31st International Conference Nuclear Energy for New Europe
NENE 2022 September 12-15 PORTOROŽ SLOVENIA

Supervised Machine Learning-based Feature Selection in the Frame of BEPU in Severe Accidents

Michela Angelucci[1], Rafael Bocanegra[2], Sandro Paci[1], Luis E. Herranz[2]

[1]University of Pisa
Largo Lucio Lazzarino
56122, Pisa, Italy
michela.angelucci@phd.unipi.it, sandro.paci@unipi.it

[2]CIEMAT
Avenida Complutense, 40
28040, Madrid, Spain
rafael.bocanegra@ciemat.es, luisen.herranz@ciemat.es

ABSTRACT

In the framework of Best Estimate Plus Uncertainty (BEPU) application to Severe Accidents (SA) – related scenarios, the assessment of the uncertainties linked to the simulation results is a mandatory step. However, Uncertainty Quantification (UQ) analyses do not provide an insight into the contribution of individual input parameters - hereon called "features" - to the calculated uncertainty band. To this end, a complementary sensitivity analysis is often needed.

In this regard, the objective of the present work is to establish a data analysis methodology allowing a deeper understanding of the features driving the uncertainty. To do so, an alternative approach for sensitivity analysis, based on the application of supervised Machine Learning (ML), is proposed. The methodology intends to exploit various regression techniques, while facing two major constraints: the high number of features involved, the reason of which can be found in the intrinsic complexity of SA phenomenology, and the small size of the database, due to the computational cost of SA codes.

As a case study, the proposed ML-based approach is applied to a database developed in the MUSA H2020 project: data coming from the simulation of the PHEBUS FPT1 test with the MELCOR code are fed to different ML regression algorithms. Preliminary results show that the use of an appropriate algorithm can actually help in shrinking the number of features, thus improving the interpretability of the model and the identification of the variables that are responsible for most of the uncertainty on the response/s. Moreover, this study suggests the need of corroborating the results with the physical meaning. In other words, expert judgement should play an instrumental role in key steps of sensitivity analyses.

1 INTRODUCTION

In the nuclear field, safety analyses rely on intensive use of simulation tools. Computer codes have been developed and improved during the last decades, and they are now able to reproduce complex systems and scenarios. However, uncertainties on code predictions are still large and need to be quantified. In this regard, Best Estimate Plus Uncertainty (BEPU) methodologies have been extensively applied in the past in conjunction with thermal-hydraulics codes [1], [2], but very few studies have been focused on their application to SAs [3], [4].

In the past three years, the "Management and Uncertainty of Severe Accidents" (MUSA) EURATOM project [5] tried to draw the attention to the problem. In particular, SA analyses codes (i.e., MELCOR [6] and ASTEC [7]) were combined with Uncertainty Quantification (UQ) tools (i.e., DAKOTA [8], RAVEN [9], URANIE [10], and so on) in the attempt to assess the uncertainties linked to the simulation results. While UQ analyses do determine (and quantify) the uncertainty band, they do not provide any insight into the parameters (features) driving it. For this reason, attempts were additionally made to perform a complementary sensitivity analysis, with the aim of identifying the input parameters with the highest influence on the selected output response/s. However, in the majority of cases [11]–[14], sensitivity analysis was limited to Correlation Coefficients (CCs), such as Pearson's and Spearman's [15].

In this framework, the present paper reports an alternative approach for sensitivity analysis, based on data analysis: more specifically, supervised Machine Learning (ML). Feature Selection (FS) techniques are explored and then applied to a database developed in the MUSA project. The database derives from the application of UQ methodologies to the PHEBUS FPT1 test [16], with the MELCOR code being the SA code selected for the simulation of the scenario.

2 METHODS

In statistics and ML, FS techniques are mostly employed to improve accuracy and interpretability when developing a predictive model. However, their core nature makes them suitable for sensitivity analysis. In fact, FS is essentially the process of selecting the features that contribute the most to the target variable/s.

Generally speaking, there are three types of FS approaches:

- *Filter methods* (fig. 1), such as CCs, evaluate the importance of the features as a pre-processing step prior to model training, independently from the ML algorithm selected to construct the model itself. They evaluate whether there is a relationship with the target variable considering each feature separately, thereby not taking into account feature dependencies;

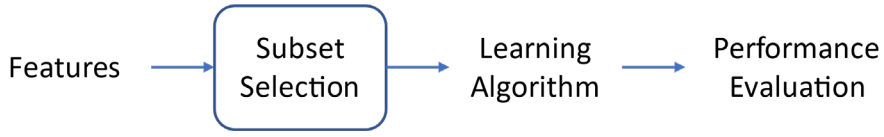

Figure 1: Feature Selection - Filter Methods.

- *Wrapper methods* (fig. 2) rely on an iterative process in which different subsets of features are tested before selection. However, this optimization process comes with a price: wrapper methods are computationally expensive. Considering that, for n features, the number of possible subsets is equal to $O(2^n)$, the search for all the subsets

seems impractical even for moderate numbers of features. In this regard, more efficient search strategies have been implemented (i.e., stepwise regression);

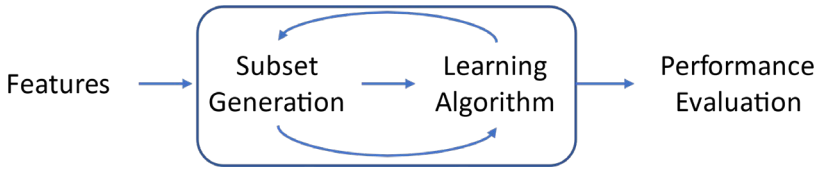

Figure 2: Feature Selection - Wrapper Methods.

- *Embedded methods* (fig.3) include FS and tuning as part of the model construction. They aim at maximizing the performance of the learning algorithm while minimizing the number of features. In addition, they make better use of the available data, while being computationally faster then wrapper methods.

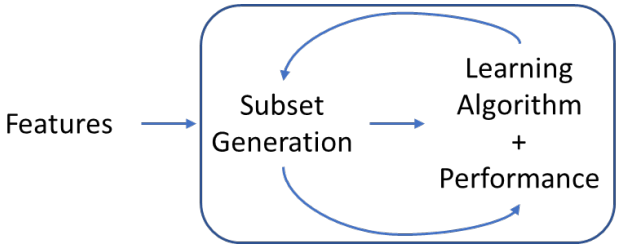

Figure 3: Feature Selection - Embedded Methods.

Considering the initial objective to propose a sensitivity analysis other than CCs, the FS techniques explored in this work belong to wrapper and embedded methods.

As already said, stepwise regression represents an efficient search strategy when dealing with wrapper methods. There are three commonly used approaches to stepwise regression: backward, forward and bidirectional. A brief description of their operating mode is reported in the following:

- Stepwise Backward Elimination: starting with a model with all the input variables, and iteratively removing the least useful variable;
- Stepwise Forward Selection: starting with a model with no input variable, and adding the variable that gives the greatest improvement to the model, one at a time;
- Stepwise Bidirectional Regression: starting with a model with no input variables, and alternating a forward and a backward step.

For what concerns embedded methods, instead, the most attractive technique with respect to FS is the LASSO regularization, whose functioning principle is presented here:

- LASSO Regularization: adding a penalty to variable coefficients, with the result of shrinking some of them to zero, thus eliminating the correspondent variable from the model.

Additional information can be found in [17]–[20].

3 APPLICATION DOMAIN

The application of the previously mentioned FS techniques to a database is subjected to the definition of two parameters: the number of features – "p" – and the size of the database – "n". The number of features is linked to the studied scenario as well as to the addressed phenomena, whilst the size of the database is related to the number of successfully performed calculations. In this regard, two situations can be encountered:

- Low Dimensional Problem: the size of the database outnumbers the number of features. All four techniques cited before can be applied;
- High Dimensional Problem: the number of features is higher than the database size. In this case, techniques involving the creation of a multiple regression model (such as Stepwise Backward Elimination and Stepwise Bidirectional Regression) do not work, as reported in [20].

When applying the previously mentioned regression techniques to SA scenarios, two major constraints are to be faced: the intrinsic complexity of the SA phenomenology and the large computational cost of SA codes. The former affects the number of involved features, which results to be high. The latter, instead, puts a limitation on the size of the database, which, on the other hand, results to be "relatively low". In fact, when considering sensitivity analysis as an additional step after UQ, it is often the case that the number of runs is in the order of 59/93 (as requested by first order Wilks formula, one-sided or two-sided [21]). In these circumstances, the data analysis problem falls very likely within the high dimensional category.

4 RESULTS & DISCUSSION

As already said, the database employed to test the proposed data analysis methodology for sensitivity analysis derives from the application of UQ methodologies to the PHEBUS FPT1 scenario. The database is characterized by a number of features equal to 88 (with features spanning from material properties and core thermal-hydraulics to aerosol behaviour) and a database size equal to 79. The selected targets for this analysis are the Cs released from the fuel bundle (as a % of the initial inventory) and the Cs retention in the circuit (as a % of the Cs released).

Being the database size lower than the features' number, it was possible to apply only two techniques, namely the Stepwise Forward Selection (wrapper class) and the LASSO Regularization (embedded method).

Table 1 reports the features selected when considering the Cs released from the fuel bundle as target variable. As it can be seen, both Forward Selection and LASSO Regularization selected similar features, with the latter singling out a slightly higher number of them. As easily understandable, selected features are mostly related to core behaviour, and in particular to:

- material properties (i.e., density, specific heat, thermal conductivity);
- heat transfer modes (i.e., Laminar Nusselt number for rod bundle);
- core inventory;
- core geometry (i.e., core pitch).

However, it has to be highlighted that some of the selected features are not "physically correlated" with the target variable. In fact, parameters such as the resuspension fraction for the

surfaces of the vault top and for the wet part of the condensers are related to containment behaviour and hardly relatable with the Cs released from the fuel bundle in the core region. This seems to suggest the need for a post-processing screening of the selected feature list in order to ensure consistency with the physical meaning of both features and target variable.

Table 1: Feature Selection for Cs release from fuel bundle.

Feature	Forward Selection	LASSO Regularization	Feature Meaning
C1212_2	☑	☑	Laminar Nusselt n.
rinp1_cor_CD	☑	☑	Core inventory, class Cd
tfscal_139	☑	☑	CS thermal conductivity
tfscal_102	☑	☑	Zr specific heat
tfscal_138	☑	☑	CS specific heat
fractResuspend_CONDENS_WET	☑	☑	Aerosol resuspension fract.
rinp1_cor_TE	☑	☑	Core inventory, class Te
tfscal_104	☑	☑	Zr density
fractResuspend_VAULT_TOP		☑	Aerosol resuspension fract.
pitch		☑	Core pitch
tfscal_166		☑	Spray-coat specific heat
tfscal_119		☑	Inconel-600 thermal cond.

Table 2, instead, reports the features selected when considering the Cs retention in the circuit as target variable. In this case, Forward Selection proved to be greedier with respect to LASSO Regularization, with almost twice the parameters selected. Few observations can be drawn when looking at these results:

- some features (such as “chi”, “deldif”, “fractResuspend_RISER_WALL”, “fractResuspend_LINE-G”, “fractResuspend_SG-U-TUBE”, AISI 304 and 616 specific heat) are directly related to the target variable. They, in fact, have an influence on what happens in the circuit;
- some others (such as the ones related to material properties and core inventory) are indirectly related to the target variable. They influence the core behaviour, that in turn influences the circuit behaviour (and therefore the selected target). This outcome shows an inheritance effect, that propagates as the scenario evolves;
- as for the features selected for the Cs released from fuel bundle, some features are not “physically correlated” with the target variable: it is the case of aerosol resuspension fraction for surfaces in the containment (i.e., VAULT BOTTOM, VAULT TOP, LID TOP, CONTAINMENT WALL, SUMP BOTTOM). Even in this case a post-processing screening seems necessary.

Table 2: Feature Selection for Cs retention in the circuit.

Feature	Forward Selection	LASSO Regularization	Feature Meaning
chi	☑	☑	Dynamic shape factor
fractResuspend_VAULT_BOTTOM	☑	☑	Aerosol resuspension fract.
tfscal_150	☑	☑	AISI-616-L specific heat
fractResuspend_VAULT_TOP	☑	☑	Aerosol resuspension fract.
deldif	☑	☑	Diffusion boundary layer thickness
fractResuspend_RISER_WALL	☑		Aerosol resuspension fract.
fractResuspend_LINE-G	☑	☑	Aerosol resuspension fract.
fractResuspend_LID_TOP	☑	☑	Aerosol resuspension fract.
fractResuspend_SG-U-TUBE	☑	☑	Aerosol resuspension fract.
rinp1_cor_BA	☑	☑	Core inventory, class Ba
tfscal_140	☑		CS density
tfscal_131	☑		Zirconia thermal cond.
tfscal_144	☑		SS density
fractResuspend_CONTAIN_WALL_	☑		Aerosol resuspension fract.
tfscal_148	☑		AISI-304-L density
fractResuspend_SUMP_BOTTOM	☑		Aerosol resuspension fract.

5 CONCLUSIONS

In the framework of BEPU application to SA-related scenarios, this paper proposes an alternative approach for sensitivity analysis, based on supervised ML. Feature Selection techniques have been explored to establish a data analysis methodology allowing a deeper understanding of the features driving the uncertainty. In addition, FS techniques have been tested against a database deriving from the application of UQ methodologies to the PHEBUS FPT1 in the framework of the MUSA project.

Considering the work done, few outcomes can be highlighted:

- The intrinsic complexity of SAs strongly conditions sensitivity analysis. The combination of a high number of involved parameters and of a usually small database size results in some FS techniques being screened out;
- The use of appropriate algorithms can actually help in shrinking the number of features, thus improving the interpretability of the model and the identification of the variables that responsible for most of the uncertainty on the response/s;
- The study suggests the need of corroborating the results with the physical meaning. In other words, expert judgment might play a key role in both feature selection and in the understanding of the results from the sensitivity analysis.

In short, this prospective study highlights the high relevance of carefully select features to make the sensitivity analysis efficient and accurate in determining the governing parameters responsible for the FOMs uncertainties. However, notwithstanding the promising outcomes, further studies have to be carried out in order to confirm the potential shown in this initial work, and to deeper investigate the strengths (and weaknesses) of the FS techniques applied in the exercise.

ACKNOWLEDGMENTS

This project has received funding from the Euratom research and training programme 2014-2018 under grant agreement No 847441.
The paper reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] F. Reventós, "Major Results of the OECD BEMUSE (Best Estimate Methods, Uncertainty and Sensitivity Evaluation) Programme", THICKET 2008, 2008.

[2] J. Baccou *et al.*, "SAPIUM: A Generic Framework for a Practical and Transparent Quantification of Thermal-Hydraulic Code Model Input Uncertainty", Nucl. Sci. Eng., vol. 194, no. 8–9, 2020.

[3] U.S. NRC, "State-of-the-Art Reactor Consequence Analyses (SOARCA) Report (NUREG-1935)", 2012.

[4] M. L. Ang *et al.*, "A risk-based evaluation of the impact of key uncertainties on the prediction of severe accident source terms—STU", Nucl. Eng. Des., vol. 209, no. 1–3, 2001.

[5] L. E. Herranz *et al.*, "The EC MUSA Project on Management and Uncertainty of Severe Accidents: Main Pillars and Status", Energies, vol. 14, no. 15, 2021.

[6] L. L. Humphries, B. A. Beeny, F. Gelbard, D. L. Louie, J. Phillips, and H. Esmaili, "MELCOR Computer Code Manuals Vol. 1: Primer and Users' Guide Version 2.2.9541, SAND2017-0455 O", 2017.

[7] P. Chatelard *et al.*, "Main modelling features of the ASTEC V2.1 major version", Ann. Nucl. Energy, vol. 93, pp. 83–93, 2016.

[8] B. M. Adams *et al.*, "DAKOTA, A Multilevel Parallel Object Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis", 2021.

[9] C. Rabiti *et al.*, "RAVEN User Manual", 2017.

[10] F. Gaudier, "URANIE: The CEA/DEN Uncertainty and Sensitivity platform", Procedia - Soc. Behav. Sci., vol. 2, no. 6, 2010.

[11] N. Elsalamouny and T. Kaliatka, "Uncertainty Quantification of the PHEBUS FPT-1 Test Modelling Results", Energies, vol. 14, no. 21, 2021.

[12] F. Mascari *et al.*, "Phebus Fpt1 Uncertainty Application With the Melcor 2.2 Code", NURETH19, 2022.

[13] R. Bocanegra and L. E. Herranz, “CIEMAT’s outcomes from the PHEBUS-FPT1 uncertainty analysis in the framework of the EU-MUSA project”, ERMSAR22, 2022.

[14] A. Stakhanova, F. Gabrielli, V. Sanchez-Espinoza, E. Pauli, and A. Hoefer, “UNCERTAINTY AND SENSITIVITY ANALYSIS OF THE ASTEC SIMULATIONS RESULTS OF A MBLOCA SCENARIO IN A GENERIC KONVOI PLANT USING THE FSTC TOOL”, ERMSAR22, 2022.

[15] S. Boslaugh and P. A. Watters, Statistics in a nutshell, O’Reilly, 2008.

[16] D. Jacquemain, S. Bourdon, A. de Braemaeker, and A. Barrachin, “PHEBUS FTP1 Final Report”, 2000.

[17] I. Guyon and A. Eliseeff, “An Introduction to Variable and Feature Selection”, J. Mach. Learn. Res., vol. 3, 2003.

[18] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics”, Bioinformatics, vol. 23, no. 19, 2007.

[19] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning Data Mining, Inference, and Prediction, Springer, 2009.

[20] T. Hastie, R. Tibshirani, G. James, and D. Witten, An introduction to statistical learning (2nd ed.), 2021.

[21] N. W. Porter, “Wilks’ formula applied to computational tools: A practical discussion and verification”, Ann. Nucl. Energy, vol. 133, 2019.