14 [495.47, 537.96]	22.74 [22.26, 23.19]	18.21 [17.86, 18.57]	-0.00 [-0.01, -0.00]
		shuffled	train	576.36 [564.86, 587.40]	24.01 [23.77, 24.24]	18.86 [18.67, 19.04]	0.00 [-0.00, 0.00]
			validation	511.60 [491.17, 532.29]	22.62 [22.16, 23.07]	18.16 [17.80, 18.51]	0.00 [-0.00, 0.00]
			test	516.01 [494.71, 536.87]	22.71 [22.24, 23.17]	18.17 [17.82, 18.53]	-0.00 [-0.01, 0.00]
	CLEF-M	regular	train	148.24 [143.44, 152.79]	12.18 [11.98, 12.36]	8.32 [8.21, 8.42]	0.74 [0.74, 0.75]
			validation	146.86 [136.13, 158.70]	12.12 [11.67, 12.60]	8.20 [7.98, 8.43]	0.71 [0.70, 0.73]
			test	144.85 [134.71, 155.46]	12.03 [11.61, 12.47]	8.22 [7.98, 8.46]	0.72 [0.70, 0.73]
		mean	train	576.99 [565.74, 588.00]	24.02 [23.79, 24.25]	18.92 [18.74, 19.10]	-0.00 [-0.00, -0.00]
			validation	512.85 [492.31, 533.53]	22.65 [22.19, 23.10]	18.21 [17.86, 18.55]	-0.00 [-0.01, -0.00]
			test	517.14 [495.47, 537.96]	22.74 [22.26, 23.19]	18.21 [17.86, 18.57]	-0.00 [-0.01, -0.00]
		shuffled	train	575.39 [564.07, 585.99]	23.99 [23.75, 24.21]	18.85 [18.66, 19.02]	0.00 [0.00, 0.00]
			validation	511.14 [490.42, 531.85]	22.61 [22.15, 23.06]	18.14 [17.77, 18.49]	0.00 [-0.00, 0.00]
			test	515.82 [494.90, 536.47]	22.71 [22.25, 23.16]	18.15 [17.80, 18.51]	-0.00 [-0.01, 0.00]
	CLEF-L	regular	train	150.78 [146.14, 155.26]	12.28 [12.09, 12.46]	8.51 [8.40, 8.61]	0.74 [0.73, 0.74]
			validation	144.87 [135.23, 155.09]	12.03 [11.63, 12.45]	8.29 [8.07, 8.53]	0.72 [0.70, 0.73]
			test	145.87 [135.70, 156.52]	12.08 [11.65, 12.51]	8.36 [8.12, 8.59]	0.72 [0.70, 0.73]
		mean	train	576.99 [565.74, 588.00]	24.02 [23.79, 24.25]	18.92 [18.74, 19.10]	-0.00 [-0.00, -0.00]
			validation	512.85 [492.31, 533.53]	22.65 [22.19, 23.10]	18.21 [17.86, 18.55]	-0.00 [-0.01, -0.00]
			test	517.14 [495.47, 537.96]	22.74 [22.26, 23.19]	18.21 [17.86, 18.57]	-0.00 [-0.01, -0.00]
		shuffled	train	575.15 [563.91, 586.01]	23.98 [23.75, 24.21]	18.85 [18.66, 19.02]	0.00 [0.00, 0.00]
			validation	511.26 [490.89, 532.03]	22.61 [22.16, 23.07]	18.14 [17.78, 18.48]	0.00 [-0.00, 0.00]
			test	514.98 [493.45, 535.85]	22.69 [22.21, 23.15]	18.13 [17.77, 18.49]	0.00 [-0.00, 0.00]

Table 22: Results for predicting RR interval from different CLEF representations and input formats. Results are reported as point estimates with 95% confidence intervals indicated in brackets. The row with the lowest test MSE per CLEF input format is highlighted in grey, and the global best result is emphasized in bold.

Results: RR Interval prediction from various CLEF representations, using XGBoost

Input	Repr.	Run	Dataset	MSE	RMSE	MAE	R2
single	CLEF-S	regular	train	11754.98 [11277.86, 12322.34]	108.41 [106.20, 111.01]	75.43 [74.50, 76.48]	0.69 [0.68, 0.70]
			validation	12982.61 [11149.82, 15565.79]	113.84 [105.59, 124.76]	75.45 [73.26, 77.77]	0.65 [0.61, 0.68]
			test	12879.38 [11586.93, 14245.34]	113.45 [107.64, 119.35]	75.88 [73.66, 78.13]	0.66 [0.63, 0.68]
		mean	train	37947.86 [37110.69, 38868.00]	194.80 [192.64, 197.15]	154.50 [152.90, 155.93]	-0.00 [-0.00, -0.00]
			validation	37305.23 [35104.23, 40286.12]	193.12 [187.36, 200.71]	151.74 [148.80, 154.84]	-0.00 [-0.00, -0.00]
			test	37595.35 [35729.88, 39422.52]	193.88 [189.02, 198.55]	152.62 [149.41, 155.79]	-0.00 [-0.00, -0.00]
		shuffled	train	37889.61 [37040.38, 38807.15]	194.65 [192.46, 197.00]	154.03 [152.41, 155.46]	0.00 [-0.00, 0.00]
			validation	37132.23 [34942.12, 40083.94]	192.67 [186.93, 200.21]	151.18 [148.14, 154.27]	0.00 [0.00, 0.01]
			test	37412.69 [35601.73, 39223.87]	193.41 [188.68, 198.05]	152.04 [148.84, 155.27]	0.00 [-0.00, 0.01]
	CLEF-M	regular	train	16881.80 [16336.78, 17473.14]	129.93 [127.82, 132.19]	94.22 [93.12, 95.34]	0.56 [0.55, 0.56]
			validation	18146.52 [16250.51, 20736.83]	134.64 [127.48, 144.00]	94.18 [91.74, 96.68]	0.51 [0.48, 0.54]
			test	18112.86 [16808.50, 19547.97]	134.56 [129.65, 139.81]	96.02 [93.45, 98.63]	0.52 [0.49, 0.54]
		mean	train	37947.86 [37110.69, 38868.00]	194.80 [192.64, 197.15]	154.50 [152.90, 155.93]	-0.00 [-0.00, -0.00]
			validation	37305.23 [35104.23, 40286.12]	193.12 [187.36, 200.71]	151.74 [148.80, 154.84]	-0.00 [-0.00, -0.00]
			test	37595.35 [35729.88, 39422.52]	193.88 [189.02, 198.55]	152.62 [149.41, 155.79]	-0.00 [-0.00, -0.00]
		shuffled	train	37910.15 [37063.92, 38819.11]	194.70 [192.52, 197.03]	154.17 [152.53, 155.60]	0.00 [-0.00, 0.00]
			validation	37100.91 [34886.31, 40038.33]	192.59 [186.78, 200.10]	151.13 [148.18, 154.16]	0.00 [0.00, 0.01]
			test	37468.98 [35651.38, 39320.13]	193.55 [188.82, 198.29]	152.20 [149.02, 155.42]	0.00 [-0.00, 0.00]
	CLEF-L	regular	train	17207.54 [16649.16, 17791.83]	131.17 [129.03, 133.39]	95.53 [94.36, 96.63]	0.55 [0.54, 0.55]
			validation	18987.02 [16999.01, 21826.50]	137.72 [130.38, 147.74]	97.00 [94.34, 99.69]	0.49 [0.46, 0.52]
			test	18848.57 [17291.12, 20432.07]	137.26 [131.50, 142.94]	96.79 [94.14, 99.61]	0.50 [0.47, 0.52]
		mean	train	37947.86 [37110.69, 38868.00]	194.80 [192.64, 197.15]	154.50 [152.90, 155.93]	-0.00 [-0.00, -0.00]
			validation	37305.23 [35104.23, 40286.12]	193.12 [187.36, 200.71]	151.74 [148.80, 154.84]	-0.00 [-0.00, -0.00]
			test	37595.35 [35729.88, 39422.52]	193.88 [189.02, 198.55]	152.62 [149.41, 155.79]	-0.00 [-0.00, -0.00]
		shuffled	train	38032.53 [37175.42, 38928.86]	195.02 [192.81, 197.30]	154.53 [152.92, 155.94]	-0.00 [-0.00, -0.00]
			validation	37297.23 [35079.20, 40281.80]	193.09 [187.29, 200.70]	151.63 [148.68, 154.77]	-0.00 [-0.00, 0.00]
			test	37571.08 [35735.11, 39422.39]	193.82 [189.04, 198.55]	152.53 [149.33, 155.76]	-0.00 [-0.00, 0.00]
double	CLEF-S	regular	train	11263.93 [10762.93, 11837.48]	106.12 [103.74, 108.80]	73.36 [72.39, 74.36]	0.70 [0.69, 0.71]
			validation	12450.85 [10605.64, 15136.48]	111.47 [102.98, 123.03]	73.73 [71.66, 75.99]	0.67 [0.62, 0.70]
			test	12228.42 [10920.24, 13540.51]	110.54 [104.50, 116.36]	74.01 [71.67, 76.15]	0.67 [0.65, 0.70]
		mean	train	37947.86 [37110.69, 38868.00]	194.80 [192.64, 197.15]	154.50 [152.90, 155.93]	-0.00 [-0.00, -0.00]
			validation	37305.23 [35104.23, 40286.12]	193.12 [187.36, 200.71]	151.74 [148.80, 154.84]	-0.00 [-0.00, -0.00]
			test	37595.35 [35729.88, 39422.52]	193.88 [189.02, 198.55]	152.62 [149.41, 155.79]	-0.00 [-0.00, -0.00]
		shuffled	train	37863.45 [37026.43, 38786.76]	194.58 [192.42, 196.94]	153.97 [152.38, 155.42]	0.00 [0.00, 0.00]
			validation	37144.91 [34937.12, 40117.46]	192.70 [186.91, 200.29]	151.11 [148.14, 154.18]	0.00 [0.00, 0.01]
			test	37449.48 [35620.92, 39260.85]	193.50 [188.74, 198.14]	152.02 [148.77, 155.23]	0.00 [-0.00, 0.00]
	CLEF-M	regular	train	19261.58 [18671.55, 19890.05]	138.78 [136.64, 141.03]	102.73 [101.54, 103.87]	0.49 [0.49, 0.50]
			validation	20531.18 [18485.73, 23321.53]	143.23 [135.96, 152.71]	102.87 [100.30, 105.53]	0.45 [0.42, 0.48]
			test	20477.63 [19143.01, 21865.69]	143.08 [138.36, 147.87]	104.73 [102.04, 107.37]	0.45 [0.43, 0.47]
		mean	train	37947.86 [37110.69, 38868.00]	194.80 [192.64, 197.15]	154.50 [152.90, 155.93]	-0.00 [-0.00, -0.00]
			validation	37305.23 [35104.23, 40286.12]	193.12 [187.36, 200.71]	151.74 [148.80, 154.84]	-0.00 [-0.00, -0.00]
			test	37595.35 [35729.88, 39422.52]	193.88 [189.02, 198.55]	152.62 [149.41, 155.79]	-0.00 [-0.00, -0.00]
		shuffled	train	37923.50 [37090.74, 38839.72]	194.74 [192.59, 197.08]	154.23 [152.61, 155.66]	0.00 [-0.00, 0.00]
			validation	37194.50 [34998.24, 40166.23]	192.83 [187.08, 200.42]	151.41 [148.49, 154.48]	0.00 [-0.00, 0.00]
			test	37564.63 [35742.45, 39414.57]	193.80 [189.06, 198.53]	152.46 [149.28, 155.64]	-0.00 [-0.00, 0.00]
	CLEF-L	regular	train	16826.67 [16311.48, 17386.54]	129.71 [127.72, 131.86]	94.48 [93.34, 95.60]	0.56 [0.55, 0.56]
			validation	18455.08 [16639.01, 20952.15]	135.79 [128.99, 144.75]	95.85 [93.31, 98.54]	0.51 [0.47, 0.53]
			test	18359.88 [16801.91, 19915.81]	135.47 [129.62, 141.12]	95.75 [93.17, 98.43]	0.51 [0.49, 0.53]
		mean	train	37947.86 [37110.69, 38868.00]	194.80 [192.64, 197.15]	154.50 [152.90, 155.93]	-0.00 [-0.00, -0.00]
			validation	37305.23 [35104.23, 40286.12]	193.12 [187.36, 200.71]	151.74 [148.80, 154.84]	-0.00 [-0.00, -0.00]
			test	37595.35 [35729.88, 39422.52]	193.88 [189.02, 198.55]	152.62 [149.41, 155.79]	-0.00 [-0.00, -0.00]
		shuffled	train	38024.68 [37181.57, 38927.04]	195.00 [192.83, 197.30]	154.52 [152.92, 155.95]	-0.00 [-0.00, -0.00]
			validation	37233.63 [35021.29, 40203.71]	192.93 [187.14, 200.51]	151.51 [148.62, 154.66]	0.00 [-0.00, 0.00]
			test	37541.51 [35700.81, 39378.75]	193.74 [188.95, 198.44]	152.50 [149.27, 155.68]	-0.00 [-0.00, 0.00]
interp.	CLEF-S	regular	train	16755.76 [16212.65, 17381.55]	129.44 [127.33, 131.84]	93.08 [91.93, 94.20]	0.56 [0.55, 0.57]
			validation	17979.97 [16047.35, 20712.13]	134.02 [126.68, 143.92]	92.83 [90.32, 95.45]	0.52 [0.48, 0.55]
			test	17643.97 [16370.98, 18983.57]	132.81 [127.95, 137.78]	93.82 [91.21, 96.54]	0.53 [0.51, 0.55]
		mean	train	37947.86 [37110.69, 38868.00]	194.80 [192.64, 197.15]	154.50 [152.90, 155.93]	-0.00 [-0.00, -0.00]
			validation	37305.23 [35104.23, 40286.12]	193.12 [187.36, 200.71]	151.74 [148.80, 154.84]	-0.00 [-0.00, -0.00]
			test	37595.35 [35729.88, 39422.52]	193.88 [189.02, 198.55]	152.62 [149.41, 155.79]	-0.00 [-0.00, -0.00]
		shuffled	train	37843.37 [36998.17, 38747.24]	194.53 [192.35, 196.84]	153.95 [152.36, 155.38]	0.00 [0.00, 0.00]
			validation	37165.82 [34992.01, 40126.34]	192.75 [187.06, 200.32]	151.14 [148.16, 154.29]	0.00 [0.00, 0.01]
			test	37476.11 [35650.71, 39293.03]	193.57 [188.81, 198.22]	152.07 [148.87, 155.28]	0.00 [-0.00, 0.00]
	CLEF-M	regular	train	10275.35 [9765.65, 10861.93]	101.36 [98.82, 104.22]	66.41 [65.46, 67.36]	0.73 [0.72, 0.74]
			validation	12020.36 [10101.62, 14787.99]	109.50 [100.51, 121.61]	67.45 [65.29, 69.80]	0.68 [0.63, 0.71]
			test	**11809.72 [10393.74, 13321.43]**	**108.62 [101.95, 115.42]**	**68.08 [65.83, 70.35]**	**0.69 [0.66, 0.71]**
		mean	train	37947.86 [37110.69, 38868.00]	194.80 [192.64, 197.15]	154.50 [152.90, 155.93]	-0.00 [-0.00, -0.00]
			validation	37305.23 [35104.23, 40286.12]	193.12 [187.36, 200.71]	151.74 [148.80, 154.84]	-0.00 [-0.00, -0.00]
			test	37595.35 [35729.88, 39422.52]	193.88 [189.02, 198.55]	152.62 [149.41, 155.79]	-0.00 [-0.00, -0.00]
		shuffled	train	37818.22 [36976.17, 38731.63]	194.47 [192.29, 196.80]	153.89 [152.26, 155.32]	0.00 [0.00, 0.00]
			validation	37121.53 [34928.67, 40098.78]	192.64 [186.89, 200.25]	151.04 [148.04, 154.14]	0.00 [0.00, 0.01]
			test	37464.93 [35638.71, 39317.03]	193.54 [188.78, 198.29]	152.13 [148.91, 155.33]	0.00 [-0.00, 0.00]
	CLEF-L	regular	train	10484.80 [10018.51, 11050.72]	102.39 [100.09, 105.12]	69.75 [68.86, 70.69]	0.72 [0.71, 0.73]
			validation	11889.04 [10091.32, 14447.75]	108.92 [100.46, 120.20]	70.30 [68.24, 72.55]	0.68 [0.64, 0.71]
			test	12013.39 [10639.24, 13513.92]	109.55 [103.15, 116.25]	71.10 [68.81, 73.35]	0.68 [0.65, 0.71]
		mean	train	37947.86 [37110.69, 38868.00]	194.80 [192.64, 197.15]	154.50 [152.90, 155.93]	-0.00 [-0.00, -0.00]
			validation	37305.23 [35104.23, 40286.12]	193.12 [187.36, 200.71]	151.74 [148.80, 154.84]	-0.00 [-0.00, -0.00]
			test	37595.35 [35729.88, 39422.52]	193.88 [189.02, 198.55]	152.62 [149.41, 155.79]	-0.00 [-0.00, -0.00]
		shuffled	train	37728.10 [36896.85, 38644.87]	194.23 [192.09, 196.58]	154.02 [152.40, 155.43]	0.01 [0.00, 0.01]
			validation	36993.60 [34800.46, 39969.84]	192.31 [186.55, 199.92]	151.16 [148.18, 154.26]	0.01 [0.00, 0.01]
			test	37319.89 [35478.16, 39162.57]	193.17 [188.36, 197.90]	152.07 [148.84, 155.24]	0.01 [0.00, 0.01]

Table 23: Results for predicting QT interval from different CLEF representations and input formats. Results are reported as point estimates with 95% confidence intervals indicated in brackets. The row with the lowest test MSE per CLEF input format is highlighted in grey, and the global best result is emphasized in bold.

Results: QT interval prediction from various CLEF representations, using XGBoost

Input	Repr.	Run	Dataset	MSE	RMSE	MAE	R2
single	CLEF-S	regular	train	4636.14 [4544.34, 4724.69]	68.09 [67.41, 68.74]	53.22 [52.64, 53.77]	0.43 [0.42, 0.43]
			validation	4591.69 [4373.42, 4808.49]	67.76 [66.13, 69.34]	52.09 [50.87, 53.25]	0.38 [0.36, 0.39]
			test	4561.50 [4354.49, 4761.98]	67.53 [65.99, 69.01]	52.11 [50.82, 53.33]	0.37 [0.35, 0.38]
		mean	train	8096.02 [7953.28, 8229.29]	89.98 [89.18, 90.72]	72.31 [71.59, 72.97]	-0.00 [-0.00, -0.00]
			validation	7394.48 [7089.35, 7690.62]	85.99 [84.20, 87.70]	68.59 [67.15, 70.03]	-0.00 [-0.01, -0.00]
			test	7262.67 [6983.14, 7542.36]	85.22 [83.57, 86.85]	68.19 [66.77, 69.61]	-0.00 [-0.01, -0.00]
		shuffled	train	8163.14 [8018.90, 8296.77]	90.35 [89.55, 91.09]	72.63 [71.92, 73.31]	-0.01 [-0.01, -0.01]
			validation	7446.88 [7137.17, 7748.32]	86.29 [84.48, 88.02]	68.89 [67.45, 70.34]	-0.01 [-0.02, -0.01]
			test	7314.91 [7028.93, 7599.57]	85.52 [83.84, 87.18]	68.47 [67.08, 69.92]	-0.01 [-0.02, -0.01]
	CLEF-M	regular	train	5369.38 [5267.74, 5474.83]	73.28 [72.58, 73.99]	57.30 [56.72, 57.89]	0.34 [0.33, 0.34]
			validation	5249.61 [5001.97, 5492.14]	72.45 [70.72, 74.11]	55.97 [54.64, 57.20]	0.29 [0.27, 0.30]
			test	5192.48 [4949.68, 5431.31]	72.05 [70.35, 73.70]	55.63 [54.28, 56.91]	0.28 [0.27, 0.30]
		mean	train	8096.02 [7953.28, 8229.29]	89.98 [89.18, 90.72]	72.31 [71.59, 72.97]	-0.00 [-0.00, -0.00]
			validation	7394.48 [7089.35, 7690.62]	85.99 [84.20, 87.70]	68.59 [67.15, 70.03]	-0.00 [-0.01, -0.00]
			test	7262.67 [6983.14, 7542.36]	85.22 [83.57, 86.85]	68.19 [66.77, 69.61]	-0.00 [-0.01, -0.00]
		shuffled	train	8117.45 [7975.40, 8251.55]	90.10 [89.31, 90.84]	72.35 [71.64, 73.03]	-0.00 [-0.00, -0.00]
			validation	7402.83 [7098.88, 7702.31]	86.04 [84.25, 87.76]	68.57 [67.14, 70.02]	-0.00 [-0.01, -0.00]
			test	7267.15 [6986.29, 7551.99]	85.24 [83.58, 86.90]	68.16 [66.75, 69.60]	-0.01 [-0.01, -0.00]
	CLEF-L	regular	train	5505.11 [5402.21, 5609.10]	74.20 [73.50, 74.89]	58.32 [57.71, 58.89]	0.32 [0.31, 0.33]
			validation	5548.00 [5295.56, 5787.95]	74.48 [72.77, 76.08]	57.81 [56.42, 59.11]	0.25 [0.23, 0.26]
			test	5499.22 [5252.38, 5737.81]	74.15 [72.47, 75.75]	57.82 [56.51, 59.12]	0.24 [0.22, 0.25]
		mean	train	8096.02 [7953.28, 8229.29]	89.98 [89.18, 90.72]	72.31 [71.59, 72.97]	-0.00 [-0.00, -0.00]
			validation	7394.48 [7089.35, 7690.62]	85.99 [84.20, 87.70]	68.59 [67.15, 70.03]	-0.00 [-0.01, -0.00]
			test	7262.67 [6983.14, 7542.36]	85.22 [83.57, 86.85]	68.19 [66.77, 69.61]	-0.00 [-0.01, -0.00]
		shuffled	train	8130.16 [7986.60, 8264.84]	90.17 [89.37, 90.91]	72.47 [71.76, 73.16]	-0.00 [-0.01, -0.00]
			validation	7413.64 [7106.06, 7712.37]	86.10 [84.30, 87.82]	68.70 [67.22, 70.16]	-0.01 [-0.01, -0.00]
			test	7293.34 [7011.06, 7572.94]	85.40 [83.73, 87.02]	68.32 [66.93, 69.73]	-0.01 [-0.01, -0.01]
double	CLEF-S	regular	train	4557.60 [4469.82, 4647.13]	67.51 [66.86, 68.17]	52.72 [52.17, 53.26]	0.44 [0.43, 0.44]
			validation	4471.65 [4248.85, 4683.91]	66.87 [65.18, 68.44]	51.42 [50.15, 52.61]	0.39 [0.38, 0.41]
			test	**4452.80 [4246.82, 4656.34]**	**66.72 [65.17, 68.24]**	**51.35 [50.15, 52.53]**	**0.38 [0.37, 0.40]**
		mean	train	8096.02 [7953.28, 8229.29]	89.98 [89.18, 90.72]	72.31 [71.59, 72.97]	-0.00 [-0.00, -0.00]
			validation	7394.48 [7089.35, 7690.62]	85.99 [84.20, 87.70]	68.59 [67.15, 70.03]	-0.00 [-0.01, -0.00]
			test	7262.67 [6983.14, 7542.36]	85.22 [83.57, 86.85]	68.19 [66.77, 69.61]	-0.00 [-0.01, -0.00]
		shuffled	train	8147.32 [8003.39, 8281.11]	90.26 [89.46, 91.00]	72.55 [71.84, 73.23]	-0.01 [-0.01, -0.01]
			validation	7424.14 [7116.89, 7723.74]	86.16 [84.36, 87.88]	68.77 [67.32, 70.21]	-0.01 [-0.01, -0.00]
			test	7300.80 [7017.35, 7583.67]	85.44 [83.77, 87.08]	68.35 [66.96, 69.80]	-0.01 [-0.01, -0.01]
	CLEF-M	regular	train	5708.09 [5602.58, 5808.09]	75.55 [74.85, 76.21]	59.63 [59.00, 60.20]	0.29 [0.29, 0.30]
			validation	5608.99 [5342.64, 5867.36]	74.89 [73.09, 76.60]	58.26 [56.85, 59.57]	0.24 [0.23, 0.25]
			test	5519.48 [5261.21, 5744.00]	74.29 [72.53, 75.79]	57.93 [56.59, 59.28]	0.24 [0.22, 0.25]
		mean	train	8096.02 [7953.28, 8229.29]	89.98 [89.18, 90.72]	72.31 [71.59, 72.97]	-0.00 [-0.00, -0.00]
			validation	7394.48 [7089.35, 7690.62]	85.99 [84.20, 87.70]	68.59 [67.15, 70.03]	-0.00 [-0.01, -0.00]
			test	7262.67 [6983.14, 7542.36]	85.22 [83.57, 86.85]	68.19 [66.77, 69.61]	-0.00 [-0.01, -0.00]
		shuffled	train	8107.85 [7965.89, 8242.30]	90.04 [89.25, 90.79]	72.35 [71.65, 73.02]	-0.00 [-0.00, -0.00]
			validation	7394.54 [7087.60, 7692.43]	85.99 [84.19, 87.71]	68.55 [67.11, 70.00]	-0.00 [-0.01, -0.00]
			test	7266.70 [6989.00, 7546.56]	85.24 [83.60, 86.87]	68.18 [66.76, 69.61]	-0.01 [-0.01, -0.00]
	CLEF-L	regular	train	5432.67 [5331.17, 5542.79]	73.71 [73.01, 74.45]	57.80 [57.21, 58.39]	0.33 [0.32, 0.34]
			validation	5485.43 [5228.74, 5741.56]	74.06 [72.31, 75.77]	57.36 [55.99, 58.68]	0.26 [0.24, 0.27]
			test	5434.59 [5192.90, 5677.69]	73.72 [72.06, 75.35]	57.33 [56.02, 58.69]	0.25 [0.23, 0.26]
		mean	train	8096.02 [7953.28, 8229.29]	89.98 [89.18, 90.72]	72.31 [71.59, 72.97]	-0.00 [-0.00, -0.00]
			validation	7394.48 [7089.35, 7690.62]	85.99 [84.20, 87.70]	68.59 [67.15, 70.03]	-0.00 [-0.01, -0.00]
			test	7262.67 [6983.14, 7542.36]	85.22 [83.57, 86.85]	68.19 [66.77, 69.61]	-0.00 [-0.01, -0.00]
		shuffled	train	8121.22 [7978.59, 8255.35]	90.12 [89.32, 90.86]	72.42 [71.72, 73.10]	-0.00 [-0.00, -0.00]
			validation	7412.85 [7100.31, 7701.96]	86.09 [84.26, 87.76]	68.67 [67.22, 70.12]	-0.01 [-0.01, -0.00]
			test	7280.80 [6993.60, 7565.23]	85.32 [83.63, 86.98]	68.28 [66.83, 69.73]	-0.01 [-0.01, -0.00]
interp.	CLEF-S	regular	train	4836.52 [4737.97, 4933.23]	69.54 [68.83, 70.24]	53.86 [53.26, 54.40]	0.40 [0.40, 0.41]
			validation	4647.06 [4413.48, 4887.84]	68.16 [66.43, 69.91]	51.96 [50.62, 53.20]	0.37 [0.35, 0.38]
			test	4721.45 [4514.17, 4935.35]	68.71 [67.19, 70.25]	52.78 [51.54, 54.00]	0.35 [0.33, 0.36]
		mean	train	8096.02 [7953.28, 8229.29]	89.98 [89.18, 90.72]	72.31 [71.59, 72.97]	-0.00 [-0.00, -0.00]
			validation	7394.48 [7089.35, 7690.62]	85.99 [84.20, 87.70]	68.59 [67.15, 70.03]	-0.00 [-0.01, -0.00]
			test	7262.67 [6983.14, 7542.36]	85.22 [83.57, 86.85]	68.19 [66.77, 69.61]	-0.00 [-0.01, -0.00]
		shuffled	train	8162.06 [8017.65, 8296.81]	90.34 [89.54, 91.09]	72.62 [71.92, 73.31]	-0.01 [-0.01, -0.01]
			validation	7439.97 [7135.69, 7742.29]	86.25 [84.47, 87.99]	68.85 [67.41, 70.29]	-0.01 [-0.01, -0.01]
			test	7306.23 [7023.63, 7586.74]	85.47 [83.81, 87.10]	68.39 [66.96, 69.81]	-0.01 [-0.02, -0.01]
	CLEF-M	regular	train	4621.69 [4530.52, 4709.59]	67.98 [67.31, 68.63]	53.28 [52.73, 53.80]	0.43 [0.42, 0.44]
			validation	4515.23 [4289.45, 4739.97]	67.19 [65.49, 68.85]	51.86 [50.57, 52.99]	0.39 [0.37, 0.40]
			test	4645.99 [4430.39, 4849.80]	68.16 [66.56, 69.64]	52.66 [51.46, 53.87]	0.36 [0.34, 0.37]
		mean	train	8096.02 [7953.28, 8229.29]	89.98 [89.18, 90.72]	72.31 [71.59, 72.97]	-0.00 [-0.00, -0.00]
			validation	7394.48 [7089.35, 7690.62]	85.99 [84.20, 87.70]	68.59 [67.15, 70.03]	-0.00 [-0.01, -0.00]
			test	7262.67 [6983.14, 7542.36]	85.22 [83.57, 86.85]	68.19 [66.77, 69.61]	-0.00 [-0.01, -0.00]
		shuffled	train	8109.58 [7969.17, 8242.75]	90.05 [89.27, 90.79]	72.42 [71.73, 73.10]	-0.00 [-0.00, -0.00]
			validation	7412.22 [7109.22, 7713.73]	86.09 [84.32, 87.83]	68.79 [67.37, 70.24]	-0.01 [-0.01, -0.00]
			test	7284.64 [7002.49, 7567.94]	85.35 [83.68, 86.99]	68.37 [66.93, 69.79]	-0.01 [-0.01, -0.00]
	CLEF-L	regular	train	4594.71 [4502.71, 4678.53]	67.78 [67.10, 68.40]	53.43 [52.88, 53.94]	0.43 [0.43, 0.44]
			validation	4587.67 [4352.52, 4828.90]	67.73 [65.97, 69.49]	52.35 [51.12, 53.49]	0.38 [0.36, 0.39]
			test	4606.23 [4391.30, 4819.53]	67.86 [66.27, 69.42]	52.52 [51.34, 53.70]	0.36 [0.35, 0.38]
		mean	train	8096.02 [7953.28, 8229.29]	89.98 [89.18, 90.72]	72.31 [71.59, 72.97]	-0.00 [-0.00, -0.00]
			validation	7394.48 [7089.35, 7690.62]	85.99 [84.20, 87.70]	68.59 [67.15, 70.03]	-0.00 [-0.01, -0.00]
			test	7262.67 [6983.14, 7542.36]	85.22 [83.57, 86.85]	68.19 [66.77, 69.61]	-0.00 [-0.01, -0.00]
		shuffled	train	8116.76 [7975.32, 8253.42]	90.09 [89.30, 90.85]	72.40 [71.71, 73.09]	-0.00 [-0.00, -0.00]
			validation	7409.11 [7098.35, 7706.46]	86.07 [84.25, 87.79]	68.66 [67.23, 70.09]	-0.01 [-0.01, -0.00]
			test	7269.27 [6988.08, 7553.80]	85.26 [83.59, 86.91]	68.24 [66.81, 69.66]	-0.01 [-0.01, -0.00]

Table 24: Results for predicting PR interval from different CLEF representations and input formats. Results are reported as point estimates with 95% confidence intervals indicated in brackets. The row with the lowest test MSE per CLEF input format is highlighted in grey, and the global best result is emphasized in bold.

Results: PR interval prediction from various CLEF representations, using XGBoost

Input	Repr.	Run	Dataset	MSE	RMSE	MAE	R2
single	CLEF-S	regular	train	1297.24 [1258.83, 1337.54]	36.02 [35.48, 36.57]	26.81 [26.50, 27.11]	0.47 [0.46, 0.47]
			validation	1266.95 [1192.21, 1350.15]	35.59 [34.53, 36.74]	26.78 [26.17, 27.44]	0.43 [0.41, 0.45]
			test	1435.05 [1347.73, 1525.74]	37.88 [36.71, 39.06]	28.00 [27.31, 28.72]	0.42 [0.41, 0.44]
		mean	train	2432.46 [2362.76, 2500.36]	49.32 [48.61, 50.00]	36.98 [36.55, 37.38]	-0.00 [-0.00, -0.00]
			validation	2231.44 [2097.89, 2369.69]	47.23 [45.80, 48.68]	35.97 [35.12, 36.82]	-0.00 [-0.01, -0.00]
			test	2489.17 [2339.29, 2649.43]	49.89 [48.37, 51.47]	37.55 [36.65, 38.47]	-0.00 [-0.00, -0.00]
		shuffled	train	2423.00 [2353.25, 2491.59]	49.22 [48.51, 49.92]	36.87 [36.44, 37.28]	0.00 [0.00, 0.01]
			validation	2219.00 [2086.57, 2357.92]	47.10 [45.68, 48.56]	35.84 [34.99, 36.70]	0.00 [-0.00, 0.01]
			test	2477.99 [2330.25, 2636.21]	49.77 [48.27, 51.34]	37.45 [36.57, 38.38]	0.00 [0.00, 0.01]
	CLEF-M	regular	train	1541.72 [1499.62, 1586.27]	39.26 [38.72, 39.83]	29.28 [28.94, 29.62]	0.37 [0.36, 0.37]
			validation	1554.28 [1462.70, 1654.07]	39.42 [38.25, 40.67]	29.51 [28.80, 30.20]	0.30 [0.28, 0.32]
			test	1725.34 [1619.18, 1834.31]	41.53 [40.24, 42.83]	30.70 [29.92, 31.50]	0.31 [0.29, 0.32]
		mean	train	2432.46 [2362.76, 2500.36]	49.32 [48.61, 50.00]	36.98 [36.55, 37.38]	-0.00 [-0.00, -0.00]
			validation	2231.44 [2097.89, 2369.69]	47.23 [45.80, 48.68]	35.97 [35.12, 36.82]	-0.00 [-0.01, -0.00]
			test	2489.17 [2339.29, 2649.43]	49.89 [48.37, 51.47]	37.55 [36.65, 38.47]	-0.00 [-0.00, -0.00]
		shuffled	train	2439.48 [2370.59, 2507.50]	49.39 [48.69, 50.07]	37.02 [36.59, 37.43]	-0.00 [-0.00, -0.00]
			validation	2234.78 [2103.13, 2372.31]	47.27 [45.86, 48.71]	36.00 [35.14, 36.88]	-0.00 [-0.01, -0.00]
			test	2490.57 [2342.87, 2651.54]	49.90 [48.40, 51.49]	37.57 [36.68, 38.49]	-0.00 [-0.00, 0.00]
	CLEF-L	regular	train	1824.16 [1773.19, 1875.66]	42.71 [42.11, 43.31]	32.24 [31.87, 32.60]	0.25 [0.24, 0.26]
			validation	1887.71 [1772.18, 2005.95]	43.44 [42.10, 44.79]	32.91 [32.12, 33.74]	0.15 [0.14, 0.17]
			test	2116.03 [1989.22, 2252.70]	45.99 [44.60, 47.46]	34.53 [33.72, 35.39]	0.15 [0.14, 0.16]
		mean	train	2432.46 [2362.76, 2500.36]	49.32 [48.61, 50.00]	36.98 [36.55, 37.38]	-0.00 [-0.00, -0.00]
			validation	2231.44 [2097.89, 2369.69]	47.23 [45.80, 48.68]	35.97 [35.12, 36.82]	-0.00 [-0.01, -0.00]
			test	2489.17 [2339.29, 2649.43]	49.89 [48.37, 51.47]	37.55 [36.65, 38.47]	-0.00 [-0.00, -0.00]
		shuffled	train	2436.82 [2366.82, 2504.70]	49.36 [48.65, 50.05]	36.99 [36.56, 37.39]	-0.00 [-0.00, -0.00]
			validation	2232.15 [2098.97, 2371.73]	47.24 [45.81, 48.70]	35.97 [35.14, 36.83]	-0.00 [-0.01, -0.00]
			test	2488.99 [2337.31, 2649.83]	49.88 [48.35, 51.48]	37.56 [36.67, 38.49]	-0.00 [-0.00, 0.00]
double	CLEF-S	regular	train	1268.30 [1231.49, 1305.82]	35.61 [35.09, 36.14]	26.47 [26.15, 26.77]	0.48 [0.47, 0.49]
			validation	1231.12 [1157.50, 1311.16]	35.08 [34.02, 36.21]	26.50 [25.88, 27.14]	0.45 [0.43, 0.46]
			test	1394.61 [1311.84, 1482.22]	37.34 [36.22, 38.50]	27.72 [27.08, 28.41]	0.44 [0.42, 0.46]
		mean	train	2432.46 [2362.76, 2500.36]	49.32 [48.61, 50.00]	36.98 [36.55, 37.38]	-0.00 [-0.00, -0.00]
			validation	2231.44 [2097.89, 2369.69]	47.23 [45.80, 48.68]	35.97 [35.12, 36.82]	-0.00 [-0.01, -0.00]
			test	2489.17 [2339.29, 2649.43]	49.89 [48.37, 51.47]	37.55 [36.65, 38.47]	-0.00 [-0.00, -0.00]
		shuffled	train	2420.05 [2350.41, 2488.03]	49.19 [48.48, 49.88]	36.85 [36.43, 37.25]	0.01 [0.00, 0.01]
			validation	2218.40 [2087.16, 2358.50]	47.09 [45.69, 48.56]	35.84 [35.00, 36.70]	0.00 [-0.00, 0.01]
			test	2478.68 [2330.92, 2637.25]	49.78 [48.28, 51.35]	37.43 [36.54, 38.36]	0.00 [0.00, 0.01]
	CLEF-M	regular	train	1668.98 [1622.43, 1717.47]	40.85 [40.28, 41.44]	30.58 [30.23, 30.93]	0.31 [0.31, 0.32]
			validation	1661.67 [1559.82, 1771.65]	40.76 [39.49, 42.09]	30.65 [29.92, 31.39]	0.25 [0.24, 0.27]
			test	1859.60 [1742.81, 1974.78]	43.12 [41.75, 44.44]	32.09 [31.26, 32.90]	0.25 [0.24, 0.27]
		mean	train	2432.46 [2362.76, 2500.36]	49.32 [48.61, 50.00]	36.98 [36.55, 37.38]	-0.00 [-0.00, -0.00]
			validation	2231.44 [2097.89, 2369.69]	47.23 [45.80, 48.68]	35.97 [35.12, 36.82]	-0.00 [-0.01, -0.00]
			test	2489.17 [2339.29, 2649.43]	49.89 [48.37, 51.47]	37.55 [36.65, 38.47]	-0.00 [-0.00, -0.00]
		shuffled	train	2433.69 [2363.87, 2501.27]	49.33 [48.62, 50.01]	36.97 [36.55, 37.37]	-0.00 [-0.00, 0.00]
			validation	2230.22 [2098.35, 2368.51]	47.22 [45.81, 48.67]	35.97 [35.12, 36.82]	-0.00 [-0.01, 0.00]
			test	2488.86 [2339.51, 2648.77]	49.88 [48.37, 51.47]	37.53 [36.63, 38.46]	-0.00 [-0.00, 0.00]
	CLEF-L	regular	train	1816.34 [1768.20, 1867.09]	42.62 [42.05, 43.21]	32.16 [31.80, 32.53]	0.25 [0.25, 0.26]
			validation	1875.97 [1767.11, 1991.68]	43.31 [42.04, 44.63]	32.77 [32.00, 33.57]	0.16 [0.14, 0.17]
			test	2116.02 [1991.22, 2250.23]	45.99 [44.62, 47.44]	34.36 [33.54, 35.25]	0.15 [0.14, 0.16]
		mean	train	2432.46 [2362.76, 2500.36]	49.32 [48.61, 50.00]	36.98 [36.55, 37.38]	-0.00 [-0.00, -0.00]
			validation	2231.44 [2097.89, 2369.69]	47.23 [45.80, 48.68]	35.97 [35.12, 36.82]	-0.00 [-0.01, -0.00]
			test	2489.17 [2339.29, 2649.43]	49.89 [48.37, 51.47]	37.55 [36.65, 38.47]	-0.00 [-0.00, -0.00]
		shuffled	train	2437.25 [2367.73, 2504.73]	49.37 [48.66, 50.05]	36.99 [36.56, 37.41]	-0.00 [-0.00, -0.00]
			validation	2233.09 [2099.18, 2370.80]	47.25 [45.82, 48.69]	35.95 [35.10, 36.79]	-0.00 [-0.01, -0.00]
			test	2490.30 [2341.68, 2648.47]	49.90 [48.39, 51.46]	37.55 [36.66, 38.49]	-0.00 [-0.00, 0.00]
interp.	CLEF-S	regular	train	1179.82 [1143.30, 1217.88]	34.35 [33.81, 34.90]	25.14 [24.83, 25.44]	0.51 [0.51, 0.52]
			validation	1153.28 [1086.24, 1226.03]	33.96 [32.96, 35.01]	25.26 [24.66, 25.86]	0.48 [0.46, 0.50]
			test	**1295.50 [1217.28, 1380.75]**	**35.99 [34.89, 37.16]**	**26.29 [25.65, 26.96]**	**0.48 [0.46, 0.50]**
		mean	train	2432.46 [2362.76, 2500.36]	49.32 [48.61, 50.00]	36.98 [36.55, 37.38]	-0.00 [-0.00, -0.00]
			validation	2231.44 [2097.89, 2369.69]	47.23 [45.80, 48.68]	35.97 [35.12, 36.82]	-0.00 [-0.01, -0.00]
			test	2489.17 [2339.29, 2649.43]	49.89 [48.37, 51.47]	37.55 [36.65, 38.47]	-0.00 [-0.00, -0.00]
		shuffled	train	2418.40 [2349.68, 2486.26]	49.18 [48.47, 49.86]	36.84 [36.41, 37.24]	0.01 [0.00, 0.01]
			validation	2218.60 [2087.96, 2355.21]	47.10 [45.69, 48.53]	35.86 [35.04, 36.73]	0.00 [-0.00, 0.01]
			test	2479.16 [2330.70, 2636.99]	49.78 [48.28, 51.35]	37.45 [36.56, 38.37]	0.00 [0.00, 0.01]
	CLEF-M	regular	train	1491.38 [1451.40, 1533.97]	38.62 [38.10, 39.17]	28.92 [28.59, 29.25]	0.39 [0.38, 0.39]
			validation	1493.97 [1405.13, 1589.09]	38.65 [37.49, 39.86]	29.18 [28.49, 29.89]	0.33 [0.31, 0.34]
			test	1690.63 [1581.78, 1803.43]	41.11 [39.77, 42.47]	30.40 [29.64, 31.19]	0.32 [0.30, 0.34]
		mean	train	2432.46 [2362.76, 2500.36]	49.32 [48.61, 50.00]	36.98 [36.55, 37.38]	-0.00 [-0.00, -0.00]
			validation	2231.44 [2097.89, 2369.69]	47.23 [45.80, 48.68]	35.97 [35.12, 36.82]	-0.00 [-0.01, -0.00]
			test	2489.17 [2339.29, 2649.43]	49.89 [48.37, 51.47]	37.55 [36.65, 38.47]	-0.00 [-0.00, -0.00]
		shuffled	train	2428.28 [2359.15, 2496.98]	49.28 [48.57, 49.97]	36.91 [36.48, 37.31]	0.00 [0.00, 0.00]
			validation	2226.95 [2094.70, 2363.12]	47.19 [45.77, 48.61]	35.93 [35.09, 36.79]	-0.00 [-0.00, 0.00]
			test	2483.67 [2336.05, 2642.77]	49.83 [48.33, 51.41]	37.52 [36.62, 38.44]	0.00 [-0.00, 0.00]
	CLEF-L	regular	train	1617.50 [1573.40, 1660.81]	40.22 [39.67, 40.75]	30.17 [29.83, 30.52]	0.33 [0.33, 0.34]
			validation	1652.48 [1556.73, 1756.15]	40.65 [39.46, 41.91]	30.69 [29.95, 31.43]	0.26 [0.24, 0.27]
			test	1885.60 [1770.12, 2010.51]	43.42 [42.07, 44.84]	32.09 [31.30, 32.92]	0.24 [0.23, 0.26]
		mean	train	2432.46 [2362.76, 2500.36]	49.32 [48.61, 50.00]	36.98 [36.55, 37.38]	-0.00 [-0.00, -0.00]
			validation	2231.44 [2097.89, 2369.69]	47.23 [45.80, 48.68]	35.97 [35.12, 36.82]	-0.00 [-0.01, -0.00]
			test	2489.17 [2339.29, 2649.43]	49.89 [48.37, 51.47]	37.55 [36.65, 38.47]	-0.00 [-0.00, -0.00]
		shuffled	train	2432.21 [2362.28, 2500.68]	49.32 [48.60, 50.01]	36.94 [36.51, 37.36]	0.00 [-0.00, 0.00]
			validation	2228.92 [2097.21, 2367.09]	47.21 [45.80, 48.65]	35.96 [35.11, 36.81]	-0.00 [-0.01, 0.00]
			test	2490.76 [2342.69, 2650.56]	49.90 [48.40, 51.48]	37.57 [36.67, 38.49]	-0.00 [-0.00, 0.00]

Table 25: Results for predicting T peak amplitude from different CLEF representations and input formats. Results are reported as point estimates with 95% confidence intervals indicated in brackets. The row with the lowest test MSE per CLEF input format is highlighted in grey.

Results: T peak amplitude prediction from various CLEF representations, using XGBoost

Input	Repr.	Run	Dataset	MSE	RMSE	MAE	R2
single	CLEF-S	regular	train	0.00 [0.00, 0.01]	0.07 [0.07, 0.07]	0.05 [0.05, 0.05]	0.54 [0.53, 0.55]
			validation	0.01 [0.00, 0.01]	0.07 [0.07, 0.08]	0.05 [0.05, 0.05]	0.48 [0.44, 0.51]
			test	0.01 [0.00, 0.01]	0.07 [0.07, 0.08]	0.05 [0.05, 0.05]	0.47 [0.41, 0.52]
		mean	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	-0.00 [-0.00, -0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
		shuffled	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	0.00 [-0.00, 0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	0.00 [-0.00, 0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, 0.00]
	CLEF-M	regular	train	0.01 [0.01, 0.01]	0.07 [0.07, 0.07]	0.05 [0.05, 0.05]	0.49 [0.48, 0.50]
			validation	0.01 [0.01, 0.01]	0.08 [0.07, 0.08]	0.05 [0.05, 0.06]	0.42 [0.38, 0.45]
			test	0.01 [0.00, 0.01]	0.08 [0.07, 0.08]	0.05 [0.05, 0.06]	0.42 [0.37, 0.46]
		mean	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	-0.00 [-0.00, -0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
		shuffled	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	-0.00 [-0.00, 0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.00 [-0.01, 0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, 0.00]
	CLEF-L	regular	train	0.01 [0.01, 0.01]	0.08 [0.08, 0.08]	0.06 [0.06, 0.06]	0.38 [0.37, 0.39]
			validation	0.01 [0.01, 0.01]	0.08 [0.08, 0.09]	0.06 [0.06, 0.06]	0.32 [0.29, 0.34]
			test	0.01 [0.01, 0.01]	0.08 [0.08, 0.09]	0.06 [0.06, 0.06]	0.31 [0.27, 0.34]
		mean	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	-0.00 [-0.00, -0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
		shuffled	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	-0.00 [-0.01, -0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, 0.00]
double	CLEF-S	regular	train	0.00 [0.00, 0.01]	0.07 [0.07, 0.07]	0.05 [0.05, 0.05]	0.54 [0.53, 0.55]
			validation	0.00 [0.00, 0.01]	0.07 [0.07, 0.08]	0.05 [0.05, 0.05]	0.49 [0.45, 0.53]
			test	0.01 [0.00, 0.01]	0.07 [0.07, 0.08]	0.05 [0.05, 0.05]	0.47 [0.42, 0.52]
		mean	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	-0.00 [-0.00, -0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
		shuffled	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	0.00 [-0.00, 0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.00 [-0.01, 0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, 0.00]
	CLEF-M	regular	train	0.01 [0.01, 0.01]	0.08 [0.08, 0.08]	0.06 [0.06, 0.06]	0.41 [0.40, 0.42]
			validation	0.01 [0.01, 0.01]	0.08 [0.08, 0.08]	0.06 [0.06, 0.06]	0.35 [0.33, 0.38]
			test	0.01 [0.01, 0.01]	0.08 [0.08, 0.09]	0.06 [0.06, 0.06]	0.35 [0.30, 0.38]
		mean	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	-0.00 [-0.00, -0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
		shuffled	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	-0.00 [-0.00, 0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.01 [-0.01, -0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, 0.00]
	CLEF-L	regular	train	0.01 [0.01, 0.01]	0.08 [0.08, 0.08]	0.06 [0.06, 0.06]	0.38 [0.37, 0.39]
			validation	0.01 [0.01, 0.01]	0.08 [0.08, 0.09]	0.06 [0.06, 0.06]	0.33 [0.30, 0.35]
			test	0.01 [0.01, 0.01]	0.08 [0.08, 0.09]	0.06 [0.06, 0.06]	0.31 [0.27, 0.34]
		mean	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	-0.00 [-0.00, -0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
		shuffled	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	-0.00 [-0.00, 0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.00 [-0.01, 0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, 0.00]
interp.	CLEF-S	regular	train	0.00 [0.00, 0.01]	0.07 [0.07, 0.07]	0.05 [0.05, 0.05]	0.52 [0.51, 0.53]
			validation	0.01 [0.00, 0.01]	0.07 [0.07, 0.08]	0.05 [0.05, 0.05]	0.48 [0.44, 0.51]
			test	0.01 [0.00, 0.01]	0.07 [0.07, 0.08]	0.05 [0.05, 0.05]	0.46 [0.40, 0.50]
		mean	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	-0.00 [-0.00, -0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
		shuffled	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	-0.00 [-0.00, 0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.00 [-0.01, 0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
	CLEF-M	regular	train	0.01 [0.00, 0.01]	0.07 [0.07, 0.07]	0.05 [0.05, 0.05]	0.51 [0.50, 0.51]
			validation	0.01 [0.00, 0.01]	0.07 [0.07, 0.08]	0.05 [0.05, 0.05]	0.46 [0.42, 0.49]
			test	0.01 [0.00, 0.01]	0.08 [0.07, 0.08]	0.05 [0.05, 0.05]	0.44 [0.38, 0.48]
		mean	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	-0.00 [-0.00, -0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
		shuffled	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	0.00 [-0.00, 0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	0.00 [-0.00, 0.01]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	0.00 [-0.00, 0.00]
	CLEF-L	regular	train	0.01 [0.01, 0.01]	0.07 [0.07, 0.08]	0.05 [0.05, 0.05]	0.48 [0.47, 0.49]
			validation	0.01 [0.00, 0.01]	0.07 [0.07, 0.08]	0.05 [0.05, 0.05]	0.45 [0.42, 0.49]
			test	0.01 [0.00, 0.01]	0.08 [0.07, 0.08]	0.05 [0.05, 0.05]	0.43 [0.38, 0.48]
		mean	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	-0.00 [-0.00, -0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, -0.00]
		shuffled	train	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.08, 0.08]	-0.00 [-0.00, 0.00]
			validation	0.01 [0.01, 0.01]	0.10 [0.10, 0.10]	0.08 [0.07, 0.08]	-0.00 [-0.01, 0.00]
			test	0.01 [0.01, 0.01]	0.10 [0.10, 0.11]	0.08 [0.07, 0.08]	-0.00 [-0.01, 0.00]

J CLEF Representations: UMAPs and Correlation Heatmaps

J.1 Single

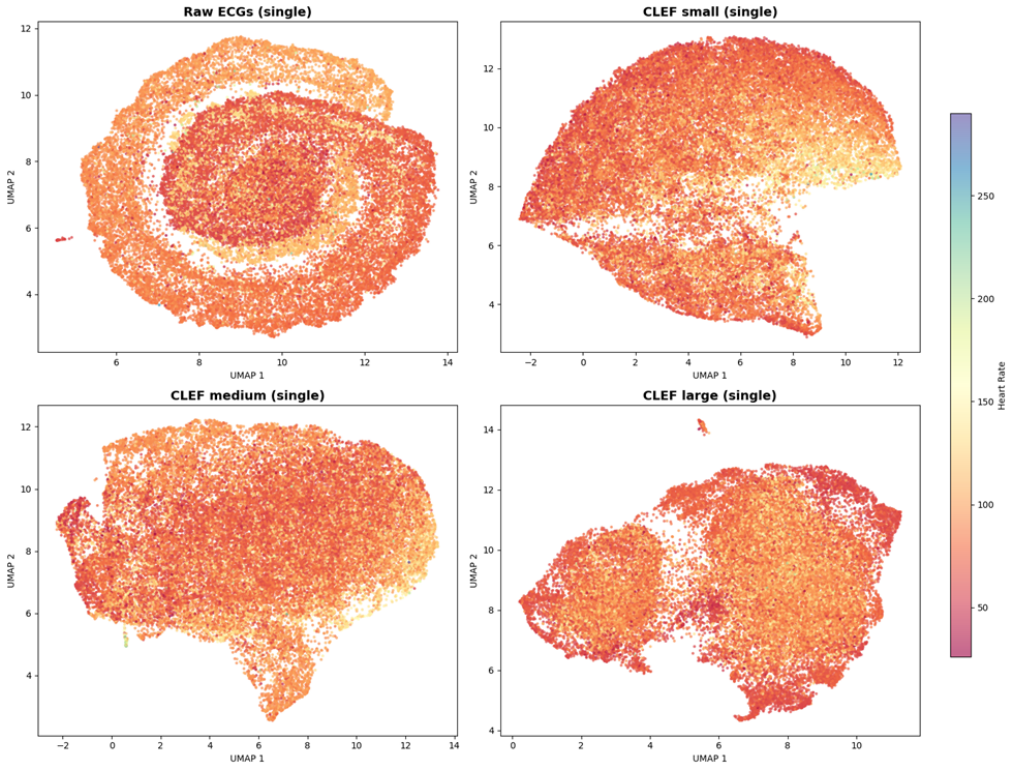

Figure 15: Comparison of UMAP representations between raw ECG inputs and CLEF-S, CLEF-M, and, CLEF-L representations, using the single input option. Points are colour coded by HR value and stem from the MIMIC-IV HR dataset.

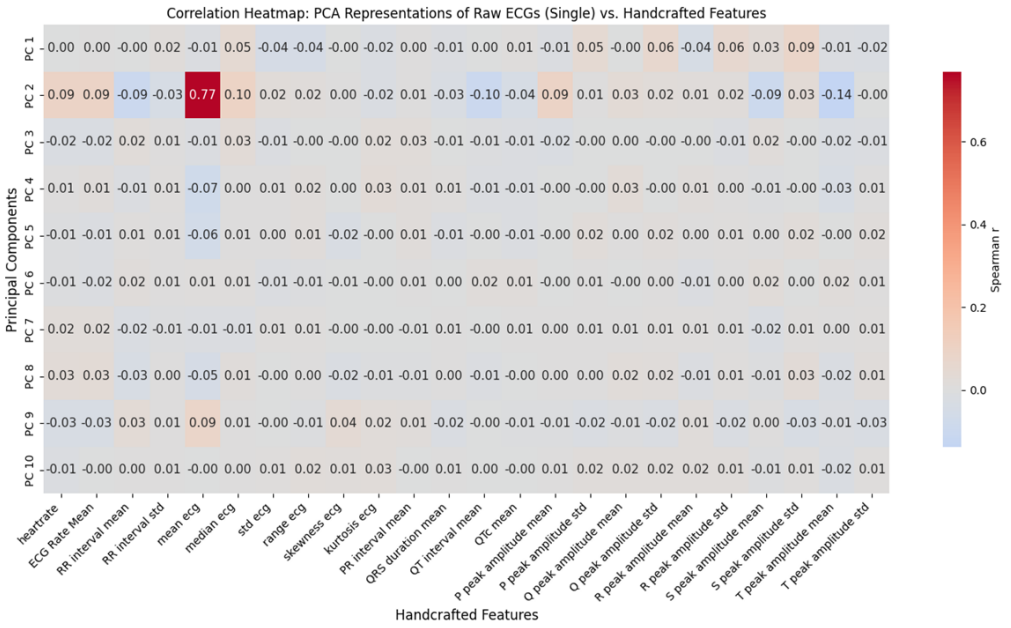

Figure 16: Spearman correlation heatmap between ten PCs extracted from raw ECG representation with single inputs from the MIMIC-IV HR dataset and handcrafted features, as well as HR.

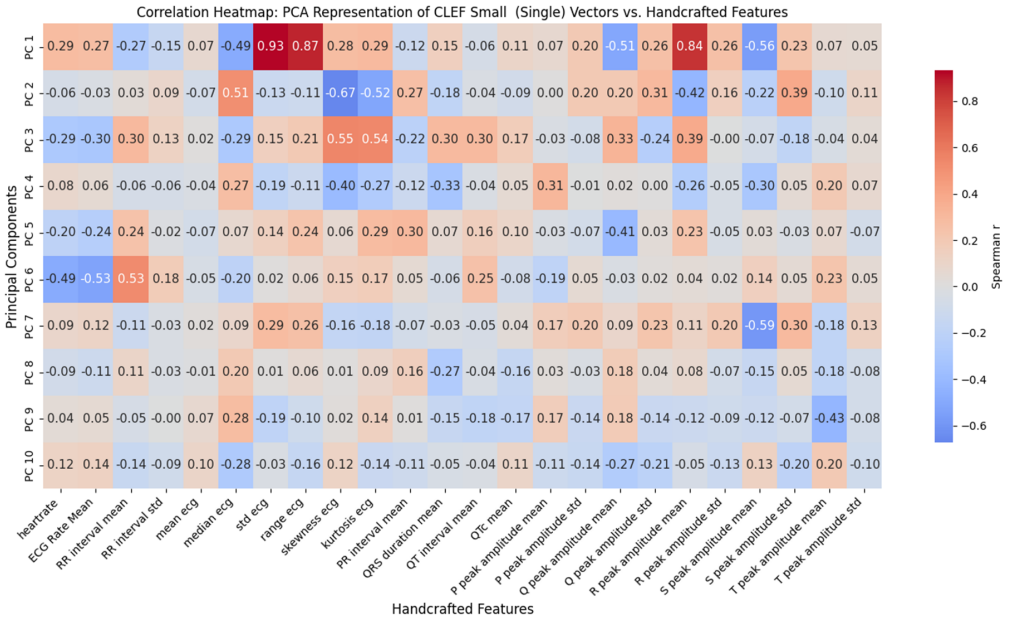

Figure 17: Spearman correlation heatmap between ten PCs extracted from CLEF-S representation with single inputs from the MIMIC-IV HR dataset and handcrafted features, as well as HR.

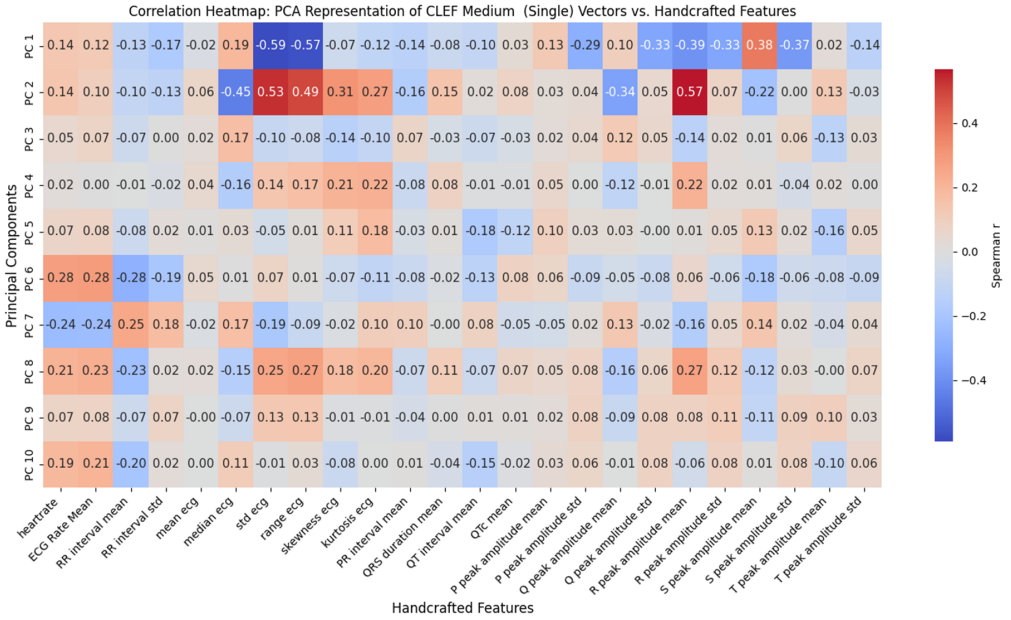

Figure 18: Spearman correlation heatmap between ten PCs extracted from CLEF-M representation with single inputs from the MIMIC-IV HR dataset and handcrafted features, as well as HR.

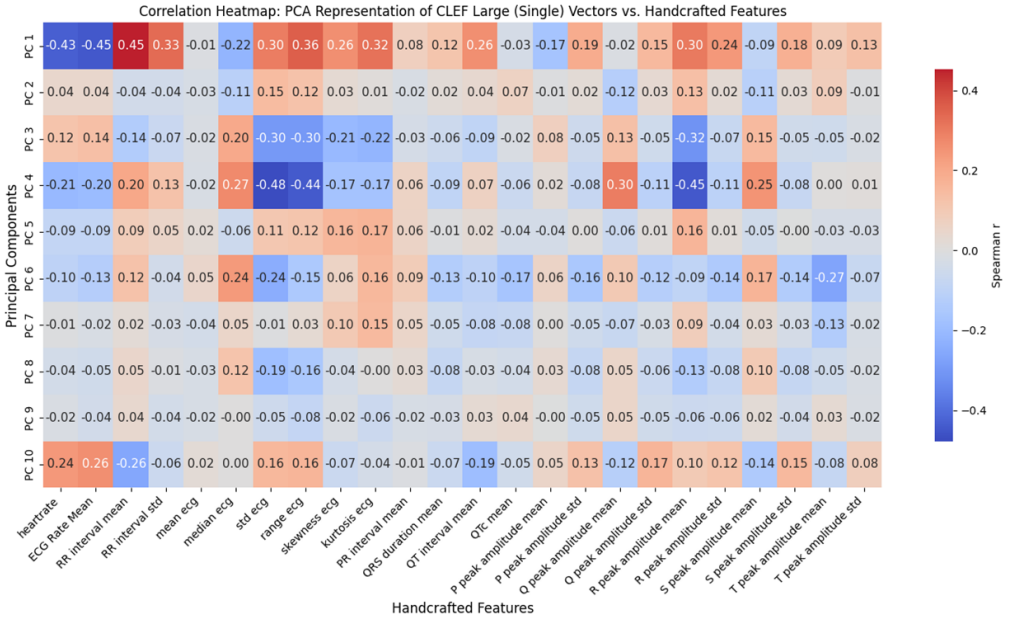

Figure 19: Spearman correlation heatmap between ten PCs extracted from CLEF-L representation with single inputs from the MIMIC-IV HR dataset and handcrafted features, as well as HR.

J.2 Double

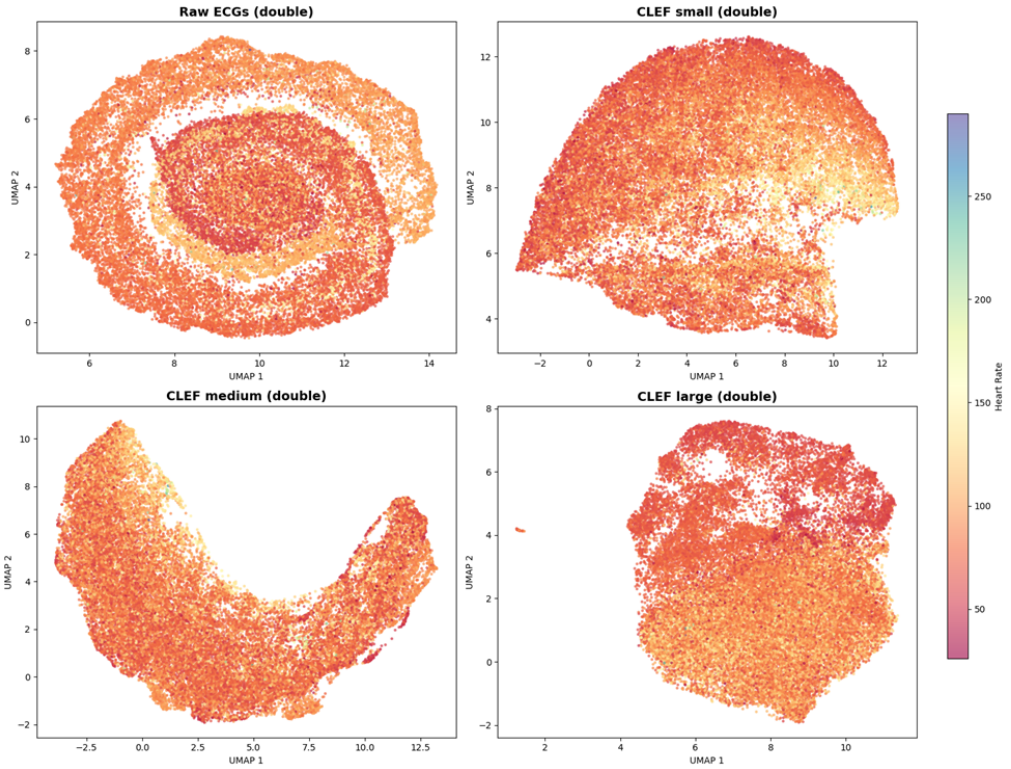

Figure 20: Comparison of UMAP representations between raw ECG inputs and CLEF-S, CLEF-M, and, CLEF-L representations, using the double input option. Points are colour coded by HR value and stem from the MIMIC-IV HR dataset.

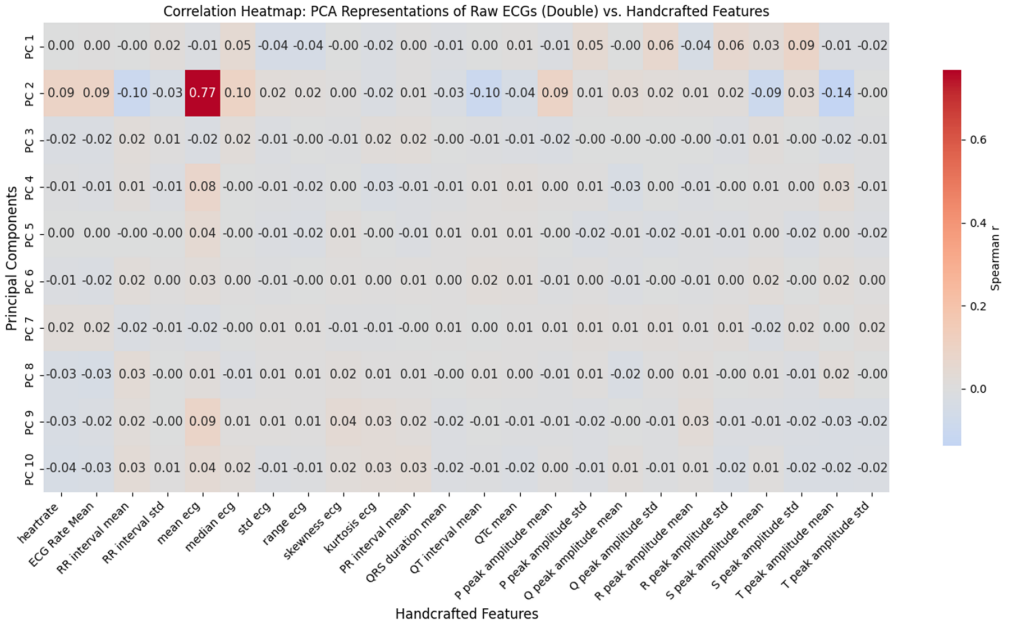

Figure 21: Spearman correlation heatmap between ten PCs extracted from raw ECG representation with double inputs from the MIMIC-IV HR dataset and handcrafted features, as well as HR.

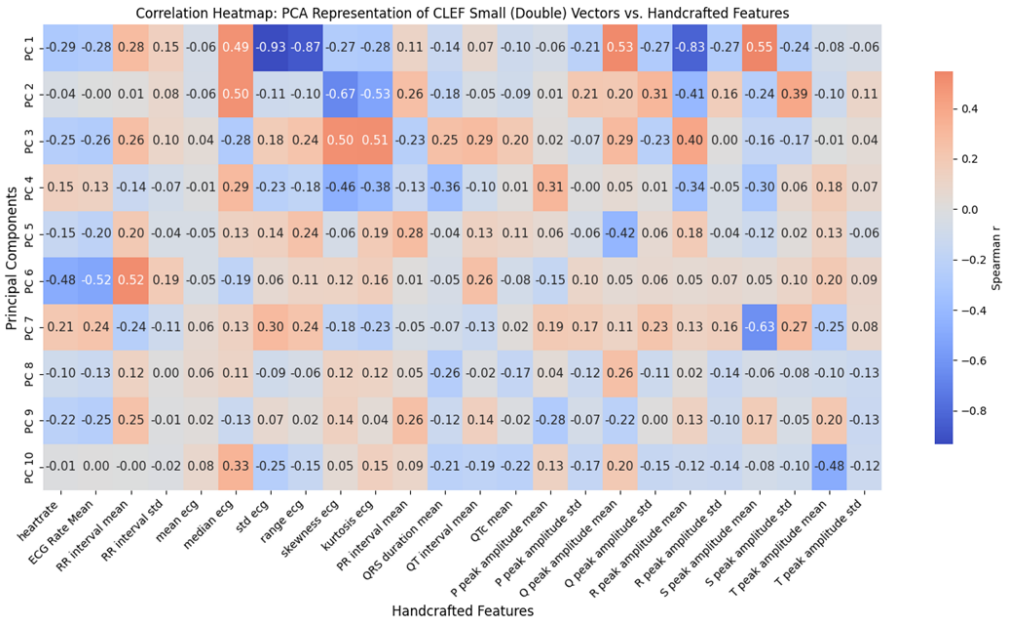

Figure 22: Spearman correlation heatmap between ten PCs extracted from CLEF-S representation with double inputs from the MIMIC-IV HR dataset and handcrafted features, as well as HR.

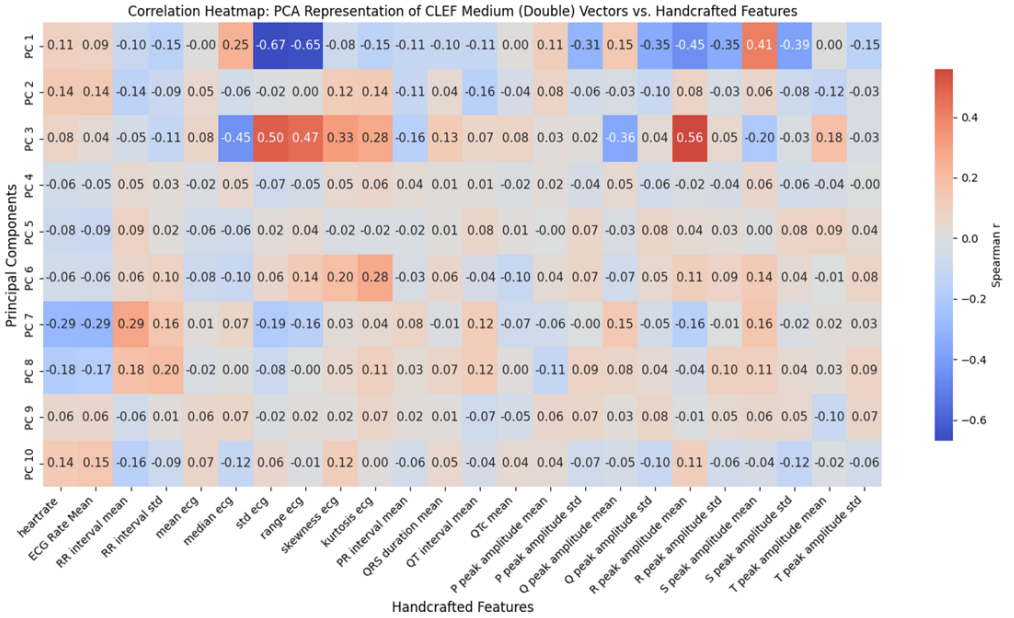

Figure 23: Spearman correlation heatmap between ten PCs extracted from CLEF-M representation with double inputs from the MIMIC-IV HR dataset and handcrafted features, as well as HR.

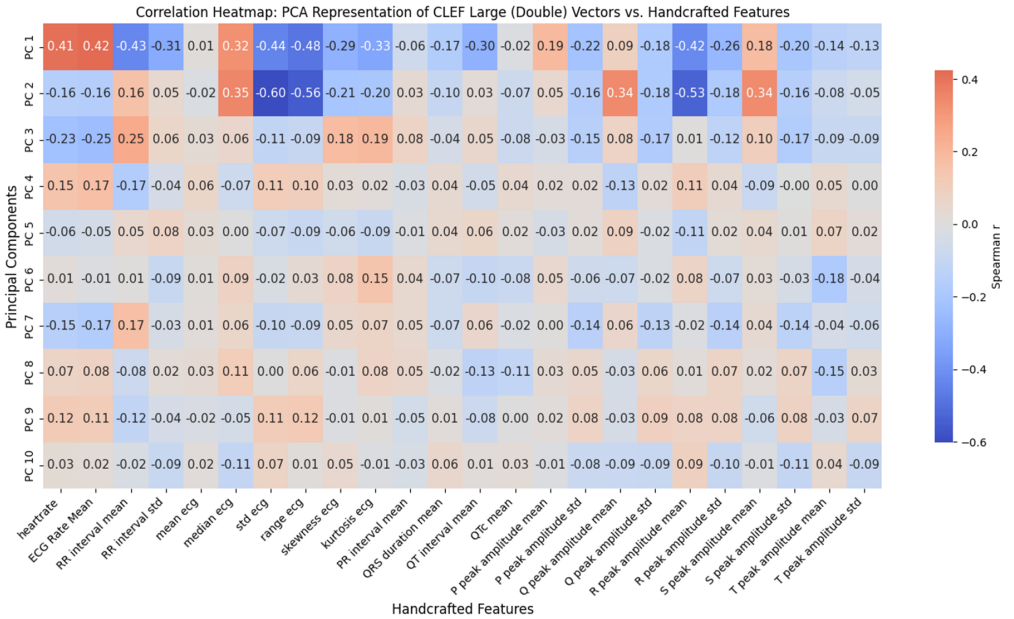

Figure 24: Spearman correlation heatmap between ten PCs extracted from CLEF-L representation with double inputs from the MIMIC-IV HR dataset and handcrafted features, as well as HR.

J.3 Interpolated

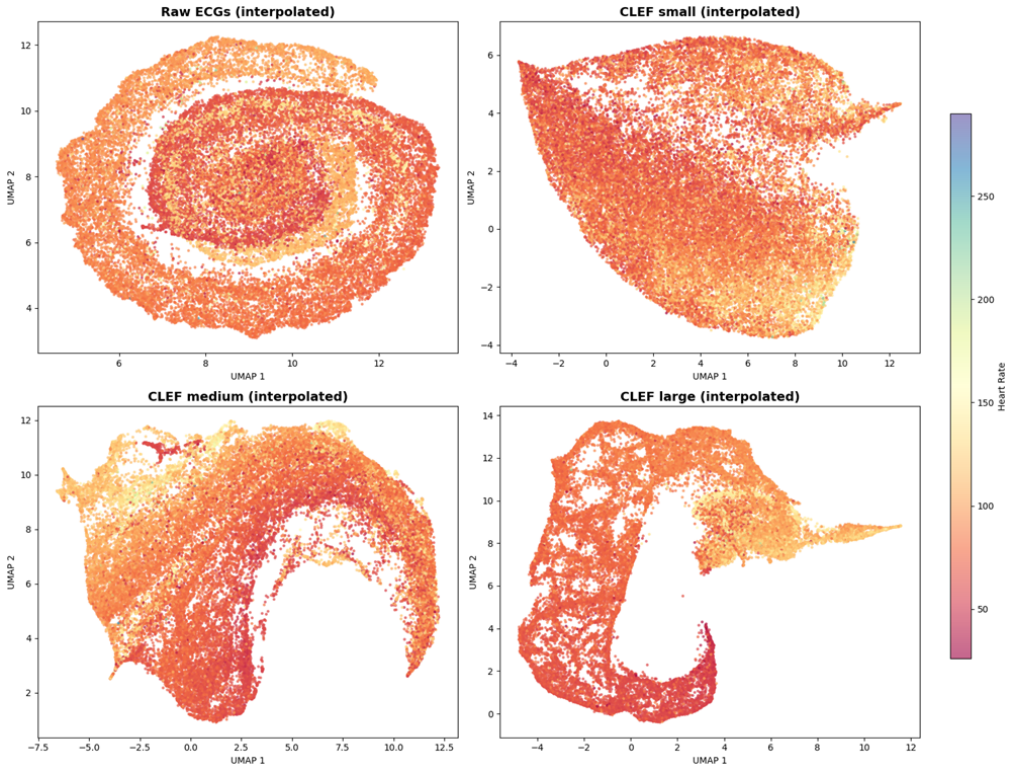

Figure 25: Comparison of UMAP representations between raw ECG inputs and CLEF-S, CLEF-M, and, CLEF-L representations, using the interpolated input option. Points are colour coded by HR value and stem from the MIMIC-IV HR dataset.

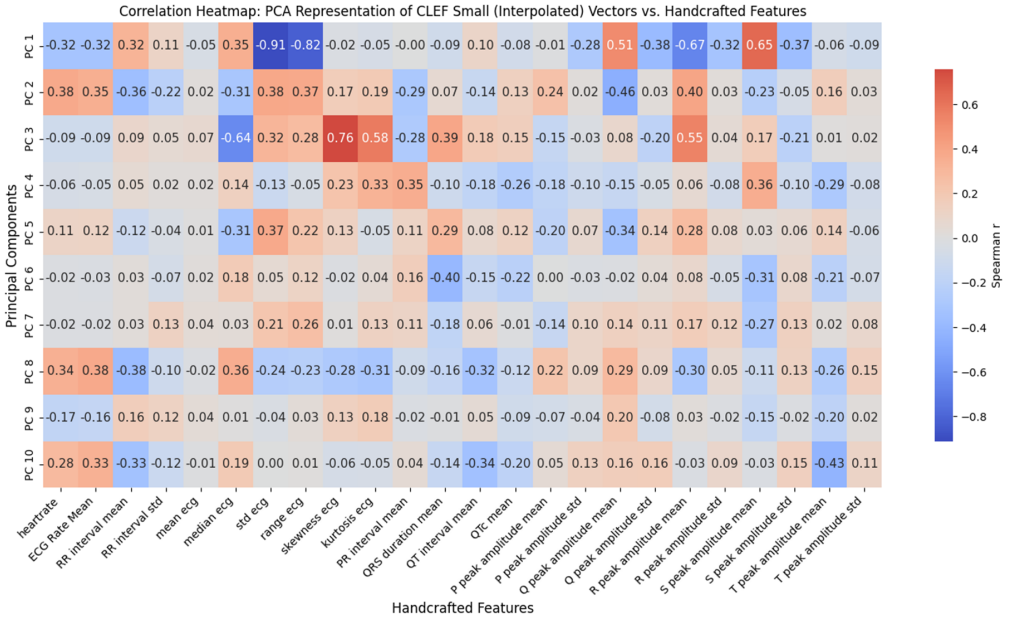

Figure 26: Spearman correlation heatmap between ten PCs extracted from CLEF small representation with interpolated inputs from the MIMIC-IV HR dataset and handcrafted features, as well as HR.

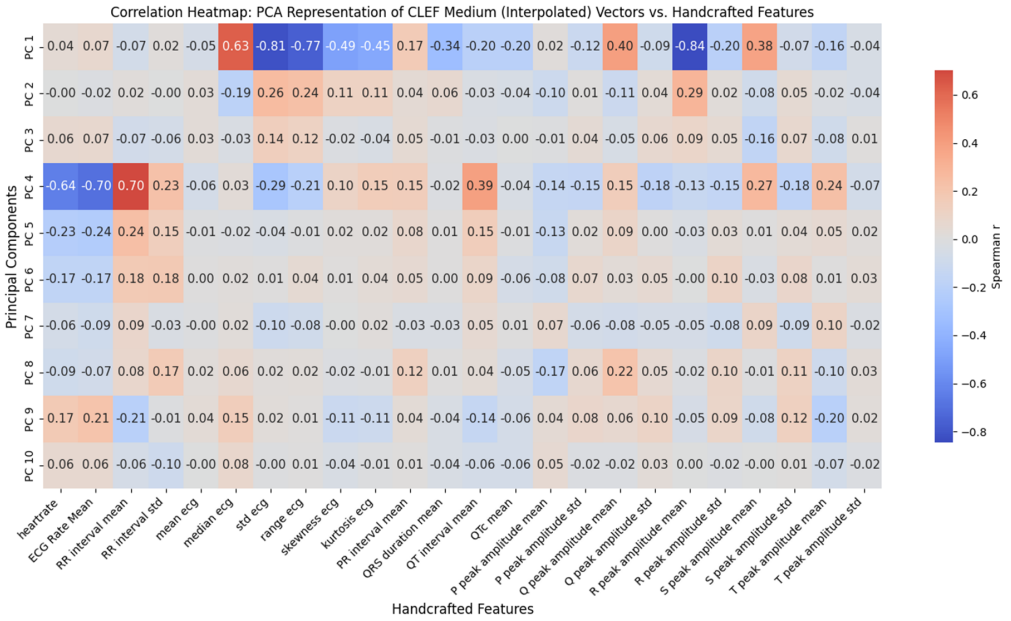

Figure 27: Spearman correlation heatmap between ten PCs extracted from CLEF-M representation with interpolated inputs from the MIMIC-IV HR dataset and handcrafted features, as well as HR.

K Epochs and Computational Load

The computational requirements for training and evaluation varied significantly across the three model architectures, reflecting their underlying complexity and input processing methods. Detailed CPU/GPU runtimes and epoch statistics are provided in tables 26 , 27 and 28.

Using the ResNet architecture, for the MIMIC-IV dataset at 200Hz, a regular training run for creatinine prediction, which lasted 14 epochs before early stopping was triggered, required a total time of 01:47:17 for training and testing. For HR prediction ResNet required nearly 19 minutes of total time on the same datasource, and also with 14 epochs of training and validation. Training and validation times for predicting creatinine with XGBoost were typically under 2 minutes for almost all configurations. For HR prediction with XGBoost training and testing times lasted round 1 minute.

On the TCC dataset for creatinine, CLEF models utilizing a linear head often completed training in abround 30 seconds. The choice of model head significantly influenced this load. Using a simple CNN head generally doubled or tripled training times compared to the one-layer linear head, leading to one to two minutes of training time. Even in HR experiments, where CLEF was trained for 50 epochs to ensure convergence, the total time required was approximately 3:30 minutes.

The models also differed in their speed to convergence. ResNet exhibited the longest per-epoch training time. It typically reached its best epoch with optimal parameters between epochs 5 and 8 for HR prediction, suggesting that the model requires multiple passes to optimize its deep convolutional layers. The CLEF models reached convergence much faster for creatinine prediction, often identifying the best epoch within 1 to 9 iterations. While the HR tasks required a longer training window of 50 epochs, the computational cost per epoch remained low, so that the total time was still a fraction of a single ResNet run. The best epoch for HR prediction using CLEF was epoch 50 for almost all runs using the linear prediction head.

Table 26: Training and evaluation time statistics and epoch information for XGBoost and ResNet.

Model	Target	Datasource	ECG Hz	Run	Trained epochs	Best epoch	Train/Val time	Test time	Total time
XGBoost	Creatinine	TCC	200	Regular			0:01:04	0:00:00	0:01:05
				Baseline: Mean			0:00:09	0:00:00	0:00:09
				Baseline: Shuffled			0:00:38	0:00:00	0:00:38
		MIMIC-IV	200	Regular			0:02:13	0:00:01	0:02:14
				Baseline: Mean			0:00:28	0:00:01	0:00:29
				Baseline: Shuffled			0:01:39	0:00:01	0:01:40
	Heart Rate	TCC	200	Regular			0:00:51	0:00:00	0:00:51
				Baseline: Mean			0:00:08	0:00:00	0:00:09
				Baseline: Shuffled			0:00:47	0:00:00	0:00:48
		MIMIC-IV	200	Regular			0:00:55	0:00:00	0:00:55
				Baseline: Mean			0:00:06	0:00:00	0:00:06
				Baseline: Shuffled			0:00:42	0:00:00	0:00:42
		MIMIC-IV	500	Regular			0:01:03	0:00:00	0:01:04
				Baseline: Mean			0:00:05	0:00:00	0:00:06
				Baseline: Shuffled			0:00:45	0:00:01	0:00:46
ResNet	Creatinine	TCC	200	Regular	7	1	0:15:47	0:00:14	0:16:02
				Baseline: Mean	11	5	0:23:08	0:00:14	0:23:23
				Baseline: Shuffled	12	6	0:25:49	0:00:14	0:26:04
		MIMIC-IV	200	Regular	14	8	1:46:21	0:00:56	1:47:17
				Baseline: Mean	13	7	1:37:25	0:00:56	1:38:22
				Baseline: Shuffled	8	2	1:05:42	0:00:55	1:06:37
	Heart Rate	TCC	200	Regular	11	5	0:23:51	0:00:14	0:24:06
				Baseline: Mean	8	2	0:16:54	0:00:14	0:17:08
				Baseline: Shuffled	7	1	0:15:36	0:00:14	0:15:50
		MIMIC-IV	200	Regular	14	8	0:18:44	0:00:10	0:18:54
				Baseline: Mean	8	2	0:10:24	0:00:09	0:10:34
				Baseline: Shuffled	10	4	0:12:54	0:00:09	0:13:04
		MIMIC-IV	500	Regular	12	6	0:30:32	0:00:13	0:30:46
				Baseline: Mean	12	6	0:30:22	0:00:13	0:30:35
				Baseline: Shuffled	11	5	0:27:47	0:00:13	0:28:01

Table 27: Training summary regarding epochs and CPU times for predicting creatinine using the CLEF backbone and different model heads.

Data	ECG Hz	Input	Model Head	Repr.	Trained Epochs	Best Epoch	Train/Val Time	Test Time	Total Time
TCC	200	single	CNN	raw	6	1	00:00:51	00:00:02	00:00:54
				CLEF-S	14	9	00:00:58	00:00:01	00:01:00
				CLEF-M	9	4	00:00:48	00:00:02	00:00:50
				CLEF-L	14	9	00:02:00	00:00:02	00:02:02
			Linear	raw	24	19	00:01:39	00:00:01	00:01:41
				CLEF-S	9	4	00:00:35	00:00:01	00:00:37
				CLEF-M	7	2	00:00:28	00:00:01	00:00:30
				CLEF-L	8	3	00:00:32	00:00:01	00:00:34
		double	CNN	raw	9	4	00:02:49	00:00:02	00:02:52
				CLEF-S	8	3	00:00:33	00:00:01	00:00:34
				CLEF-M	10	5	00:00:54	00:00:02	00:00:56
				CLEF-L	6	1	00:00:51	00:00:02	00:00:54
			Linear	raw	14	9	00:01:02	00:00:01	00:01:04
				CLEF-S	7	2	00:00:27	00:00:01	00:00:29
				CLEF-M	8	3	00:00:31	00:00:01	00:00:33
				CLEF-L	8	3	00:00:33	00:00:01	00:00:35
		interpolated	CNN	raw	8	3	00:02:31	00:00:02	00:02:34
				CLEF-S	6	1	00:00:25	00:00:01	00:00:27
				CLEF-M	9	4	00:00:48	00:00:02	00:00:50
				CLEF-L	7	2	00:01:00	00:00:02	00:01:02
			Linear	raw	19	14	00:01:23	00:00:02	00:01:25
				CLEF-S	8	3	00:00:31	00:00:01	00:00:33
				CLEF-M	6	1	00:00:24	00:00:01	00:00:26
				CLEF-L	6	1	00:00:25	00:00:01	00:00:27
MIMIC-IV	200	single	CNN	raw	26	21	00:11:05	00:00:04	00:11:09
				CLEF-S	6	1	00:01:05	00:00:03	00:01:08
				CLEF-M	6	1	00:01:27	00:00:03	00:01:31
				CLEF-L	7	2	00:02:58	00:00:04	00:03:02
			Linear	raw	10	5	00:01:43	00:00:03	00:01:46
				CLEF-S	41	36	00:06:40	00:00:03	00:06:43
				CLEF-M	18	13	00:03:03	00:00:03	00:03:06
				CLEF-L	12	7	00:02:04	00:00:03	00:02:07
		double	CNN	raw	12	7	00:11:48	00:00:06	00:11:54
				CLEF-S	11	6	00:01:59	00:00:03	00:02:02
				CLEF-M	7	2	00:01:41	00:00:03	00:01:45
				CLEF-L	10	5	00:04:14	00:00:04	00:04:18
			Linear	raw	20	15	00:03:49	00:00:03	00:03:52
				CLEF-S	50	50	00:08:12	00:00:03	00:08:15
				CLEF-M	19	14	00:03:16	00:00:03	00:03:19
				CLEF-L	11	6	00:01:54	00:00:03	00:01:58
		interpolated	CNN	raw	16	11	00:15:46	00:00:06	00:15:52
				CLEF-S	7	2	00:01:13	00:00:03	00:01:16
				CLEF-M	10	5	00:02:25	00:00:03	00:02:29
				CLEF-L	8	3	00:03:23	00:00:04	00:03:27
			Linear	raw	15	10	00:02:48	00:00:03	00:02:51
				CLEF-S	23	18	00:03:44	00:00:03	00:03:48
				CLEF-M	28	23	00:04:43	00:00:03	00:04:46
				CLEF-L	12	7	00:02:03	00:00:03	00:02:06

Table 28: Training summary regarding epochs and CPU times for predicting HR using the CLEF backbone and different model heads.

Data	ECG Hz	Input	Model Head	Repr.	Trained Epochs	Best Epoch	Train/Val Time	Test Time	Total Time
TCC	200	single	CNN	raw	20	15	00:02:46	00:00:02	00:02:48
				CLEF-S	12	7	00:00:48	00:00:01	00:00:50
				CLEF-M	31	26	00:02:40	00:00:01	00:02:42
				CLEF-L	33	28	00:04:35	00:00:02	00:04:37
			Linear	raw	50	50	00:03:20	00:00:01	00:03:22
				CLEF-S	50	50	00:03:09	00:00:01	00:03:11
				CLEF-M	50	50	00:03:13	00:00:01	00:03:15
				CLEF-L	50	50	00:03:18	00:00:01	00:03:20
		double	CNN	raw	21	16	00:06:22	00:00:02	00:06:25
				CLEF-S	18	13	00:01:12	00:00:01	00:01:14
				CLEF-M	16	11	00:01:23	00:00:02	00:01:25
				CLEF-L	49	44	00:06:47	00:00:02	00:06:50
			Linear	raw	50	50	00:03:39	00:00:01	00:03:41
				CLEF-S	50	50	00:03:11	00:00:01	00:03:13
				CLEF-M	50	50	00:03:18	00:00:01	00:03:20
				CLEF-L	50	50	00:03:21	00:00:01	00:03:23
		interpolated	CNN	raw	21	16	00:06:24	00:00:02	00:06:27
				CLEF-S	22	17	00:01:30	00:00:01	00:01:32
				CLEF-M	24	19	00:02:04	00:00:02	00:02:07
				CLEF-L	18	13	00:02:29	00:00:02	00:02:32
			Linear	raw	50	50	00:03:37	00:00:01	00:03:39
				CLEF-S	50	50	00:03:10	00:00:01	00:03:12
				CLEF-M	50	50	00:03:17	00:00:01	00:03:19
				CLEF-L	50	50	00:03:19	00:00:01	00:03:21
MIMIC-IV	200	single	CNN	raw	14	9	00:01:17	00:00:01	00:01:19
				CLEF-S	14	9	00:00:41	00:00:01	00:00:42
				CLEF-M	13	8	00:00:47	00:00:01	00:00:48
				CLEF-L	31	26	00:02:52	00:00:01	00:02:54
			Linear	raw	50	50	00:02:24	00:00:01	00:02:26
				CLEF-S	50	50	00:02:18	00:00:01	00:02:20
				CLEF-M	50	50	00:02:21	00:00:01	00:02:23
				CLEF-L	50	50	00:02:22	00:00:01	00:02:24
		double	CNN	raw	18	13	00:03:27	00:00:02	00:03:29
				CLEF-S	19	14	00:00:55	00:00:01	00:00:57
				CLEF-M	14	9	00:00:50	00:00:01	00:00:52
				CLEF-L	27	22	00:02:29	00:00:01	00:02:31
			Linear	raw	50	50	00:02:34	00:00:01	00:02:36
				CLEF-S	50	50	00:02:17	00:00:01	00:02:19
				CLEF-M	50	50	00:02:20	00:00:01	00:02:22
				CLEF-L	50	49	00:02:24	00:00:01	00:02:26
		interpolated	CNN	raw	17	12	00:03:17	00:00:02	00:03:19
				CLEF-S	13	8	00:00:38	00:00:01	00:00:39
				CLEF-M	43	38	00:02:35	00:00:01	00:02:37
				CLEF-L	23	18	00:02:07	00:00:01	00:02:09
			Linear	raw	50	50	00:02:34	00:00:01	00:02:36
				CLEF-S	50	50	00:02:18	00:00:01	00:02:20
				CLEF-M	50	50	00:02:23	00:00:01	00:02:25
				CLEF-L	50	50	00:02:26	00:00:01	00:02:27
MIMIC-IV	500	single	CNN	raw	18	13	00:03:31	00:00:02	00:03:33
				CLEF-S	14	9	00:00:40	00:00:01	00:00:42
				CLEF-M	27	22	00:01:39	00:00:01	00:01:40
				CLEF-L	30	25	00:02:50	00:00:01	00:02:51
			Linear	raw	50	50	00:02:35	00:00:01	00:02:37
				CLEF-S	50	50	00:02:16	00:00:01	00:02:17
				CLEF-M	50	50	00:02:21	00:00:01	00:02:22
				CLEF-L	50	50	00:02:22	00:00:01	00:02:24

L Additional Run: CLEF HR prediction using Batch Size 16

Table 29: Model performance metrics (200Hz) across input types, model heads (CNN and linear), and CLEF representations, reducing batch size to 16 for HR prediction using the TCC dataset. Results are reported as point estimates with 95% confidence intervals indicated in brackets. Grey rows indicate the best representation (lowest MSE) for each category. Bolded values highlight the globally best performing configuration.

Input	Head	Repr.	Dataset	MSE	RMSE	MAE	R2
single	CNN	raw	train	137.47 [133.34, 141.67]	11.72 [11.55, 11.90]	8.04 [7.96, 8.12]	0.64 [0.63, 0.65]
			validation	156.80 [149.61, 164.37]	12.52 [12.23, 12.82]	8.71 [8.53, 8.90]	0.61 [0.60, 0.63]
			test	172.15 [163.96, 180.16]	13.12 [12.80, 13.42]	9.36 [9.18, 9.53]	0.55 [0.53, 0.57]
		CLEF-S	train	216.26 [210.97, 221.36]	14.71 [14.52, 14.88]	11.18 [11.08, 11.27]	0.44 [0.43, 0.45]
			validation	258.51 [248.02, 269.69]	16.08 [15.75, 16.42]	12.29 [12.07, 12.51]	0.36 [0.35, 0.38]
			test	240.13 [231.35, 248.85]	15.50 [15.21, 15.78]	12.06 [11.86, 12.24]	0.38 [0.36, 0.39]
		CLEF-M	train	176.76 [172.48, 181.28]	13.29 [13.13, 13.46]	9.39 [9.30, 9.48]	0.54 [0.53, 0.55]
			validation	222.78 [211.39, 234.75]	14.92 [14.54, 15.32]	10.44 [10.21, 10.67]	0.45 [0.44, 0.47]
			test	199.96 [191.16, 208.39]	14.14 [13.83, 14.44]	10.20 [10.00, 10.39]	0.48 [0.47, 0.49]
		CLEF-L	train	340.91 [333.58, 348.01]	18.46 [18.26, 18.66]	13.43 [13.31, 13.55]	0.11 [0.10, 0.12]
			validation	390.76 [374.39, 408.95]	19.77 [19.35, 20.22]	14.20 [13.92, 14.51]	0.04 [0.02, 0.06]
			test	359.32 [345.19, 373.65]	18.95 [18.58, 19.33]	13.96 [13.70, 14.21]	0.07 [0.04, 0.09]
	Linear	raw	train	379.95 [373.37, 386.79]	19.49 [19.32, 19.67]	15.09 [14.97, 15.21]	0.01 [0.01, 0.01]
			validation	415.03 [400.64, 431.64]	20.37 [20.02, 20.78]	15.61 [15.34, 15.89]	-0.02 [-0.03, -0.02]
			test	393.27 [379.96, 406.66]	19.83 [19.49, 20.17]	15.51 [15.25, 15.78]	-0.02 [-0.03, -0.02]
		CLEF-S	train	173.84 [168.94, 178.69]	13.18 [13.00, 13.37]	9.33 [9.24, 9.42]	0.55 [0.54, 0.56]
			validation	192.90 [184.30, 202.54]	13.89 [13.58, 14.23]	9.75 [9.55, 9.97]	0.53 [0.51, 0.54]
			test	200.66 [192.80, 208.91]	14.16 [13.89, 14.45]	10.34 [10.15, 10.54]	0.48 [0.46, 0.50]
		CLEF-M	train	106.96 [103.74, 109.89]	10.34 [10.19, 10.48]	7.13 [7.05, 7.19]	0.72 [0.72, 0.73]
			validation	141.30 [122.55, 172.20]	11.87 [11.07, 13.12]	7.98 [7.80, 8.18]	0.65 [0.58, 0.69]
			test	130.26 [125.48, 135.65]	11.41 [11.20, 11.65]	8.48 [8.33, 8.64]	0.66 [0.65, 0.67]
		CLEF-L	train	123.56 [120.81, 126.67]	11.12 [10.99, 11.25]	8.03 [7.96, 8.11]	0.68 [0.67, 0.68]
			validation	158.27 [151.17, 165.42]	12.58 [12.29, 12.86]	9.12 [8.95, 9.29]	0.61 [0.60, 0.62]
			test	164.34 [157.64, 170.97]	12.82 [12.56, 13.08]	9.63 [9.45, 9.81]	0.57 [0.56, 0.59]
double	CNN	raw	train	108.99 [105.69, 112.55]	10.44 [10.28, 10.61]	7.24 [7.17, 7.31]	0.72 [0.71, 0.72]
			validation	132.06 [125.24, 139.31]	11.49 [11.19, 11.80]	8.05 [7.88, 8.23]	0.68 [0.66, 0.69]
			test	144.42 [137.60, 151.04]	12.02 [11.73, 12.29]	8.88 [8.71, 9.04]	0.62 [0.61, 0.64]
		CLEF-S	train	454.83 [448.14, 462.19]	21.33 [21.17, 21.50]	18.07 [17.96, 18.18]	-0.19 [-0.20, -0.17]
			validation	517.00 [501.59, 534.34]	22.74 [22.40, 23.12]	19.02 [18.76, 19.28]	-0.27 [-0.31, -0.24]
			test	461.97 [449.28, 474.36]	21.49 [21.20, 21.78]	18.12 [17.90, 18.35]	-0.20 [-0.23, -0.17]
		CLEF-M	train	462.90 [454.75, 470.67]	21.52 [21.32, 21.70]	17.36 [17.24, 17.48]	-0.21 [-0.22, -0.19]
			validation	518.96 [499.75, 539.65]	22.78 [22.36, 23.23]	17.93 [17.64, 18.25]	-0.28 [-0.30, -0.25]
			test	475.70 [459.32, 490.76]	21.81 [21.43, 22.15]	17.38 [17.12, 17.64]	-0.24 [-0.27, -0.21]
		CLEF-L	train	165.30 [160.86, 169.56]	12.86 [12.68, 13.02]	9.03 [8.94, 9.11]	0.57 [0.56, 0.58]
			validation	221.72 [210.81, 233.75]	14.89 [14.52, 15.29]	10.39 [10.16, 10.62]	0.45 [0.44, 0.47]
			test	209.24 [199.89, 218.73]	14.46 [14.14, 14.79]	10.35 [10.15, 10.56]	0.46 [0.44, 0.47]
	Linear	raw	train	382.19 [375.78, 388.79]	19.55 [19.39, 19.72]	15.13 [15.00, 15.24]	0.00 [0.00, 0.01]
			validation	418.11 [403.40, 434.60]	20.45 [20.08, 20.85]	15.68 [15.40, 15.95]	-0.03 [-0.03, -0.02]
			test	398.84 [385.44, 412.32]	19.97 [19.63, 20.31]	15.61 [15.35, 15.88]	-0.04 [-0.05, -0.03]
		CLEF-S	train	175.53 [170.74, 180.38]	13.25 [13.07, 13.43]	9.38 [9.29, 9.47]	0.54 [0.53, 0.55]
			validation	188.73 [180.12, 197.81]	13.74 [13.42, 14.06]	9.69 [9.49, 9.90]	0.54 [0.52, 0.55]
			test	198.95 [190.86, 207.22]	14.10 [13.82, 14.40]	10.29 [10.10, 10.49]	0.48 [0.46, 0.50]
		CLEF-M	train	108.86 [105.63, 112.03]	10.43 [10.28, 10.58]	7.18 [7.11, 7.25]	0.72 [0.71, 0.72]
			validation	140.95 [125.31, 164.90]	11.86 [11.19, 12.84]	8.03 [7.84, 8.21]	0.65 [0.60, 0.69]
			test	129.22 [124.30, 134.66]	11.37 [11.15, 11.60]	8.41 [8.26, 8.56]	0.66 [0.65, 0.68]
		CLEF-L	train	126.58 [123.53, 129.64]	11.25 [11.11, 11.39]	8.25 [8.17, 8.33]	0.67 [0.66, 0.68]
			validation	159.50 [152.41, 166.69]	12.63 [12.35, 12.91]	9.23 [9.07, 9.41]	0.61 [0.59, 0.62]
			test	171.98 [165.54, 178.14]	13.11 [12.87, 13.35]	10.07 [9.89, 10.23]	0.55 [0.54, 0.57]
interp.	CNN	raw	train	151.96 [147.84, 156.11]	12.33 [12.16, 12.49]	8.71 [8.63, 8.79]	0.60 [0.60, 0.61]
			validation	161.20 [153.39, 170.00]	12.70 [12.39, 13.04]	9.02 [8.83, 9.21]	0.60 [0.59, 0.62]
			test	170.55 [163.06, 178.15]	13.06 [12.77, 13.35]	9.66 [9.47, 9.86]	0.56 [0.54, 0.57]
		CLEF-S	train	202.98 [199.33, 206.35]	14.25 [14.12, 14.36]	11.29 [11.21, 11.37]	0.47 [0.46, 0.48]
			validation	245.49 [237.44, 254.51]	15.67 [15.41, 15.95]	12.37 [12.17, 12.58]	0.40 [0.38, 0.41]
			test	247.49 [239.81, 254.62]	15.73 [15.49, 15.96]	12.50 [12.29, 12.70]	0.36 [0.34, 0.37]
		CLEF-M	train	358.30 [352.11, 364.33]	18.93 [18.76, 19.09]	15.27 [15.16, 15.38]	0.07 [0.06, 0.08]
			validation	410.29 [395.22, 426.48]	20.25 [19.88, 20.65]	16.01 [15.76, 16.29]	-0.01 [-0.03, 0.01]
			test	381.80 [367.82, 396.21]	19.54 [19.18, 19.90]	15.41 [15.18, 15.65]	0.01 [-0.02, 0.03]
		CLEF-L	train	279.64 [276.04, 283.32]	16.72 [16.61, 16.83]	13.95 [13.86, 14.03]	0.27 [0.26, 0.28]
			validation	301.13 [292.16, 310.72]	17.35 [17.09, 17.63]	14.36 [14.16, 14.56]	0.26 [0.24, 0.28]
			test	351.99 [341.79, 362.89]	18.76 [18.49, 19.05]	15.48 [15.26, 15.72]	0.08 [0.05, 0.11]
	Linear	raw	train	385.27 [378.56, 392.03]	19.63 [19.46, 19.80]	15.20 [15.08, 15.31]	-0.00 [-0.01, 0.00]
			validation	421.09 [406.88, 437.51]	20.52 [20.17, 20.92]	15.74 [15.47, 16.02]	-0.04 [-0.04, -0.03]
			test	399.05 [385.84, 412.87]	19.98 [19.64, 20.32]	15.59 [15.34, 15.85]	-0.04 [-0.05, -0.03]
		CLEF-S	train	224.91 [220.02, 230.02]	15.00 [14.83, 15.17]	11.29 [11.20, 11.39]	0.41 [0.41, 0.42]
			validation	251.09 [239.30, 262.74]	15.84 [15.47, 16.21]	11.73 [11.50, 11.95]	0.38 [0.36, 0.40]
			test	253.39 [243.86, 263.13]	15.92 [15.62, 16.22]	12.07 [11.85, 12.29]	0.34 [0.32, 0.36]
		CLEF-M	train	96.89 [93.25, 100.88]	9.84 [9.66, 10.04]	6.74 [6.66, 6.80]	0.75 [0.74, 0.76]
			validation	121.65 [114.85, 129.13]	11.03 [10.72, 11.36]	7.64 [7.49, 7.82]	0.70 [0.69, 0.71]
			test	**125.29 [118.82, 131.78]**	**11.19 [10.90, 11.48]**	**8.01 [7.86, 8.17]**	**0.67 [0.66, 0.69]**
		CLEF-L	train	115.91 [112.76, 119.02]	10.77 [10.62, 10.91]	7.59 [7.51, 7.67]	0.70 [0.69, 0.70]
			validation	154.43 [146.90, 162.64]	12.43 [12.12, 12.75]	8.82 [8.65, 9.01]	0.62 [0.61, 0.63]
			test	162.83 [155.25, 170.95]	12.76 [12.46, 13.07]	9.03 [8.86, 9.22]	0.58 [0.56, 0.59]