Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

EXPLAINABLE AI FOR DATA FARMING OUTPUT ANALYSIS: A USE CASE FOR KNOWLEDGE GENERATION THROUGH BLACK-BOX CLASSIFIERS

Niclas Feldkamp
Jonas Genath
Steffen Strassburger

Group for Information Technology in Production and Logistics
Technische Universität Ilmenau
Max-Planck-Ring 12
Ilmenau, 98693, GERMANY

ABSTRACT

Data farming combines large-scale simulation experiments with high performance computing and sophisticated big data analysis methods. The portfolio of analysis methods for those large amounts of simulation data still yields potential to further development, and new methods emerge frequently. Among the most interesting are methods of explainable artificial intelligence (XAI). Those methods enable the use of black-box-classifiers for data farming output analysis, which has been shown in a previous paper. In this paper, we apply the concept for XAI-based data farming analysis on a complex, real world case study to investigate the suitability of such concept in a real world application, and we also elaborate on which black-box classifiers are actually the most suitable for large-scale simulation data that accumulates in a data farming project.

1 INTRODUCTION

For complex simulation models, conducting a traditional simulation study usually aims to achieve a specific, pre-determined goal, such as conducting a scenario-based analysis or even a simulation-based optimization. This can still leave a lot of room for actually understanding the behavior of the model in terms of relationships between factors and outcomes (Feldkamp et al. 2015; Painter et al. 2006). Sometimes, the discovery of new and interesting relationships that were previously unknown and outside the previously defined scope of a simulation project can actually improve decision making. In this context, the method of data farming has been developed (Horne and Meyer 2005). Data farming refers to the method of extracting data from the simulation model by using large-scale experimental design, high-performance computers for massively parallelized experiments to focus on more complete coverage of possible system responses and machine-assisted analysis (Horne and Schwierz 2008). Data farming problems fall into a category of problem solving that Lempert et al. define as long-term policy analysis that requires consideration of large ensembles of scenarios (Lempert et al. 2003). Data farming research has also always been concerned with the application of advanced data analysis methods to process these large volumes of simulation output data and produce appropriate and adequate insights (Lucas et al. 2015; Sanchez 2014). The concept of Knowledge Discovery in Simulation Data (KDS) was developed to dive deep into the analytics side of data farming and provide a detailed process model for applying data mining methods along with appropriate visualization and interaction methods to large-scale simulation data (Feldkamp et al. 2020). One aspect of the KDS process model is the use of model building algorithms that can approximate relationships between simulation input and output data. Rules from these models can then be used to derive knowledge about the system (Feldkamp et al. 2015). Initially, only white-box models such as mining frequent patterns, decision trees, or Bayesian networks have been used for this, since those expose their internal set of rules and mappings in order to derive knowledge from it. However, in a recent paper, we introduced a concept for using methods of explainable artificial intelligence (Feldkamp 2021). Those overcome the lack of explainability and transparency of black-box algorithms such as artificial neural networks, which are known for being very precise at approximation even complex relations and response surfaces. In our previous paper, we evaluated the concept using an academic case study, which left little room for testing different black-box algorithms given its simplicity. In this paper, we use a complex real-world case study to evaluate the suitability of different black-box algorithms in detail, as well as to examine the applicability of the concept to a real-world data farming project in general. The remainder of this paper is structured as follows: In section two, we give an overview of the related work, namely being XAI, data farming, and the combination of both. In Section 3, we present the case study, by first introducing the scenario and simulation model, followed by an evaluation of suitable black-box classifiers. This is then followed by the application of the XAI-based concept for data farming output analysis. The paper ends with some final remarks and a discussion of possible future work in Section 4.

2 RELATED WORK

2.1 Explainable AI

The transparency of decisions made by artificial intelligence and machine learning algorithms is becoming increasingly important, especially when people's everyday lives are directly affected. A decision explained with "because the computer said so" is no longer sufficient and consequently can also cause legal problems when suspected of unfairness and discrimination, for example in areas such as recruitment or credit granting (Dosilovic et al. 2018; Guidotti et al. 2019). Some lawmakers even think of a "right to an explanation" (Edwards and Veale 2017). On the other hand, black-box methods of machine learning and artificial intelligence, such as artificial neural networks, are among the most powerful algorithms for prediction and classification tasks. Due to the complexity of these algorithms in contrast to their white-box counterparts, such as decision trees and linear regression, there is a trade-of between performance and interpretability. Therefore, XAI has become a popular field of research recently in an effort to make black-box algorithms transparent. The term XAI actually encompasses a very wide range of diverse methods and so are the efforts to catalog and categorize those methods (Barredo Arrieta et al. 2020; Ras et al. 2018; Tjoa and Guan 2020). The most straight-forward distinction is between global and local explanations. While local explanations try to explain one individual predication, global explanation methods aim to explain the underlying model and its internal relations in a more general way, but these usually come at higher computational cost. Among the most commonly used and frequently cited methods are Local Interpretable Model-Agnostic Explanations (Lime) (Ribeiro et al. 2016), Anchors High-Precision Model-Agnostic Explanations (Anchors) (Ribeiro et al. 2018), and SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017).

2.2 Data Farming Output Analysis Using Explainable AI

Data farming combines large-scale simulation experiments with high-performance computing and sophisticated big data analysis techniques. The portfolio of analysis methods for these large volumes of simulation data still offers development potential, and new methods are frequently emerging. As an extension, the concept of knowledge discovery in simulation data was developed to take a deep dive into the analysis side of data farming by providing a process model and workflow for using data mining methods and suitable, interactive visualizations (Feldkamp et al. 2015). This is especially helpful for models with a large number of relevant outputs that exhibit a complex, multidimensional response surface. When outputs are congregated into groups of disparate system behavior by using unsupervised data mining algorithms such as clustering, the relation between factors and outputs can then be investigated to conclude knowledge about the system (Feldkamp et al. 2020). For this purpose, supervised learning algorithms can create models that represent the relations between simulation input data and previously created clusters, from which in turn we can derive rules that can contribute to knowledge creation through human interpretation. For a supervised algorithm, each simulation experiment acts as a training record (Feldkamp et al. 2020). In other words, a classification problem needs to be solved. Extracting the underlying decision rules for the classification from white-box algorithms such as decision trees is relatively easy, as those are explicitly visible. However, other classification algorithms like for example artificial neural networks or random forests are very good at approximating a given input-output-relation for predictions, but less good at actually revealing the underlying model regarding its rules and relations. However, this is what we need in order to derive analyses and knowledge from it. To overcome the lack of understanding of such black-box algorithms, XAI can help to make those rules and relations visible and explain them in an understandable and comprehensible way. The workflow that we proposed in (Feldkamp 2021) is shown in Figure 1.

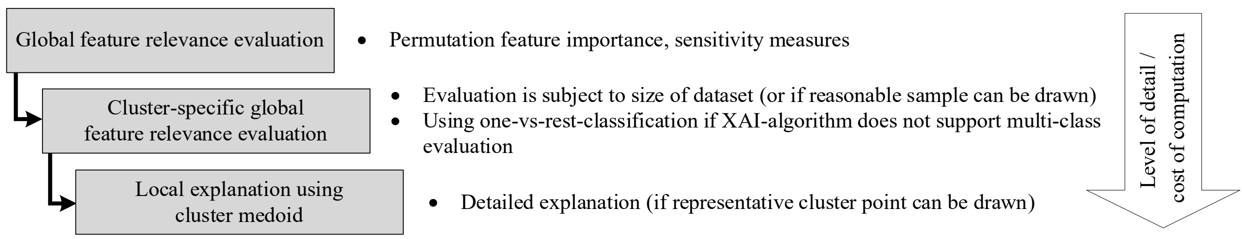

Figure 1: Workflow for application of XAI-based methods for data farming output analysis (Feldkamp 2021).

The starting point are easy-to-calculate relevance measures, such as the importance of permutation features. These measures evaluate the overall importance of a factor for the output of the model, but they can only provide a rough overview. The next step is to extend the analysis to the actual classifications (i.e., the assignment of factors to clusters) at the global level. Using these methods, we can draw conclusions about the differences between clusters, i.e., which factor scores contribute to which cluster assignments. It should be noted that many XAI methods only support explanations for binary classifications. However, when using the KDS workflow, the data is usually classified into multiple clusters, creating a multi-class classification problem. This problem can be circumvented by converting the analysis into a one-versus-rest classification. This means that we compare each cluster of interest with the rest of the data. The drawback of this approach is that each of these comparisons requires a separately trained classification model and subsequent XAI evaluation. The final step is to use local explanation methods, that can be used to analyze selected representative points, such as the cluster medoid (representing the average of a given cluster).

3 USE CASE SCENARIO

3.1 Scenario and Simulation Model

To investigate the suitability of XAI methods for data farming output analysis in real world applications, we used the discrete-event simulation model of a production line from the field of high voltage engineering. More precisely, in our case study insulators for high-voltage plants are manufactured. The process is schematically shown in Figure 2. The simulation model was built using the simulation software Siemens Plant Simulation.

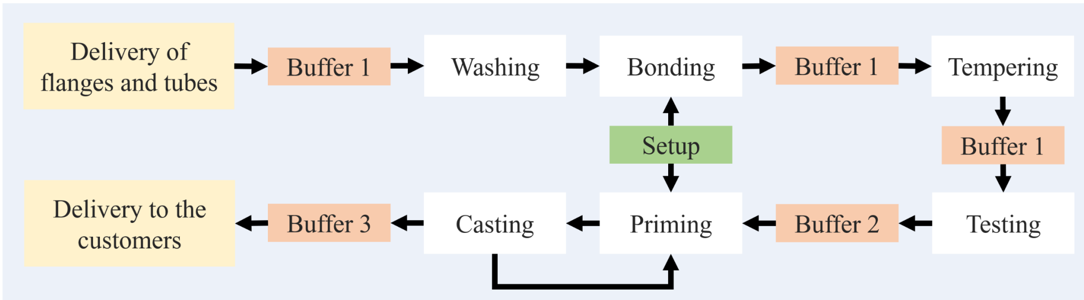

Figure 2: Flowchart of the production process.

First, the delivered insulator tubes and flanges must be washed using a washing system so that the subsequent bonding, priming and casting processes can be performed. The tubes are then joined together with the flanges. Subsequently the tube is coated with the insulator in several priming and casting processes. A big challenge of this production line are the different sizes and shapes of the insulators depending on the customer requirements and a large number of special processes. For this reason, many setup operations have to be performed and several specially trained workers are available, all of whom have their own set of workflow rules. Therefore a key goal of optimizing this use case is to reduce setup times and maximize worker utilization. The quality of these manufacturing processes is ensured by several testing and control procedures. After the complete insulator is manufactured, it is ready to be delivered to customers.

In the simulation model of the described use case, the following adjustable factors are available. First of all, it is possible to vary the number of certain available workers. For worker ten (*Worker 10 count*), the number can be set between one and four, which all perform the same work. For other workers it is possible to vary their availability, this includes the workers two (*Worker 2 available*), nine (*Worker 9 available*) and a second worker eleven (*Worker 11-2 available*). These all have procedures specifically set for them and therefore presumably have an impact on the performance of the system. The second group of factors relates to the washing system, which includes the volume of the washing basket (*Volume washing basket*), the maximum parts in the washing basket (*Maximum parts washing basket*), the setup time (*Setup time washing basket*), and the residual running time of the washing system (*Residual running time washing machine*). With regards to the insulators, the maximum laytime after the washing process (*Max laytime after washing*) and the buffer size (*Buffer size joining*) in which the insulators can be temporarily stored before being transported for joining can be varied. Finally, the orders for manufacturing the insulators can be sorted according to six different sorting strategies (*Sorting strategy*). These include earliest due date, random sorting, some strategies which aim at optimizing setup times and sorting according to the size of the order quantity (larger orders first). For the experimental design, an LHS with the size of ten thousand was created, which was crossed with the six sorting strategies, as it was suspected that these would have a larger impact on the output data. This sums up to sixty thousand experiments that have been conducted in the corresponding data farming study. For this study, this was the best tradeoff between performance and information gain. After performing these experiments the most relevant outputs for this simulation model according to their magnitude of variance are the cycle time, the worker utilization (average for all workers) and the amount of insulators which exceeded their maximum laytime. To group the generated data farming output, we used a Gaussian mixture model with the mentioned output for which we first transformed the data to the interval zero and one to improve the training process. Three components achieved the best separation of the data and the result is shown in Figure 3 as a scatterplot matrix where each dot represents one simulation experiment. The orange cluster contains the bad performing experiments where the staff utilization is low, the number of insulators exceeding the laytime is high, and the cycle time is above average. The green cluster, on the other hand, collects the good performing experiments, with the best worker utilization, low cycle time, and the fewest insulators that exceed the laytime.

On this basis, we performed an analysis of the data using typical data farming methods in order to subsequently investigate whether XAI methods can confirm these findings or even find new knowledge.

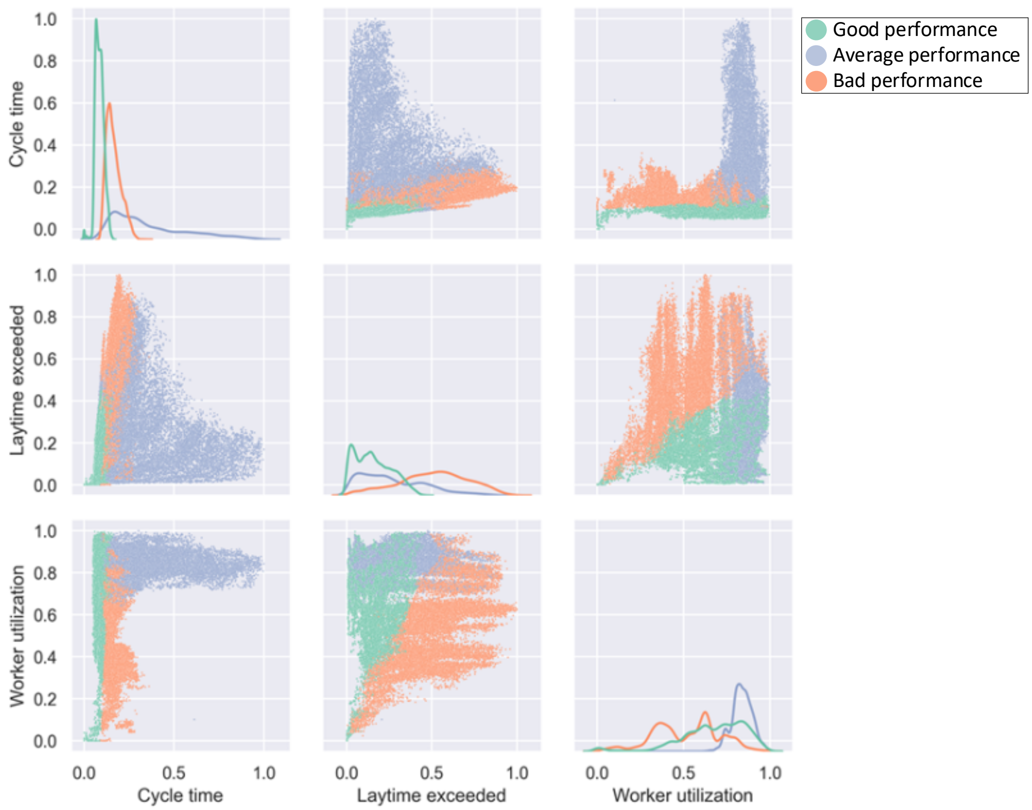

Figure 3: Scatterplot matrix of the results of the clustering based on three outputs.

An important finding of the analyses is that the number of parts in the washing basket has a significant impact on the output of the system. It can be suspected that after the washing system, the production line mainly lacks small parts and the throughput and the worker utilization decreases. Therefore, the number of parts seems to limit the system more than the volume of the washing basket. Another important finding is that allowing a high maximum laytime leads to high worker utilization. Therefore, our management recommendations include the following: Try to maximize the number of parts that are washed at the same time and to check if there are possibilities to extend the allowed laytime of the insulators (e. g. dust protection). A deeper insight into the results through typical data farming methods would exceed the scope of this paper. The goal here is to investigate if XAI-based analysis can be applied on a real world case study, and if we can confirm the previous findings or even find new insights.

3.2 Preliminary Considerations on Black Box Classifiers

Before using XAI-algorithms for analyzing the data farming output, we need to train an AI-model in the first place. This then raises the question which type of classification algorithm is best suited for modeling the response surface of an discrete event simulation model in the context of production and logistics as introduced in the previous section. Large quantities of farmed simulation data tend to expose some special features that the regarding AI-model needs to reflect on. The nature of the input space is usually packed very densely. For the response, we cluster multiple outputs that are relevant into different clusters, resulting into a one-dimensional response that corresponds with the cluster allocation. Since the underlying outputs that are used for the clustering also tends to be very dense and packed, we apply partitioning clustering algorithms like k-means that sharply partition the response surface into multiple cluster areas, since this has turned out to produce the best separation between clusters. In some cases, an clustering can even be improved by using Gaussian mixture models. Those work very similar to k-means, but add a variance component to each cluster, which basically means that the cluster may overlap in some of their underlying output dimensions. This can render very problematic for some classification algorithms, since it means that in some rare cases the final cluster allocation could be different although the input factor values are nearly similar for some datasets. This can furthermore be problematic if the model contains strong stochastic influences. This is usually intercepted by replications. However, it must be taken into account that in view of limited computing capacity, the number of actual experiments must be reduced by increasing the number of replications. Which number of replications is considered appropriate is quite controversial if the generated data are to be further analyzed with other mathematical methods. Santos and Santos were able to show that a lower number of replications in favor of a higher number of experiments does not increase the accuracy of regression models trained on them (Santos and Santos 2009). However, MacDonald and Gunn, in turn, were able to show that this is not true for non-parametric regressions. According to this, a good coverage of the result space is clearly more important than the statistical accuracy of a single data point (MacDonald and Gunn 2012). Different methods can react differently and sensitively to the properties of data sets generated by data farming, and can therefore behave differently with regard to their accuracy in terms of mapping the relationship between factors and cluster allocation. We carried out a comparison of accuracy for the most commonly used black-box classification algorithms. For this purpose, we let each of them fit a model of the farmed simulation data in terms of mapping factors to the cluster allocation as shown in Section 3.1. We then calculated the accuracy by letting them predict the complete data set. Note that hyperparameters for each algorithm were tuned by pre-testing to achieve the highest possible accuracy. Hyperparameters will not be discussed further here, as these must be adjusted each time anew depending on the individual model and use case via testing in order to achieve a good performance. There was no splitting in training and test data, because overfitting is actually desired. This is because the trained model is not utilized to make predictions but rather to learn from its internal input/output mapping and approximation of rules. Because we already have a large and mostly complete representation of the response surface through data farming, a training phase as thoroughly as possible is desired, so that the existing data can be fitted as smoothly as possible. The result of this comparison is shown in Figure 4. We can see that with the exception of artificial neural network and support vector machine, all classification algorithms were able to fit the model with an accuracy of almost 100%. Since the dataset used for training contains 60 000 records, the reason for the lower accuracy of the artificial neural network may be that these data points are insufficient to overtrain the network. For the support vector machine, the overlaps between data points are probably too large, as can be seen in Figure 4. This makes it impossible for the support vectors to separate the data 100% correctly. The same presumably applies for the artificial neural network. These overlaps result from our choice of a Gaussian mixture model for the clustering, as explained above.

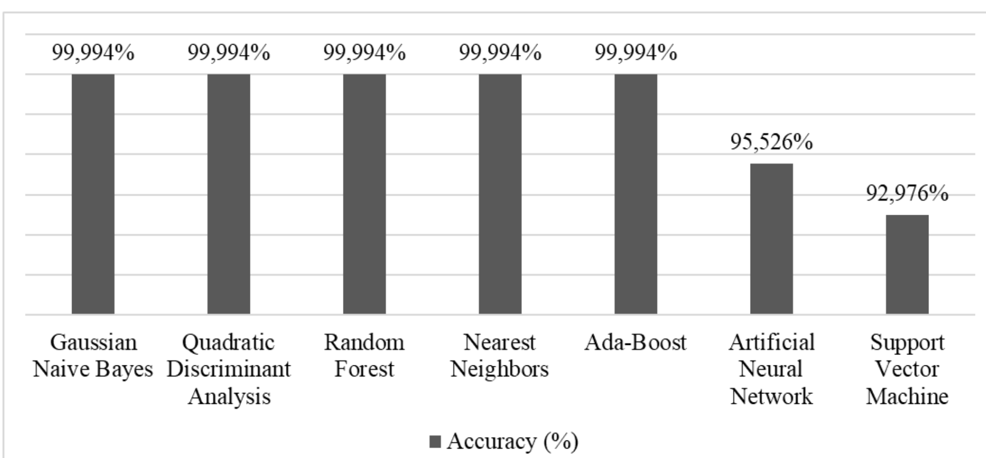

Figure 4: Comparison of classifier accuracy

In addition to prediction accuracy, we also measured the time for training the model, as shown in Figure 5 (left side). Here, we see some dramatic differences in training time, with artificial neural network and support vector machine again among the worst performing contenders. As a third important metric, we measured the average prediction time, which is the time needed to predict the total dataset once, as shown in Figure 5 (right side).

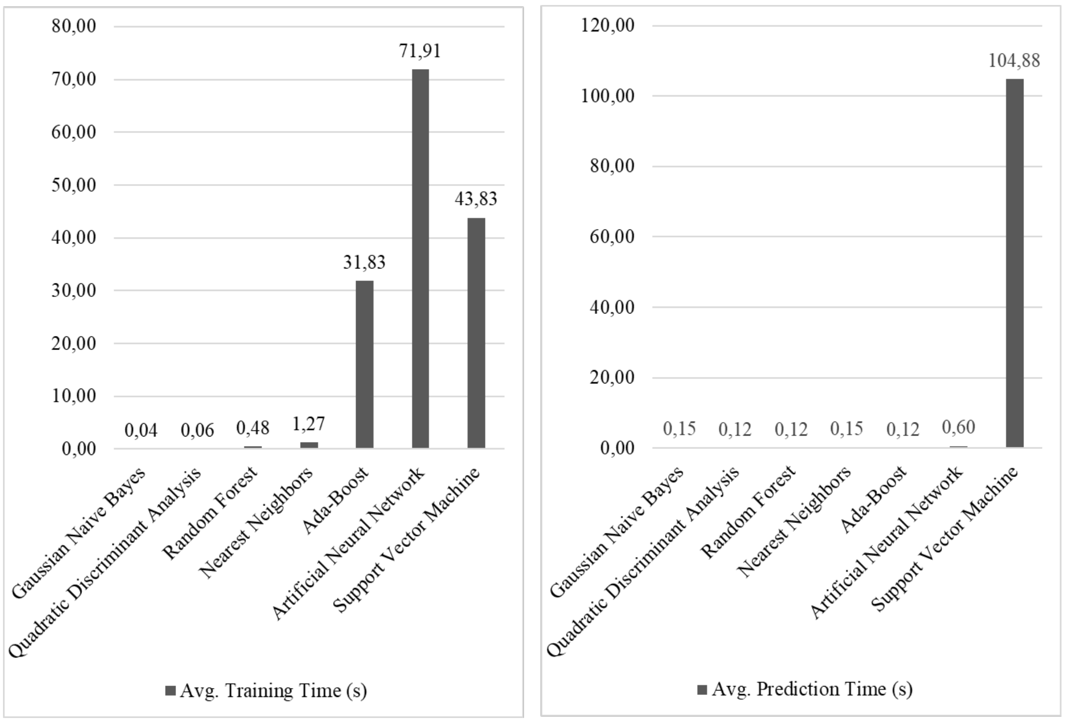

Figure 5: Comparison of average classifier training time (left) and average prediction time (right).

XAI algorithms are usually based on making predictions using the underlying model and deriving the respective explanations through a large number of pairwise comparisons. The computation of XAI explanations can be extremely computationally intensive and thus lengthy, in most cases significantly longer than the actual training time of the underlying model. The prediction time of a model therefore also contributes significantly to how long the XAI algorithm needs to explain the model. The measured prediction time is nearly the same for classification algorithms except once again artificial neural networks which is about 4 times slower than the other contenders and support vector machine being several hundred times slower. In summary, the comparison shows that quadratic discriminant analysis and random forest should be favored for using them for subsequent XAI analysis of the data farming output. Both show the highest accuracy, fastest prediction time, and adequate training time. As a result, the differences in the suitability of black-box classifiers between simple academic case studies (Feldkamp 2021) and complex real-world applications can be identified. While in the academic case studies no differences in accuracy could be observed, surprisingly the more traditional machine learning algorithms (random forest and naïve bayes) perform better in a real world application than the ones that get more attention in scientific and public perception nowadays (artificial neural networks and support vector machines). Reinforced by the results in training time and prediction time, the assessment can be made that for real world applications machine learning models that have fewer hyperparameters are better suited for XAI-based evaluations of data farming datasets, such as random forest.

3.3 Results Using XAI-based Analysis

In this section, we discuss the results of applying XAI-based data farming output analysis in order to see if we can confirm findings from the previous study or if we even can find new insights. Figure 6 shows the results of an XAI method called permutation feature importance. This method evaluates the contribution of factors to the cluster allocation in general, i.e., the contribution of an allocation to any of the three clusters. The figure confirms the finding that the factors *max laying time after washing* and *maximum parts washing basket* have the most influence on cluster allocation. The factor *worker 10 count* also has a recognizable influence, since the availability of this employee can be varied between 1 and 4. This was therefore to be expected in previous analyses and is confirmed here. Noteworthy, but also expected, is that the sorting strategy 2 (random sorting) has no influence on the cluster allocation, whereas sorting strategy 3 has the most influence on the cluster allocation of all sorting strategies.

A new finding, however, is the high importance of the availability of employee 2 (*worker 2 available*), which did not emerge to the same extent in the previous analyses. In further detailed analyses, we are going to evaluate which concrete factor values lead to which cluster assignment regarding this factor.

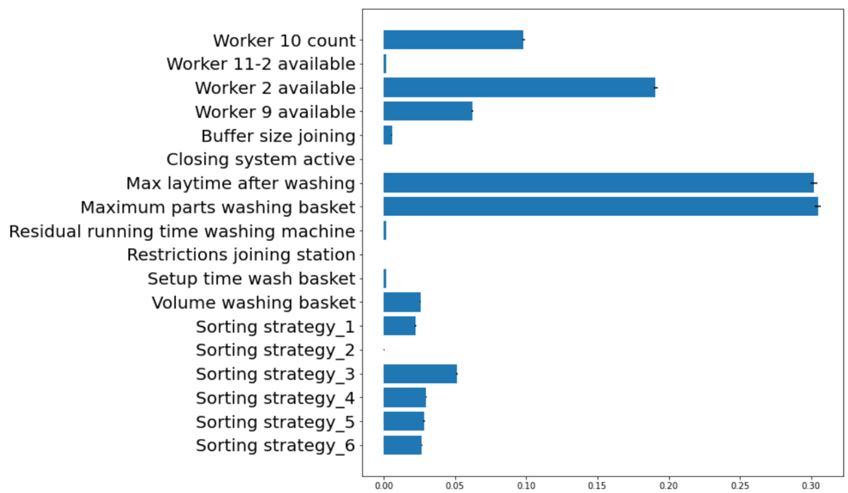

Figure 6: Results of feature relevance evaluation using permutation feature importance.

To perform a more in-depth analysis, we used the SHAP package for computation of so-called SHAP values (Lundberg et al. 2020; Lundberg and Lee 2017), which is shown in Figure 7. The concept of SHAP values was adapted from game theory, where the Shapley value quantifies the contribution of each player to the outcome of a game. The SHAP value for XAI picks up this idea by treating every factor of a black-box model as an individual player, thus calculating their contribution to the overall outcome of a prediction (representing the outcome of one game). This is done by aggregating the marginal contribution of each factor in every possible combination with the other factors present. This makes the calculation of SHAP values very costly in terms of computation time, because computation time grows exponentially with the number of factors. The individual SHAP values for the factors can be interpreted relative to the base effect, comparable to how we would interpret coefficients relative to the intercept in a linear regression model. We again resort to a binary one-vs-rest classification, so a value of 1 indicates a classification towards the corresponding good performance cluster, and a value of 0 indicates a classification towards not being in those cluster, hence being in one of the other clusters. Figure 7 shows a random selection of experiments, represented by the vertical lines. For each line, the explained contribution of each factor to the classification is shown, where a value greater than 0.5 indicates a contribution to good performance classification, while a value smaller than 0.5 indicates a contribution to the other classification. The color of the lines indicates the actual classification. Note that the mean value is not 0.5 but less than 0.4. This is because there are significantly more experiments in the other cluster and the average contribution shifts accordingly.

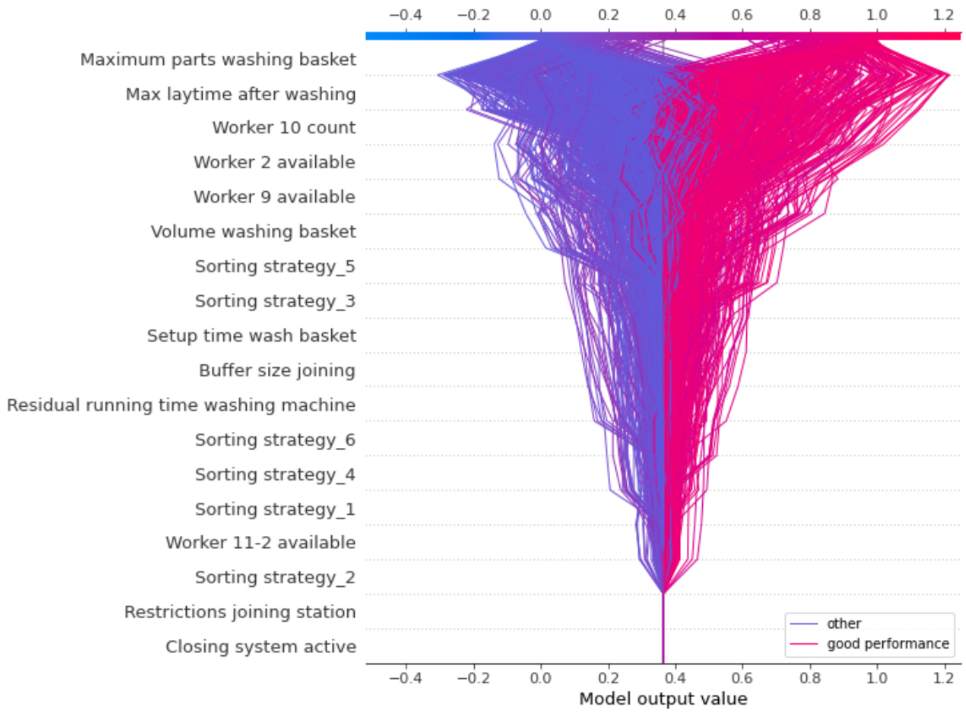

Figure 7: SHAP values plot for summarizing all factors sorted by importance in descending order.

Again, we can see that the factors *maxium parts washing basket*, *max laytime after washing*, and *worker 10 count* have the strongest influence on the cluster allocation regarding good performance, with *maximum parts washing basket* consistently being the largest influencing factor, but some simulation experiments show some extreme contributions for *max laytime after washing* in one direction or the other, strongly influencing the final classification. In addition, there are plots available that can show the SHAP values for individual factors for the range of their factor values. One example of this is shown in Figure 8 for the factor *max laytime after washing*. Interestingly, we can see that high factor values strongly contribute to good performance, as we already knew from previous analysis, but nevertheless we see that even factor values below 45.000 seconds contribute to good performance in most cases, although the contribution is not quite as strong.

Figure 8: SHAP values plot for the factor *maximum laytime after washing*. The X-axis represents the factor value (in seconds), the y-axis represents the SHAP values.

For an even more detailed explanation of results, we used the Anchors-Package (Ribeiro et al. 2018), as shown in Figure 9. First, we again analyzed the good performance cluster by conducting a binary good-vs-other classification. Because Anchors is a local explanation method, which means that it is used to explain individual samples, we took the cluster medoid for explanation. The medoid is the point that is the closest to the average of all points in the cluster, therefore representing the majority of all points. The result of this explanation is shown in the left side of Figure 9. We can see a rather complex rule with a lot of components, but nonetheless confirming our assumptions form the previous analyses. However, we can see that the factor *worker 2 available* seems not to be relevant here, although it was the third most important factor according to the relevance evaluation over all three clusters (Figure 6) and the fourth most important factor according to the good-vs-rest evaluation (Figure 7). We therefore looked for samples that were not predicted as good performance, and it became obvious that this factor becomes relevant in a certain combination with the factor *maximum parts washing basket* (Figure 9 right side), where the outcome is not good performance but rather one of the two other clusters in almost 100% of the cases. This is another finding that was not discovered in previous analyses before using XAI-methods.

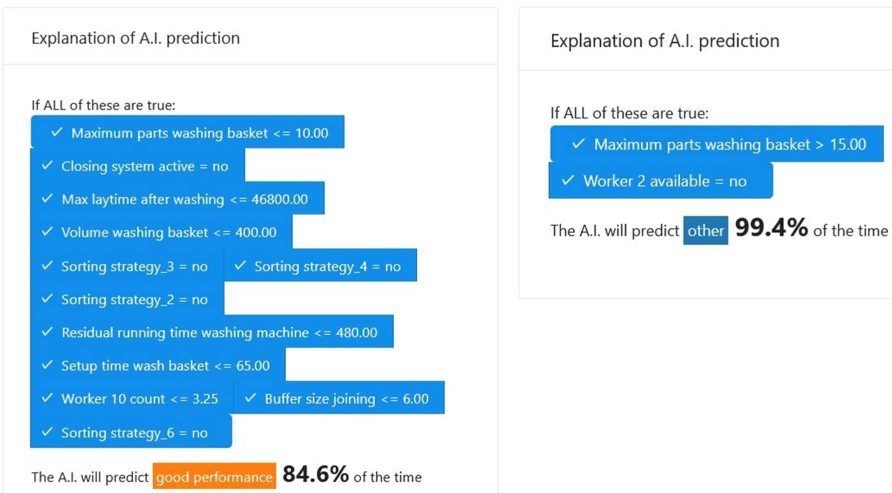

Figure 9: Local explanations using Anchors showing explanations for good performance vs other clusters comparison.

We also conducted a binary bad-vs-other classification and used Anchors to explain the medoid of the bad cluster, which is shown in Figure 10.

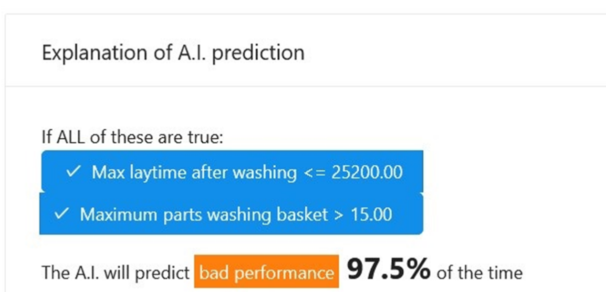

Figure 10: Local explanations using Anchors showing explanations for bad performance vs other clusters comparison.

We see the rule for this cluster allocation is much simpler than for the good performance case, and it also has a much higher probability. However, since we are only looking at the cluster medoid, we can most likely assume that other combinations of values of those two factors would also lead to bad performance allocation, if we look at points that are located more towards the boundary of the cluster.

4 CONCLUSIONS AND FUTURE WORK

In this paper, we successfully applied methods of explainable artificial intelligence for the output analysis of data farming projects in a real world case study. We were able to confirm previously known relations in the model and also could investigate some of those findings in even more detail. Furthermore, we were also able to see findings that were previously unknown and could not be extracted with traditional analysis methods. Therefore, the application of XAI-based output analysis can be worthwhile, especially if a level of detail is needed that cannot be reached with other analysis methods. Another advantage is that these methods usually have a very intuitive, graphical approach and are therefore easy to interpret, even for users that are not simulation-experts.

However, most of these methods are very computational intensive and therefore are not suitable for time-critical tasks. However, XAI is a very current research topic. Many further developments and optimizations of the methods and algorithms are to be expected, which in turn could also be of interest for use in data farming. Therefore, developments in XAI should be monitored closely for future work.

REFERENCES

Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. 2020. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". *Information Fusion* 58:82–115.

Dosilovic, F. K., M. Brcic, and N. Hlupic. 2018. "Explainable Artificial Intelligence: A Survey". In *41st International Convention on Information and Communication Technology, Electronics and Microelectronics,* edited by K. Skala, 210–215. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Edwards, L., and M. Veale. 2017. "Slave to the Algorithm? Why a Right to Explanation is Probably Not the Remedy You are Looking for". *SSRN Electronic Journal* 16(1):18–84.

Feldkamp, N. 2021. "Data Farming Output Analysis Using Explainable AI". In *Proceedings of the 2021 Winter Simulation Conference,* edited by S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Feldkamp, N., S. Bergmann, and S. Strassburger. 2020. "Knowledge Discovery in Simulation Data". *ACM Transactions on Modeling and Computer Simulation* 30(4):1–25.

Feldkamp, N., S. Bergmann, and S. Strassburger. 2015. "Knowledge Discovery in Manufacturing Simulations". In *Proceedings of the 3rd ACM SIGSIM Conference on Principles of Advanced Discrete Simulation,* edited by S. J. E. Taylor, N. Mustafee, and Y.-J. Son, 3–12. New York, New York: ACM Press.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. 2019. "A Survey of Methods for Explaining Black Box Models". *ACM Computing Surveys* 51(5):1–42.

Horne, G. E., and K.-P. Schwierz. 2008. "Data Farming Around The World Overview". In *Proceedings of the 2008 Winter Simulation Conference,* edited by S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, 1442–1447. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Horne, G. E., and T. E. Meyer. 2005. "Data Farming: Discovering Surprise". In *Proceedings of the 2005 Winter Simulation Conference,* edited by M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 1082–1087. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lempert, R. J., S. W. Popper, and S. C. Bankes. 2003. "Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis". RAND Technical Report MR 1626 RPC, Santa Monica,CA, USA.

Lucas, T. W., W. D. Kelton, P. J. Sánchez, S. M. Sanchez, and B. L. Anderson. 2015. "Changing the paradigm: Simulation, now a method of first resort". *Naval Research Logistics (NRL)* 62(4):293–303.

Lundberg, S. M., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions". In *Advances in Neural Information Processing Systems 30,* edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–4774. Curran Associates, Inc.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. 2020. "From local explanations to global understanding with explainable AI for trees". *Nature Machine Intelligence* 2(1):2522–5839.

MacDonald, C., and E. A. Gunn. 2012. "Allocation of Simulation Effort For Neural Network Vs. Regression Metamodels". In *Proceedings of the 2012 Winter Simulation Conference,* edited by C. Laroque, R. Himmelspach, R. Pasupathy, O. Rose, and A. Uhrmacher. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Painter, M. K., M. Erraguntla, G. L. Hogg, and B. Beachkofski. 2006. "Using Simulation, Data Mining, and Knowledge Discovery Techniques for Optimized Aircraft Engine Fleet Management". In *Proceedings of the 2006 Winter Simulation Conference*: *Monterey, CA, U.S.A., Dec 3-6, 2006,* edited by L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ras, G., M. van Gerven, and P. Haselager. 2018. "Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges". In *Explainable and Interpretable Models in Computer Vision and Machine Learning,* edited by H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, and M. van Gerven, 19–36. Cham: Springer International Publishing.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2018. "Anchors: High-Precision Model-Agnostic Explanations". In *Proceedings of the 32nd Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence,* edited by S. Zilberstein, S. McIlraith, and K. Weinberger. Palo Alto, CA: Association for the Advancement of Artificial Intelligence Press.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. ""Why Should I Trust You?"". In *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining,* edited by B. Krishnapuram, M. Shah, A. Smola, C. Aggarwal, D. Shen, and R. Rastogi, 1135–1144. New York, NY, USA: *Association for Computing Machinery*.

Sanchez, S. M. 2014. "Simulation Experiments: Better Data, Not Just Big Data". In *Proceedings of the 2014 Winter Simulation Conference,* edited by A. Tolk, S. D. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 805–816. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Santos, P., and I. Santos. 2009. "Design Experiments for the Construction of Simulation Metamodels". In *Proceedings of the 23rd European Conference on Modelling and Simulation*: *June 9th - 12th, 2009, Madrid, Spain ; 23rd European Conference on Modelling and Simulation,* edited by J. Otamendi, 338–344. ECMS.

Tjoa, E., and C. Guan. 2020. "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI". *Institute of Electrical and Electronics Engineers Transactions on Neural Networks and Learning Systems* :1–21.

AUTHOR BIOGRAPHIES

NICLAS FELDKAMP is a research associate in the Group for Information Technology in Production and Logistics at the Ilmenau University of Technology. He received his Doctoral degree in business information systems from the Ilmenau University of Technology, and holds M.Sc. from the Ilmenau University of Technology and B.Sc. from the University of Cologne. His research interests include data science, business analytics, and industrial simulations. His email address is niclas.feldkamp@tu-ilmenau.de.

JONAS GENATH has been working since the beginning of 2020 as a research associate in the Group for Information Technology in Production and Logistics at the Ilmenau University of Technology. He received his M. Sc. from the Ilmenau University of Technology in 2019. His research topics focus on the automation of knowledge discovery in simulation data, data science and visual analytics. His email address is jonas.genath@tu-ilmenau.de.

STEFFEN STRASSBURGER is a professor at the Ilmenau University of Technology and head of the Group for Information Technology in Production and Logistics. Previously he was head of the "Virtual Development" department at the Fraunhofer Institute in Magdeburg, Germany and a researcher at the Daimler Chrysler Research Center in Ulm, Germany. He holds a Doctoral and a Diploma degree in Computer Science from the University of Magdeburg, Germany. His research interests include distributed simulation, automatic simulation model generation, and general interoperability topics within the digital factory and Industry 4.0 context. His email address is steffen.strassburger@tu-ilmenau.de.