Expert RuleFit: Complementing Rule Ensembles with Expert Knowledge

Luisa Ebner[1], Malte Nalenz[2][0000−0003−3439−4469], Annette ten Teije[2][0000−0002−9771−8822], Frank van Harmelen[1][0000−0002−7913−0048], and Thomas Augustin[2][0000−0002−1854−6226]

[1] Vrije Universiteit Amsterdam, de Boelelaan 1081a, Amsterdam, NL
[2] University of Munich, Ludwigstr. 33, Munich, Germany
ebner.luisa@gmx.de
{Frank.van.Harmelen,Annette.ten.Teije}@vu.nl
{Malte.Nalenz,Thomas.Augustin}@stat.uni-muenchen.de

Abstract. Machine learning algorithms have great potential to enhance clinical diagnosis and treatment. Yet, their overall performance is limited by the quality and quantity of available training data, while their adoption is limited by the level of trust ascribed by human experts. Injecting additional knowledge obtained from existing literature or from human expertise into the machine learning algorithm is widely seen as a solution to both of these problems. Yet, few implementations of expert-guided machine learning exist to date. We present Expert RuleFit (ERF), an approach to integrate expert knowledge in the form of rules and linear terms into an existing method for rule learning (RuleFit). A customized regularization strategy allows us to consider the different strengths of expert knowledge. For an empirical evaluation, we trained ERF models on a diabetes dataset for which we acquired expert rules from medical guidelines and expert interviews. We show that our ERF method enriches or replaces potentially spurious correlations learned from a patient sample with expert-derived, validated domain knowledge without sacrificing predictive performance. The integration of different knowledge sources makes the ERF model a promising tool for learning accurate, explainable and trustworthy medical decision rules.

Keywords: Decision rules · Rule learning · Explainable Machine Learning.

1 Introduction

Machine Learning (ML) systems offer great potential in medicine to provide healthcare improvements. Their ability to learn from data without explicit human guidance provides an attractive solution to the problems of manual knowledge acquisition encountered in the development of rule-based expert systems [16]. Considering the complexity and dynamics of medical knowledge, inductive learning is essential for successful decision support systems [16]. However, it is

not to be forgotten that rule-based expert systems [2] have two clear advantages over ML systems.

First, expert systems allow for the integration of and reasoning with various sources of expert knowledge, ranging from personal assessments to factual textbook knowledge. ML algorithms, to the contrary, are dependent on training examples as the only source of information. To generalize well to unseen cases, ML requires sufficient data to represent the population as a whole. Besides the number of training examples, this depends on the amount of information present in the recorded attributes, the amount of noise and the presence of hidden confounders. Due to the high cost and effort of information acquisition, privacy concerns and an intrinsic uncertainty of medical data, clinical datasets are often characterized by few examples, many missing values and insufficient task-relevant input attributes. Then, ML models suffer from limited generalizability. Indeed, significant performance decline is often observed when a model trained on data from e.g. one hospital is used to predict patient outcomes from another.

Second, expert systems draw upon expert-derived knowledge (e.g. in form of rules) to perform reasoning. As a result, expert system recommendations come with explanations that resemble human knowledge and reasoning in structure and vocabulary. The state-of-the-art in ML, to the contrary, often trades explainability for predictive accuracy. In safety-critical applications, this may diminish human trust and chances for system adoption. Without the possibility of expert validation, high performance on test sets derived from the same distribution as the training set is often considered as evidence that real knowledge has been captured by a model. This is dangerous because, in practice, ML cannot guarantee reasonable patterns. Lacking any general domain knowledge, it cannot be ruled out that ML algorithms make mistakes that would appear trivial to a human [11].

A solution to both problems of generalizability and explainability is to incorporate prior knowledge. As an additional source of information, it allows ML algorithms to better generalize to unseen cases while allowing human experts to better understand and validate recommendations. We meet this challenge proposing Expert RuleFit, a classification method that combines the strengths of inductive ML and expert rule-based reasoning. Expert RuleFit injects expert knowledge in the form of rules and linear terms into the existing rule ensemble method RuleFit [9]. We use the term expert knowledge to refer to any form of knowledge that experts consider state of the art and that they formulate to the best of their knowledge. As such, expert knowledge is somehow validated, e.g. through expert reasoning and academic studies or practically from experience or usage. The rules allow to stratify a patient population into task-relevant subpopulations, while the linear terms allow to express correlations between patient attributes and the target. Furthermore, by means of a tailored regularization strategy, our approach allows experts to specify *confirmed* knowledge to certainly enter the final prediction model as well as *optional* knowledge to be promoted over data rules through a customized penalization strategy. By adding expert knowledge in the form of rules to a data-generated rule set, we increase

the explainability and trustworthiness of ML results, to meet the high demands on medical decision support systems.

The contribution of this paper is a novel approach to combine rule *learning* with rule based *expertise*, implemented as an extension of the existing rule ensemble RuleFit and illustrated by a use case of diabetes diagnosis. We start in section 2 with a brief discussion of related work. Section 3 describes the existing RuleFit method and discusses its limitations in medical application contexts. Section 4 presents Expert RuleFit as an expert-guided RuleFit extension to obtain the benefits of expert knowledge inclusion. Section 5 compares Expert RuleFit with the conventional RuleFit method on a use case of diabetes diagnosis. Finally, section 6 concludes and points out research paths for future work.

2 Related Work

Knowledge representation in the form of rules has a long tradition in medical AI, in particular in rule-based expert systems [2]. More recently, the integration of symbolic prior knowledge into the process of learning from examples has been considered in hybrid systems [18]. In pursuit of theory-guided data science and scientific consistency in machine learning, research interest is increasingly devoted to combinations of data- and knowledge driven approaches. Under the umbrella term *informed machine learning*, the recent survey paper [17] provides a structured overview on many different ML learning algorithms that can be enriched with prior knowledge. A taxonomy classifies them according to the *source* of knowledge, its *representation* and its *integration* into the ML process. The majority of research work is concerned with the use of symbolic knowledge in neural networks. [20] and [4] use logical formulas to guide the output of deep neural networks as logical constraints in loss functions. In [14], knowledge graphs enhance deep neural networks with rules about relations between instances. In contrast, Expert RuleFit does not add rule-based knowledge to deep learning, but to a rule learning engine. Using the terminology from the informed ML taxonomy [17], we use expert knowledge as knowledge source, rules as knowledge representation and integrate knowledge directly into the learning algorithm. One particularly related approach is Expert-Augmented Machine Learning (EAML) [11], where domain experts use an online platform to assess the relative risk of subpopulations defined by RuleFit rules and the difference between their assessment and the empirical risk is considered as part of a novel regularization strategy for RuleFit. Whereas EAML first derives knowledge from data and then has the same evaluated by experts post hoc, our approach allows to formulate knowledge a priori which is then taken into account by the rule-learning algorithm. This allows human knowledge to be explicitly integrated to co-define model components.

3 Context

RuleFit is a rule ensemble method proposed by Friedman and Popescu [9]. In general, ML ensembles solve prediction problems by combining the predictions of

several classifiers. That is to say, multiple classifiers are trained to solve the same problem and consequently, their individual results are aggregated in the form of a generalized, linear model to form one joint prediction model [1]. The ensemble members – commonly referred to as base learners – are potentially different functions of different subsets of predictor attributes derived from the training data [9]. The rationale of ensembling is performance improvement by variance reduction. Given a labelled dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, RuleFit derives regression and classification ensembles whose accuracy is competitive with state-of-the-art methods such as random forests [13]. The rule learning ability avoids knowledge acquisition, enables massive data inputs and allows for knowledge discovery. The use of rules and linear terms as base learners makes RuleFit models potentially comprehensible for humans [10]. RuleFit has a better accuracy-complexity trade-off than most of the state-of-the art in ML [11]. This makes it a promising candidate to provide decision support in safety-critical domains with high demands on both accuracy and explainability [21]. The RuleFit algorithm proceeds in two stages: Ensemble Generation and Regularisation.

Stage 1: Ensemble Generation RuleFit models are linear combinations of rules and linear terms, whose predictive relevance is to be defined by respective coefficients. Following Friedman's stochastic gradient boosting strategy of rule learning [8], RuleFit generates an overly large set of candidate rules from boosted tree ensembles. As products of attribute-value tests from the root node to every other node in the tree, rules act as binary classifiers $r(\mathbf{x}) \in \{0, 1\}$, where $\mathbf{x}$ is the covariate vector, indicating whether observations match their conditions. To help illustrate the idea of the rule generation process, Figure 1 depicts an exemplary, simple decision tree generated from the Pima Indian Diabetes (PID) dataset available from the UCI ML Repository [5]. The rules listed in Table 1 correspond to the paths to all nodes of the tree. Note that in RuleFit only the conditions are kept as decision rules, not the predictions in the leaf nodes. The rationale however is that decision rule specify subgroups that are predictive with respect to the target attribute. Rule r_3 in Table 1 specifies patients that are at least 29 years and have a BMI of less than 27. From Fig.1. we can see that the risk for diabetes is lower in this group.

Table 1. Rules corresponding to the decision tree in Fig. 1.

Rules	Conditions
r_1	Age $<$ 29
r_2	Age $\geq$ 29
r_3	Age $\geq$ 29 & BMI $<$ 27
r_4	Age $\geq$ 29 & BMI $\geq$ 27

This process is repeated for each tree from the boosted tree ensemble and the extracted decision rules are concatenated to a large set of rules. The resulting rule set is cleaned according to sufficient support, colinearity and duplicates. To

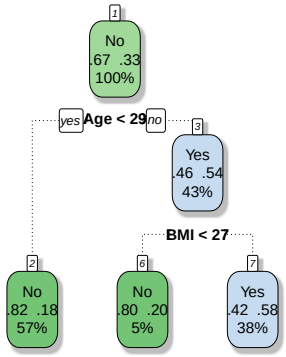

Fig. 1. Decision tree extracted from the PID dataset.

better capture linear effects, numeric attributes $\mathbf{x}_j$ are preprocessed (see [9]) and added as linear terms $l(x_j)$ in the ensemble:

$$F(\mathbf{x}) = \alpha_0 + \sum_{d=1}^{D} \alpha_d r_d(\mathbf{x}) + \sum_{j=1}^{p} \beta_j l(x_j). \tag{1}$$

Stage 2: Regularization To boil the overly large set of candidate rules down to the truly informative ones, Lasso regularized regression is applied to learn the regression coefficients α_d and β_j [19]. The least absolute shrinkage and selection operator "Lasso" is widely considered in ML literature and -practice for sparsification problems. It is easy to implement with a number of efficient solvers available and known for its selective nature when confronted with high dimensional data. Accordingly, the model coefficients $\gamma_{RF} = (\alpha_0, \{\alpha_d\}_1^D, \{\beta_j\}_1^p)$ are learned as:

$$\gamma_{RF} = \arg\min_{\gamma_{RF}} \sum_{i=1}^{N} L\left(y_i, F(\mathbf{x})\right) + \lambda \cdot \left(\sum_{d=1}^{D} |\alpha_d| + \sum_{j=1}^{p} |\beta_j| \right), \tag{2}$$

where L is an appropriately chosen loss function (typically sum of squared for linear regression and negative log-likelihood with sigmoid link-function on $F(\mathbf{x})$ for binary classification) [13]. The result are relatively sparse prediction models, where the majority of coefficient estimates is set to zero [7].

RuleFit's suitability in medical contexts is limited by its dependency on sufficient data and human acceptance. Similar to the majority of ML algorithms, model generalizability is constrained by the quantity and quality of the training set. At the same time, human acceptance of model results is constrained by the number and complexity of learned decision rules as well as their consistency with domain knowledge and expert assessments. Without reference to any general knowledge of the domain, RuleFit rules may often combine conditions

that contradict expert assessments. In this regard, RuleFit offers no possibility to remove some rules and include others. This diminishes RuleFit's chances for regular use and consultation in clinical practice [11].

4 Method

A solution to both the problems of limited generalizability and limited trust lies in an incorporation of expert knowledge into the RuleFit algorithm. Expert knowledge is a natural way to counter the problem of insufficient training data, and in medicine it is widely available. Whereas expert knowledge commonly refers to normal cases, typical symptoms and causal relationships, data-derived patterns reflect real patients with comorbidities, confounding factors and individual differences [11]. Therefore, training data can extend the coverage of expert knowledge through exceptional cases or unknown patterns while expert knowledge can compensate the effects of spurious patterns learned from poor, atypical training examples with medical regularities and consensual knowledge. Furthermore, the inclusion of expert knowledge is likely to increase human trust in model results. After all, physicians formulate patient conditions according to their understanding of the human physiology and task-relevant symptoms and effects, while ML algorithms learn only correlations from the empirical distribution of a patient sample. We therefore present *Expert* RuleFit (ERF) as a classification method, derived in 3 stages: Knowledge Acquisition, Combined Ensemble Generation and Knowledge-Aware Regularization.

Stage 1: Expert Knowledge Acquisition. Prior to the learning process, expert knowledge regarding the learning task may be formulated as rules and linear effects. Similar to the knowledge acquisition strategy used to develop rule-based expert systems, this involves manual knowledge acquisition from domain experts, medical guidelines, study results and textbooks. This information is then translated into rules and linear effects. Rules separate patients into subpopulations with respect to their target values. For example, in diabetes diagnosis the expert rule BMI > 40 & Age ≥ 60 & BP > 120 defines a subpopulation of obese, elderly people with increased blood pressure. A physician might specify this subpopulation to have a high incidence rate of diabetes compared to the whole population. Particularly favourable for rule formulation are clinical practice guidelines, whose recommendations on the diagnosis and treatment of patients with specific clinical conditions are either directly formulated as rules or as structured, rule-like statements. To distinguish different degrees to which expert knowledge is validated, ERF allows users to declare some rules and linear terms as *confirmed* and others as *optional* knowledge.

Stage 2: Combined Ensemble Generation. Consequently, at most 4 different sets of expert knowledge enter the ERF model together with the given dataset. These are the sets of confirmed expert rules $r_c, c \in \mathcal{I}_c$ and linear terms

$l_{c_l}, c_l \in \mathcal{I}_{c_l}$ as well as the sets of their optional counterparts $r_o, o \in \mathcal{I}_o$ and $l_{o_l}, o_l \in \mathcal{I}_{o_l}$, where $\mathcal{I}_c$, $\mathcal{I}_{c_l}$, $\mathcal{I}_o$ and $\mathcal{I}_{o_l}$ are disjoint index sets that are also disjoint with $\{1, ..., D\}$. Based on the given dataset and the encoded expert knowledge, data rules r_d are generated using the RuleFit method. This results in one common, enlarged set of base learners to enter the linear predictor

$$F(\mathbf{x}) = \alpha_0 + \sum_{d=1}^{D} \alpha_d r_d(\mathbf{x}) + \sum_{c \in \mathcal{I}_c} \alpha_c r_c(\mathbf{x}) + \sum_{o \in \mathcal{I}_o} \alpha_o r_o(\mathbf{x}) + \sum_{c_l \in I_{cl}} \beta_{c_l} l(x_{c_l}) + \sum_{o_l \in I_{ol}} \beta_{o_l} l(x_{o_l}) \tag{3}$$

of the ERF model. In difference to RuleFit, linear terms are not included by default, but according to expert knowledge on their respective relevance. Since expert knowledge is included before learning the weight coefficients that specify the importance of the base learners, redundant and non-informative expert knowledge can be recognised and assessed as such while expert knowledge incompleteness may be compensated by the rules learned from the training data. ERF models cover the entire spectrum from purely data-driven RuleFit models to models that include only expert knowledge and no data rules.

Stage 3: Knowledge-Aware Regularization. To learn the coefficients, we developed a tailored regularization strategy, where adaptive *penalty factors* serve to guarantee an inclusion of confirmed expert knowledge and allow for a promotion of optional expert knowledge over data-generated predictors in the final model. The term penalty factors refers to multiplicative weight vectors applied to the Lasso penalty term λ, which allow to adjust penalization differently for every coefficient, e.g. to put discount on the inclusion of selected model terms [22]. The optimization problem for estimating the model coefficients $\gamma_{ERF} = (\alpha_0, \{\alpha_d\}_1^D, \{\alpha_c\}_{c \in \mathcal{I}_c}, \{\alpha_o\}_{o \in \mathcal{I}_o}, \{\beta_{c_l}\}_{c_l \in \mathcal{I}_{c_l}}, \{\beta_{o_l}\}_{o_l \in \mathcal{I}_{o_l}})$ extends to

$$\gamma_{ERF} = \underset{\gamma_{ERF}}{\arg\min} \sum_{i=1}^{N} L\left(y_i, F(\mathbf{x})\right) + \lambda \cdot \left[\sum_{d=1}^{D} |\alpha_d| + \sum_{o \in \mathcal{I}_o} \nu_o |\alpha_o| + \sum_{o_l \in \mathcal{I}_{o_l}} \eta_{o_l} |\beta_{o_l}| \right]. \tag{4}$$

Data rules are fully penalized using **1** as penalty factor. Confirmed expert rules and linear terms are exempted from penalization using **0** as penalty factor: They are certainly included in the final model and therefor do not appear in the penalty term above. For optional expert rules and linear terms, the user may specify customized vectors $\boldsymbol{\nu}$ with $\nu_o \in [0, 1]$ and $\boldsymbol{\eta}$ with $\eta_{o_l} \in [0, 1]$ as penalty factors to prefer each respective base learner to a customized degree over the data rules. The smaller $\boldsymbol{\nu}$ and $\boldsymbol{\eta}$ are chosen, the cheaper it is for the model to include the corresponding covariates. Setting all components of $\boldsymbol{\nu}$ and $\boldsymbol{\eta}$ to 1 leads to an equal treatment of optional terms and data generated terms. This approach loosely resembles the adaptive lasso [22], but with the penalty factors chosen in accordance to medical expertise. Our promotion of expert knowledge

Table 2. An excerpt of expert rules collected from medical guidelines and expert interviews including the degree to which rule accordance is regarded as diabetes indicator.

Expert Rule	Source	Diabetes Prevalence	Type
Age >= 60 & BP >= 81 & BMI > 40	SMC-D	very high	confirmed
Glucose > 110 & BP > 90	NHG-D	mid	optional
Age <= 42 & BP <= 80 & BMI <= 29	Expert 1	low	confirmed
Age >= 45 & BP >= 90 & BMI >= 35 & Glucose >= 130	Expert 2	high	confirmed

through diminished penalties is designed to balance the data rules that precisely fit an empirical distribution and thus help to achieve more robust, generalizable models.

5 Experiments

We evaluate the performance of ERF on the diagnosis of Type 2 diabetes.

Experimental Setting. We use the aforementioned PID dataset, obtained from the UCI repository [5], where the learning task is to diagnose diabetes patients. For 768 adult women, information is recorded regarding the number of pregnancies, age, BMI, triceps skinfold thickness, blood pressure (BP), insulin- and glucose levels, a genetic predisposition to diabetes and the diabetes test result.
As expert knowledge, we manually extracted rules and linear terms from two diabetes guidelines, the *Standards of Medical Care in Diabetes* [3] and the *National healthcare guideline – Diabetes Mellitus Type 2* [15]. Both sources reference attributes in the PID dataset and present knowledge in the form of patient conditions. Guideline information about the extent to which specified conditions indicate the presence of diabetes was used to classify expert knowledge as indicators of *minor*, *moderate*, *strong* and *very strong* diabetes risks. In addition, we conducted two expert interviews with practicing physicians, who specified a set of task-relevant patient subpopulations based on their diagnostic understanding and experience. According to the rationale that indicators of minor or (very) strong diabetes risk are more reliable separators than moderate indicators, we defined 20 confirmed expert rules, 2 confirmed linear terms, 34 optional expert rules and 3 optional linear terms.

Experimental Protocol. We train four different versions of our proposed ERF model, one existing implementation of the conventional RuleFit model and one Random Forest model, whereby the latter two serve as baselines. The four proposed ERF models are as follows: First, a standard ERF model called **ERF** includes data rules together with confirmed and optional expert knowledge with full penalty put on all optional expert knowledge ($\boldsymbol{\nu} = \boldsymbol{\eta} = \mathbf{1}$). Second, the model **ERF prio** is the same, but with optional expert knowledge preferred over data rules using $\boldsymbol{\nu} = \boldsymbol{\eta} = \mathbf{0.5}$. The penalty factors were chosen as equal

because the guidelines and expert assessments did not provide a finer subdivision to justify different penalty values. Third, the model **ERF only** includes only expert knowledge and no data rules. Fourth, the model **RuleFit** – as our implementation of the conventional RuleFit method – contains only data rules but no expert knowledge. In addition, we use the model **PRE** as an existing implementation of the RuleFit method [3] as well as a **Random Forest** model [4] as baselines. We evaluate the models on AUC (area under the ROC curve) and classification accuracy. With regard to explainability, we use the size of the final ensemble as an indicator of model complexity. Furthermore, we consider the proportion of expert knowledge in the final model as an indicator of medically coherent predictors, supporting explainability and trustworthiness of model results. To investigate training data dependence, all models are trained on 4 different sized subsamples of the PID data set. Finally, every individual model was made subject to 10-fold cross validation (CV) to provide balanced accuracy measures [13]. As usual, we derive 10-fold-CV estimates from splitting the original training data into 10 random, equally sized subsets or folds. For each fold k, the model is retrained, using the observations in the other 9 folds and evaluated using the observations in fold k. Eventually, the final performance is calculated as the average performance over the 10 folds [7].

Results. AUC and classification accuracy results (Fig. 2) are similar for all model settings, especially on the full dataset and the sample set of 400 patients. This shows that expert knowledge is task-relevant and often able to replace data rules without sacrificing predictive performance. For the larger data set sizes, the ERF models comprising both data- and expert knowledge are most competitive. For the smaller samples, **ERF** models achieve the same accuracy while including expert-validated patient conditions as predictors. The results of the **ERF only** models, which do not include any predictors learned from the dataset they are evaluated on, suggest that the expert knowledge contains as much task-relevant information as 400 training examples. Looking at the performance of **RuleFit** and **PRE**, we were not able to show a performance gain through the inclusion of expert knowledge. We presume this is partly because our expert knowledge is not complete and partly because the validation set has been randomly subsampled from the same empirical distribution as the training data. Eventually, **ERF** and **ERF prio** are slightly outperformed by the **Random Forest** model on the larger dataset sizes and cleary outperformed on the 200 sample. Our **RuleFit** implementation is rather competitive with **RandomForest** on all dataset sizes and significantly better than **PRE** on the 200 sample.

Final model sizes (Fig. 3) – ranging from 10 to 25 – are similar throughout the competing model settings and dataset sizes, indicating a high interpretability

[3] We use PRE, as the original R-implementation by [9] is no longer available. We adapted our penalization strategy to make results comparable with PRE, by using λ_{1se}, the largest λ within one standard error of the minimal one, to produce a more sparse solution. This was found to produce a better accuracy-complexity tradeoff.

[4] We use the default settings of the R-package randomForest

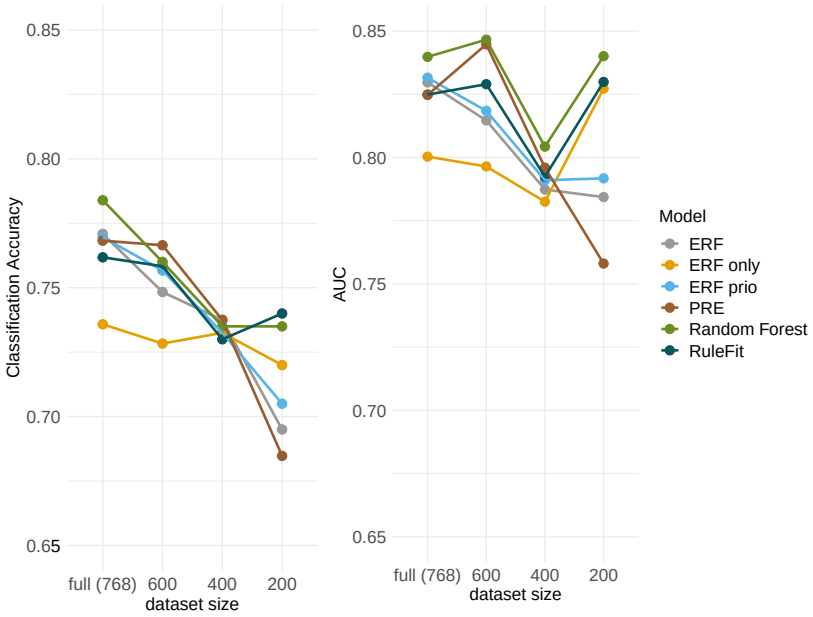

Fig. 2. Cross-validated (10-fold) results on classification accuracy (left) and AUC (right) of ERF- and RuleFit models and a Random Forest model trained on different sized samples of the PID dataset.

of the RuleFit algorithm and its variants, **PRE** and **ERF**, on this dataset. Although we initially entered a total of 59 expert knowledge terms, about half of the confirmed- and about 20 of the optional expert rules were removed due to insufficient support on the dataset or perfect correlation with other expert- or data rules. We see that the inclusion of expert knowledge decreases the ensemble size compared to our implementation of **RuleFit**, but remains slightly above the **PRE** version of RuleFit. Finally, the integration of expert knowledge in the form of additional base learners did not significantly influence the size of the final model.

Results on the proportion of expert knowledge in the final models as well as among their 10 most important base learners (Fig. 4) show high expert knowledge accordance across all ERF models. In particular, 50-75% of all base learners that remain in the final model and 8-10 out of the 10 most important terms (i.e. the terms with highest coefficients) correspond to expert knowledge. Of course, this is partly due to the concept of expert knowledge-aware regularisation, where confirmed expert knowledge is exempted from penalisation. Yet, the results show the value of expert knowledge for adequate predictions. Finally, an exemption from penalisation does not automatically make a model coefficient large and the associated base learner important. Thus, ERF models largely base their results on validated, medically coherent predictors instead of correlations derived from a patient sample.

We conclude that the ERF method yields explainable and more medically coherent models than RuleFit without sacrificing predictive accuracy or adding to model complexity. ERF's potential to yield increased accuracy at decreased model complexity was shown in an associated simulation study in [6]. As such, ERF promises accurate and yet simple models, including a large fraction of

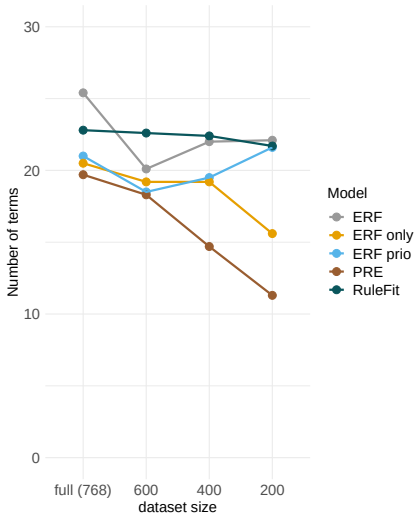

Fig. 3. Cross-validated (10-fold) results on the ensemble size of ERF- and RuleFit models trained on different sized samples of the PID dataset.

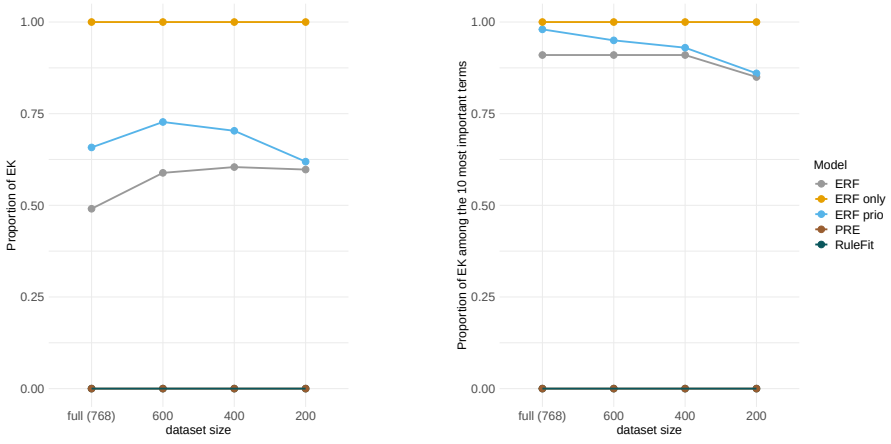

Fig. 4. 10-fold CV results on the proportion of expert knowledge in the final models, among the 10 most important base learners and overall. Importance of a term is defined as the absolute size of its coefficient.

validated, causal knowledge as important predictors and thus lowering the risk of including spurious relationships.

6 Conclusion and Future Work

We presented ERF as an expert-guided ML model for binary classification. Our approach combines the strengths of both ML and rule-based expert systems. While making use of RuleFit's rule learning ability, ERF allows human experts to complement an automatically generated knowledge base with knowledge they themselves work with to make decisions. In addition, ERF allows users to vary between purely expert knowledge-driven and purely data-driven models, depending on which information sources they trust most. This turns machine learning into a tool to enhance human reasoning, instead of overwriting it [12]. Finally, the increased level of human involvement promotes human trust in model results, which in turn raises the chances of adoption in clinical practice.

An inherent limitation is the constraint of expert knowledge to attributes in the dataset. Since ERF learns the weight coefficients of expert knowledge from corresponding data examples, a reference in the data is necessary to empirically evaluate the predictive influence of an expert rule or linear term.

Future work on ERF opens up several research paths. In the first instance, we would like to conduct larger scale experiments with more diverse data sets. Using the PID dataset, all models were evaluated using a validation set that has been randomly subsampled from the empirical distribution. This is risky when the set of patient examples is not representative of the whole population of interest as it is the case with many clinical datasets. However, the test set generally contains similar correlations as the training set [11]. To further evaluate and compare the out-of-sample performance of ERF and RuleFit models, test sets from different health institutions or different survey dates could help to investigate whether the inclusion of general, causal expert knowledge reduces performance decline over time or makes models more robust to changes in underlying variable distributions. If a certain patient group is underrepresented in the dataset used to train the model, expert rules concerning this patient group could help to make the model more generalizable to the entire patient population. To support our assumptions on increased explainability and human acceptance, models should be evaluated and compared by domain experts. On another note, the possibility to evaluate hypotheses or theories on empirical data suggests the use of ERF as an exploratory tool in medical research or even the strongly theory driven social sciences. Finally, the strengths of the ERF method are currently associated with the efforts of manual expert knowledge acquisition and -formulation. Even though ERF facilitates and restricts the latter, good results demand for thought-out development of rules and linear predictors. Especially with regard to scalability, manual knowledge acquisition is suboptimal and may sometimes speak against ERF usage. A promising aspect of future work lies in an integration of ERF with methods for automated knowledge acquisition, such that experts

could point to medical text from which rules and linear effects are extracted automatically.

Acknowledgements

We thank Dr. Waltraud Hoellein and Dr. Peter Hoellein for sharing their diabetes expertise in personal interviews. In addition, we would like to thank the anonymous reviewers for their constructive criticism and detailed comments on a previous version that helped to improve the quality of this paper.

References

1. Błaszczyński, J., Dembczyński, K., Kotłowski, W., Słowiński, R., Szelag, M.: Ensembles of decision rules for solving binary classification problems in the presence of missing values. In: International Conference on Rough Sets and Current Trends in Computing. pp. 224–234. Springer (2006)
2. Buchanan, B.G., Shortliffe, E.H.: Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project. Addison-Wesley (1984)
3. Care, F.: Standards of medical care in diabetes 2019. Diabetes Care **42**(Suppl 1), S124–S138 (2019)
4. Diligenti, M., Roychowdhury, S., Gori, M.: Integrating prior knowledge into deep learning. In: 2017 16th IEEE ICMLA conference. pp. 920–923. IEEE (2017)
5. Dua, D., Graff, C.: UCI machine learning repository (2017), http://archive.ics.uci.edu/ml
6. Ebner, L.: Expert RuleFit – Complementing Rule Ensembles with Expert Knowledge. Master's thesis, Faculty of Science, Vrije Universiteit Amsterdam (2021), https://www.ub.vu.nl/en/university-library-for-students/vu-thesis-database/index.aspx
7. Fokkema, M.: Fitting prediction rule ensembles with R package pre. Journal of Statistical Software **92**, 1–30 (2020)
8. Friedman, J.H.: Stochastic gradient boosting. Computational Statistics & Data analysis **38**(4), 367–378 (2002)
9. Friedman, J.H., Popescu, B.E.: Predictive learning via rule ensembles. The Annals of Applied Statistics **2**(3), 916–954 (2008)
10. Fürnkranz, J., Gamberger, D., Lavrač, N.: Foundations of rule learning. Springer (2012)
11. Gennatas, E.D., Friedman, J.H., et al.: Expert-augmented machine learning. Proceedings of the National Academy of Sciences **117**(9), 4571–4577 (2020)
12. Giraud-Carrier, C.: Flare: Induction with prior knowledge. Proceedings of Expert Systems 1996 **13**, 173–181 (1996)
13. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction. Springer (2009)
14. Marino, K., Salakhutdinov, R., Gupta, A.: The more you know: Using knowledge graphs for image classification. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 20–28 (2017)
15. Nationale Versorgungs-Leitlinie: Diabetes Mellitus Typ 2. Nationales Programm für Versorgungs-Leitlinien bei der Bundesärztekammer **7** (2002)

16. Ravuri, M., Kannan, A., Tso, G.J., Amatriain, X.: Learning from the experts: From expert systems to machine-learned diagnosis models. In: Machine Learning for Healthcare Conference. pp. 227–243. PMLR (2018)
17. von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., et al.: Informed machine learning–a taxonomy and survey of integrating knowledge into learning systems. arXiv:1903.12394 (2019)
18. Sun, R.: Connectionist implementationalism and hybrid systems. Encyclopedia of cognitive science (2006)
19. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996)
20. Xu, J., Zhang, Z., Friedman, T., Liang, Y., Broeck, G.: A semantic loss function for deep learning with symbolic knowledge. In: ICML. pp. 5502–5511. PMLR (2018)
21. Yang, W., Zhang, S., Chen, Y., Chen, Y., Li, W., Lu, H.: Mining diagnostic rules of breast tumor on ultrasound image using cost-sensitive rulefit method. In: 2008 3rd International Conference on Intelligent System and Knowledge Engineering. vol. 1, pp. 354–359. IEEE (2008)
22. Zou, H.: The adaptive Lasso and its oracle properties. Journal of the American Statistical Association **101**(476), 1418–1429 (2006)