Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 58-24 (2024) 129–134

Retrospective Classification of ARDS in ICU Time-series data using Random Forest with a focus on Data Pre-processing

Simon Fonck [*,**] **Sebastian Fritsch** [**,***,****] **Hannes Pieper** [*] **Alexander Baron** [*] **Stefan Kowalewski** [*] **André Stollenwerk** [*,**]

[*] *Informatik 11 - Embedded Software, RWTH Aachen University, 52074 Aachen, Germany (**Corresponding Author**: fonck@embedded.rwth-aachen.de)*
[**] *Center for Advanced Simulation and Analytics (CASA), Forschungszentrum Jülich, 52428 Jülich, Germany*
[***] *Department of Intensive Care Medicine, University Hospital RWTH Aachen, 52074 Aachen, Germany*
[****] *Jülich Supercomputing Centre, Forschungszentrum Jülich, 52428 Jülich, Germany*

Abstract: Acute Respiratory Distress Syndrome (ARDS) is a severe lung injury associated with high mortality. Epidemiological studies have shown that ARDS is often diagnosed too late or not at all. Artificial intelligence (AI) can help clinicians identify ARDS and initiate appropriate therapy earlier. Various data must be collected and processed for the training of such AI methods. It is particularly important to consider the data basis and describe the pre-processing steps of the data, as this has a major influence on the results of an AI model. A random forest algorithm is proposed to automatically assess a patient's condition for compatibility with an ARDS using time-series data (like vital signs, laboratory values and other parameters). We emphasize the data preparation and its influence on the results. The model achieved moderate to excellent results depending on the preparation and dataset.

Keywords: Medical application, Acute Respiratory Distress Syndrome, Artificial Intelligence, Secondary data, Random Forest

1. INTRODUCTION

Acute Respiratory Distress Syndrome (ARDS) is a severe medical condition first described by Ashbaugh et al. (1967). The syndrome affects around 10 % of mechanically ventilated patients in intensive care units (ICU) (Bellani et al. (2016)). It is characterized by a rapidly evolving, massively impaired gas exchange due to extensive inflammation of the lung tissue. During the Covid-19 pandemic, ARDS has become increasingly relevant for research and treatment because an underlying infection can lead to an onset of the syndrome (Gorman et al. (2022)). According to Bellani et al. (2016), the in-hospital mortality of all ARDS patients across all severities is approximately 40 %. Since 2012, the condition has been diagnosed according to the Berlin Definition (BD) by Ranieri et al. (2012), which includes four criteria that must be met:

1. **Timing**: Onset within one week of a known clinical insult or new or worsening respiratory symptoms
2. **Origin of edema**: Respiratory failure that cannot fully be explained by cardiac failure or fluid overload
3. **Chest imaging**: Bilateral opacities, which cannot be explained by effusions, lung collapse or nodules
4. **Oxygenation**: Horovitz-quotient (or P/F ratio) lower than 300 mmHg with a PEEP higher or equal than 5 cmH_2O. The P/F ratio is calculated as follows:

$$P/F\ ratio = \frac{P_aO_2}{F_iO_2}, \tag{1}$$

where P_aO_2 is the oxygen arterial partial pressure and F_iO_2 denotes the inspiratory oxygen fraction set in the mechanical ventilator.

Despite its high relevance for critical care and its clear defined diagnostic criteria, Bellani et al. (2016, 2020) discovered that ARDS is significantly under-recognized in daily routine. In a large clinical study, Bellani et al. (2016) found out, that approximately one third of all ARDS cases were not diagnosed. In mildly affected patients, the diagnosis rate was as low as 50 %, leaving half of all mild ARDS cases unrecognized. Other studies showed even worse detection rates in earlier years (Ferguson et al. (2005); Fröhlich et al. (2013)). To improve this situation, several artificial intelligence (AI) methods have been proposed in the literature for the detection and assisted treatment of ARDS (Rashid et al. (2022)). In particular, the third (Sjoding et al. (2021); Fonck et al. (2023)) and fourth (Rashid et al. (2022)) criteria are often addressed promoting the use of AI-assisted detection methods. Training and evaluation of such AI methods requires a large amount of data, whose collection is time-consuming and poses regulatory challenges. Therefore, published secondary databases, i.e. data that is collected and reused for another purpose, have become increasingly important in recent years as

10.1016/j.ifacol.2024.11.024

they allow for better reproducibility and are in line with the FAIR principles by Wilkinson et al. (2016). Despite the necessity and benefits of using secondary databases, several challenges arise. First, information on patients' diagnoses is limited. In some databases, only the International Classification of Diseases (ICD) codes are published without a timestamp during the admission time indicating a specific onset of the diagnosis; other databases do not contain any diagnostic information at all, making it difficult to pre-select patients for supervised learning. Second, data from large healthcare-related databases inherit the undetected ARDS cases as missing diagnosis annotations. Third, there are many missing data samples because not all parameters are collected with the same granularity and sensor errors can lead to inconsistent values. To address these challenges, we propose a data processing concept to implement a Random Forest (RF) algorithm for the retrospective classification of ARDS in intensive care secondary data.

2. STATE OF THE ART

The importance and status of AI methods for the detection of ARDS is shown by the systematic review by Rashid et al. (2022) in which a total of 19 studies on this topic are identified and presented. 53 % of the found articles dealt with the diagnosis of ARDS in time series and radiological images. All developed models showed good to excellent results on commonly used performance metrics such as area under the receiver operating characteristic curve (AUC) or sensitivity. Although no RF algorithm was used to directly detect ARDS, the predictive ability of the method was demonstrated in a study by Ding et al. (2019). In their work, an RF was implemented for an early prediction of ARDS on the first day of hospitalization. The model incorporated various vital signs and laboratory data from 296 patients collected between 2011 and 2014. The average AUC values were 0.82 and the predictive accuracy was 0.83. In 2022, Pai et al. (2022) used an RF algorithm in an ensemble method for assisted diagnosis of ARDS in intensive care time series data as well as radiographs. In their work, three different machine learning (ML) algorithms, eXtreme Gradient Boosting, RF and Logistic Regression, were combined with a convolutional neural network to determine a probability for the presence of ARDS. A total of 1,577 patients from the Taichung Veterans General Hospital in Taiwan with vital signs, laboratory data and chest radiographs were used for final data analysis. The ensemble method of all models together achieved AUC values of 0.925. The RF algorithm achieved accuracy of 0.840, sensitivity of 0.791, specificity of 0.855 and AUC of 0.910. These studies demonstrate the high potential of using RF for the classification of ARDS. Unfortunately, the models and data from each study are not available, making it difficult to extend or reproduce the results. For this reason, we present the implementation of an RF using primarily publicly available databases. In the following section, we concentrate on the pre-processing concept of the data in order to increase reproducibility and describe the influence of this pre-processing on the results of the RF algorithm.

3. DESIGN & IMPLEMENTATION

Even though RF has already been used in the context of ARDS classification, we would like to emphasize the data preparation as an addition to support reproducibility. We decided to use an RF algorithm because it provides promising results in the literature and can be used well for binary classification. RF is an ensemble method, that constructs multiple decision trees during training and outputs the average prediction of the individual trees (Breiman (2001)). Each tree is trained on a random subset of the training features (Bagging), which reduces the risk of overfitting. The implementation is structured into two main parts: First, the data preparation including extraction, preprocessing, filtering and feature selection, and second, the model learner, which encompasses the training and evaluation of the models using various datasets (see Fig. 1). Each of the substeps are described in the following subsections.

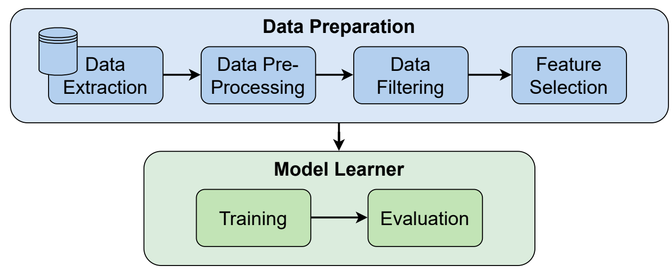

Fig. 1. Workflow for the RF algorithm with data preparation (blue) and the model learner (green).

3.1 Data Extraction

In order to test the generalizability and transferability of the proposed RF algorithm, we used three different intensive care databases for training and evaluation: MIMIC-IV, eICU, and a database provided by the University Hospital RWTH Aachen (UHA). The first two databases are published by the PhysioNet (Goldberger et al. (2000)).

Johnson et al. (2021) The MIMIC-IV database contains data from admissions to the Beth Israel Deaconess Medical Center in Boston, USA collected from 2008 to 2019, including information on patients, admissions, laboratory measurements, diagnoses, and other data. There are 523,740 distinct admissions in the database.

Pollard et al. (2018) The eICU database contains data from admissions to 208 different hospitals in the United States. The data were collected by the Philips eICU program between 2014 and 2015, including information on 200,859 distinct admissions.

UHA The UHA database is provided by the Data Integration Center of the University Hospital Aachen and contains data of patients admitted to the intensive care units of this hospital, including 13,067 distinct admissions. The data were collected and processed as part of the ASIC study (Marx et al. (2021)).

3.2 Data Pre-Processing

To select relevant parameters for ARDS detection, the Catalog of Items from the ASIC study (Marx et al.

(2021)) was used. In this catalog, health care professionals defined a large number of vital signs, laboratory values, medications and diagnoses that could be relevant for the detection of ARDS. In total, 107 parameters were extracted from the databases (see Supplementary Material S1: Selected Parameters). It is important to note that not all parameters were available in all three databases. However, some of them could be calculated using existing data. Because ARDS does not have a single, unique occurrence, a period of up to 10 days was considered for each hospital admission. For this purpose, a possible onset was determined by using the lowest mean P/F ratio calculated using a sliding window. Values for all available parameters from the 7 days prior and 3 days after were then included to capture the possible cause and outcome of ARDS. Since most of the parameters are available with different frequencies, there are many missing values when combining the parameters into a multivariate time series as model input. To fill these gaps, time windows were defined depending on the frequency of the data (from every 4 hours for higher granularity data to once per analyzed window for very low granularity data - see Supplementary Material S1: Selected Parameters). Within these time windows, the arithmetic mean for the parameter is calculated to downsample the data (see Fig. 2). As the data granularity also varies greatly between patients, data points may remain empty during this process. Given that RF cannot handle undefined data points, these missing values were imputed with the implausible value -100,000.

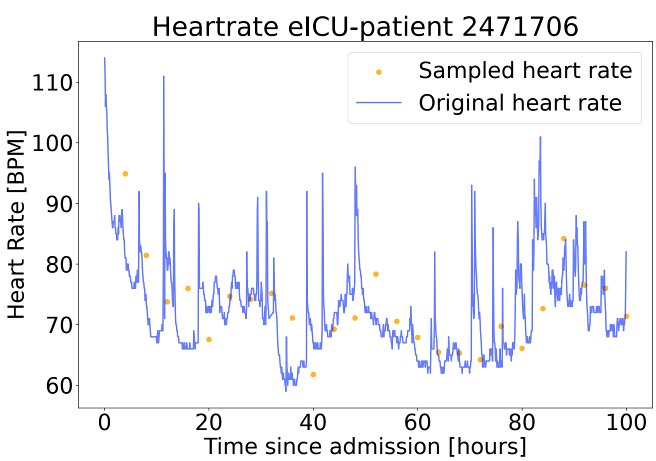

Fig. 2. Data Sampling for the heart rate of patient 2471706, where the arithmetic mean is calculated for the heart rate every 4 hours. Blue is the raw data and Orange are the calculated parameters, that are used as input for the algorithm.

3.3 Data Filtering

After data pre-processing, various filters are applied to the data to remove possible outliers and inconsistencies (see Fig. 3). Initially, all patients under 18 and who have been in the ICU for less than 2 hours are filtered in accordance to the ASIC study (Marx et al. (2021)). In addition, all admissions for which no ICD-10 codes, P/F ratios and PEEPs are available/calculable are excluded, as they cannot be clearly classified. The resulting group of admissions is divided into potential ARDS and Non-ARDS cases using the ICD-10 code "J80". As mentioned in the introduction, there is some uncertainty in the labeling of Non-ARDS patients due to underrecognition of ARDS. To address this issue, additional filters were applied on the data to remove admissions where there is a potential uncertainty. First we excluded all admissions from the ARDS group, where there is no P/F ratio lower than 300 mmHg, as this would contradict the BD by Ranieri et al. (2012) (**BD filter**). For the Non-ARDS group we implemented two filters: a **Lite** (excludes a smaller amount of Non-ARDS patients) and a **Strict** (excludes a larger amount of Non-ARDS patients) one. The **Strict filter** excludes all Non-ARDS admissions, where the patient has a P/F ratio below 200 mmHg, an according PEEP value above 5 cmH_2O, as these patients suffer from hypoxemia, which would be consistent with an ARDS. The **Lite filter** also checks whether one of the exclusion diagnoses listed in the second criterion of the BD is present, as an ARDS would be likely in such cases. These filters are evaluated and discussed later on. This filtering results in 3 different datasets for each database (see Table 1).

Database	Filter	ARDS	Non-ARDS
UHA	No Filter	1.001	11.733
	BD + Lite	1.000	9.151
	BD + Strict	1.000	1.977
eICU	No Filter	606	6.949
	BD + Lite	584	3.912
	BD + Strict	584	3.193
MIMIC-IV	No Filter	173	7.534
	BD + Lite	163	5.932
	BD + Strict	163	4.475

Table 1. Number of ARDS and Non-ARDS patients based on the according filter (see Subsection 3.3).

3.4 Feature Selection

The next step is an optional feature selection, which reduces the feature space and can therefore improve model accuracy, reduce overfitting and speed up the training process. After data pre-processing and down sampling, a total of over 2500 features is available to serve as input for the RF algorithm. Four different options were tested for feature selection: Without feature selection, with mutual information (the top 30 % and the top 60 % of features) and the embedded feature importance from the RF. After the relevant features are identified, all other features are dropped from the dataset.

3.5 Model Learner: Training & Evaluation

The model learner is split into the training and evaluation phase (see Fig. 1). Before the training phase, the described datasets from all three databases (see Table 1) are split into a training and a test dataset (75 %/25 %). The training dataset of each database is then used to perform a 5-fold cross-validation. The best resulting model is applied to the test datasets of all three databases. This allows for an internal validation and furthermore for an external validation to check the generalizability of the algorithm. The RF is implemented using the sk-learn library (Pedregosa et al. (2011)). The results are described and discussed in the following sections.

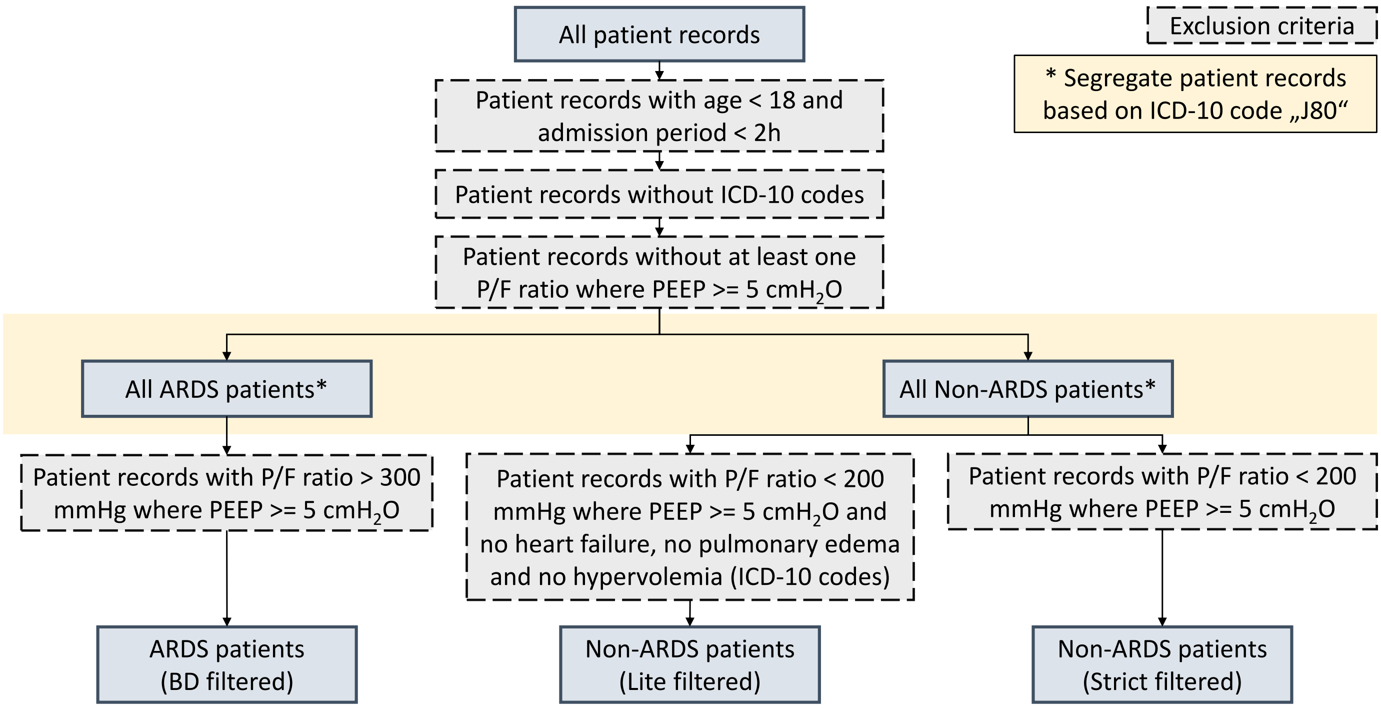

Fig. 3. Exclusion and filter criteria for data pre-processing. The first three exclusion criteria are used for all patient admissions. On the bottom left, the filter **BD** (for Berlin-Definition) is described. On the Non-ARDS side, the **Lite** and **Strict** filter are displayed. The yellow box encompasses the segregation by ICD-10 codes and the admission groups used for the training without filters.

4. RESULTS

The RF algorithm trained with the datasets from Table 1 achieved varying results based on the according feature selection and filtering. Exemplary results for the cross-validation of the RF algorithm using the embedded feature selection can be seen in Table 2. The according ROC curve for the UHA dataset is presented in Fig. 4 (See Appendix A.1 and Appendix A.2 for the MIMIC-IV and eICU results). For all datasets used, the model with all filter options has a very high specificity, which shows that almost all negative cases were correctly identified as such. The algorithm performs slightly worse for the positive cases. In particular, without filtering and with the light filter, the sensitivity is lower. Overall, the rather high AUC indicates that the classifier achieves a reliable classification. The best model for the 5-fold cross-validation (Filter BD + Strict) is then used for an external validation on the test datasets of all databases (see Table 3). The external validation shows that the respective models can be transferred well to the test data of each database and are therefore generalizable.

Database	**Filter**	**Sen.**	**Spec.**	**AUC**
UHA	No Filter	0.384	0.976	0.895
	BD + Lite	0.449	0.964	0.885
	BD + Strict	0.973	1.000	0.999
eICU	No Filter	0.363	0.951	0.848
	BD + Lite	0.625	0.953	0.933
	BD + Strict	0.932	0.988	0.985
MIMIC-IV	No Filter	0.584	0.922	0.892
	BD + Lite	0.672	0.958	0.945
	BD + Strict	0.884	0.981	0.985

Table 2. Mean Sensitivity (Sen.), Specificity (Spec.) and AUC for the internal cross-validation for each database using the different filtered datasets for the RF feature selection.

5. DISCUSSION

In general, our results are consistent with those published in the related literature, where an RF algorithm has already been used for the classification and early prediction for ARDS in time-series data. However, using the unfiltered data we achieve only moderate sensitivity, which may be due to the homogeneity or imbalance of the two patient groups. When the data are filtered however, the algorithm produces excellent results in all considered metrics. The question remains to what extent the population in the non-ARDS group still corresponds to reality. These research results show the importance of data pre-processing as it has a major impact on the training and performance of ML algorithms. Often, the principle of "garbage in, garbage out" is used to address the poor data quality that also affects the results of AI methods (Kilkenny and Robinson (2018)). To generate truly robust results from AI methods in the medical context, we need to focus more on the quality of the data and its annotations. Apart from the dependence of the results on the dataset, our work shows that RF is a suitable method for retrospectively classifying ARDS patients. Furthermore, the trained models can be used for analysis of diagnoses in intensive care datasets to improve annotation. This would require external validation by medical experts to ensure efficient use of secondary data. Also, the transferability to other datasets without

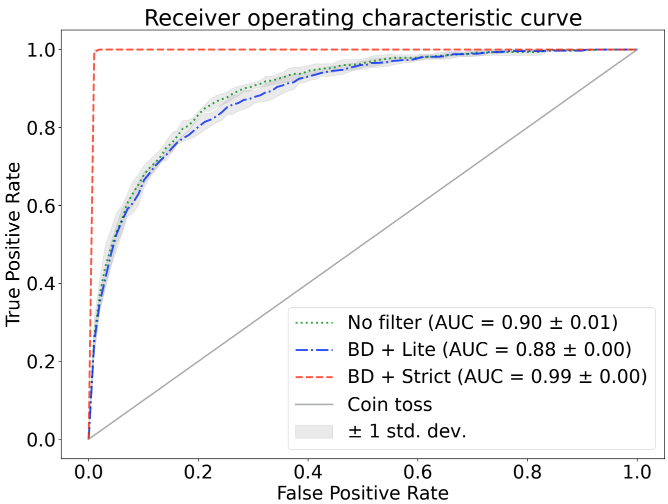

Fig. 4. ROC curves for the internal cross-validation for the UHA database without filters and with RF feature selection.

Model	Testset	Sen.	Spec.	AUC
UHA	UHA	0.968	1.000	1.000
	eICU	0.875	1.000	0.981
	MIMIC-IV	0.825	0.992	0.989
eICU	UHA	1.000	0.994	0.999
	eICU	0.943	0.987	0.990
	MIMIC-IV	0.900	0.989	0.987
MIMIC-IV	UHA	1.000	0.964	0.999
	eICU	0.943	0.971	0.983
	MIMIC-IV	0.900	0.978	0.992

Table 3. Sensitivity (Sen.), Specificity (Spec.) and AUC for the models evaluated with the testdata of all databases (RF feature selection).

much loss of performance indicates the generalizability of the algorithm. Surprisingly, the model initially trained on the MIMIC-IV data performed even better on the other test data than on the internal test set. This may be due to better data quality and higher resolution of the relevant data, which should be investigated in further research. Another aspect that should be addressed in future work is the subdivision of ARDS patients into different severity levels in order to analyze whether a multiclass classification is possible with the proposed algorithm.

Our research also has limitations: The problem of filtering the data has already been addressed and should be improved in further work using external physician validation. In addition, the database for ARDS patients, especially the MIMIC-IV database, is very small. With the help of other intensive care databases, the training and evaluation process could be further extended. The UHA dataset used is not publicly available at the time of this publication, which limits reproducibility. We hope to remedy this in the future. An easy understanding of the algorithm is not given at the moment, as the model acts as a black box. By incorporating explanatory methods, we hope to extend the application in the future and make it more comprehensible, as already shown by Raab et al. (2023) for the use case of seizure detection in EEG time series data.

6. CONCLUSIONS

ARDS is a critical lung disease with a high mortality rate. One reason for this is the late and often missed diagnosis. AI can assist physicians in the diagnostic process. These methods can be trained and evaluated using (published) secondary databases. However, there are some challenges that need to be openly described and discussed. Using three intensive care datasets, we implemented an RF algorithm for the retrospective detection of ARDS and explained the data processing procedure. Our results are comparable to and, depending on the dataset, even better than those in the related literature. In this work, we wanted to focus on the data to show the importance of sufficiently addressing the processing steps. In further work, we would like to validate our approach with more datasets and use the models to test or possibly improve the uncertainty of the annotations. In addition, other ML methods such as Bayesian Networks, Support Vector Machines or Neural Networks can be used for ARDS classification to extend this classification. In the medical context, the comprehensibility of these algorithms is of particular importance and should be supported by explanatory methods. The results should also be validated externally by integrating doctors or other datasets.

7. CODE & DATA AVAILABILITY

The code and supplementary material used for this research is available on the following github: https://github.com/embedded-software-laboratory/RF-for-ARDS-classification. The data can be accessed using the provided references: Johnson et al. (2021); Pollard et al. (2018). The data collected during the ASIC project is currently not available, but is planned to be published in the future.

ACKNOWLEDGEMENTS

Conflict of interest: Authors state no conflict of interest. Ethical approval: The conducted research is not related to either human or animals use. Simulations were performed with computing resources granted by RWTH Aachen University under project rwth1474.

REFERENCES

Ashbaugh, D.G., Bigelow, D.B., Petty, T.L., and Levine, B.E. (1967). Acute respiratory distress in adults. *The Lancet*. doi:10.1016/s0140-6736(67)90168-7.

Bellani, G., Laffey, J.G., Pham, T., Fan, E., Brochard, L., et al. (2016). Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *Journal of the American Medical Association*, 315(8), 788–800. doi:10.1001/jama.2016.0291.

Bellani, G., Pham, T., and Laffey, J.G. (2020). Missed or delayed diagnosis of ards: a common and serious problem. *Intensive care medicine*, 46(6), 1180–1183. doi:10.1007/s00134-020-06035-0.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. doi:10.1023/A:1010933404324.

Ding, X.F., Li, J.B., Liang, H.Y., Wang, Z.Y., et al. (2019). Predictive model for acute respiratory distress syndrome events in icu patients in china using machine learning algorithms: a secondary analysis of a cohort study. *Journal of translational medicine*, 17(1), 326. doi: 10.1186/s12967-019-2075-0.

Ferguson, N.D., Frutos-Vivar, F., Esteban, A., Fernández-Segoviano, P., et al. (2005). Acute respiratory distress syndrome: underrecognition by clinicians and diagnostic accuracy of three clinical definitions. *Critical care medicine*, 33(10), 2228–2234. doi: 10.1097/01.ccm.0000181529.08630.49.

Fonck, S., Fritsch, S., Nottenkämper, G., and Stollenwerk, A. (2023). Implementation of resnet-50 for the detection of ards in chest x-rays using transfer-learning. *Proceedings on Automation in Medical Engineering*, 2(1), 741–742.

Fröhlich, S., Murphy, N., Doolan, A., Ryan, O., and Boylan, J. (2013). Acute respiratory distress syndrome: underrecognition by clinicians. *Journal of critical care*, 28(5), 663–668. doi:10.1016/j.jcrc.2013.05.012.

Goldberger, A.L., Amaral, L.A.N., Glass, L., Peng, C.K., Stanley, H.E., et al. (2000). Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23), e215–e220. doi: 10.1161/01.CIR.101.23.e215.

Gorman, E., O'Kane, C., and McAuley, D. (2022). Acute respiratory distress syndrome in adults: diagnosis, outcomes, long-term sequelae, and management. *The Lancet*, 400(10358), 1157–1170. doi:10.1016/S0140-6736(22)01439-8.

Johnson, A., Bulgarelli, L., Pollard, T., Celi, L.A., Mark, R., and Horng, S. (2021). MIMIC-IV-ED. doi: 10.13026/77Z6-9W59.

Kilkenny, M.F. and Robinson, K.M. (2018). Data quality:"garbage in–garbage out". doi: 10.1177/1833358318774357.

Marx, G., Bickenbach, J., Fritsch, S.J., Riedel, M., Stollenwerk, A., Fonck, S., et al. (2021). Algorithmic surveillance of icu patients with acute respiratory distress syndrome (asic): protocol for a multicentre stepped-wedge cluster randomised quality improvement strategy. *BMJ Open*. doi:10.1136/bmjopen-2020-045589.

Pai, K.C., Chao, W.C., Huang, Y.L., Sheu, R.K., Chen, L.C., et al. (2022). Artificial intelligence–aided diagnosis model for acute respiratory distress syndrome combining clinical data and chest radiographs. *Digital Health*, 8. doi:10.1177/20552076221120317.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12, 2825–2830.

Pollard, T., Johnson, A., Raffa, J., Celi, L., Mark, R., and Badawi, O. (2018). The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5, 180178. doi: 10.1038/sdata.2018.178.

Raab, D., Theissler, A., and Spiliopoulou, M. (2023). Xai4eeg: spectral and spatio-temporal explanation of deep learning-based seizure detection in eeg time series. *Neural Computing and Applications*, 35(14), 10051–10068. doi:10.1007/s00521-022-07809-x.

Ranieri, V., Rubenfeld, G., Thompson, B., Ferguson, N., Caldwell, E., et al. (2012). Acute respiratory distress syndrome: the berlin definition. *Jama*, 307(23), 2526–2533. doi:10.1001/jama.2012.5669.

Rashid, M., Ramakrishnan, M., Chandran, V.P., Nandish, S., et al. (2022). Artificial intelligence in acute respiratory distress syndrome: A systematic review. *Artificial intelligence in medicine*, 131. doi: 10.1016/j.artmed.2022.102361.

Sjoding, M.W., Taylor, D., Motyka, J., Lee, E., Claar, D., et al. (2021). Deep learning to detect acute respiratory distress syndrome on chest radiographs: a retrospective study with external validation. *The Lancet Digital Health*, 3(6), e340–e348. doi:10.1016/S2589-7500(21)00056-X.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1), 1–9. doi:sdata201618.

Appendix A. ADDITIONAL ROC CURVES

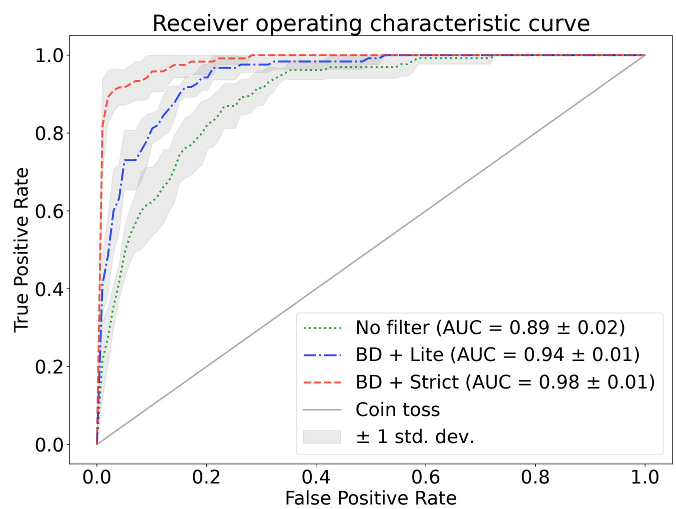

Fig. A.1. ROC curves for the internal cross-validation for the MIMIC-IV database without filters and with RF feature selection.

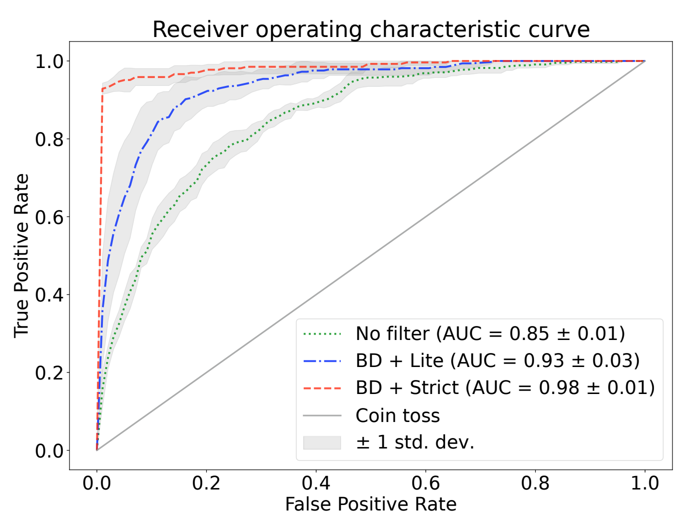

Fig. A.2. ROC curves for the internal cross-validation for the eICU database without filters and with RF feature selection.