

Comparing Humans and Algorithms in Feature Ranking: A Case-Study in the Medical Domain

Jonas Hanselle^{1,2,*,†}, Jaroslaw Kornowicz^{3,†}, Stefan Heid^{3,†}, Kirsten Thommes³ and Eyke Hüllermeier^{1,2}

¹LMU Munich, Germany

²Munich Center for Machine Learning, Germany

³Paderborn University, Germany

Abstract

The selection of useful, informative, and meaningful features is a key prerequisite for the successful application of machine learning in practice, especially in knowledge-intensive domains like decision support. Here, the task of feature selection, or ranking features by importance, can, in principle, be solved automatically in a data-driven way but also supported by expert knowledge. Besides, one may of course, conceive a combined approach, in which a learning algorithm closely interacts with a human expert. In any case, finding an optimal approach requires a basic understanding of human capabilities in judging the importance of features compared to those of a learning algorithm. Hereto, we conducted a case study in the medical domain, comparing feature rankings based on human judgment to rankings automatically derived from data. The quality of a ranking is determined by the performance of a decision list processing features in the order specified by the ranking, more specifically by so-called probabilistic scoring systems.

Keywords

Feature Ranking, Feature Selection, Scoring System, Machine Learning, Decision Support

1. Introduction

With the increasing access to technology, computational resources, and massive amounts of data, the idea of taking advantage of machine learning (ML) methodology to optimize decision support is becoming more and more feasible. Automated or partially automated decision-making with data-driven models is appealing for various reasons, especially as it is potentially more rational, objective, and accurate than decision-making by humans alone, which may be subjective or error-prone. For example, think of decisions in the context of employee recruitment, such as hiring or placement decisions [1], or the construction of individualized treatment rules in personalized medicine [2].

LWDA'23: Learning, Knowledge, Data, Analysis. October 09–11, 2023, Marburg, Germany

*Corresponding author.


†These authors contributed equally.

✉ jonas.hanselle@ifi.lmu.de (J. Hanselle); jaroslaw.kornowicz@upb.de (J. Kornowicz); stefan.heid@upb.de (S. Heid); kirsten.thommes@upb.de (K. Thommes); eyke@lmu.de (E. Hüllermeier)

🆔 0000-0002-1231-4985 (J. Hanselle); 0000-0002-5654-9911 (J. Kornowicz); 0000-0002-9461-7372 (S. Heid);

0000-0002-8057-7162 (K. Thommes); 0000-0002-9944-4108 (E. Hüllermeier)

© 2023 Copyright by the paper's authors. Copying permitted only for private and academic purposes. In: M. Leyer, Wichmann, J. (Eds.): Proceedings of the LWDA 2023 Workshops: BIA, DB, IR, KDML and WM. Marburg, Germany, 09.-11. October 2023, published at <http://ceur-ws.org>

 CEUR Workshop Proceedings (CEUR-WS.org)

That said, decision models constructed in a data-driven way will not be accepted by human experts [3] — and hence not be used in practice — unless these models are comprehensible, meaningful, and interpretable. In this regard, the selection and prioritization of decision criteria, or *features* in machine learning jargon, appears to be of major importance: The features on which a decision is based need to be semantically meaningful; features deemed relevant by the expert should be included in the model, while irrelevant features should be omitted.

Needless to say, these properties are not necessarily guaranteed when selecting features in a purely data-driven way. As another extreme, one may think of letting the human expert preselect the features by hand. For various reasons, however, this might be suboptimal either, for example, because the expert might be subjectively biased, or her knowledge might not be perfect. Presumably, the best approach is somewhere in-between, namely, *hybrid* in the sense that the human expert and the machine learning algorithm select features jointly in the course of an interactive process. Either way, these considerations beg an essential question: How capable are human experts in selecting the most important features or in ranking features in descending order of importance, and how do human experts compare to ML algorithms selecting features in a data-driven manner [4, 5, 6]?

This is the question addressed by the current paper. We conducted a case study in the medical domain, comparing feature (importance) rankings based on human judgment to feature rankings derived from data. The quality of a ranking is determined by the performance of a decision list processing features in the order specified by the ranking. In a decision list, features are considered incrementally, one by one. In each stage of the process, there are two options: either a final decision is made based on the feature values seen so far, or the process is continued by observing the next feature. Features should be ranked in decreasing order of importance to make well-informed decisions as quickly as possible. We implement this approach with so-called scoring systems, specifically appealing from an interpretability perspective and commonly used in the medical domain [7, 8].

Previous research suggests that data-driven methods generally surpass knowledge-driven methods in performance, though these findings are not entirely unambiguous. Our study contributes to resolving this continuing debate and extends the current literature by assessing these methods within the context of interpretable machine learning models. In high-stakes environments such as in the medical domain, the *constructor* of the decision model can be a significant factor for decision-makers, influencing their trust and reliance on the system. Consequently, evaluating the quality of various feature selection methods on such models is vital.

Our study shows that while data-driven feature ranking exhibits superior performance in identifying patterns unseen by human actors, the risk of overfitting, especially in small or biased datasets, necessitates the incorporation of human judgment for optimal results. We suggest an interactive, co-constructive approach, merging human expertise with algorithmic analytics, as a potential solution to offset overfitting effects while enhancing user acceptance of decision models. We encourage future research to leverage our findings, specifically targeting the inclusion of more domain professionals in the dataset, to further enrich and generalize these insights across various fields.

2. Data- and Knowledge-Driven Feature Selection

In the realm of supervised machine learning, most algorithms assume a representation of data objects (instances) in terms of feature vectors, which means that each object is specified by its values on a predefined number of features, also known as independent variables, dimensions, or inputs. The latter are supposed to carry important information for predicting the outcome or target variable [9]. Careful feature selection is a crucial step in the modeling process and a key prerequisite for learning accurate predictors [10]. Selecting a manageable number of meaningful features also facilitates interpretability and explainability [6].

Feature selection has been researched intensively in the past, with a specific focus on data-driven approaches. Here, an algorithm autonomously ranks or selects features based on the properties of the data. In contrast, knowledge-driven approaches determine a feature subset through literature review [11, 12, 13] or by consulting domain experts [4, 14]. Interactive machine learning fosters a combination of these approaches [15]. For instance, experts might underscore highly relevant observations and features that a data-driven algorithm can subsequently focus on [16]. Alternatively, experts might vote on different feature subsets, indirectly revealing their subjective preferences [17]. It is also possible to aggregate multiple selection and ranking methods into a single approach [18, 4, 19, 20].

Choosing the optimal method for a specific dataset and problem domain is inherently challenging. Guyon and Elisseeff [21] and Li et al. [6] advocate for including domain knowledge in the selection process. Conversely, Filippova et al. [5] find human intervention to be less beneficial than expected, while McKay [22] demonstrate that, for the same classification problem, a model with merely four features based on social science knowledge can rival models involving 10,000 features. On the other side, Cheng et al. [4] find that the features chosen by individual cardiologists, or an aggregation of their selections, can enhance accuracy compared to a baseline of all features, although they are still outperformed by data-driven methods. In their experimental study, Corrales et al. [11] observe that, in certain combinations of datasets and learning algorithms, expert knowledge can outperform data-driven methods. They conclude that expert knowledge can be especially beneficial under limited computational resources, for example, when working with high-dimensional datasets.

3. Probabilistic Scoring Lists

A so-called *scoring system* is a simple decision model that checks a set of features, adds (or subtracts) a certain number of points to a total score for each feature that is satisfied, and finally makes a decision by comparing the total score to a threshold. Scoring systems have a long history of active use in safety-critical domains such as healthcare [23] and justice [24], where they provide guidance for making objective and accurate decisions.

Hanselle et al. [25] propose an extension of scoring systems, called probabilistic scoring list (PSL). First, to increase uncertainty-awareness, a PSL produces predictions in the form of probability distributions (instead of making deterministic decisions). Second, to increase cost-efficiency, a PSL is conceptualized as a *decision list*: At prediction time, features are being evaluated one by one. The procedure may be stopped as soon as the practitioner decides that the

confidence in the predictions is high enough for the application context at hand. In the example in Table 1, the relevant information for an evaluation at stage 3 is highlighted in boldface. All features with their accompanying scores up to that stage need to be evaluated. The probabilities for the positive class are obtained by looking up the value corresponding to the total sum of the selected scores T . Here, the task is to diagnose a patient as COVID-19 positive or negative, given information about various features. In the concrete case, “Fatigue” would be determined as a first feature, and if present, contributes a score of 2. Fever would then be determined as the next feature, contributing a score of 1 if present, and this process continues with the remaining features. At stage 2, the probability of the positive class is predicted as 0 if the total score is 0, 0.1 if the total score is 1, etc. Note that adding a feature with a corresponding score of 0 is practically equivalent with ignoring said feature. Thus, we only consider score sets excluding 0.

Table 1
Example of a probabilistic scoring list for the COVID-19 use case

Stage	Feature	Score	T=-1	T=0	T=1	T=2	T=3	T=4	T=5	T=6
0	-	-	-	0.1	-	-	-	-	-	-
1	Fatigue	+2	-	0.1	-	0.3	-	-	-	-
2	Fever	+1	-	0.0	0.1	0.2	0.4	-	-	-
3	Cough	+2	-	0.0	0.1	0.1	0.2	0.2	0.5	-
4	Loss of smell	+1	-	0.0	0.1	0.1	0.2	0.2	0.4	1.0
5	Contact w/ inf. person	-1	0.0	0.0	0.1	0.1	0.2	0.4	0.4	1.0

The learning algorithm introduced in Hanselle et al. [25] constructs PSLs incrementally in a greedy manner. Starting with the empty list, one additional feature with a corresponding score (taken from a predefined set of scores) is added to the list in each stage. To this end, each feature/score pair is tentatively added as a candidate, and the resulting model is evaluated in terms of the *expected entropy* as performance measure:

$$E = \sum_{T \in \Sigma} \frac{N_T}{N} \cdot H(\hat{q}(T)), \quad (1)$$

where Σ is the set of total scores that can be produced by the decision list, $N = |\mathcal{D}|$ is the total number of training examples, and N_T the number of training examples with total score T . Moreover, $\hat{q}(T)$ is the estimated probability of the positive class given total score T , and H is the Shannon entropy

$$H(q) = -q \cdot \log(q) - (1 - q) \log(1 - q).$$

The feature/score combination leading to the highest performance is eventually added to the list, and the algorithm proceeds to the next stage (unless all features are used or the gain in terms of expected entropy is negative). The probabilities $\hat{q}(T)$ are estimated in terms of relative frequencies, rectified by isotonic regression to guarantee monotonicity (the probability of the positive class increases with an increasing total score).

Note that the expected entropy (1) is a meaningful measure of informedness at every stage of the decision process: The information provided by the prediction of a probability distribution \hat{q}

is quantified in terms of Shannon entropy, which is an established measure of information, and weighted by the (estimated) probability that this prediction is delivered.

The PSL produced by the above algorithm also suggests a ranking of features in the sense that features appearing earlier in the list seem to be more important in terms of performance than features queried only later on (or possibly not at all, if a decision is made before). With a straightforward modification, the algorithm can also be used to learn scoring systems for a predefined ranking of features: In each stage, it then adopts the corresponding feature and only optimizes over the set of possible scores, instead of optimizing over all features/score pairs.

4. Evaluation

In the following, we compare PSLs constructed solely in a data-driven fashion to PSLs in which the evaluated features are ordered according to human choices.

4.1. COVID-19 Dataset

We employed a non-public medical dataset, based on the work of Hufner et al. [26]. A minor deviation from the original dataset in our study pertains to the exclusion of a single observation that contained a missing value. Consequently, our dataset has a total of 696 patient observations.

According to the medical tests conducted in the original study, 633 patients (90.95%), tested negative for COVID-19. This dataset is comprised of 11 binary features, which, apart from information regarding patient contact with an infected individual, include all patient symptoms. Figure 1 shows all features, their respective distributions across the entire dataset, and the distributions for both positive and negative cases. While our dataset does not include additional demographic information, Hufner et al. [26] state in their study that 51.1% of the patients were female and the average age was 55.2 years.

Figure 2 shows the correlation between all features. Quite remarkably, the feature ‘‘Contact with an infected person’’ is negatively correlated to the target variable. Intuitively, contact with an infected person and the associated risk of exposure to the virus should have a positive correlation with an infection. One possible explanation for this peculiarity might be, that people who know that they had contact with an infected person may have higher awareness

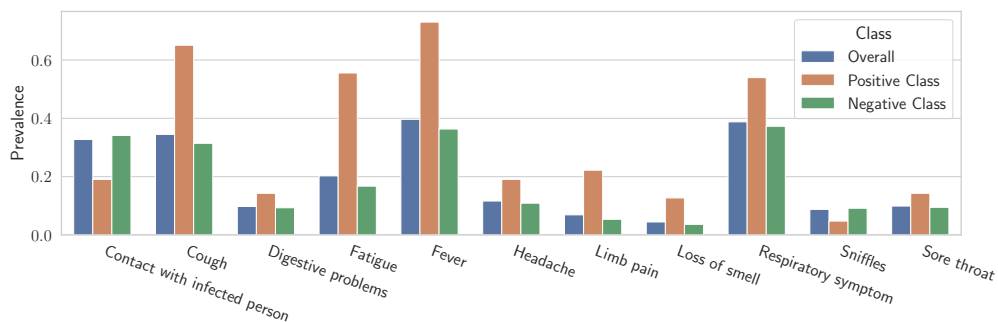


Figure 1: Feature prevalence overall and split between positive and negative class.

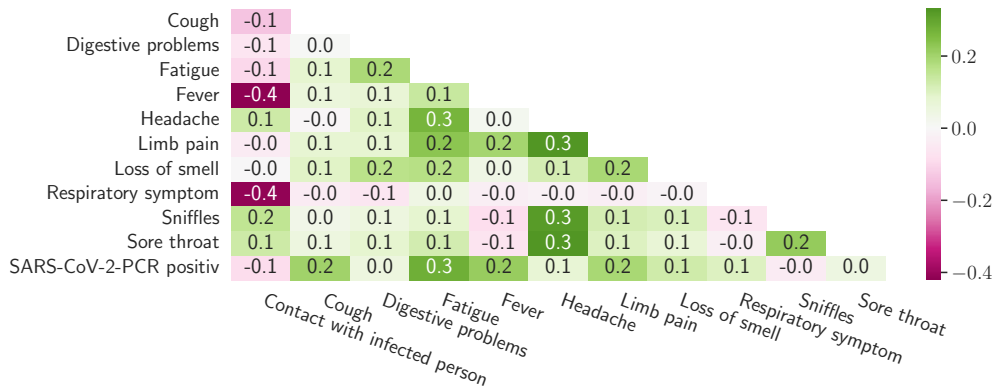


Figure 2: Heatmap showing features Pearson correlation. The last row shows the correlation with the target variable.

and hence be tempted to ask for a medical examination more quickly, even when showing no clear symptoms. This trend is further observable in the first column of the heatmap, where the correlations with symptoms such as respiratory issues and fever also exhibit a negative association.

4.2. Experimental Setup

In our experimental evaluation, we use PSLs constructed in five different manners. First, we consider PSLs derived from training data using the algorithm described in Section 3. These are called PSL.

Second, we compare them against PSLs built from expert input, specifically the original *Covid Score* system proposed by Hufner et al. [26] (EXPERT-PSL). The *Covid Score* was compiled as a consensus of medical experts. It was evaluated on the proposed dataset, however, it has not been used in the process of deriving the score. Note that EXPERT-PSL is a probabilistic scoring list and thus conceptually different from the original scoring system, which always evaluates the entire feature set and uses a constant threshold of 5 as a decision rule.

Two further approaches are derived based on a recent incentivized behavioral experiment conducted by Kornowicz and Thommes [27]. In this study, 234 subjects, recruited from the Prolific.co¹ platform, were requested to rank features based on their perceived importance for the classification task. Despite these subjects lacking specific medical field expertise, it remains plausible that the aggregate of their rankings might approximate the quality of expert opinions, as suggested by research in the field of expert elicitation [28, 29, 30]. We primarily utilized the rankings generated individually by subjects (SUBJECT-PSL), along with a method of consensus ranking referred to as Behavioral Aggregation (SUBJECTBA-PSL). For this method, 90 subjects were grouped into sets of three to agree upon a collective ranking. As there are no specified scores attached to the latter, we chose the scores associated with the features in the same greedy, data-driven manner as the first approach to allow for a fair comparison.

¹<https://www.prolific.co/>

Table 2

All considered PSLs in the experimental evaluation

Approach	Feature sequence chosen algorithmically	Scores chosen algorithmically
PSL	✓	✓
EXPERT-PSL	✗	✗
SUBJECT-PSL	✗	✓
SUBJECTBA-PSL	✗	✓
RANDOM-PSL	✗	✓

Lastly, as a baseline, we consider PSLs constructed from random feature permutations, for which the scores have been chosen in the same manner (RANDOM-PSL). We chose $\mathcal{S} = \{\pm 1, \pm 2, \pm 3\}$ as the set of possible scores for all methods except the expert method. The expert method’s scores are taken from the scoring system by Hübner et al. [26] and hence constrained to $\mathcal{S} = \{+1, +2, +3\}$. An overview of the considered constructions is depicted in Table 2.

We evaluated the individual PSLs in terms of a Monte Carlo cross-validation (MCCV) with 10 repetitions. In each repetition, we use a fraction of two-thirds of the available data as training data and one-third as test data. We report the expected entropy as a neutral measure of informativeness at each stage of the decision model in order to compare the approaches. Additionally, we evaluate the decision models in terms of expected loss minimization. In the domain of medical decision-making, it is common that a false negative prediction, i.e., not isolating and treating a COVID-19-infected patient, has far more severe consequences than a false positive. To capture this, we employ an asymmetric loss function that assigns a loss of 1 to false positives and a loss of $M \gg 1$ to false negatives. Given the PSLs probabilistic prediction \hat{p} for the positive class, the risk-minimizing decision is

$$\hat{y} = \begin{cases} 1 & \text{if } 1 - \hat{p} < M \cdot \hat{p} \\ 0 & \text{otherwise} \end{cases},$$

and the (estimated) expected loss itself by $\mathbb{E}(\hat{y}) = \min\{1 - \hat{p}, M \cdot \hat{p}\}$. For the experiments, we chose $M := 10$, i.e., penalizing false negatives ten times as much as false positives.

4.3. Results

In the following, we compare the five different PSL constructions against each other. Figure 3 shows the mean expected entropy and expected loss of the PSLs for each stage, i.e., after evaluating the stated number of features.

We observe that PSL achieves the best mean expected entropy throughout all stages. The SUBJECT-PSL and SUBJECTBA-PSL constructions perform very similar. Up until stage 3, they exhibit a higher mean expected entropy than the RANDOM-PSL baseline before consistently outperforming it as off stage 5. The EXPERT-PSL construction also performs worse than the random baseline within the first stages, even deteriorating when evaluating the first two features, both in terms of expected entropy as well as expected loss. This is due to the fact that the first two features selected by EXPERT-PSL are “Contact w/ inf. person” and “Respiratory symptom”.

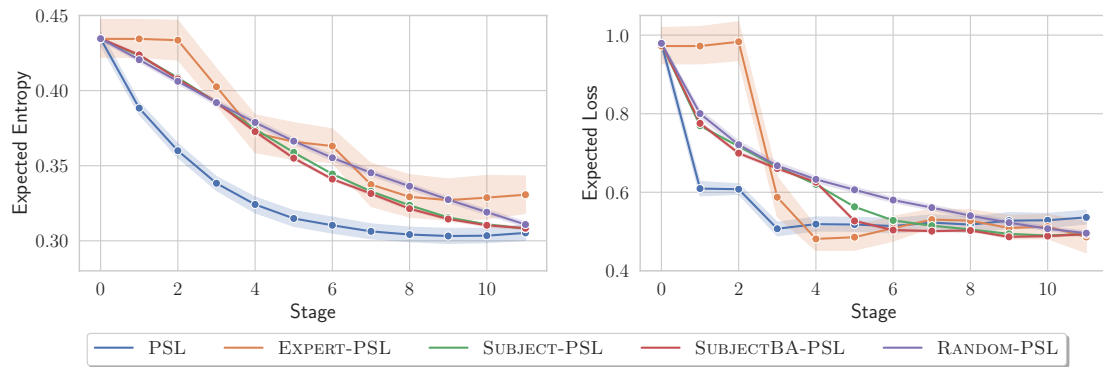


Figure 3: Mean expected entropy and expected loss of all considered PSL variants trained on the full training data. Error bands indicate the 95% confidence interval.

As already discussed in Section 4.1, the “Contact w/ inf. person” is negatively correlated with the target “SARS-CoV-2 positive” and the respiratory symptom is only weakly positively correlated to it. These two features both receive a score of +3 in the EXPERT-PSL construction, yielding poor performances early on and even deteriorating over the performance at stage 0 in which no feature is considered. The fact that these two features, which seem quite indicative for the human eye, do not have a strong positive influence on the outcome remains undiscovered for the experts. Here, the data-driven approach PSL takes advantage of having access to training data, placing it on average at a rank of 9.

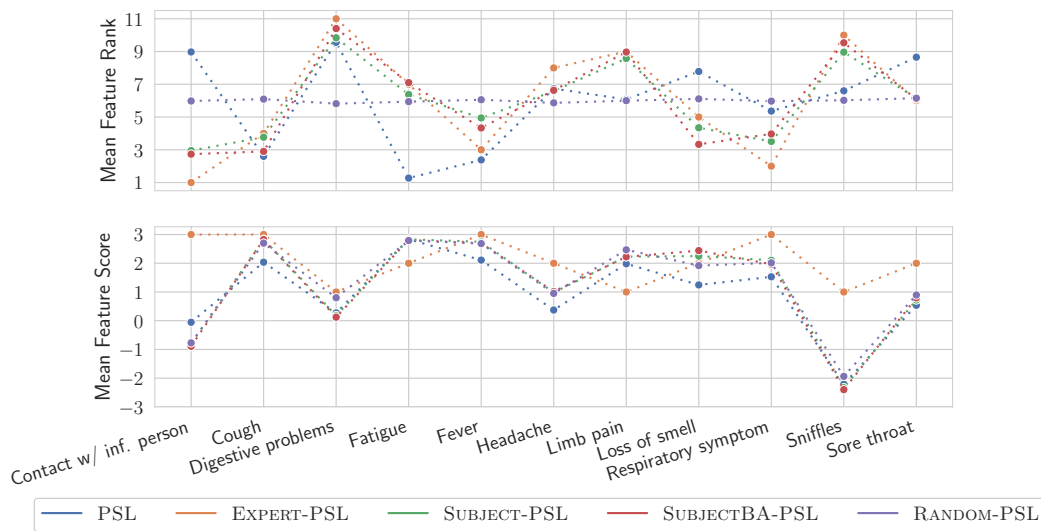


Figure 4: Average rank and average score of each feature across the methods.

Figure 4 shows an overview of the average ranks and scores of all features across the considered methods. For many features, the average ranks of the different approaches are quite similar, with the exception of the “Fatigue” and the “Contact w/ inf. person”. Since the scores

are optimized to the data in all approaches except for the EXPERT-PSL, the scores are really similar. This holds true regardless of the average rank of the feature. Note that the expert scores are selected according to Hufner et al. [26], constraining them to only positive scores.

Reducing available training data As discussed in the previous section, the data-driven approach PSL manages to unveil specifics from the data that are not taken into account by human actors. To make this feasible, it makes use of training data whose availability is a necessary condition for applying such methods. To investigate how much the data-driven approaches are dependent on the availability of data, we restricted them to 20% of the original training data by drawing subsamples from the original data without replacement and repeated the experiments 10 times. Figure 5 shows the expected entropy of the different PSLs when training them on these reduced training datasets.

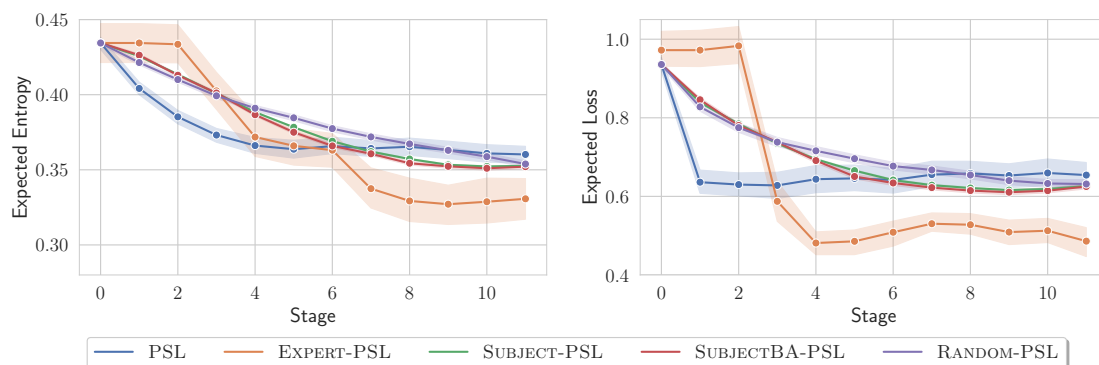


Figure 5: Mean expected entropy and expected loss of all considered PSL variants trained on a reduced set of 20% of the original training data. Error bands indicate the 95% confidence interval. Scales have been chosen in accordance with Figure 3

We observe that the PSL is outperformed from stage 7 on by the EXPERT-PSL and also by the SUBJECT-PSL and SUBJECTBA-PSL as of stage 9 in terms of expected entropy. When it comes to the expected loss, PSL is already beaten by EXPERT-PSL at stage 3 and the SUBJECT-PSL and SUBJECTBA-PSL methods at stage 7. In the end, even the RANDOM-PSL baseline exhibits a slightly lower mean expected error than the PSL. As expected, data-driven approaches become less reliable once access to data is restricted. In such scenarios, human expertise and common sense achieve better results than automated methods.

5. Conclusion

This paper has explored the comparative effectiveness of humans and algorithms in feature ranking for decision support. A case study in the medical domain was conducted, in which we compared feature rankings based on human judgment to rankings automatically derived from data. It was observed that the data-driven approach can identify patterns and specifics that remained hidden from human actors, leading to better performances in our experimental evaluation. On the other hand, feature rankings solely derived in an algorithmic manner bear

the risk of being overfitted to the available training data, resulting in poor generalization performance. This becomes especially important when training datasets are small or significantly biased. In this case, human knowledge and common sense may be a good countermeasure to compensate for such effects.

An interactive feature ranking procedure that combines the strengths of human and data-driven approaches constitutes an interesting direction for future work. Harnessing the benefits of human expertise and computational analytics in a co-constructive approach potentially leads to more accurate decision models while mitigating the risk of overfitting. Additionally, including humans in the learning procedure may also increase the practitioner's acceptance of the obtained decision model, as purely algorithmically constructed models are often faced with distrust [31].

As machine learning-based decision support systems continue to gain traction, our findings offer valuable insights to researchers in this emerging field. Future research efforts could potentially build upon and generalize our findings by employing different datasets and extending the scope to various domains. One of the key strengths of our dataset is the high volume of human rankings; however, these subjects notably lack significant domain experience, with the exception of the utilization of the *Covid Score* system of Hufner et al. [26]. While the recruitment of a larger number of domain professionals presents a challenge, pursuing this could undoubtedly yield more insightful findings in future research.

Acknowledgments

We gratefully acknowledge funding by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG): TRR 318/1 2021 – 438445824.

References

- [1] D. Pessach, G. Singer, D. Avrahamia, H. C. Ben-Gal, E. Shmueli, I. Ben-Gala, Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming, *Decision Support Systems* 134 (2020).
- [2] Y. Zhao, D. Zeng, A. Rush, M. Kosorok, Estimating individualized treatment rules using outcome weighted learning, *Journal of the American Statistical Association* 107 (2012) 1106–1118. doi:10.1080/01621459.2012.695674.
- [3] M. Ashoori, J. D. Weisz, In ai we trust? factors that influence trustworthiness of ai-infused decision-making processes, arXiv:1912.02675 [cs] (2019). URL: <http://arxiv.org/abs/1912.02675>, arXiv: 1912.02675.
- [4] T.-H. Cheng, C.-P. Wei, V. Tseng, Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches, in: *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, 2006, p. 165–170. doi:10.1109/CBMS.2006.87.
- [5] A. Filippova, C. Gilroy, R. Kashyap, A. Kirchner, A. C. Morgan, K. Polimis, A. Usmani, T. Wang, Humans in the loop: Incorporating expert and crowd-sourced knowledge for predictions using survey data, *Socius* 5 (2019) 2378023118820157. doi:10.1177/2378023118820157.

- [6] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, H. Liu, Feature selection: A data perspective, *ACM Computing Surveys* 50 (2017) 94:1–94:45. doi:10.1145/3136625.
- [7] A. G. Rapsang, D. C. Shyam, Scoring systems in the intensive care unit: A compendium, *Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine* 18 (2014) 220–228. doi:10.4103/0972-5229.130573.
- [8] B. Ustun, C. Rudin, Supersparse linear integer models for optimized medical scoring systems, *Machine Learning* 102 (2016) 349–391. doi:10.1007/s10994-015-5528-6.
- [9] G. James, D. Witten, T. Hastie, R. Tibshirani, *An introduction to statistical learning*, volume 112, Springer, 2013.
- [10] S. Studer, T. B. Bui, C. Drescher, A. Hanuschkin, L. Winkler, S. Peters, K.-R. Müller, Towards crisp-ml(q): A machine learning process model with quality assurance methodology, *Machine Learning and Knowledge Extraction* 3 (2021) 392–413. doi:10.3390/make3020020.
- [11] D. C. Corrales, E. Lasso, A. Ledezma, J. C. Corrales, Feature selection for classification tasks: Expert knowledge or traditional methods?, *Journal of Intelligent & Fuzzy Systems* 34 (2018) 2825–2835. doi:10.3233/JIFS-169470.
- [12] J. Nahar, T. Imam, K. S. Tickle, Y.-P. P. Chen, Computational intelligence for heart disease diagnosis: A medical knowledge driven approach, *Expert Systems with Applications* 40 (2013) 96–104. doi:10.1016/j.eswa.2012.07.032.
- [13] J. Wang, J. Oh, H. Wang, J. Wiens, Learning credible models, in: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18*, Association for Computing Machinery, New York, NY, USA, 2018, p. 2417–2426. URL: <https://doi.org/10.1145/3219819.3220070>. doi:10.1145/3219819.3220070.
- [14] S. Moro, P. Cortez, P. Rita, A divide-and-conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing, *Expert Systems* 35 (2018) e12253. doi:10.1111/exsy.12253.
- [15] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop?, *Brain Informatics* 3 (2016) 119–131. doi:10.1007/s40708-016-0042-6.
- [16] A. H. C. Correia, F. Lecue, Human-in-the-loop feature selection, *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (2019) 2438–2445. doi:10.1609/aaai.v33i01.33012438.
- [17] F. Bianchi, L. Piroddi, A. Bemporad, G. Halasz, M. Villani, D. Piga, Active preference-based optimization for human-in-the-loop feature selection, *European Journal of Control* 66 (2022) 100647. doi:10.1016/j.ejcon.2022.100647.
- [18] V. Bolón-Canedo, A. Alonso-Betanzos, Ensembles for feature selection: A review and future trends, *Information Fusion* 52 (2019) 1–12. doi:10.1016/j.inffus.2018.11.008.
- [19] D. J. Dittman, T. M. Khoshgoftaar, R. Wald, A. Napolitano, Classification performance of rank aggregation techniques for ensemble gene selection, in: *The twenty-sixth international FLAIRS conference*, 2013.
- [20] R. Wald, T. M. Khoshgoftaar, D. Dittman, W. Awada, A. Napolitano, An extensive comparison of feature ranking aggregation techniques in bioinformatics, in: *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*, 2012, p. 377–384. doi:10.1109/IRI.2012.6303034.
- [21] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, *Journal of machine*

- learning research 3 (2003) 1157–1182.
- [22] S. McKay, When $4 \approx 10,000$: The power of social science knowledge in predictive performance, *Socius* 5 (2019) 2378023118811774.
 - [23] A. Six, B. Backus, J. Kelder, Chest pain in the emergency room: value of the heart score, *Netherlands Heart Journal* 16 (2008) 191–196.
 - [24] C. Wang, B. Han, B. Patel, C. Rudin, In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction, *Journal of Quantitative Criminology* (2022) 1–63.
 - [25] J. Hanselle, F. Fürnkranz, E. Hüllermeier, Probabilistic scoring lists for interpretable machine learning, in: *Proc. DS, 23rd Int. Conference on Discovery Science*, Springer, Porto, Portugal, 2023.
 - [26] A. Hüfner, D. Kiefl, M. Baacke, R. Zöllner, E. Loza Mencía, O. Schellein, N. Avan, S. Pemmerl, Risikostratifizierung durch implementierung und evaluation eines covid-19-scores, *Medizinische Klinik - Intensivmedizin und Notfallmedizin* 115 (2020) 132–138. doi:10.1007/s00063-020-00754-4.
 - [27] J. Kornowicz, K. Thommes, Aggregating human domain knowledge for feature ranking, in: H. Degen, S. Ntoa (Eds.), *Artificial Intelligence in HCI, Lecture Notes in Computer Science*, Springer Nature Switzerland, Cham, 2023, p. 98–114. doi:10.1007/978-3-031-35891-3_7.
 - [28] D. Önkal, J. F. Yates, C. Simga-Mugan, Ş. Öztin, Professional vs. amateur judgment accuracy: The case of foreign exchange rates, *Organizational Behavior and Human Decision Processes* 91 (2003) 169–185. doi:10.1016/S0749-5978(03)00058-X.
 - [29] M. Nofer, *Are Crowds on the Internet Wiser than Experts? – The Case of a Stock Prediction Community*, Springer Fachmedien, Wiesbaden, 2015, p. 27–61. URL: https://doi.org/10.1007/978-3-658-09508-6_3. doi:10.1007/978-3-658-09508-6_3.
 - [30] E. Vul, H. Pashler, Measuring the crowd within: Probabilistic representations within individuals, *Psychological Science* 19 (2008) 645–647. doi:10.1111/j.1467-9280.2008.02136.x.
 - [31] H. Mahmud, A. K. M. N. Islam, S. I. Ahmed, K. Smolander, What influences algorithmic decision-making? a systematic literature review on algorithm aversion, *Technological Forecasting and Social Change* 175 (2022) 121390. doi:10.1016/j.techfore.2021.121390.