An Initial Study of Machine Learning Underspecification using Feature Attribution Explainable AI Algorithms: a COVID-19 Virus Transmission Case Study

James Hinns[1], Xiuyi Fan[1], Siyuan Liu[1], Veera Raghava Reddy Kovvuri[1], Mehmet Orcun Yalcin[2], Markus Roggenbach[1]

[1] Computer Science Department, Swansea University
[2] Department of Data Science and Knowledge Engineering, Maastricht University

Abstract. From a dataset, one can construct different machine learning (ML) models with different parameters and/or inductive biases. Although these models give similar prediction performances when tested on data that are currently available, they may not generalise equally well on unseen data. The existence of multiple equally performing models exhibits *underspecification* of the ML pipeline used for producing such models. In this work, we propose identifying underspecification using feature attribution algorithms developed in Explainable AI. Our hypothesis is: **by studying the range of explanations produced by ML models, one can identify underspecification**. We validate this by computing explanations using the Shapley additive explainer and then measuring statistical correlations between them. We experiment our approach on multiple datasets drawn from the literature, and in a COVID-19 virus transmission case study.

Keywords: Underspecification, Explainable AI, COVID-19

1 Introduction

Underspecification has been identified as a major challenge in machine learning (ML) research. Roughly speaking, an ML pipeline is underspecified *"when it can return many predictors with equivalently strong held-out performance in the training domain."* [4] Having multiple different predictors is problematic in real-world applications as the current practice often treats such predictors as equivalent (based on their training performances), while they usually give different behaviours in deployment. Thus, we see that ML models sometimes exhibit unexpectedly poor behaviours when they are used in real-world applications when such multi-predictor phenomenon occurs.

The first step of addressing underspecification is to identify it. To this end, stress tests measuring prediction performances - evaluations that probe a predictor by observing its outputs on specifically designed inputs - have been reported in the literature [4]. However, with a few exceptions, as we discuss in Section 4,

Table 1. Two simple string datasets, D_1, and D_2 for underspecification illustration.

	Data	`POS` Explanation Pattern(s)
D_1	`POS`: 01101, 11101, 11111, 01111 `NEG`: 00000, 00010, 10010, 10000	$\cdot 1 \cdots$, $\cdot\cdot 1 \cdot\cdot$, $\cdot\cdot\cdot\cdot 1$
D_2	`POS`: 01101, 11101, 11111, 01111, 01001, 11100 `NEG`: 00000, 00010, 10010, 10000, 00001, 10111	$\cdot 1 \cdots$

existing approaches identify underspecification solely with traditional prediction metrics such as accuracy and root mean square error, which will make underspecification not fully identified in many situations.

In this work, we present an alternative approach: identifying underspecification with explanations. In a nutshell, given a dataset, we construct a set of predictors and study explanations generated using a feature attribution algorithm [14] from these predictors. We identify underspecification when observing "too many" different explanations form such predictors on the dataset. We observe that: **if a dataset can be explained in multiple ways, then a ML pipeline built from it is likely underspecified.**

Our core idea can be illustrated with the following example. Consider two binary classification datasets, D_1 and D_2, shown in Table 1. D_1 and D_2 contain eight and twelve 5-bit strings as data instances, respectively, on the alphabet $\{0, 1\}$. Each string is labelled either `POS` (positive) or `NEG` (negative). D_2 contains all strings of D_1 and four additional strings. Both datasets are balanced with each containing the same number of `POS` and `NEG` strings. If we consider each bit in a string representing a feature, which can be a potential explanation for a string's positivity, then there are three "1-bit explanations" for the positivity of strings in D_1 as follows:

- $\cdot 1 \cdots$: *a string is* `POS` *because its second bit is 1*,
- $\cdot\cdot 1 \cdot\cdot$: *a string is* `POS` *because its third bit is 1*, and
- $\cdot\cdot\cdot\cdot 1$: *a string is* `POS` *because its fifth bit is 1*.

There are no reasons to prefer any one of these explanations to the others given the dataset D_1. However, with the four additional strings introduced in D_2, we see that both explanations $\cdot\cdot 1 \cdot\cdot$ and $\cdot\cdot\cdot\cdot 1$ are ruled out, as 10111 and 00001 are both `NEG`. So there is a single explanation left for all strings in D_2:

- $\cdot 1 \cdots$: *a string is* `POS` *because its second bit is 1*.

Thus, we observe that D_2 with more data yields fewer 1-bit explanations than D_1 and can better specify prediction models than D_1.

Various explanation construction techniques have been developed in Explainable AI (XAI) [18]. These techniques produce explanations of different types, see e.g., [19] for an overview. In this work, we use a feature attribution explanation method, SHapley Additive exPlanations (SHAP) [14], which computes explanations to data instances in the form of "feature weights", to facilitate underspecification identification. SHAP is chosen in this work for its sound mathematical foundation and its ease of implementation.

SHAP is based on the coalitional game theory concept *Shapley value*, which is assigned to each feature of a data instance. Shapley values are defined to answer the question: "What is the fairest way for a coalition to divide its payout among the players"? It assumes that payouts should be assigned to players in a game depending on their contribution towards total payout. In a machine learning context, feature values are "players"; and the prediction is the "total payout". In this setting, the Shapley value of a feature represents its contribution to the prediction and thus explains the prediction. SHAP is model-agnostic and thus independent of underlying prediction models. For a data instance x, SHAP computes the marginal contribution of each feature to the prediction of x.

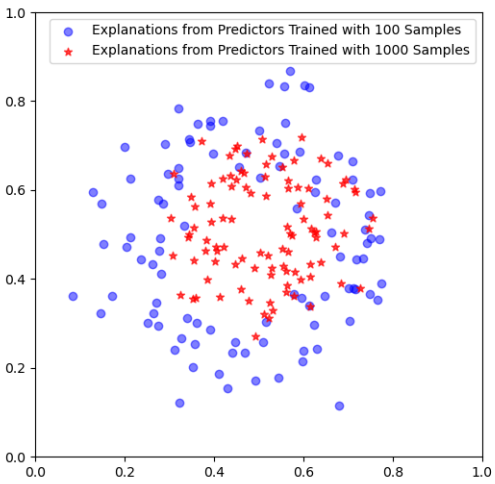

Fig. 1. An illustration of explanations from predictors trained with different sample sizes. Predictors trained with more data - hence less underspecified ML pipelines - produce more agreeable explanations. Red stars are placed closer to each other than blue dots are.[1]

Given a prediction model $P \in \mathcal{P}$, where $\mathcal{P}$ is the set of models, let $\mathbf{y} = P(\mathbf{x})$ be the prediction made by P on the input $\mathbf{x} = \langle x_1, \ldots, x_M \rangle \in \mathbb{R}^M$, SHAP gives an explanation $\langle \phi_1, \ldots, \phi_M \rangle \in \mathbb{R}^M$ (for $\mathbf{y} = P(\mathbf{x})$); ϕ_i can be viewed as the contribution of x_i for this prediction. We can think SHAP as a function $\mathit{\Pi} : \mathcal{P} \times \mathbb{R}^M \mapsto \mathbb{R}^M$. From a dataset, we train a set of models $\mathcal{P} = \{P_1, \ldots, P_n\}$. For the same input $\mathbf{x}$, we compute a set of explanations $\mathit{\Phi} = \{\mathit{\Pi}(P_i, \mathbf{x}) | P_i \in \mathcal{P}\}$.

By looking at how "compact" $\mathit{\Phi}$ is, we identify underspecification of the ML pipeline - if explanations in $\mathit{\Phi}$ are close to each other, that means models in $\mathcal{P}$ are agreeable with each other, thus less underspecified. Otherwise, explanations in $\mathit{\Phi}$ are apart from each other, then models in $\mathcal{P}$, although might be making the same prediction $\mathbf{y}$, make predictions for different reasons, hence more underspecified.

To put things into a concrete setting, we study how underspecification occurs in the context of predicting COVID-19 virus transmission. To this end, we construct a dataset containing daily confirmed cases between March 2020 and January 2021 and non-pharmaceutical control measures used in the UK and predict whether the infectious rate is growing on a given day. As illustrated in

[1] Blue dots and red stars represent explanations obtained from predictors trained with 100 and 1000 randomly selected samples in the COVID-19 dataset respectively. Within each set, the coordinates $\mathbf{x}_i$ are computed with a stochastic hill climbing algorithm that solves $\arg\min_{\mathbf{x}_i, \mathbf{x}_j} \sum |L_2(\mathbf{x}_i, \mathbf{x}_j) - D_\tau(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j)|$, where L_2 is the L2 norm, D_τ is the Kendall distance of each pair of explanations $(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j)$.

Figure 1, underspecification is observed when explanations generated from models are far apart from each other; whereas when explanations are close to each other and form compact clusters, there is less underspecification.

Overall, the proposed approach to identifying underspecification with explanations has the following advantages:

1. It is model-agnostic and applicable to any data types and ML models as long as such a model can be analysed with a model-agnostic explainer.
2. It is self-contained and does not require any additional information such as domain knowledge or human expert inputs.
3. It is simple and does not require any special treatment to the dataset, e.g., stratification or alteration, to estimate underspecification.

Our contributions in this work are as follows:

- We formulate underspecification identification as a problem of measuring correlations between explanations.
- We perform the explanation distance measurement using a well studied statistical metric, Kendall Rank Correlation Coefficient.
- We demonstrate our approach on both existing datasets in the literature and a real-world COVID-19 dataset.

The rest of this paper is organised as follows. Section 2 introduces our main approach with results produced from a synthesised dataset. Section 3 introduces the virus transmission case study in detail. Section 4 discusses some related work. We conclude in Section 5.

2 Our Approach

As introduced in [4], we consider underspecification in a supervised learning setting. Specifically, we consider an ML pipeline with a dataset D that produces a model (predictor) P, drawn from a set of predictors $\mathcal{P}$. Regardless of the method used to construct P, it is evaluated with some performance measures such as accuracy or root mean square error on D. An ML pipeline is *underspecified* if it can return multiple different predictors such that they give similar performances, while encoding substantially different inductive biases that can result in different generalisation behaviours on datasets beyond D (Out-of-Distribution).

Since predictors can contain a vast amount of parameters and/or have different internal structures, it is not straightforward to directly compare two predictors and determine how similar they are. Thus, in order to determine whether an ML pipeline is underspecified, we study explanations obtained from predictors produced by the ML pipeline, and use those as a proxy to estimate the differences between predictors.

Our core assumption is that:

> *If two predictors give the same explanation to a prediction, then they encode the same inductive bias; hence they should be considered the same.*

In this setting, given predictors $\mathcal{P} = \{P_1, \ldots, P_K\}$ produced by an ML pipeline with dataset D, we first use the SHAP explainer Π to compute *global* explanations $\mathbf{\Phi}_P$ for each predictor $P \in \mathcal{P}$ on the entire dataset D:

$$\mathbf{\Phi}_P = \sum_{\mathbf{x} \in D} \Pi(P, \mathbf{x}). \tag{1}$$

The rank of explanations from P is the ranked list calculated over $\mathbf{\Phi}_P$. For example, if SHAP values $\mathbf{\Phi}_P$ were [0.1,0.2,0.4,0.3] the ranked list would be [4,3,1,2]. This process of generating models and then computing their rank of explanations is shown in Algorithm 1. Note that the parameter θ used in line 3 is to ensure that all predictors trained in $\mathcal{P}$ have similar and high performances. K is the parameter that controls the number of predictors in experiments.

Algorithm 1 GenModels(D, K, θ) **return R**

Input: The number of models K, Dataset D, Prediction Performance Threshold θ
Output: Global Explanation ranks $\mathbf{R}$

1: $\mathbf{R} = []$
2: **while** $|\mathbf{R}| < K$ **do**
3: Train a predictor P with D such that the performance of P is greater than θ
4: $\mathbf{\Phi}_P = \langle 0, \ldots, 0 \rangle$ with $|\mathbf{\Phi}_P|$ the number of features in D
5: **for** each $\mathbf{x} \in D$ **do**
6: $\mathbf{\Phi}_P = \mathbf{\Phi}_P + \Pi(P, \mathbf{x})$
7: Append the ranked list of $\mathbf{\Phi}_P$ to $\mathbf{R}$
8: **return R**

With explanation rank lists $\mathbf{R}$ computed for all predictors, to identify underspecification, we compute

$$\mathcal{T} = \frac{2}{K(K-1)} \sum_{i=0}^{K} \sum_{j>i}^{K} \tau_{i,j}, \tag{2}$$

where $\tau_{i,j}$ is the pair-wise Kendall rank correlation coefficient over ranks of explanations generated from predictors P_i and P_j. $\mathcal{T}$ is the average Kendall rank correlation coefficient between all explanation pairs in $\mathcal{P}$. We can see that:

- $-1 \leq \mathcal{T} \leq 1$ for any ML pipelines and datasets; and
- the larger $\mathcal{T}$ is, the closer explanations are, hence less underspecification.

We test our approach on four datasets found in the literature, string classification [31], house price [3], abalone age [5] and mushroom [5]. Characteristics of these four datasets are summarised in Table 2.

Table 2. Datasets for Experiments.

Dataset	# of Samples	# of POS Samples	# of Feature	Type of Features
String [31]	9,623	4,410	20	Categorical
House Price [3]	1,461	728	79	Mixed
Abalone [5]	4,178	2,081	8	Mixed
Mushroom [5]	8,124	3,916	22	Categorical

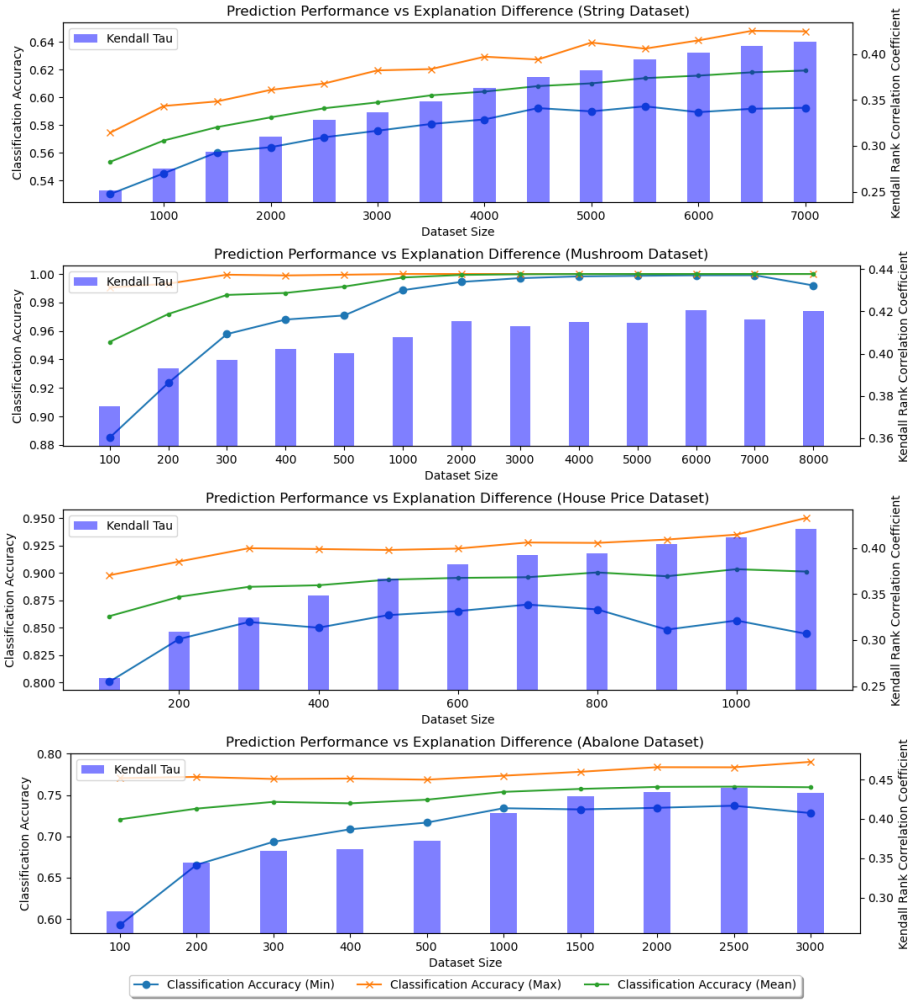

Fig. 2. Explanation Correlation vs Dataset Sizes.

To investigate how underspecification changes with different dataset sizes, we stratify each dataset into multiple smaller datasets in different sizes. For each of these smaller dataset lengths we trained $K = 100$ random forest predictors and test their performances on the whole dataset, comparing their explanation correlations with classification accuracy. This experiment was then repeated 10

Table 3. Non-pharmaceutical COVID Control Measures.

Meeting Friends / Family (Indoor)	Meeting Friends/Family (Outdoor)
Domestic Travel Control	International Travel Control
Cafes and Restaurants Control	Pubs and Bars Control
Sports and Leisure Closure	Hospitals / Care and Nursing Home Visits
Non-Essential Shops Closure	School Closure

times with averages shown in Figure 2. In this figure, we can see that for all four datasets, as we increase the dataset size, the explanation correlation increases. This means that with a larger dataset, explanations become more similar. Both the explanation correlations and classification accuracy plateau for larger dataset sizes indicating that once the dataset size reaches a certain threshold, introducing more samples does not reduce underspecification.

3 COVID-19 Virus Transmission Case Study

In this section, we apply our approach to a coronavirus virus transmission case study. This case study can be viewed as a realistic experiment modelled after the epidemiological model that demonstrates underspecification in [4]. In a nutshell, the model in [4] illustrates that at early stages of an epidemic, there is insufficient amount of data to fully specify an accurate prediction model; so multiple prediction trajectories can be formed based on the insufficient training data, consequentially the predictions becomes largely arbitrary.

From the Public Health England website[2], we collected daily infection numbers reported across 12 regions in UK: East Midlands, East of England, London, North East, North West, Northern Ireland, Scotland, South East, South West, Wales, West Midlands as well as Yorkshire and The Humber. Non-pharmaceutical control measure data was composed based on UK's COVID policies as summarised in Table 3. Data was corrected from various sources including Wikipedia and major news agencies. Control Measures were coded based on level of severity (e.g., "High", "Moderate", "Low") for all control measures excluding Non-essential shops and School closures, which are coded as binary choices ("Open" and "Closed"). Data points for temperature and humidity were extracted from the weather website Raspisaniye Pogodi Ltd[3]. In total 4,257 data points were collected between February 2020 and February 2021.

From daily infection numbers, we estimate R_t using the method reported in [7, 30]. R_t is one of the most important quantities used to measure the epidemic spread. If $R_t > 1$, then the epidemic is expanding at time t, whereas if $R_t < 1$, then it is shrinking at time t. A *serial interval distribution*, which is a Gamma distribution g(τ) with mean 7 and standard deviation 4.5, is used to model the time between a person getting infected and them subsequently infecting another

[2] https://www.gov.uk/government/organisations/public-health-england
[3] https://rp5.ru/Weather_in_the_world

person on day τ. The number of new infections c_t on a day t is computed as:

$$c_t = R_t \sum_{\tau=0}^{t-1} c_\tau g_{t-\tau}, \tag{3}$$

where c_τ is the number of new infections on day τ,

$$g_1 = \int_{\tau=0}^{1.5} \mathrm{g}(\tau) d\tau,$$

and for $s = 2, 3, \ldots$,

$$g_s = \int_{\tau=s-0.5}^{s+0.5} \mathrm{g}(\tau) d\tau.$$

From Equation 3, we have:

$$R_t = \frac{c_t}{\sum_{\tau=0}^{t-1} c_\tau g_{t-\tau}} \tag{4}$$

For $x = t$ and τ, c_x is the difference between the confirmed case on day x and the confirmed case on day $x - 1$, which is available from the dataset directly.

With this data, we pose a simple classification question:

> *Given the infection number and control measures implemented on a day t, is $R_t \geq 1$?*

To account the fact that control measures take time to affect the infection rate, we expand the dataset to include the duration of control measure implementation for all control measures. For example, *"Meeting Indoors (High) = 5"* means that *"it is the 5th day that meeting indoors has been banned completely"*. Similarly, *International Travel (Low) = 0* means that *"there is no restriction implemented on international travel"*. We also drop instances before March 15, 2020 across all 12 regions in our dataset due to the low number of infections.[4] In this way, we form a data file with 25 features and 3,937 instances with 2,288 positive ones.

To demonstrate the effect of underspecification, we stratify the dataset D into 11 random groups with sizes 100 to 3500, respectively. We train 100 random forest predictors with each group in D and compute explanation correlations using the process described in Section 2. In addition, we also calculate the classification accuracy over the remaining dataset. Figure 3 shows the results from these experiments. We observe that as the dataset size increases, both the classification performance and explanation correlation increase, as expected.

[4] As can be seen from Equation 4, when c_x is small, R_t can flatulate in a unrealistically large range and generate noises in the dataset.

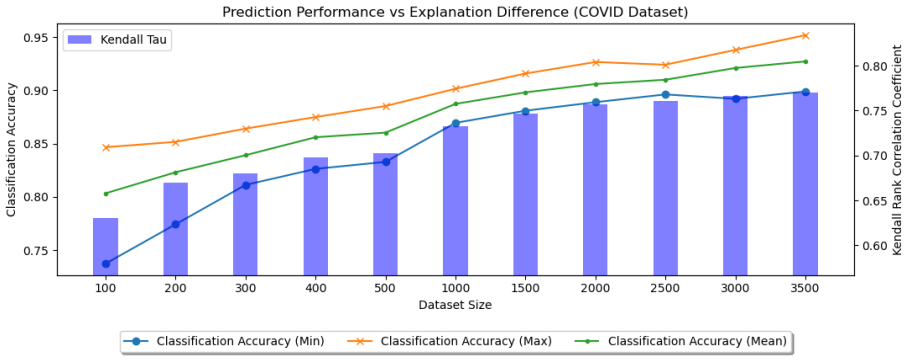

Fig. 3. COVID-19 R_t classification case study.

4 Related Work

As briefly discussed in the Introduction, stress tests have been used to identify underspecification [4]. In particular, stratified performance evaluations, testing whether different strata of a dataset give similar performance on a predictor (see e.g., [1, 21]), shifted performance evaluation, testing whether the average performance of a predictor generalises when the test distribution differs in a specific way from the training distribution (see e.g., [11, 28]), and contrastive evaluation, testing whether a particular modification of the input causes the output of the model to change in unexpected ways (see e.g., [24, 10]) are notable approaches. Comparing with these, our work studies underspecification from a different angle.

Underspecification has been studied in the ML literature in different notions. In deep learning, the discussion focuses on the local geometric properties of objective functions [2], and the geometry of loss surfaces in model averaging and network pruning [13, 8, 29, 9]. Recently there have been analyses of overparameterisation in theoretical and real deep learning models, where underspecificaiton is considered to be caused by potential more degrees of freedom than datapoints induce [17, 20]. In [6, 26, 15, 25], underspecificaiton is treated as different near-optimal solutions for a single learning problem specifications having different properties such as interpretability or fairness.

Our idea of looking at underspecification from the explanation dimension is highly relevant but also orthogonal to the line of recent works on "right for the right reason", for example [25] and [16]. In [25], domain knowledge capturing "right explanations" and human experts are introduced in an ML pipeline to directly assist the prediction and select the most suitable predictor from a group of predictors based on their explanations, respectively. In [16], predictors for natural language inference tasks are tested against a set of common but sometimes wrong reasons, encoded as learning heuristics benchmarks. Comparing with these, we do not attempt to increase prediction performance or develop

datasets for benchmarking; instead, we focus on studying the relation between explanations and underspecification and show that the number of "distinct" explanations, or the "average distance" between explanations, generated from different predictors is a good indicator for the degree of underspecification.

5 Conclusion

In this work, we present an alternative approach that identifies underspecification by investigating explanation correlation. Simply put, given a set of equally high performing predictors trained from an ML pipeline, if they produce highly correlated explanations to their predictions, then the ML pipeline is not underspecified; otherwise, the pipeline is underspecified. We illustrate our approach in multiple classification tasks and in a real-world case study. Our results show that having more data usually helps to address underspecification.

As an early work in studying underspecification, there are several limitations of this work we plan to address in the future. Firstly, we believe that the concept of *underspecification* must be further refined. The current state-of-the-art as represented by [4] suggests underspecification is a qualitative concept without precise quantification. However, to advance this field, measurable quantification is needed so researchers can compare two different ML pipelines and compare their degrees of underspecification quantitatively so "improvement" can be discussed meaningfully. We believe explanation correlation suggested in this work could be such a metric, yet a deeper study is needed.

Secondly, additional explanation generation algorithms should be considered. As feature attribution algorithms are in rapid development, there are techniques other than SHAP, e.g., LIME [23], that also compute feature weights. Although SHAP shows certain superiority over LIME as found in some studies [14, 12, 22, 27], it would be interesting to see whether our SHAP-based results can be reproduced with LIME, or some other interesting behaviours can be discovered.

Thirdly, other forms of machine learning should be studied. This work has focused on classification tasks in supervised learning. We need to consider regression and unsupervised learning tasks. We believe some of the techniques introduced in this work could be carried over to a regression setting. However, carefully planned experiments are necessary to validate such approaches. For analysing underspecification in unsupervised learning, some theoretical work is needed to clearly define and scope the problem.

Lastly, this work focuses solely on identifying underspecification. Ultimately, we would like to have a technique that addresses underspecification with data that is currently available. To this end, the technique needs to select predictors with the "correct" inductive bias. We would like to explore whether explanation properties can be used for such identification.

Acknowledgements

This work is supported by the Welsh Government Office for Science, Ser Cymru III programme – Tackling Covid-19.

References

1. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A., Wilson, C. (eds.) FAT. Proceedings of Machine Learning Research, vol. 81, pp. 77–91. PMLR (2018)
2. Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., Zecchina, R.: Entropy-sgd: Biasing gradient descent into wide valleys. Journal of Statistical Mechanics: Theory and Experiment **2019**(12), 124018 (2019)
3. Cock, D.D.: Ames, iowa: Alternative to the boston housing data as an end of semester regression project. Journal of Statistics Education **19**(3) (2011)
4. D'Amour, A., et al.: Underspecification presents challenges for credibility in modern machine learning. CoRR **abs/2011.03395** (2020)
5. Dua, D., Graff, C.: UCI machine learning repository (2017), http://archive.ics.uci.edu/ml
6. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research **20**(177), 1–81 (2019)
7. Flaxman, S., Mishra, S., Gandy, A., Unwin, H., Coupland, H., Mellan, T., Zhu, H., Berah, T., Eaton, J., Perez Guzman, P., et al.: Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on covid-19 in 11 european countries. Tech. rep., Imperial College London (2020)
8. Fort, S., Hu, H., Lakshminarayanan, B.: Deep ensembles: A loss landscape perspective. arXiv preprint arXiv:1912.02757 (2019)
9. Frankle, J., Dziugaite, G.K., Roy, D., Carbin, M.: Linear mode connectivity and the lottery ticket hypothesis. In: International Conference on Machine Learning. pp. 3259–3269. PMLR (2020)
10. Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E.H., Beutel, A.: Counterfactual fairness in text classification through robustness. In: Conitzer, V., Hadfield, G.K., Vallor, S. (eds.) Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019. pp. 219–226. ACM (2019)
11. Hendrycks, D., Dietterich, T.G.: Benchmarking neural network robustness to common corruptions and perturbations. CoRR **abs/1903.12261** (2019)
12. Honegger, M.: Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions. CoRR **abs/1808.05054** (2018)
13. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., Wilson, A.G.: Averaging weights leads to wider optima and better generalization. arXiv preprint arXiv:1803.05407 (2018)
14. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA. pp. 4765–4774 (2017)
15. Marx, C., Calmon, F., Ustun, B.: Predictive multiplicity in classification. In: International Conference on Machine Learning. pp. 6765–6774. PMLR (2020)
16. McCoy, T., Pavlick, E., Linzen, T.: Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In: Korhonen, A., Traum, D.R.,

Màrquez, L. (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. pp. 3428–3448. Association for Computational Linguistics (2019)
17. Mei, S., Montanari, A.: The generalization error of random features regression: Precise asymptotics and double descent curve. arXiv preprint arXiv:1908.05355 (2019)
18. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
19. Molnar, C.: Interpretable Machine Learning (2019), https://christophm.github.io/interpretable-ml-book/
20. Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., Sutskever, I.: Deep double descent: Where bigger models and more data hurt. arXiv preprint arXiv:1912.02292 (2019)
21. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (10 2019)
22. Rathi, S.: Generating counterfactual and contrastive explanations using SHAP. CoRR **abs/1906.09293** (2019)
23. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 1135–1144. ACM (2016)
24. Ribeiro, M.T., Wu, T., Guestrin, C., Singh, S.: Beyond accuracy: Behavioral testing of NLP models with checklist. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. pp. 4902–4912. Association for Computational Linguistics (2020)
25. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. In: Sierra, C. (ed.) Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017. pp. 2662–2670. ijcai.org (2017)
26. Semenova, L., Rudin, C., Parr, R.: A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. arXiv preprint arXiv:1908.01755 (2019)
27. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: Markham, A.N., Powles, J., Walsh, T., Washington, A.L. (eds.) Proc. of AIES. pp. 180–186. ACM (2020)
28. Wang, H., Ge, S., Lipton, Z.C., Xing, E.P.: Learning robust global representations by penalizing local predictive power. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 10506–10518 (2019)
29. Wilson, A.G., Izmailov, P.: Bayesian deep learning and a probabilistic perspective of generalization. arXiv preprint arXiv:2002.08791 (2020)
30. Wu, J.T., Leung, K., Bushman, M., Kishore, N., Niehus, R., de Salazar, P.M., Cowling, B.J., Lipsitch, M., Leung, G.M.: Estimating clinical severity of covid-19 from the transmission dynamics in wuhan, china. Nature Medicine pp. 1–5 (2020)
31. Yalcin, O., Fan, X., Liu, S.: Evaluating the correctness of explainable ai algorithms for classification. CoRR **abs/2105.09740** (2021)