Camelia Oprea*, Amrutha Kannan, Lena Olivier, Mark Schoberer, and André Stollenwerk

Comparing a rule-based decision tree to an LSTM network for the detection of ineffective efforts during expiration in neonates

https://doi.org/10.1515/cdbme-2025-0194

Abstract: Detecting Ineffective Efforts during Expiration (IEE) in invasively ventilated neonates could help to improve the ventilation process and lower the discomfort of the patient. We propose an intrinsically explainable rule-based decision tree to detect IEE in high-frequency airway flow data. The algorithm is evaluated and compared to a previously developed Long Short-Term Memory (LSTM) neural network. The decision tree achieves a lower yet comparable performance to the LSTM, while being more stable during training and needing considerably less time for fitting and inferring. Further, the decision tree achieved good results on only 20% of the training data, which in future could reduce annotation labor. Overall, the results encourage us not to disregard rule-based algorithms as outdated, especially when explainability is of importance and annotated data is scarce and time-intensive to acquire.

Keywords: rule-based, LSTM, high-frequency timeseries, IEE, patient-ventilator asynchrony, neonates

1 Introduction

During the mechanical ventilation of neonates different patient-ventilator asynchronies (PVA) may occur. Ineffective efforts during expiration (IEE) are such an asynchrony, in which the patient's inspiratory efforts during the expiration phase are not detected and supported by the ventilator. Multiple occurrences of IEE are stressful for the patient and can lead to a prolonged stay on the ICU [1]. A manual detection is very time- and resource-intensive. Algorithms using ventilator timeseries such as airway flow and pressure can detect IEEs and monitor the amount of asynchronies. For adult patients, various algorithms have been presented, employing both rule-based methods [2, 3] and neural networks [3, 4]. For neonates, however, considerably fewer algorithms are available, such as [5, 6], both employing neural networks. The choice between rule-based algorithms using knowledge-based features and neural networks using learned features is something to consider when designing a new detection algorithm, especially for a high-stake field, such as mechanical ventilation of neonates. For such an application, the explainability and a good generalizability of the algorithm become extremely relevant. Intrinsically explainable algorithms, such as rule-based or simpler machine learning algorithms are often discarded as not being state-of-the-art, which is either motivated by them being limited to expert knowledge or not being able to capture underlying processes well enough. In contrast, neural networks can uncover features embedded in the data, yet they work as black-boxes. Further, the results of deep learning heavily depend on the quality and quantity of the used data. For neonatal applications, the data quantity itself represents a problem, as high-frequency ventilator data is scarce and annotating it is a time-intensive process requiring qualified medical staff.

In this paper, we propose a rule-based decision tree, which combines domain knowledge-based features with a shallow decision tree to form transparent rules for the detection of IEE in invasively ventilated neonates. We evaluate and compare it to a previously developed Long Short-Term Memory (LSTM) neural network, trained on the same data. We employed measures for the classification performance, needed computation time and needed amount of data. Our intent to do this comparison is to highlight the advantages and disadvantages of the two methods, to encourage more careful considerations before developing a new algorithm.

2 Methods

Previously presented rule-based algorithms [2, 3] were developed for adults. Due to the differences in the dynamics of adult and neonatal ventilation data, the algorithms are not easily transferable. Therefore, we developed a new algorithm to identify IEEs using airway flow measurements (125Hz) from ten invasively ventilated neonates. The description of the used data and its annotation can be found in the previous work on the LSTM model [6]. The data set comprises 1,428 breaths ex-

***Corresponding author: Camelia Oprea,** Embedded Software (Informatik 11), RWTH Aachen, Aachen, Germany, oprea@embeddedrwth-aachen.de
Amrutha Kannan, André Stollenwerk, Embedded Software (Informatik 11), RWTH Aachen, Aachen, Germany
Lena Olivier, Mark Schoberer, Neonatology Section of the Department of Paediatric and Adolescent Medicine, RWTH Aachen University Hospital, Aachen, Germany

hibiting IEE and 1,237 non-IEE breaths (normal). The breaths in the dataset vary greatly both in magnitude and duration. To overcome these differences, the breaths were normalized and resampled. Min-max normalization was applied, mapping a breath's flow to the interval *[-1,1]*. Resampling all breaths to the same duration could distort the breath features. Therefore, we plotted the breath durations as a histogram and identified four meaningful bins and corresponding durations (0.6s, 1.12s, 1.44s, 2.16s). A breath is then resampled to one of the four durations depending on the bin it is in. Lastly, to avoid noise being classified as an ineffective effort, the breaths were smoothed using Gaussian smoothing with a factor of 0.8.

After preprocessing, we relied on domain knowledge to extract features and form rules for the detection of IEE, the entire pipeline being shown in Figure 1. Through consultation with experts, we derived that in a normal breath the absolute flow value after the peak expiratory flow (smallest negative flow) would continuously decrease until it reaches zero. In an IEE breath, the absolute flow value first decreases (often faster than in a normal breath due to the inspiratory effort) and then increases again as the remaining air is expired. Combining this theory with observations on the flow measurements, we concluded that often there is no clear single decrease phase followed by a single increase phase but rather that decreases and increases may alternate. Therefore, we first identified the peak expiratory flow and from that point on extracted from the data the segments in which the absolute flow value increases continuously. Decreases smaller than 0.5 ml/min are ignored as measurement noise. Per extracted segment, we computed the absolute flow increment, by subtracting the first flow value from the last flow value. From the resulting segments and corresponding flow increments, we extracted and tested multiple features. Plotting the feature values for the IEE and normal breaths as histograms, only two features distinguished the classes meaningfully. The first feature represents the maximum of flow increment of all segments divided by the peak expiratory flow, to obtain a measure of the ineffective effort's magnitude relative to the breath's magnitude. The second feature is the sum of all computed continuous flow increments.

The two defined features were extracted from the annotated breaths and used to form rules to discriminate between IEE and normal breaths, by comparing the computed features to thresholds. To adequately set the thresholds, we trained a decision tree of depth two taking as input the two features and used the breath annotations as labels. The train-test split of the data was kept the same as for the previously developed LSTM for better comparability. The resulting decision tree delivers simple threshold-based rules to identify IEE breaths. The decision tree and its evaluation on two breaths are shown in Figure 1.

3 Evaluation

In the following, we will present the measures used to evaluate the presented rule-based decision tree and compare it to a previously developed LSTM on the task of classifying breaths as normal or IEE.

Classification performance. The results from both a 10-fold cross validation and the evaluation of the best model from the cross validation on the test data for the two algorithms are given in Table 1. The LSTM outperforms the decision tree in all metrics, both for the mean of the cross-validation and for the best model evaluated on the test data. However, the variances of the metrics are lower for the decision tree than for the neural network, except for the AUROC metric.

Execution time. We next compared the training and inference time of the two algorithms, performed on the same workstation. The training and test data comprised 2,132 breaths and 533 breaths, respectively. The decision tree trained for 0.004 seconds and had an inference time of 0.001 seconds. The LSTM trained for 1,170.38 seconds and had an inference time of 7.80 seconds. Thus, the mean inference time per breath for the rule-based decision tree amounts to $1.87\mu s$ and for the LSTM to $1.15 * 10^4 \mu s$.

Training sample size. Lastly, we estimated the effect different training sample sizes have on the individual algorithms, by fitting the algorithms on varying fractions of the training set. For ten repetitions per fraction, we randomly sampled data from the training set, using different seeds per sample but the same set of seeds for both algorithms. When the fraction is set to one, we cannot repeatedly resample, but rather train the algorithm once on the entire train set. The resulting accuracy using the full train set is lower than the accuracy reported on the test data in Table 1, as the latter was obtained from the best performing cross validation fold. The resulting accuracies per fraction are shown in Figure 2. A clear difference is observed between the performance of the decision tree and that of the LSTM on smaller training sets. Considering the mean accuracy over the ten repetitions, the LSTM needed 60% of the data to reach the same performance the decision tree achieved on 20% of the data. The confidence intervals for the decision tree narrow down with increasing amount of training data, while the LSTM does not have any consistent trend for the confidence intervals.

4 Discussion

The rule-based decision tree's classification performance was outperformed by the LSTM. Specificity was the lowest metric for both models, hinting at a high false positive value (many normal breaths classified as IEE). The specificity should be

Tab. 1: IEE classification performance comparison between rule-based decision tree and LSTM. The first two rows show the results from the 10-fold cross validation (cv) and are given as mean ± std. The second two rows (test) show the best performing model from the cross validation evaluated on the test set.

Method	Accuracy	Precision	Recall	Specificity	F1-Score	AUROC
RB+DT (cv)	88.02 ± 1.88	86.64 ± 2.55	92.03 ± 2.55	83.33 ± 3.82	89.21 ± 1.62	91.49 ± 2.19
LSTM (cv)	89.31 ± 2.52	88.55 ± 3.76	92.21 ± 3.48	85.96 ± 5.76	90.25 ± 2.16	95.35 ± 1.55
RB+DT (test)	90.67	87.42	96.52	83.88	91.75	90.20
LSTM (test)	92.12	92.36	93.01	91.09	92.68	97.40

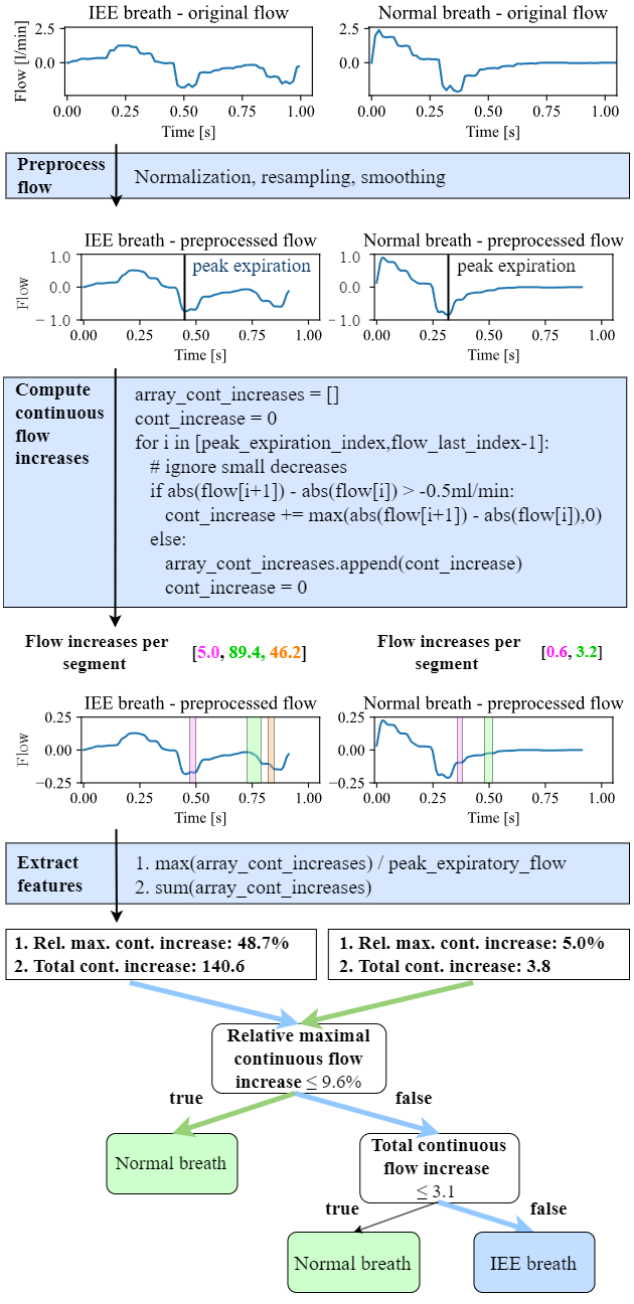

Fig. 1: Data processing pipeline traversed exemplary for an IEE and a normal breath ending with the execution of the decision tree on the extracted features.

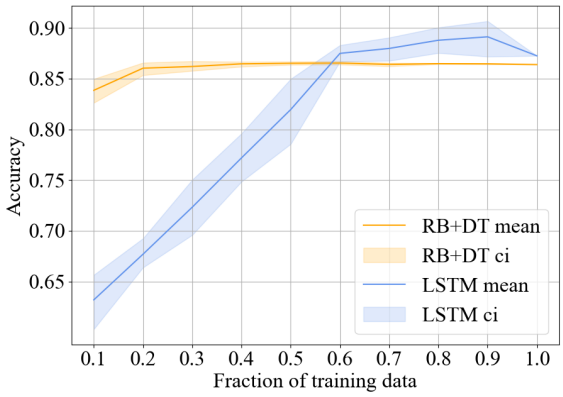

Fig. 2: Accuracy achieved on training the decision tree and LSTM on varying fractions of the training data given as mean and 95% confidence interval (ci).

improved in future work, as it could lead to alarm fatigue if IEE is repeatedly detected incorrectly. The decision tree achieved lower variances for the measured metrics during cross validation, which could imply that it generalizes better and that the neural network might be overfit. Often hand-engineered features are criticized for being potentially biased, as they base on expert-knowledge. In this application, however, we noticed the opposite effect, probably due to insufficiently large sample size for the training of the LSTM. Overall, considering the LSTM's highly optimized weights and architecture, the rule-based decision tree was not consistently outperformed by a high margin. Future work could consider an ensemble approach combining the algorithms, as their wrong classifications overlapped only partly. Previous comparisons between a rule-based method and a convolutional neural network for the detection of PVAs in adults [3], showed that the rule-based algorithm could achieve a better performance than the neural network. The exact approach was not transferable to neonates, however, as the data used (pressure-volume loops) do not exhibit the same dynamics for neonates as they do for adults.

Regarding execution time, both algorithms are able to classify a breath in a shorter time than the breath duration. The decision tree was considerably faster, which could enable not

only the inference on embedded hardware within a ventilator, but also an on-the-flight adaptation of the algorithm on new data. In previous work, a PVA detection algorithm has been developed to produce detections while a breath is taken, for immediate information [4]. This could also be achievable for the presented decision tree, as a new maximal increase in flow and the total increase in flow can be continuously computed during a breath to detect an IEE, before the breath is finished.

The LSTM needed not only more time to train but it also needed more data than the decision tree to achieve the same performance. This increased need for data when training neural networks is well known, yet it is difficult to assess the exact amount of data needed for a specific task. The work in [7] sums up different recommendations for the amount of data a neural network needs, varying from 10 times the number of weights or 50-100 times the number of classes up to 10-100 times the number of input features. Thus, the amount of available data should be taken into consideration both for the development and evaluation of a new algorithm.

Beyond the performance of the individual algorithms, we stated that explainability is an important feature for algorithms with applicability in a critical field. We consider the developed rule-based decision tree to be intrinsically explainable. For each breath input it can deliver the two computed features, an explanation on how they were computed, and show which rules fired to reach a decision. For the LSTM, on the other hand, we added attention layers to highlight the parts of a breath important for the classification [6]. We evaluated the explanations with a small user study, which showed that the concept of attention-based visualizations was intuitive, however, single explanations were not always meaningful. We did not perform a user-study to objectify the explainability of the decision tree algorithm. This would be a valuable future task.

Lastly, we observed that in a few cases similar breaths were annotated differently. While a neural network might try and learn these discrepancies, a simpler rule-based algorithm can be used to point these out. As the extracted features are interpretable by humans, they can be leveraged as a tool to gain insight on annotation quality.

5 Conclusion

A rule-based decision tree and an LSTM were compared on the same task of detecting IEE in ventilated neonates using high-frequency timeseries data. The comparison involved an estimate on the models' classification performance, their training and inference time as well as an analysis on their accuracy depending on the size of the training dataset.

The superiority of the LSTM is marginal. It comes at the cost of much higher computing load, the demand of at least three times more training data and a loss of transparency. The choice of methods in applications must therefore rely on the availability of annotated data, computing power and the demand of acceptance from the group of users.

Outlook. An external evaluation of the two algorithms could be a future task, to gain a better understanding on the models' generalization capabilities. Further, a user study could be conducted on a head-to-head explainability comparison of the two methods with different stakeholders. A more long-term goal could be an in-depth analysis of model complexity requirements, depending on the available data and the application field.

Author Statement

Research funding: Federal Ministry of Education and Research (031L0303A). Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent has been obtained from all individuals included in this study. Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the Ethics Committee of the Medical Faculty, RWTH Aachen University (EK 118-19).

References

[1] De Wit, M., Miller, K. B., Green, D. A., Ostman, H. E., Gennings, C., & Epstein, S. K. (2009). Ineffective triggering predicts increased duration of mechanical ventilation. Critical care medicine, 37(10), 2740-2745.

[2] Casagrande, Alberto, et al. "An effective pressure–flow characterization of respiratory asynchronies in mechanical ventilation." Journal of Clinical Monitoring and Computing 35 (2021): 289-296.

[3] Ang, Christopher Yew Shuen, et al. "Patient-ventilator asynchrony classification in mechanically ventilated patients: Model-based or machine learning method?." Computer Methods and Programs in Biomedicine 255 (2024): 108323.

[4] van de Kamp, Lars, et al. "Automatic patient-ventilator asynchrony detection framework using objective asynchrony definitions." IFAC Journal of Systems and Control 27 (2024): 100236.

[5] Chong, David, and Gusztav Belteki. "Detection and quantitative analysis of patient-ventilator interactions in ventilated infants by deep learning networks." Pediatric Research 96.2 (2024): 418-426.

[6] C. Oprea et al., “Detecting Ineffective Efforts during Expiration for Neonates with Attention RNNs,” Current Directions in Biomedical Engineering, vol. 10, no. 4, pp. 478–481, Dec. 2024, doi: 10.1515/cdbme-2024-2117 .

[7] A. Alwosheel, S. van Cranenburgh, and C. G. Chorus, “Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis,” Journal of Choice Modelling, vol. 28, pp. 167–182, Sep. 2018, doi: 10.1016/j.jocm.2018.07.002 .