EasyChair Preprint

№ 5049

ExTree - Explainable Genetic Feature Coupling Tree using Fuzzy Mapping for Dimensionality Reduction with Application to NACA 0012 Airfoils Self-Noise Data Set

Javier Viaña and Kelly Cohen

February 26, 2021

North American Fuzzy Information Processing Society (NAFIPS 2020)
Genetic Fuzzy Systems Special Session
(Preprint and Prereview Version)

ExTree - Explainable Genetic Feature Coupling Tree using Fuzzy Mapping for Dimensionality Reduction with Application to NACA 0012 Airfoils Self-Noise Data Set

Javier Viaña[1][0000-0002-0563-784X] and Kelly Cohen[2][0000-0002-8655-1465]

[1,2] University of Cincinnati, Cincinnati OH 45219, USA
vianajr@mail.uc.edu
cohenky@ucmail.uc.edu

Abstract. This research presents an AI-based tool (ExTree) that provides high explainability in prediction problems that involve multiple continuous inputs. The algorithm uses an input coupling tree that gradually reduces the dimension of the system. The desired dimension reduction is achieved developing a network of fuzzy inference systems (FIS) wherein in each layer of the network, two inputs get combined to yield a single outcome. These outcomes are then submitted to the same procedure at the following layer until we arrive at a single output, thereby reducing the dimensionality of the problem in every step. Hence, large scale problems with more inputs require more layers. The final outcome is that we obtain a set of FIS nodes across the network, where each FIS may be characterized using an explainable control surface. The structure of the tree is optimized using a genetic algorithm that gets the best hierarchy of fuzzy features to minimize the dispersion of the final outcome. This tool has been benchmarked using NASA's wind tunnel testing database of NACA 0012 Airfoils. The results, demonstrating accurate validation, are of value not only from the perspective of a high performing AI-based algorithm, but also because of the substantial amount of interpretability and traceability that the algorithm offers.

Keywords: ExTree, Input Coupling Tree, Genetic Algorithm, Fuzzy Feature Mapping, Dispersion, Airfoils, Dimensionality Reduction, Explainability.

1 Introduction

The complexity of the data increases every day, requiring more features to characterize properly a meaningful problem. The curse of dimensionality makes it difficult, and sometimes impossible, to visualize the data. Therefore, the algorithms that make use of this information should provide more clarity. In other words, AI has the task of shedding light on the problem, providing traceability for every prediction made.

However, this is currently not the case. Instead of doing so, the tools obscure the process, jeopardizing the understanding from the human perspective. It has even reached a point where the term "black box" is used to define the functions that predict the results. The vast majority of the algorithms that are being used lack explainability.

In the aerospace sector, as in many others, a wise decision (as understood by the machine) is not enough for humans. It is also necessary to justify the reasoning that leads to that prediction ([1]).

Neural Networks can be easily fooled ([2]), thus the outcomes of such an opaque reasoning require an update. Explainable Artificial Intelligence, XAI as identified by DARPA ([3]), seeks to find this clarification ([4],[5]).

The tool presented in this research seeks primarily explainability in the predictions of multidimensional problems. The proposed algorithm couples features in a tree fashion. The input features are grouped in pairs based on the dispersion they demonstrate in the target feature. The result of each coupling is a fuzzy output that ranges from 0 to 1. Therefore, it could be understood as a mapping of two features into a non-physical fuzzy feature that still contains the information of both its inputs. To optimize the structure of the tree, a Genetic Algorithm (GA) is used, which aims to minimize the dispersion of the fuzzy outputs obtained from each pair.

The tree is composed of layers. In each layer, the size of the problem can be reduced by half. In the last layer, all the information is contained in a single fuzzy feature. Therefore, the system is translated into a two-dimensional problem (the fuzzy variable produced in the last layer and the values of the target feature). The information of the mapping is recorded inside the tree in the form of control surfaces or matrices. These control surfaces relate the two input features to the fuzzy output feature produced, thus creating an Explainable Tree ("ExTree") with accessible visual data. The term "ExTree" will be used to refer to the whole explainable structure of couplings.

The software developed offers the possibility of filtering the data in case there is a high presence of noise. For this, the dispersion of each individual is compared in relation to the dispersion shown by the average in the last layer of the tree as this contains all the information of the original features in a single variable.

This tool has been tested and validated with one of the NASA databases. Particularly, the self-noise results of an anechoic wind tunnel testing of several NACA 0012 airfoil blade sections ([6]) provided by the UCI Machine Learning Repository ([7]). The results obtained have demonstrated that the algorithm not only has a high explainability in the decision making, but also accuracy in the prediction of values.

2 NASA Airfoil Self-Noise Data Set

This data set was obtained from a series of aerodynamic and acoustic tests carried out by NASA of two and three-dimensional airfoil blade sections conducted in an anechoic wind tunnel. The experiments were conducted in an anechoic wind tunnel with the observer's position and the span of the airfoil being constant through the whole testing.

The data acquired was multivariable, formed by 1503 instances, with no missing values, and six continuous real features. The five input features are frequency (measured in [Hz]), angle of attack (measured in [°]) , chord length (measured in [m]), wind tunnel's free-stream velocity (measured in [m/s]) and suction side displacement thickness (measured in [m]). The only target feature is noise measured with the scaled sound pressure level ([dB]).

Other authors ([8],[9]) have already studied this data set in depth using Regression tools, Random Forests, Neural Networks, Boosting or Bagging ([10],[11]).

3 Objectives of the research

The first objective is dimensionality reduction, showing the ability of this tool to gather all the information of a multidimensional problem in a visual bidimensional plot.

The second objective is explainability, creating a structure of control surfaces that map the fuzzy features together (two to one) in a tree shape, the ExTree. Furthermore, the tool has an expanding property, which fills any missing combination of features with inferred values.

The final objective is validation, proving the aforementioned objectives using the NASA Airfoil Self-Noise Data Set while still getting high efficiency in the prediction.

4 Methodology

4.1 Dimensionality Reduction, Structure of the Tree

For the dimensionality reduction, the input features $\{X_1, X_2, \dots, X_n\}$ will be grouped in pairs. From every pair, a fuzzy outcome feature will be created $\{A_1, A_2, \dots, A_{n/2}\}$ (if n is even, if n is odd then there will be $(n+1)/2$ fuzzy features). A control surface, CS_κ, will relate both input features, X_i and X_{i+1}, to the fuzzy feature, A_j. This control surface will depend on both the dispersion of the variables (D_κ),

$$D_\kappa = f(X_i, X_{i+1}) \tag{1}$$

and the values of the target feature (Y),

$$Y = f(X_1, X_2, \dots, X_i, X_{i+1}, \dots, X_n) \tag{2}$$

such that

$$CS_\kappa = f(D_\kappa, Y) \tag{3}$$

When two new values of the two features that define D_κ are considered, X'_i and X'_{i+1}, the resulting fuzzy feature, A'_j, can be extracted mapping X'_i and X'_{i+1} in CS_κ,

$$A'_j = CS_\kappa(X'_i, X'_{i+1}) \tag{4}$$

In such a way that the data can be understood as dependent on the new fuzzy features

$$Y = f(A_1, A_2, \dots, A_{n/2}) \tag{5}$$

instead of using the original features (2). Thus, obtaining a reduction of dimensionality.

The process is then repeated considering $\{A_1, A_2, \dots, A_{n/2}\}$ as the new input features; a new set of fuzzy features $\{B_1, B_2, \dots, B_{n/4}\}$ is obtained (again, if n/2 is even).

Eventually, iterations will lead to a single output fuzzy feature, Z_1 . This last feature will thus contain the information of all the inputs $\{X_1, X_2, \dots, X_n\}$ and the dispersion of each coupling. Therefore, the multidimensional data could then be represented in two dimensions as,

$$Y = f(Z_1) \tag{6}$$

For the case of the NASA Airfoil Self-Noise Data Set, the ExTree is as described in **Fig. 1.**

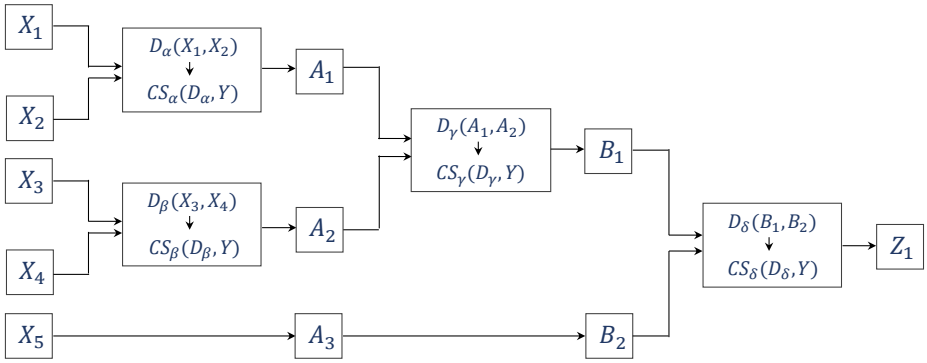

Fig. 1. ExTree structure for 5 input features, NASA Airfoil Self-Noise Data Set.

4.2 Fine Tuning, Genetic Algorithm

A GA will be used to select accordingly the couples. Other authors ([12]) have already used GAs for camber control morphing of airfoils depending on the external conditions.

But in this case, the chromosomes will refer to the order of the features in the ExTree algorithm. Each gene will contain the position of the feature to which it is referring. Therefore, it could be understood as a similar optimization to the travelling salesman problem ([13]). However, the fitness function in this case will reward those chromosomes that lead to a structure of Dispersion matrices with low values, an ExTree with low randomness.

4.3 Expansion, filling unknown entries

The control surface that arises from the combination of two features, CS_κ, is a matrix. However, the individuals of the data set might not cover all the entries of this matrix. Thus, once the known individuals are incorporated in the calculation of CS_κ, an inference process begins. Here, the main purpose is to obtain a matrix with all entries filled. The algorithm for the inference of those unknown combinations uses an interpolation of the surrounding known information.

5 Results

5.1 Genetic Algorithm

For the optimization problem, 60 generations of chromosomes were considered with a population of 20 individuals. The probability of crossover and mutation were set to 0.8 [-] and 0.1 [-] respectively. Additionally, elitism was considered with a ratio of 0.9 [-]. The fittest chromosome of the last population after running the GA is the order for which the minimum dispersion in the results. The result obtained was the following: X_1 suction side displacement thickness, X_2 frequency, X_3 angle of attack, X_4 free-stream velocity and X_5 chord length. Additionally, other parameters were optimized, such as weights for the ponderations used in the inference.

5.2 Control Surfaces

The structure of control surfaces created by the algorithm for the case studied is shown below. For each of the four couplings, a set of six plots are represented. The first two plots of every coupling (Fig. 2 for coupling 1) are scattered representations of all the individuals of the data set. The green data points refer to the real data. The blue, purple and pink data points refer to the amount of real data points that are contained within each cell of the three-dimensional space. The lighter the color and the bigger in size, the more concentrated that cell of the workspace. For the plots all the features have been normalized.

After finding the fuzzy feature A_j, of each couple (X_i, and X_{i+1},), the results are presented in a two-dimensional matrix. For the case of the coupling 1, the first matrix shown in Fig. 3 represents the preliminary version of the control surface. It has some black areas, for which there is no known information. The yellow color represents the highest value of the CS_α, whereas the dark blue has the lowest values. Higher values of CS_α imply higher dispersion on the values of Y. The algorithm is then able to expand the information to those NaN (Not a Number) areas of the matrix for which there is no prior information (second matrix of Fig. 3). This will be very helpful when a certain new value of a feature is encountered for which there is no similar test in the data set. Finally, the matrix is interpolated to have a continuous solution of the CS_α (third matrix of Fig. 3).

After substituting for the different values of the data set in CS_α, the output fuzzy feature A_1 can be obtained (Fig. 4). In this figure, the red line provides the separation of the data according to the dispersion ranges of each set. Thus, the most useful part is contained within the first interval, and those data points that refer to higher randomness are constrained in the second. The algorithm allows the possibility of modifying the number of intervals and their limits depending on the user's commands.

The process is then repeated for all the different layers; the results can be seen below for each of the remaining three couplings of the system.

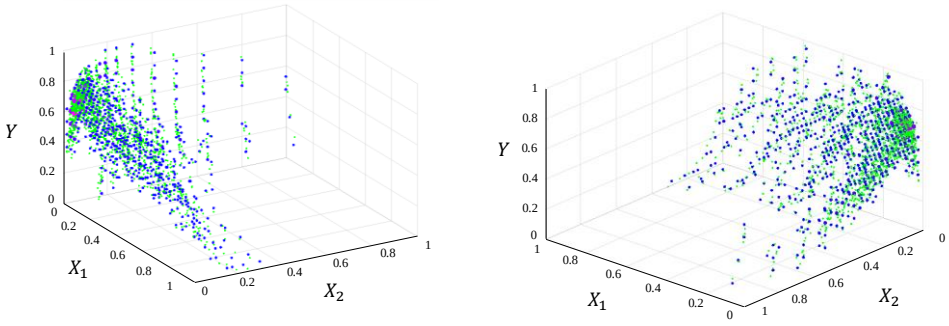

Fig. 2. Original data set plotted using X_1 and X_2 as independent features and Y as dependent feature.

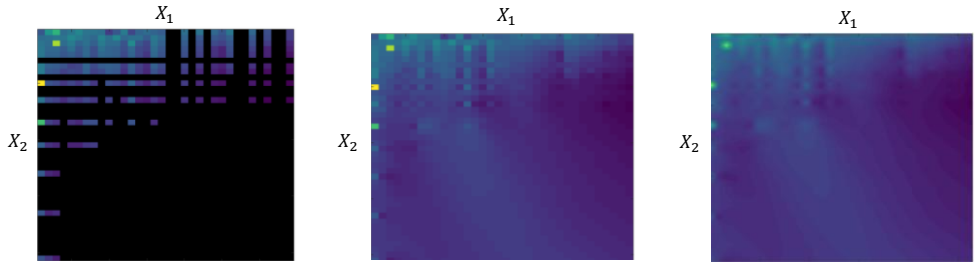

Fig. 3. Inference process to obtain the control Surface CS_α for the original data set.

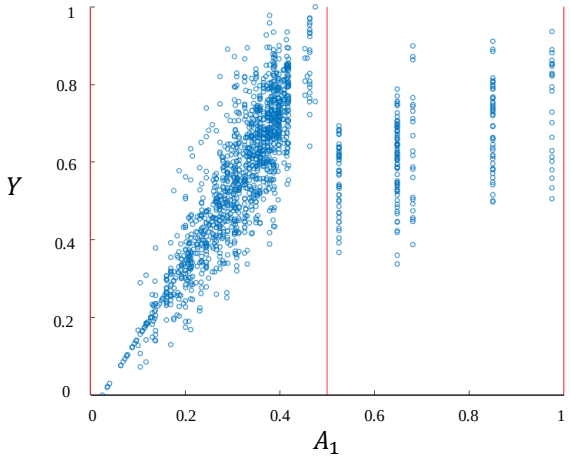

Fig. 4. Plotting of the fuzzy output variable A_1 against Y.

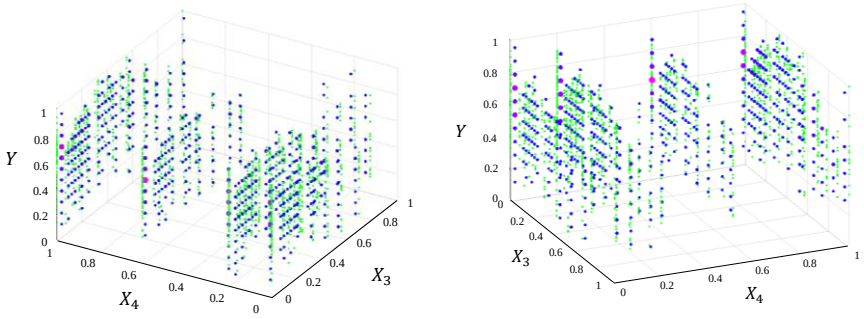

Fig. 5. Original data set plotted using X_3 and X_4 as independent features and Y as dependent feature.

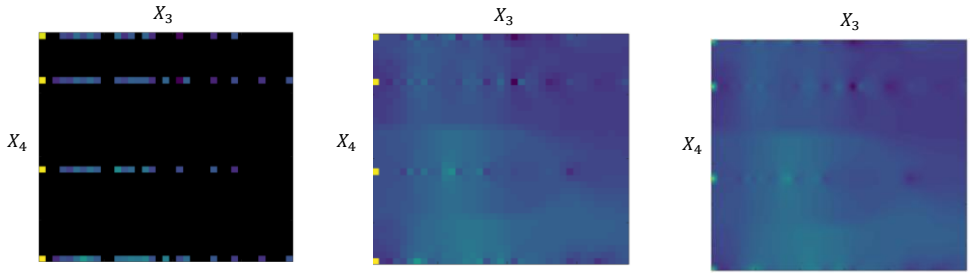

Fig. 6. Inference process to obtain the control Surface CS_β for the original data set.

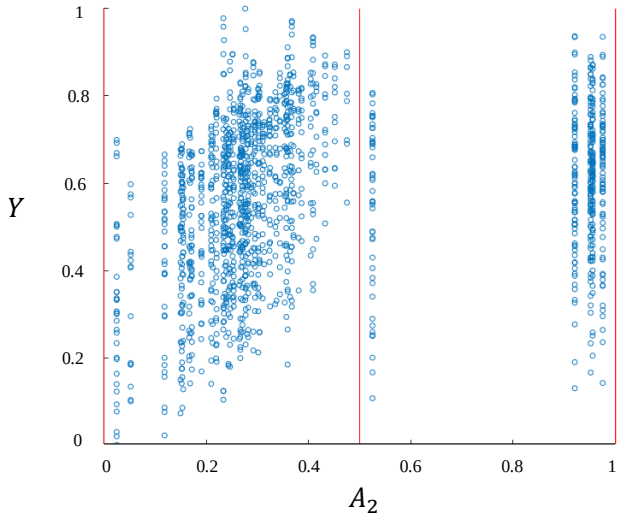

Fig. 7. Plotting of the fuzzy output variable A_2 against Y.

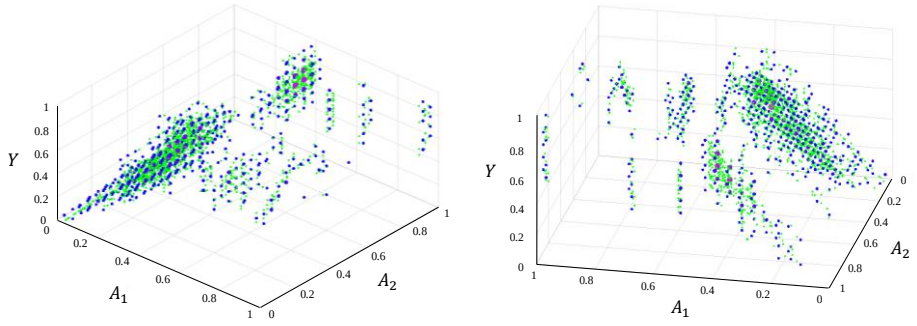

Fig. 8. Original data set plotted using A_1 and A_2 as independent features and Y as dependent feature.

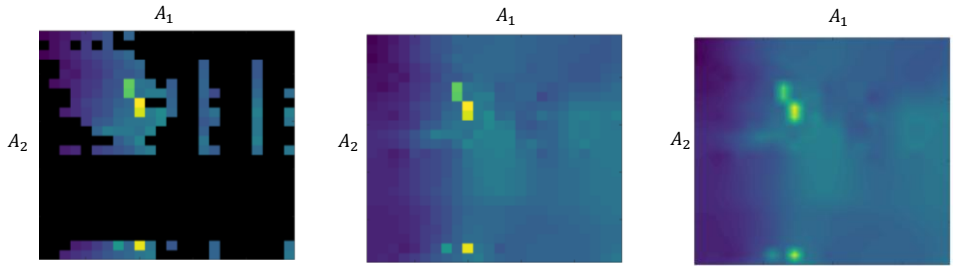

Fig. 9. Inference process to obtain the control Surface CS_Y for the original data set.

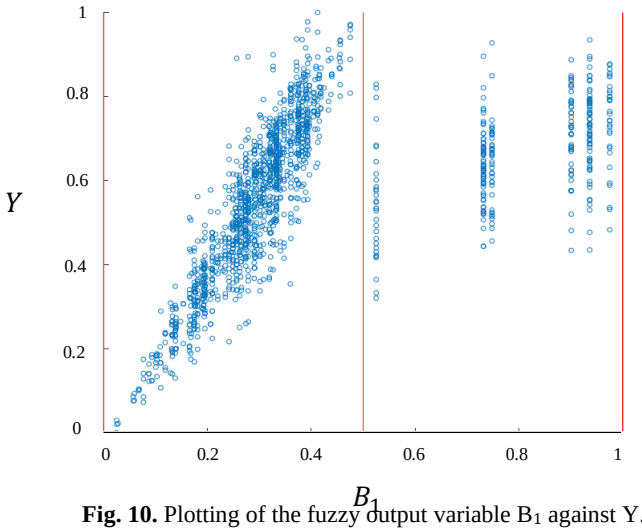

Fig. 10. Plotting of the fuzzy output variable B_1 against Y.

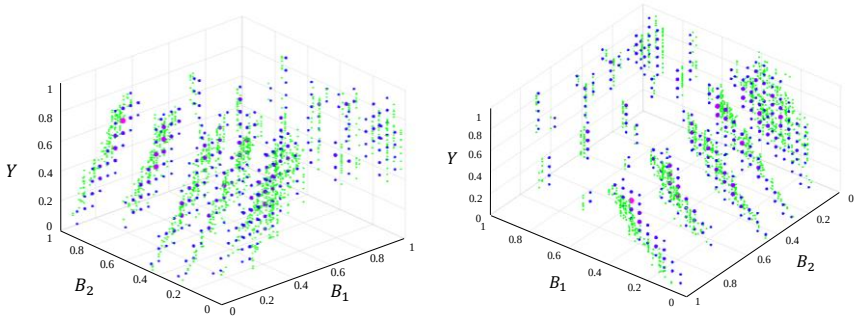

Fig. 11. Original data set plotted using B_1 and B_4 as independent features and Y as dependent feature.

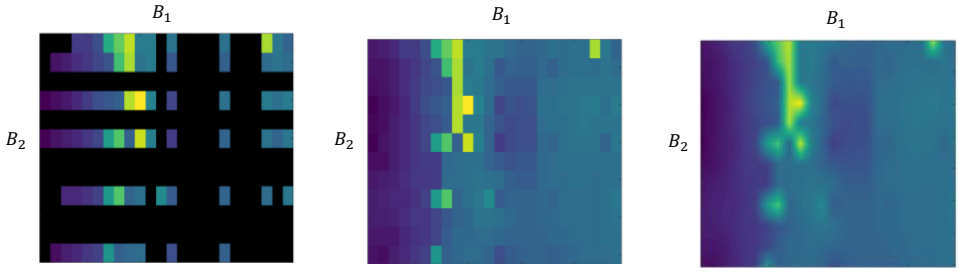

Fig. 12. Inference process to obtain the control Surface CS_δ for the original data set.

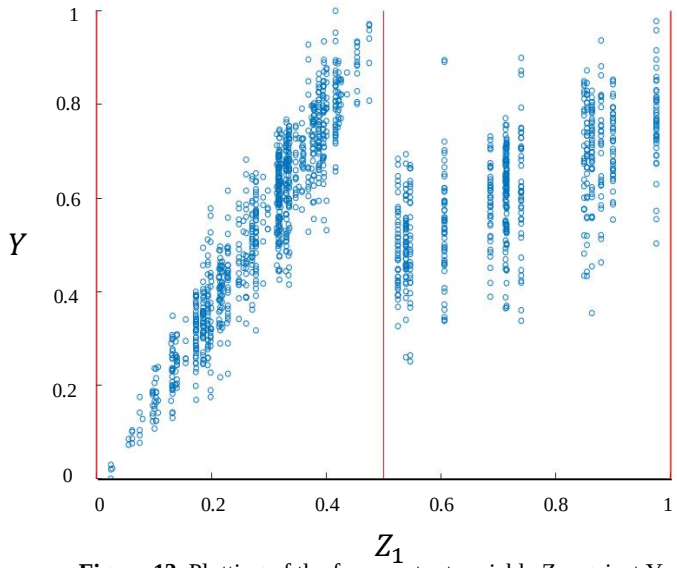

Figure 13. Plotting of the fuzzy output variable Z_1 against Y.

5.3 Predicted Values

The data was divided in 80 [%] training and 20 [%] testing. Fig. 14 shows the difference in the results between predicted and real values. The average RMSE obtained after normalization of the data is 0.0838 (using cross validation). The computational time required for the training to obtain such result was less than 1 minute in a common computer of 8GB CPU.

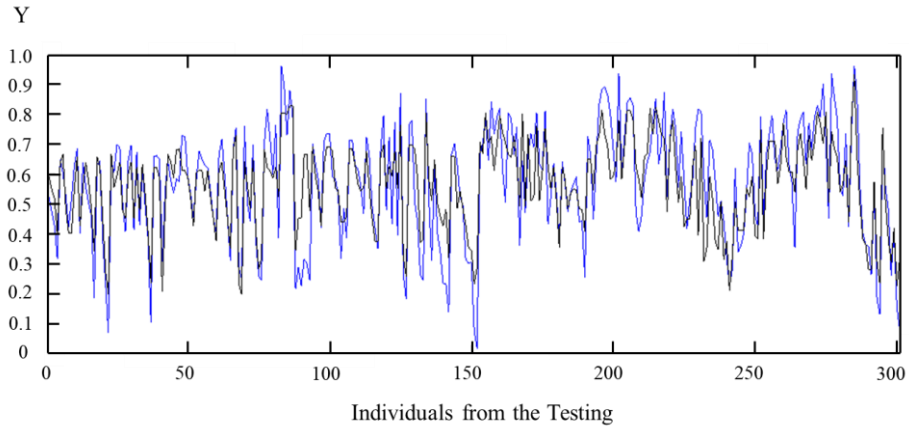

Fig. 14. Predicted in Black vs Real in Blue.

This solution has been obtained for a certain size of the workspace cells of each coupling; however, it can be improved by properly refining the lattice.

It provides a better prognosis than the Regression, and Trees techniques, while its computational cost is better. In addition, it allows extrapolating the information for unknown cases with high uncertainty (as previously mentioned for the process of matrix extension), proving accurate results. But above all, it reduces the dimensionality of the problem increasing significantly the transparency of the algorithm.

6 Conclusion

The tool proposed in this study, ExTree, allows the visualization of the data for any n-dimensional problem. Therefore, contributes significantly in the explainability of data prediction. The algorithm uses multiple inferencing systems in every layer, producing a tree shape structure with intermediate plots that provide transparency and clarity in the decision making. For the optimization problem, a GA correctly obtains the most effective combination of features for the coupling process, minimizing the dispersion in the output fuzzy features of the tree. The preliminary results obtained proved that this methodology can still obtain accurate solutions while avoiding the use of any “black box”.

References

1. Galitsky, B.: Customers' Retention Requires an Explainability Feature in Machine Learning Systems They Use. In: Proceedings of the 2018 AAAI Spring Symposium Series, Palo Alto (2018).
2. Heaven D.: Why deep-learning AIs are so easy to fool. Nature 574, 163-166 (2019).
3. Gunning, D.: Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency, DARPA (2017).
4. Adadi, A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138-52160 (2018).
5. Alonso, J.M.: From Zadeh's Computing with Words Towards eXplainable Artificial Intelligence. In: Fullér R., Giove S., Masulli F. (eds) Fuzzy Logic and Applications 2018, WILF, vol 11291. Springer, Cham (2019).
6. Brooks, T.F., Pope, D.S., Marcolini, A.M.: Airfoil self-noise and prediction. Technical report, NASA RP-1218 (1989).
7. Dua, D., Graff, C.: UCI Machine Learning Repository, http://archive.ics.uci.edu/ml, University of California, School of Information and Computer Science (2019).
8. Lau. K.: A neural networks approach for aerofoil noise prediction. Department of Aeronautics. Imperial College of Science, Technology and Medicine, London (2006).
9. Lopez. R.: Neural Networks for Variational Problems in Engineering. Technical University of Catalonia, Barcelona (2008).
10. Freund, Y., Schapire, R. E.: Experiments with a new boosting algorithm. In: Proceedings of the 13th International Conference on Machine Learning, pp. 148–156. San Francisco (1996).
11. Breiman, L.: Bagging predictors. Machine Learning 24 (2), 123–140 (1996).
12. Lafountain, C.: Matlab-based Development of Intelligent. Systems for Aerospace Applications. University of Cincinnati, College of Engineering and Applied Science (2015).
13. Braun, H.: On solving travelling salesman problems by genetic algorithms. Parallel Problem Solving from Nature, 129–133 (1991).