International Journal of Medical Informatics 175 (2023) 105086

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Generalizable calibrated machine learning models for real-time atrial fibrillation risk prediction in ICU patients

Jarne Verhaeghe [a,*,1], Thomas De Corte [b,c,1], Christopher M. Sauer [d,e], Tom Hendriks [f], Olivier W.M. Thijssens [f], Femke Ongenae [a], Paul Elbers [e,g], Jan De Waele [b,c], Sofie Van Hoecke [a]

[a] IDLab, Ghent University - imec, Technologiepark-Zwijnaarde 126, Ghent, 9052, Belgium
[b] Department of Internal Medicine and Pediatrics, Faculty of Medicine and Health Sciences, Ghent University, C. Heymanslaan 10, Ghent, 9000, Belgium
[c] Department of Intensive Care Medicine, Ghent University Hospital, C. Heymanslaan 10, Ghent, 9000, Belgium
[d] Center for Critical Care Computational Intelligence (C4i), Amsterdam Medical Data Science (AMDS), Amsterdam Public Health (APH), VU University Medical Center Amsterdam, Amsterdam, the Netherlands
[e] Department of Intensive Care Medicine, Amsterdam Cardiovascular Sciences (ACS), Amsterdam Infection and Immunity Institute (AI and II), VU University Medical Center Amsterdam, Amsterdam, the Netherlands
[f] Pacmed, Stadhouderskade 55, Amsterdam, the Netherlands
[g] Research VUmc Intensive Care (REVIVE), VU University Medical Center Amsterdam, Amsterdam, the Netherlands

ARTICLE INFO

Dataset link: https://github.com/predict-idlab/atrial_fibrillation_prediction

Dataset link: https://github.com/AmsterdamUMC/AmsterdamUMCdb

Dataset link: https://physionet.org/content/mimiciv/2.1/

Keywords:
Atrial fibrillation
Uncertainty quantification metrics
Risk score
Machine learning
ICU
Calibration

ABSTRACT

Background: Atrial Fibrillation (AF) is the most common arrhythmia in the intensive care unit (ICU) and is associated with increased morbidity and mortality. Identification of patients at risk for AF is not routinely performed as AF prediction models are almost solely developed for the general population or for particular ICU populations. However, early AF risk identification could help to take targeted preemptive actions and possibly reduce morbidity and mortality. Predictive models need to be validated across hospitals with different standards of care and convey their predictions in a clinically useful manner. Therefore, we designed AF risk models for ICU patients using uncertainty quantification to provide a risk score and evaluated them on multiple ICU datasets.
Methods: Three CatBoost models, utilizing feature windows comprising data 1.5-13.5, 6-18, or 12-24 hours before AF occurrence, were built using 2-repeat-10-fold cross-validation on AmsterdamUMCdb, the first freely available European ICU database. Furthermore, AF Patients were matched with no-AF patients for training. Transferability was validated using a direct and a recalibration evaluation on two independent external datasets, MIMIC-IV and GUH. The calibration of the predicted probability, used as an AF risk score, was measured using the Expected Calibration Error (ECE) and the presented Expected Signed Calibration Error (ESCE). Additionally, all models were evaluated across time during the ICU stay.
Results: The model performance reached Areas Under the Curve (AUCs) of 0.81 at internal validation. Direct external validation showed partial generalizability with AUCs reaching 0.77. However, recalibration resulted in performances matching or exceeding that of the internal validation. All models furthermore showed calibration capabilities demonstrating adequate risk prediction competence.
Conclusion: Ultimately, recalibrating models reduces the challenge of generalization to unseen datasets. Moreover, utilizing the patient-matching methodology together with the assessment of uncertainty calibration can serve as a step toward the development of clinical AF prediction models.

* Corresponding author.
E-mail address: jarne.verhaeghe@ugent.be (J. Verhaeghe).
URL: http://predict.idlab.ugent.be/ (S. Van Hoecke).
[1] Both authors contributed equally and share first authorship.

https://doi.org/10.1016/j.ijmedinf.2023.105086
Received 19 January 2023; Received in revised form 19 April 2023; Accepted 21 April 2023
Available online 26 April 2023

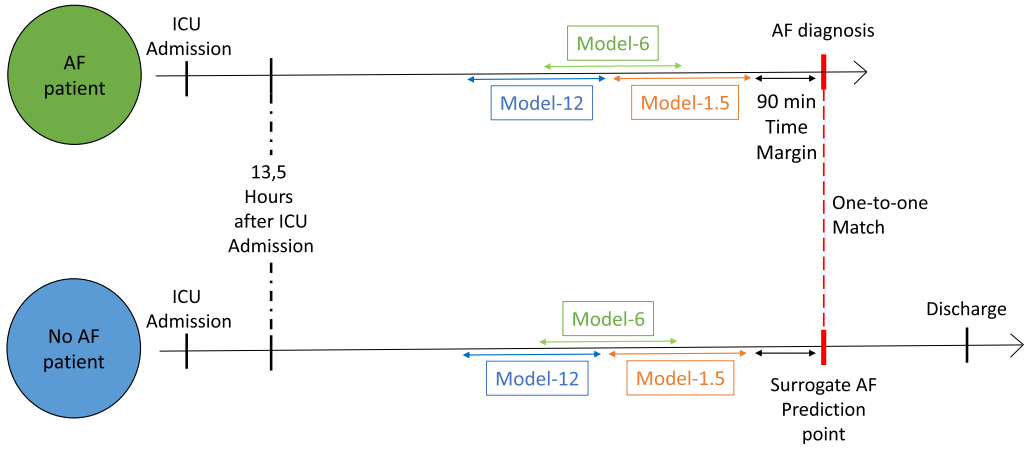

Fig. 1. Illustration of the model development and the case-control matching procedure used. All information between the event and 90 minutes prior to the event was excluded. Every AF patient (green circle) was matched with a no-AF patient (blue circle). For the no-AF patient, the surrogate AF prediction point was defined as the same time point, relative to ICU admission, as AF occurrence in the AF patient (red line). Model-1.5 is built on a time window of 1.5-13.5 hours before AF occurrence. Model-6 is built on a time window of 6-18 hours. Model-12 is built on a time window of 12-24 hours.

1. Introduction

Atrial fibrillation (AF) is a heart rhythm disorder that causes an irregular and often abnormally fast heart rate. It affects between 4.5 to 15% of patients admitted to the intensive care unit (ICU). The incidence is even higher in specific patient populations, i.e. patients admitted after cardiac surgery (35%) or patients with septic shock (40%-46%) [1–3]. Several studies have indicated that the occurrence of AF in critically ill patients is associated with poorer outcomes, including prolonged length of stay (LOS) and increased hospital mortality [1,4]. Although several risk factors for AF are non-modifiable (e.g., age), identifying patients at high risk of developing AF could allow clinicians to preemptively address modifiable risk factors (e.g., electrolyte imbalances or medication). However, clinical identification of patients at risk for AF is not routinely performed in the ICU as developed clinical risk prediction models are often limited to either the general population or to selected ICU patient populations [3,5,6]. Therefore, we aimed to build a machine learning (ML) model to predict the risk of AF occurrence in real-time for any ICU patient using calibrated uncertainty predictions. Special attention was given to developing a model that generates a meaningful, interpretable risk output for the bedside clinician by using Shapley values. Additionally, the models were validated on data from multiple ICUs across the globe to determine their generalizability. Finally, the models were evaluated in a simulated clinical situation across time to understand their behavior in clinical practice, while fully explaining the model using interpretability libraries.

2. Materials and methods

A complete overview of all conducted experiments and methods in this study is visualized in Online Supplementary (OS) Fig. 1.

2.1. Prediction model development

2.1.1. Study population and prediction window

The AmsterdamUMCdb database (v1.02) was used for model development and internal validation [7]. The outcome, i.e. the occurrence of AF, was operationalized as the first AF registration by the nursing staff after at least one registration of a sinus rhythm. The study cohort was limited to patients with a minimal LOS of 13.5 hours. These 13.5 hours are based on a feature aggregation window of 12 hours and a 1.5-hour time window between feature aggregation and event occurrence to avoid data leakage. The database and the study population characteristics are summarized in Table 1.

Three models with different feature windows, but all with an aggregation window of 12 hours, were designed. The different windows reflect the period from which collected data is used to make a risk prediction. The models respectively have a window of 1.5 – 13.5, 6 – 18, and 12 – 24 hours before AF occurrence (Model-1.5, Model-6, and Model-12).

As there were more ICU admissions without AF (16,163) than with AF (2,000), and as AF patients tend to have a longer LOS [1,4] the training set was balanced by one-to-one matching every AF admission to a no-AF admission. This was achieved by fixing the surrogate AF prediction point for the no-AF admission to the time after admission of the AF admission AF diagnosis following a case-control study design, as visualized in Fig. 1. Only the relative time after ICU admission was considered. The AF admission AF diagnosis time should be before the ICU discharge of the candidate no-AF admission. Furthermore, the no-AF admission should not have been matched already to an AF admission. Once these requirements were fulfilled, the two admissions were matched and shared their relative prediction points. This procedure was performed before train-test splitting. Afterward, all AF admissions were split into a training and a test set with an 80/20 ratio. Their respective no-AF match was then also put in their respective training or test set to avoid splitting a match. The remaining no-AF admissions without a match were given a random prediction moment between 13.5 hours (resp. 18 or 24 hours, depending on the model) after admission and before discharge. These no-AF admissions were only used in a separate imbalanced test set for evaluation on the complete cohort. Ensuing, the models were trained on the training set and optimized using 2-repeat-10-fold cross-validation as there was enough data where two repeats were sufficiently stable. Finally, the models were evaluated on two different test sets: a balanced test set containing only matched admissions, and an imbalanced test set containing the matched AF admissions and all unmatched no-AF admissions.

2.1.2. Model building and feature selection

The models were built using Python (v3.8.10). CatBoost classification models were chosen because of their interpretability, missing value handling, strong prediction performances, and automatic categorical feature processing [8]. Furthermore, the model probability output was used for uncertainty quantification (UQ) to provide trustworthiness and an AF risk score. The classification performance was measured using the precision, recall, and Area Under the Curve (AUC) metrics.

Based on expert opinion and literature review, 282 unique potential variables were selected in AmsterdamUMCdb (listed in the OS Section 2). Variables with a count < 250 or above 99% missingness in

Table 1
The study population characteristics. Mean (Q25 - Q75) for continuous variables, percentages for categorical variables. BMI = Body Mass Index. SOFA = Sequential Organ Failure Assessment score. APACHE = Acute Physiology and Chronic Health Evaluation. LOS = Length Of Stay. *Only patients admitted to the ICU were evaluated for inclusion.

	AmsterdamUMC		MIMIC- IV*		GUH	
Time period	2003 – 2016		2008 – 2019		2013 - 2020	
Total admissions	23,106		69,211		25,297	
Study cohort size	18,163		59,492		23,459	
	AF	no-AF	AF	no-AF	AF	no-AF
Study cohort size	2,000	16,163	5,350	54,142	1,005	22,454
Age (group)	71 (69–75)	64 (59–75)	71 (63–80)	61 (50–73)	71 (66-79)	60 (50-72)
Gender (Male)	62%	65%	59%	56%	/	/
BMI	26.2 (23.9 – 27.8)	26.1 (23.9 – 27.8)	29 (22.1 - 34.8)	29.0 (22.0 - 33.6)	28.20 (24.3 - 30.6)	26.1 (22.5 - 28.4)
SOFA first 24 h	8.8 (6–11)	5.7 (3–8)	5.3 (3-7)	3.4 (1-5)	7.1 (3-10)	4.3 (2-6)
APACHE II	22.0 (17–26)	16.6 (12–20)	/	/	27.8 (24-32)	20.2 (14-26)
Prediction point (Hours)	91.7 (31.3 – 91.3)	49.2 (16 – 41.0)	64.7 (27.2 - 71.06)	46.6 (19 - 48.0)	80.03 (33.9 - 82.4)	57.1 (18.4 - 51.7)
ICU LOS (days)	15.47 (3.9 - 20.4)	3.7 (0.9 - 2.9)	7.8 (2.9 - 9.0)	3.3 (1.2 - 3.5)	10.3 (3.7 - 11.7)	4.2 (1.0 - 4.0)
ICU survival	46.1%	71.8%	88.9%	94.6%	85.7%	92.8%

the whole dataset were excluded, resulting in 194 remaining variables. Subsequently, aggregate features were constructed to quantify the trend over time, such as min, mean, max, and slope. Powershap [9] was used as the feature selection method to determine the final feature set (the full procedure and final hyperparameters are available in OS Section 3). For a proper comparison across datasets, features in the feature set after model development that were not available in the external datasets were dropped. The impact of dropping these features is documented in the OS Section 4. The final list of features for all models can be found in OS Section 5.

2.2. External validation

Two datasets were used for external validation. The first one is the publicly available Medical Information Mart for Intensive Care (MIMIC-IV) [10]. The second dataset comprised all patients admitted between 2013 and 2020 to the ICU of Ghent University Hospital (GUH), a tertiary Belgian hospital with a total of 52 surgical and medical ICU beds. The study population in both datasets was also defined as described in Section 2.1.1.

Several approaches were used to evaluate different hypotheses during external validation. To test 'out-of-the-box-readiness' and robustness, the model trained on the AmsterdamUMCdb data was directly applied to the external datasets (denoted as Direct - MIMIC-IV and Direct – GUH).

The second hypothesis was that the prediction models could be transferred between hospitals and adapted to local customs without a detrimental drop in performance or the need for complete redevelopment. To test this, the model framework developed on AmsterdamUMCdb (type of model, hyperparameters, definitive feature set, etc.) was retrained using the same one-to-one matching training approach on each external dataset individually to recalibrate the model (denoted as Recalibrated – MIMIC-IV and Recalibrated – GUH, respectively). These recalibrated models were then evaluated against an unseen test set derived from this same dataset. The same approach as explained in Section 2.1.1 for AmsterdamUMCdb was used to construct the balanced training and test set and model development.

Finally, to determine if adding more data would increase performance, two datasets were created: AmsterdamUMCdb with GUH and AmsterdamUMCdb with MIMIC-IV. These datasets were then used to develop a model using the previously mentioned method.

2.3. Uncertainty calibration

To assess the UQ performance, the uncertainty prediction calibration was measured using the Expected Calibration Error (ECE) to quantify the average absolute calibration error [11]. We also propose a variant, called the Expected Signed Calibration Error (ESCE), to quantify the uncertainty bias. These metrics are the classification variant of the distribution coverage error and absolute distribution coverage error used for regression UQ evaluation [12]. These metrics are used instead of the Brier score because the Brier score does not evaluate the clinical value of diagnostic tests or prediction models [13]. To calculate the ECE and ESCE, the probability output $\hat{p}_i$, which is bounded between 0 and 1, is binned. This will result in M bins in total with a bin size equal to $1/M$. The errors are then calculated by subtracting the average confidence per bin from the accuracy within that bin using the following formulae:

$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i) \tag{1}$$

$$conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \tag{2}$$

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{3}$$

$$ESCE = \sum_{m=1}^{M} \frac{|B_m|}{n} (acc(B_m) - conf(B_m)) \tag{4}$$

B_m is the bin with the samples whose probability output falls into the half-open interval $((1-m)/M, m/M]$. $\hat{y}_i$ and y_i are the predicted resp. true class label for sample i, and n represents the number of samples. These formulae are defined for a single bin size of $1/M$. However, by only using a single bin size these metrics are susceptible to local variance and can therefore only give a limited view of the uncertainty calibration. To understand the total model calibration, the bin size was varied from 0.005 to 0.05 with 0.001 intervals. The mean of the ECE and ESCE over all these varying bin sizes are reported in this manuscript. Furthermore, this method also facilitates proper graphical visualization in calibration plots.

2.4. Prediction over time analysis

As the goal is clinical practice, it is beneficial to understand how the proposed models will work in this setting. Therefore, the models were evaluated throughout the ICU admission on all test patients to understand the model's behavior across time. Hence, every test patient received multiple prediction points throughout their ICU admission instead of one. These moments were defined as percentages of time from the original diagnostic/surrogate AF prediction point (T_AF_Event) determined after one-to-one matching: $f * T\_AF\_Event$ with f a fraction between 0 and 1. These fractions started at admission and were incremented by 0.04 until the original prediction point was reached. This resulted in 26 evaluations across time for every test patient. The label was the same across all these prediction points, i.e., AF or no-AF. With this method, it is possible to verify whether or not the models provide a risk estimate, characterized by a steady or slightly increasing performance across time. Furthermore, to evaluate and compare the performance of each model at the same time instant, every model was evaluated on an unbalanced dataset 24 to 36 hours before the prediction point.

2.5. Shapley analysis

To interpret the final models, the SHAP [14] library was used to understand the predictions and find potential correlated risk factors using Shapley values [14]. A Shapley value is always tied to a feature value of a data sample and represents the impact that feature has on the prediction, compared to the average prediction of the model across all samples.

3. Results

3.1. Internal and external validation

The Model-1.5 internal and external validation metrics are shown in Table 2. The tables for Model-6 and Model-12 can be found in the OS Section 6. The internal validation reached an AUC of 0.81 showing prediction capabilities, even when tested on all patients. Additionally, models with a time window closer to the AF occurrence performed better than models with a more distant time window (Model-1.5 AUC = 0.81 (Table 2), Model-6 AUC = 0.79 (OS Table 15), Model-12 AUC = 0.78 (OS Table 16)). Directly applying the designed models to the unseen datasets resulted in an expected performance drop, represented by the imbalanced recall metrics. The recalibrated models, however, showed comparable performance to the models developed on AmsterdamUMCdb for the MIMIC-IV dataset, and even better performance on the GUH dataset. Looking at the ECE and ESCE, the models showed sufficient uncertainty calibration with at most an average 5% error (ECE) in UQ for the internal results without a high bias (ESCE). However, the UQ performance on the GUH dataset was worse reaching up to 9% error for Recalibration-GUH. Fig. 2 visualizes this calibration performance of Model-1.5 across all bins on the internal dataset. This figure shows that the calibration is accurate but slightly conservative. For example, the figure suggests that a prediction of 90% probability for a certain class is in 95% of the predictions correct.

3.2. Evaluation over time

Fig. 3 and Fig. 4 display how Model-1.5 performs over the entire admission period. Fig. 3 demonstrates that the model can achieve a class-weighted average recall of 65% and an AUC of around 0.7 at 10% of the time to the original prediction point, which is close to ICU admission. These results improve when the prediction point reaches the original AF prediction point, verifying the prediction of risks.

Looking at the individual level, shown in Fig. 4, these trends continue. Here, the predicted probability of AF changes across time in

Table 2
The results of Model-1.5. A = All patients, B = Balanced test set.

Method	External dataset	Validation patient group	no-AF patients	AF patients	no-AF recall	AF recall	no-AF precision	AF precision	AUC	ECE	ESCE
Internal validation	/	A	14563	400	0.75	0.71	0.99	0.07	0.81	0.04	0.04
		B	400	400	0.74	0.71	0.72	0.73	0.81	0.05	0.02
Direct	MIMIC -IV	A	49863	1070	0.90	0.38	0.99	0.07	0.77	0.12	0.12
		B	1070	1070	0.90	0.38	0.59	0.79	0.76	0.09	-0.08
	GUH	A	19244	183	0.57	0.91	1.00	0.02	0.84	0.15	-0.15
		B	182	183	0.57	0.91	0.87	0.68	0.83	0.08	0.04
Recalibration	MIMIC-IV	A	49863	1070	0.69	0.73	0.99	0.05	0.79	0.02	-0.01
		B	1070	1070	0.67	0.73	0.71	0.69	0.78	0.05	0.02
	GUH	A	19244	183	0.74	0.84	1.00	0.03	0.85	0.04	-0.04
		B	182	183	0.74	0.84	0.82	0.77	0.85	0.09	0.07
Combining Amsterdam UMCdb with external dataset	MIMIC-IV	A - MIMIC-IV	49863	1070	0.70	0.72	0.99	0.05	0.79	0.02	0.00
		B - MIMIC-IV	1070	1070	0.68	0.72	0.71	0.69	0.78	0.05	0.02
		A - UMCdb	14563	400	0.70	0.76	0.99	0.06	0.81	0.02	0.01
		B - UMCdb	400	400	0.69	0.76	0.74	0.71	0.81	0.06	0.04
	GUH	A - GUH	19244	183	0.71	0.83	1.00	0.03	0.84	0.04	-0.03
		B - GUH	182	183	0.71	0.83	0.81	0.74	0.83	0.10	0.08
		A - UMCdb	14563	400	0.73	0.70	0.99	0.07	0.80	0.02	0.02
		B - UMCdb	400	400	0.75	0.70	0.71	0.73	0.81	0.06	0.02

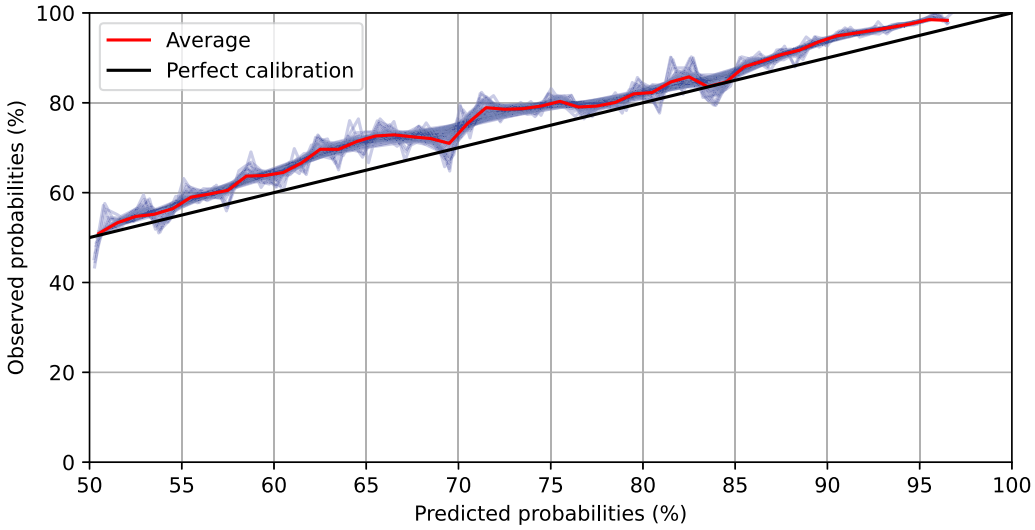

Fig. 2. The Calibration plot for Model-1.5 on all patients evaluated on the AmsterdamUMCdb. The blue lines represent the varying bin sizes from 0.005 to 0.05. The red line is the average calibration of all these bin sizes. The probabilities represent the probability for the predicted class.

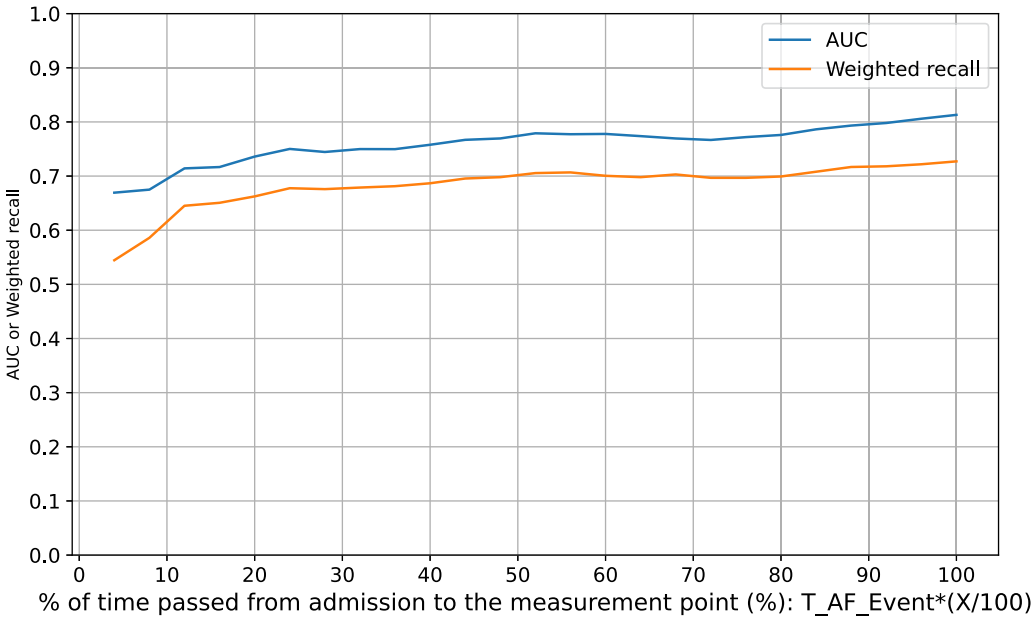

Fig. 3. Evaluation over time for Model-1.5. The weighted recall is the weighted average of the recalls of both classes. The results are calculated using the balanced test set.

response to the features in the dataset. Additionally, the Shapley values visualize the impact on the prediction resulting from these features and their changes across time, enabling a total image of the model's behavior across time. As an example, consider the time frame between 20% and 40% in Fig. 4. The maximum heart frequency rises to 180 bpm resulting in a significantly higher AF risk prediction for that period, indicated by increased Shapley values from 0.0 to 1.0. This is also visible in the AF probability sub-graph, where the risk increases to 90%, and drops to 60% when the heart frequency drops again. Accordingly, the changes in the other features and their respective Shapley values can be analyzed using the same interpretation.

The results of the comparison of all models at the same time instances are shown in Table 3. Model-1.5 has the best AF recall performance, demonstrating Model-1.5 can best find patients at risk for AF compared to the other models (Wilcoxon signed rank test p-value < 0.001).

Table 3
The results of evaluating each model on the unbalanced test set but created with data 24 to 36 hours before the prediction point to compare and evaluate the performance of the models at the same time instant, while being trained on their original dataset. These results are created by bootstrapping this dataset 1000 times and reported in the following format: mean [95% confidence interval].

Model	no-AF recall	AF recall	no-AF precision	AF precision	AUC
Model-1.5	0.68 [0.66-0.69]	0.74 [0.67-0.81]	0.98 [0.98-0.99]	0.10 [0.09-0.11]	0.77 [0.74-0.80]
Model-6	0.66 [0.65-0.68]	0.73 [0.67-0.80]	0.98 [0.98-0.98]	0.10 [0.09-0.10]	0.76 [0.73-0.80]
Model-12	0.68 [0.67-0.70]	0.71 [0.65-0.78]	0.98 [0.97-0.98]	0.10 [0.09-0.11]	0.78 [0.74-0.81]

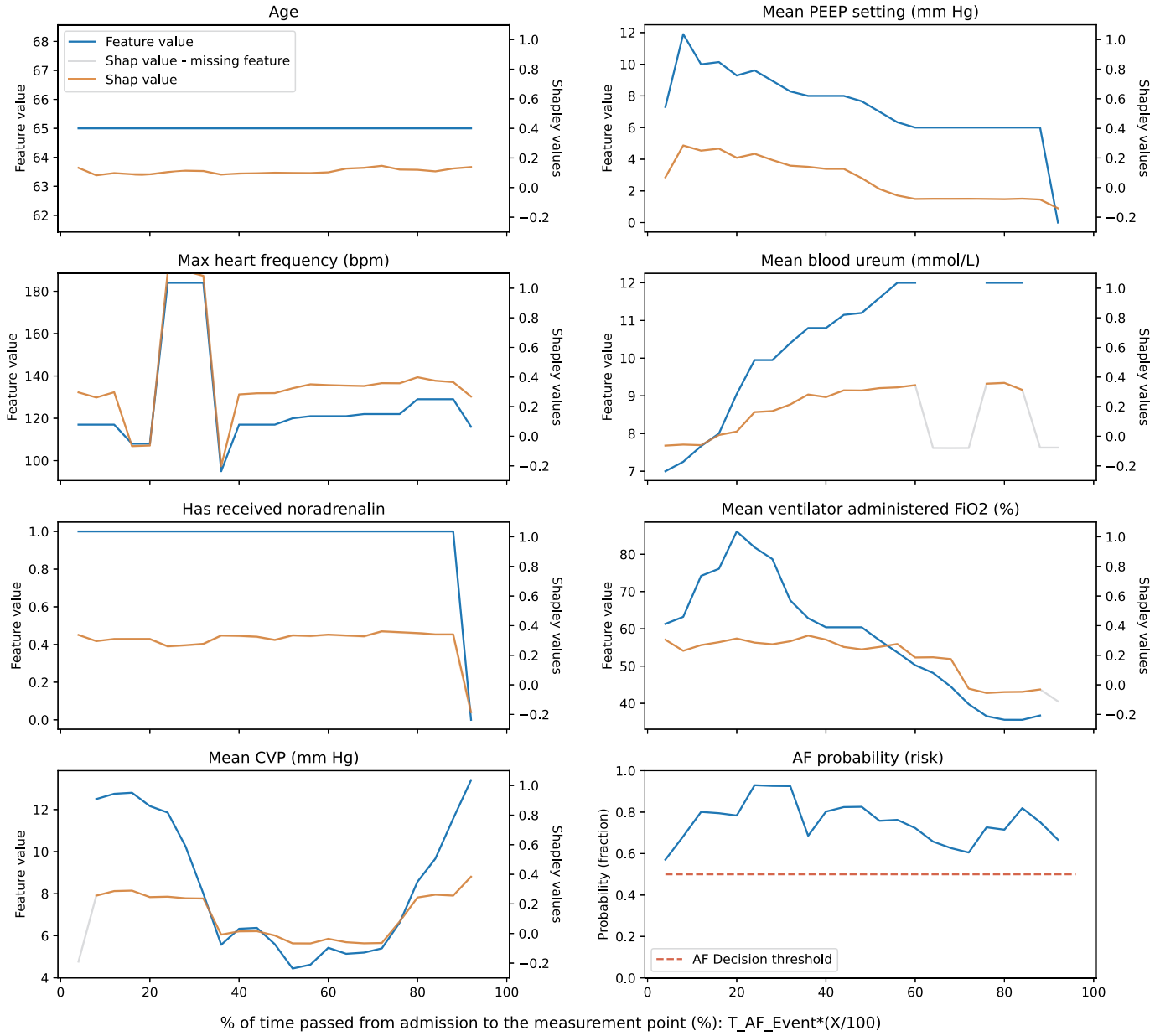

Fig. 4. Evaluation over time for a single AF patient with Model-1.5. Blue = the feature value, orange = the Shapley value, grey = the Shapley value corresponding to the missing feature value, and Red = the decision boundary. The seven most important features are visualized together with their Shapley values. The prediction probability of AF over time is also visualized, where AF is predicted when this probability is above 0.5.

3.3. Shapley analysis

Fig. 5 visualizes the Shapley analysis for the features in the model. This figure demonstrates, e.g., that giving noradrenalin is correlated with a higher risk of AF according to the model. The most important feature here is age, where the model discovered that higher age (red) is a considerable predictor for AF (positive Shapley value). This method also works on the sample prediction level and enables physicians to evaluate both input and output to make an informed decision when using the model.

4. Discussion

We developed risk prediction models for AF occurrence in critically ill patients achieving adequate internal validation performances. Furthermore, retraining the models on combined multi-centric datasets did not improve the performance metrics indicating that the maximum performance was reached with the current features, models, and methods. Previous papers have also aimed to predict the occurrence of AF using ML. For instance, Ortega-Martorell et al. used a similar approach and achieved a comparable AUC (0.836), though without verified risk prediction and external validation [15]. Others tried to leverage electrocardiography (ECG) signals in ML with promising results [16,17]. However, these models focus on diagnosing AF a short time before its occurrence and use ECG data instead of routinely collected healthcare data.

In high-stakes environments, such as the ICU, it is argued that model interpretability and transparency are paramount [18]. The use of 'white-box' models and the Shapley analysis are key factors for both interpretability and transparency. Furthermore, the transparency of the models is additionally reinforced by providing a calibrated risk score between 0 and 1 with low E(S)CE calibration errors. The uncertainty can resonate with the thought process clinicians use when making decisions and therefore contribute to the interpretability and transparency

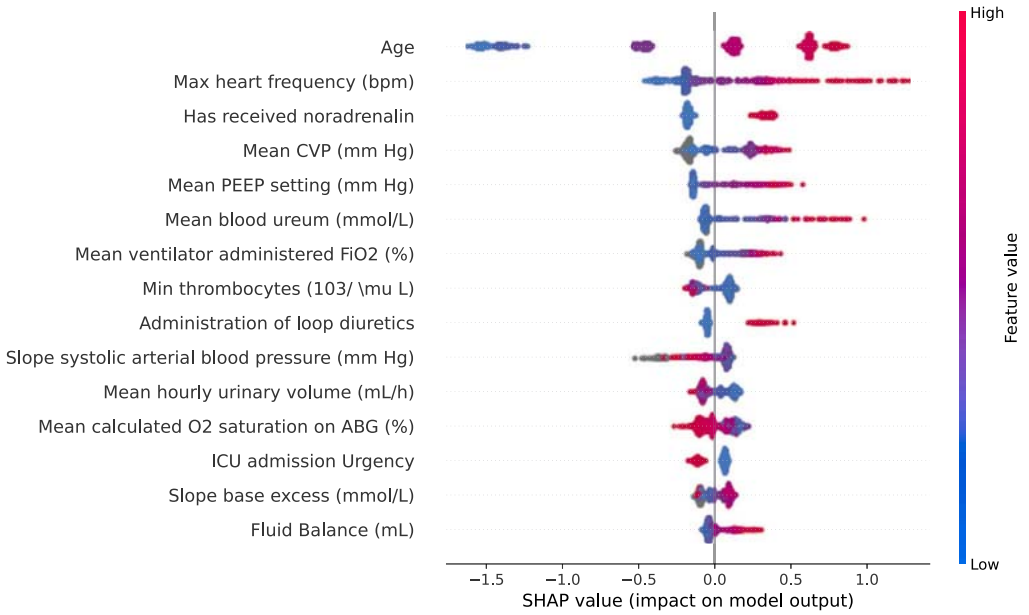

Fig. 5. Shapley analysis of Model-1.5. The grey values are NaNs (missing feature values).

of the models. Combined, all these properties transform the output of the model from a binary occurrence prediction to a continuous risk prediction that facilitates the development of decision support systems. The Shapley analysis facilitates hypothesizing which predictors are risk factors for AF and whether clinicians can influence these to lower the risk of AF. Additionally, given the calibrated UQ results, the analysis of the effects of adjustment of care delivery can be used as a steppingstone for future causal (machine learning) studies.

Although several ML models have been developed in recent years, many fail to be translated into clinical practice [19]. One of the main issues hampering their widespread integration is the lack of external validity when directly applying developed models to unseen data. This also applies to this study, as directly applying our internally validated models to an external dataset was accompanied by a drop in overall performance mainly attributable to a data shift (Experiments are available in OS Section 7). However, using the recalibration methods we could mitigate this issue. By applying this recalibration method, the external validation AUC remained consistent with the internal validation AUC. This demonstrates a suitable strategy to overcome a considerable barrier between AI research and patient care and provide generalizability for many settings.

Our study also has limitations. AF diagnosis was based on nursing charts as electrocardiograms were not available in AmsterdamUMCdb. Therefore, the study depends on the accuracy and timely recording of the diagnosis by the nurses to avoid data leakage. However, this accuracy and timely recording of AF registrations has been studied and found to be adequate [20]. Additionally, Moss et al. found that new-onset AF diagnosed by both clinicians and an ECG model was associated with increased LOS and hospital mortality, whereas new-onset AF diagnosed by only the algorithm was not [4]. Hence, using only nursing chart notes is likely sufficient to capture clinically relevant AF episodes. Classifying ML models that use ECG waveforms have already been developed to improve AF detection and timely diagnosis registration [21,22]. Although adding ECG data could increase the performance, our method enables the clinician to get a risk prediction using only routinely collected data. Finally, as medical history is not recorded in the AmsterdamUMCdb database, the focus was not to discriminate new-onset AF from an AF event in a patient already known with AF.

Table 4
Summary table.

What was already known on the topic	- Atrial Fibrillation (AF) is associated with an increase in length of stay and mortality for patients admitted to the ICU. - Clinical screening tools for broad screening of the ICU population are not routinely used. Machine Learning tools are being developed, but often only identify AF risk several minutes in advance. - In general, few prediction models are being used in clinical practice due to their drop in performance when applied to an unseen environment.
What this study adds to our knowledge	- Design of interpretable and calibrated risk prediction models for identifying patients at risk for AF well in advance and validated on three distinct datasets. - A case-control design to compensate for unbalanced data for model training and evaluation. - Recalibrating the designed models on unseen datasets can achieve an equally good performance as on original data, reducing a barrier to the widespread use of ML models across various ICUs. - Accessible uncertainty prediction complementing the classification prediction with verified calibration using the uncertainty calibration metrics ECE and ESCE.

5. Conclusion

We proposed AF Risk models that provide a calibrated risk score between 0 and 1 for ICU patients. The calibration of these risk scores was verified using the ECE and the ESCE metrics. These models were built using a case-control design for training to facilitate discriminating between AF and no-AF patients to achieve meaningful results of 0.81 AUC. Furthermore, these models were validated on multiple datasets using various validation methods. Among these, the recalibration method was identified to be a reliable method with minimal effort to achieve generalization across ICUs globally. Ultimately, the used methodologies in this article can serve as a step toward the development of clinical AF prediction models across multiple ICUs (Table 4).

CRediT authorship contribution statement

Jarne Verhaeghe and Thomas De Corte did equal work and share the first authorship. They both conceived the study design, designed the models, and wrote the manuscript. Jarne Verhaeghe wrote the code while Thomas De Corte did the medical analysis. Christopher M. Sauer helped design the study and analyze the results. Sofie Van Hoecke, Femke Ongenae, and Jan J. De Waele supervised the study. Paul Elbers enabled the study, provided the AmsterdamUMCdb, and aided in supervision. Tom Hendriks and Olivier W. M. Thijssens provided feedback during the initial parts of the study. All authors reviewed the manuscript.

Declaration of competing interest

Jarne Verhaeghe is funded by the Research Foundation Flanders (FWO, Ref. 1S59522N). Part of the research was funded by the FWO Junior Research project HEROI2C which investigates hybrid machine learning for improved infection management in critically ill patients (Ref. 1881020N). Olivier W. M. Thijssens received funding from Pacmed; he disclosed work for hire. Christopher M. Sauer is supported by the German Research Foundation funded UMEA Clinician Scientist Program (grant FU356/12-2). Jan De Waele is a senior clinical investigator funded by the Research Foundation Flanders (FWO, Ref. 1881020N).

Data availability

The code is available on GitHub using the following link: https://github.com/predict-idlab/atrial_fibrillation_prediction. The AmsterdamUMCdb dataset is available by request. More information can be found here: https://github.com/AmsterdamUMC/AmsterdamUMCdb. The GUH dataset can only be shared after the ethical approval of the GUH ethical board. The MIMIC-IV dataset is available by request on the following link: https://physionet.org/content/mimiciv/2.1/.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2023.105086.

References

[1] T. Yoshida, T. Fujii, S. Uchino, M. Takinami, Epidemiology, prevention, and treatment of new-onset atrial fibrillation in critically ill: a systematic review, J. Intensive Care 3 (1) (Apr. 2015) 19, https://doi.org/10.1186/s40560-015-0085-4, zSCC: 0000062.

[2] J.W. Greenberg, T.S. Lancaster, R.B. Schuessler, S.J. Melby, Postoperative atrial fibrillation following cardiac surgery: a persistent complication, Eur. J. Cardiothorac. Surg. 52 (4) (2017) 665–672, https://doi.org/10.1093/ejcts/ezx039, zSCC: 0000138.

[3] P.M.C. Klein Klouwenberg, J.F. Frencken, S. Kuipers, D.S.Y. Ong, L.M. Peelen, L.A. van Vught, M.J. Schultz, T. van der Poll, M.J. Bonten, O.L. Cremer, Incidence, predictors, and outcomes of new-onset atrial fibrillation in critically ill patients with sepsis. A cohort study, Am. J. Respir. Crit. Care Med. 195 (2) (2017) 205–211, https://doi.org/10.1164/rccm.201603-0618OC.

[4] T.J. Moss, J.F. Calland, K.B. Enfield, D.C. Gomez-Manjarres, C. Ruminski, J.P. DiMarco, D.E. Lake, J.R. Moorman, New-onset atrial fibrillation in the critically ill, Crit. Care Med. 45 (5) (2017) 790–797, https://doi.org/10.1097/CCM.0000000000002325, zSCC: 0000096.

[5] W. Zhang, W. Liu, S.T.H. Chew, L. Shen, L.K. Ti, A clinical prediction model for postcardiac surgery atrial fibrillation in an Asian population, Anesth. Analg. 123 (2) (2016) 283–289, https://doi.org/10.1213/ANE.0000000000001384.

[6] A. Alonso, B.P. Krijthe, T. Aspelund, K.A. Stepas, M.J. Pencina, C.B. Moser, M.F. Sinner, N. Sotoodehnia, J.D. Fontes, A.C.J.W. Janssens, R.A. Kronmal, J.W. Magnani, J.C. Witteman, A.M. Chamberlain, S.A. Lubitz, R.B. Schnabel, S.K. Agarwal, D.D. McManus, P.T. Ellinor, M.G. Larson, G.L. Burke, L.J. Launer, A. Hofman, D. Levy, J.S. Gottdiener, S. Kääb, D. Couper, T.B. Harris, E.Z. Soliman, B.H.C. Stricker, V. Gudnason, S.R. Heckbert, E.J. Benjamin, Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium, J. Am. Heart Assoc. 2 (2) (2013) e000102, https://doi.org/10.1161/JAHA.112.000102, publisher: American Heart Association.

[7] P.J. Thoral, J.M. Peppink, R.H. Driessen, E.J.G. Sijbrands, E.J.O. Kompanje, L. Kaplan, H. Bailey, J. Kesecioglu, M. Cecconi, M. Churpek, G. Clermont, M. van der Schaar, A. Ercole, A.R.J. Girbes, P.W.G. Elbers, Sharing ICU patient data responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: the Amsterdam University Medical Centers database (AmsterdamUMCdb) example, Crit. Care Med. 49 (6) (2021) e563, https://doi.org/10.1097/CCM.0000000000004916, zSCC: NoCitationData[s0].

[8] A.V. Dorogush, V. Ershov, A. Gulin, CatBoost: gradient boosting with categorical features support, arXiv:1810.11363, Oct. 2018 [cs, stat] ZSCC: 0000342.

[9] J. Verhaeghe, J. Van Der Donckt, F. Ongenae, S. Van Hoecke Powershap, A powerfull Shapley feature selection method, arXiv:2206.08394, Jun. 2022, https://doi.org/10.48550/arXiv.2206.08394 [cs, stat].

[10] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L.A. Celi, R. Mark, MIMIC-IV, version Number: 0.4 Type: dataset, https://doi.org/10.13026/A3WN-HQ05, 2022.

[11] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks, arXiv:1706.04599 [cs], Aug. 2017.

[12] J. Verhaeghe, S.A.M. Dhaese, T. De Corte, D. Vander Mijnsbrugge, H. Aardema, J.G. Zijlstra, A.G. Verstraete, V. Stove, P. Colin, F. Ongenae, J.J. De Waele, S. Van Hoecke, Development and evaluation of uncertainty quantifying machine learning models to predict piperacillin plasma concentrations in critically ill patients, BMC Med. Inform. Decis. Mak. 22 (1) (2022) 224, https://doi.org/10.1186/s12911-022-01970-y.

[13] M. Assel, D.D. Sjoberg, A.J. Vickers, The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models, Diagn. Progn. Res. 1 (1) (2017) 19, https://doi.org/10.1186/s41512-017-0020-3.

[14] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017.

[15] S. Ortega-Martorell, M. Pieroni, B.W. Johnston, I. Olier, I.D. Welters, Development of a risk prediction model for new episodes of atrial fibrillation in medical-surgical critically ill patients using the AmsterdamUMCdb, Front. Cardiovasc. Med. 9 (2022).

[16] S.K. Bashar, E.Y. Ding, A.J. Walkey, D.D. McManus, K.H. Chon, Atrial fibrillation prediction from critically ill sepsis patients, Biosensors 11 (8) (2021) 269, https://doi.org/10.3390/bios11080269.

[17] A. Narin, Y. Isler, M. Ozer, M. Perc, Early prediction of paroxysmal atrial fibrillation based on short-term heart rate variability, Phys. A, Stat. Mech. Appl. 509 (2018) 56–65, https://doi.org/10.1016/j.physa.2018.06.022.

[18] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) 206–215, https://doi.org/10.1038/s42256-019-0048-x, publisher: Nature Publishing Group.

[19] L.M. Fleuren, P. Thoral, D. Shillan, A. Ercole, P.W.G. Elbers, M. Hoogendoorn, B. Gibbison, T.L.T. Klausch, T. Guo, L.F. Roggeveen, E.L. Swart, A.R.J. Girbes, Right data right now collaborators, machine learning in intensive care medicine: ready for take-off?, Intensive Care Med. 46 (7) (2020) 1486–1488, https://doi.org/10.1007/s00134-020-06045-y.

[20] E.Y. Ding, D. Albuquerque, M. Winter, S. Binici, J. Piche, S.K. Bashar, K. Chon, A.J. Walkey, D.D. McManus, Novel method of atrial fibrillation case identification and burden estimation using the MIMIC-III electronic health data set, J. Intensive Care Med. 34 (10) (2019) 851–857, https://doi.org/10.1177/0885066619866172.

[21] C.A. Millán, N.A. Girón, D.M. Lopez, Analysis of relevant features from photoplethysmographic signals for atrial fibrillation classification, Int. J. Environ. Res. Public Health 17 (2) (Jan. 2020), https://doi.org/10.3390/ijerph17020498, zSCC: 0000005.

[22] T. Rieg, J. Frick, H. Baumgartl, R. Buettner, Demonstration of the potential of white-box machine learning approaches to gain insights from cardiovascular disease electrocardiograms, PLoS ONE 15 (12) (Dec. 2020), https://doi.org/10.1371/journal.pone.0243615.