Comparison of different ANFIS models for the condition monitoring of a rack and pinion contact using methods of explainable artificial intelligence

Tobias Biermann[1,*], Jonathan Millitzer[2], and Karsten Schmidt[1]

[1]Department of Computer Science and Engineering, Frankfurt University of Applied Sciences, Nibelungenplatz 1, 60318 Frankfurt am Main, Germany

[2]Fraunhofer Institute for Structural Durability and System Reliability LBF, Bartningstraße 47, 64289 Darmstadt, Germany

*Corresponding author: tobiermann@gmx.com

Received Month X, XXXX | Accepted Month X, XXXX | Posted Online Month X, XXXX

Abstract: This paper investigates the use of XAI and TAI methods for condition monitoring on a laser cutting machine. The focus is on the analysis of the rack and pinion contact with wear being predicted by four differently derived ANFIS models. Using both model-agnostic and model-specific parameters integrated in a weighted evaluation framework, the models are evaluated with respect to the effectiveness of explanations. This framework is based on the observation of the outputs of the individual layers of ANFIS, also focusing on aspects of two multi-valued logics, namely fuzzy logic and support logic. The results show that the introduced weighted evaluation framework makes it possible to quantify the explainability of the individual models in terms of XAI and TAI. Finally, a preselection of a model for predicting the wear of the rack and pinion contact can be made.

Keywords: ANFIS; XAI; condition monitoring; rack and pinion contact

1 Introduction

The use of AI methods is considered an innovative opportunity in many applications [1]. In Germany alone, and primarily in the manufacturing industry, around 220 billion euros were already generated in 2019 using AI applications [2]. Turnover is forecast to more than double to 488 billion euros by 2025, which would correspond to a 13% share of the gross domestic product. Despite this enormous potential, a decisive hurdle in the use of AI methods is their so-called black box behavior. This leads to a loss of acceptance and trust, as the inner workings and decisions of AI models are not intuitively comprehensible to non-experts [3, 4]. To counteract this, various approaches to Explainable Artificial Intelligence (XAI) are currently being pursued in research. The focus lies on the development of transparent and interpretable AI systems [5]. As part of a German Federal Ministry of Education and Research (BMBF) research project, the condition monitoring of the rack and pinion contact of a laser cutting machine is being investigated. Condition monitoring of rack and pinion systems is technically demanding, as these open-drive mechanisms are directly exposed to wear-inducing influences. Key damage mechanisms include material loss caused by relative motion between the gear components and abrasive particles in the lubricant, as well as tooth breakage resulting from fatigue failure or excessive mechanical loads. Feed axis failures are a major source of machine downtime and maintenance, service, and unavailability can amount to up to 18% of overall life cycle costs. Therefore, effective condition monitoring becomes both economically and technically essential [6].

The approach presented in this paper is footed on knowledge-based AI and includes the use of Adaptive-Network-Based Fuzzy Inference System(s)

(ANFIS). Thus, the interpretability of condition monitoring in the form of human understandable explanations is ensured and improves the machine in terms of availability and reliability. In particular, this will result in shorter downtimes as well as damage to the laser cutting machine. Due to the high importance of the system in terms of functional safety and economic efficiency, the Trustworthy Artificial Intelligence (TAI) approach, as set out in the German standardization roadmap for artificial intelligence, must be guaranteed during operation [3, 7].

The focus of this paper is the evaluation of ANFIS in the context of XAI and TAI in relation to four different models (M1-M4) for the condition monitoring of the rack and pinion contact. To this end, a new methodology in the form of a weighted evaluation framework is presented. This approach follows current research to derive and apply evaluation metrics for XAI methods as described in [8] and [9]. AI developers and users will be able to evaluate and quantify the effectiveness of explanations of a model both model-agnostically and model-specifically. The focus is on factors of complexity-based and semantic-based interpretability as well as local and global explanations of the models.

The paper is structured as follows. Section two presents the theoretical background. Section three contains a description of the laser cutting machine and the experimental setup. The data preprocessing and the derivation of the initial state (M1) are presented in section 4. The computational complexity of ANFIS and the data set are also discussed. Section 5 introduces the new methodology. The focus is on the derivation and training of models M2-M4 as well as the presentation of the weighted evaluation framework in the context of XAI and TAI. The sixth section contains the results and analyses. Section 7 summarizes the paper and provides an outlook on further research opportunities.

2 Theoretical Background

2.1 Support Logic

Support logic is based on the idea of modeling uncertainty in an expert system. As in fuzzy logic, multi-valued logic is also used here, as most of the information in the world cannot be adequately represented with two-valued logic. For this reason, human communication is often based on probability theory, which includes the mathematical property [10, 11]:

$$P(A) = 1 - P(\neg A) \tag{1}$$

These restrictions are relaxed in support logic by applying the following [10, 11]:

$$S(A) \leq 1 - S(\neg A) \tag{2}$$

$S(A)$ describes the support for an assertion A. An assertion corresponds to a logical statement or claim. Using this approach, it is possible to specify an assertion in a system, with both support for (SL - lower support) and support against an assertion (SU - upper support). This is translated to truth values, in a value range of [0,1]. The unsureness U stands for the ignorance of information on the assertion and can also assume a value between 0 and 1. SL is synonymous with the amount of support that can guarantee absolute certainty of the truth of an assertion. Therefore, the latter stands for the minimum confidence in an assertion. In comparison, SU defines the amount of support that possibly corresponds to the truth of the assertion. The following equations apply [11]:

$$SU(A) \geq SL(A) \tag{3}$$

$$SU(A) = 1 - SL(\neg A) \tag{4}$$

$$U(A) = SU(A) - SL(A) \tag{5}$$

For a visual and explainable representation, an assertion is required first. This could be, for example:

$$A: "Component\ B\ is\ worn\ out"$$

The transcription to support logic could be as follows.

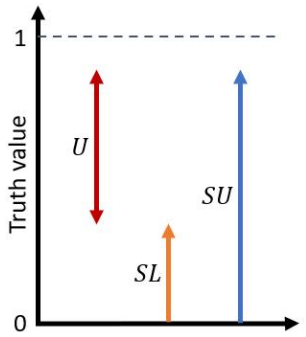

Fig. 1. Possible results of the support logic

Due to the low SL, it can be argued with a low truth value that the assertion is “definitely true”. As a

result of the high SU, however, it can be argued with an even lower truth value that the assertion is "definitely false" (1- SU). The reason for this lies in the high unsureness U of the system.

2.2 Structure and function of an ANFIS

An ANFIS is based on a hybrid neuro-fuzzy model and was developed by Jang in 1993. Here, a fuzzy system is embedded in a feedforward neural network with five layers L1-L5 [12]. In terms of training, ANFIS falls into the category of supervised learning methods [13]. Figure 2 shows an example of an ANFIS (type 3 architecture) as investigated in this paper with two input vectors $\{x_1, x_2\}$ and fuzzy sets $\{\widetilde{A}_j, \widetilde{B}_j\}$ each as well as a target vector y.

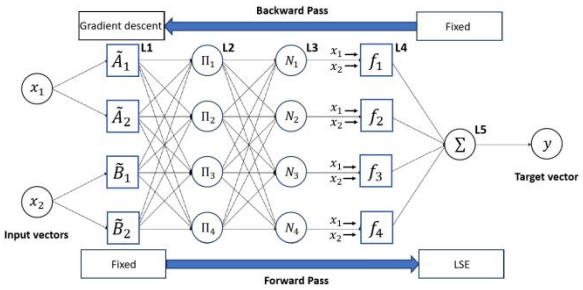

Fig. 2. Structure of the ANFIS (type 3 architecture), adapted from [14]

The functionality of the individual layers is described in more detail below. The set {1,2,3,4} applies to the index i and the set {1,2} to the index j. In the first layer L1, the node functions O_i^1 are used to fuzzify the input vectors. The crisp value x of the input vector is assigned a function value $\mu_{\widetilde{A}}(x)$ of the fuzzy set. Equation 6 [12, 14] applies here:

$$O_i^1 = \mu_{\widetilde{A}_j}(x_1), \mu_{\widetilde{B}_j}(x_2) \qquad (6)$$

For example, the notation in [12] results in equation 7 for a gaussian membership function:

$$\mu_{\widetilde{A}_j}(x) = e^{-\left(\frac{x-c_j}{a_j}\right)^2} \qquad (7)$$

The so-called premise parameters $\{ a_j , c_j \}$ are optimized during the ANFIS learning process and are used to calculate the premise of the Takagi-Sugeno rules. In layer L2, each node models the conjunctive operation (AND operation) of two degrees of membership, which describes the non-normalized firing strength of a rule. For this purpose, the function values of the membership functions are multiplied with each other [12, 14]:

$$O_i^2 = \omega_i = \mu_{\widetilde{A}_j}(x_1) * \mu_{\widetilde{B}_j}(x_2) \qquad (8)$$

In the third layer, the ratio of the i[th] firing strength of a rule to the sum of all firing strengths is calculated. The outputs of the nodes are also referred to as normalized firing strengths [12]:

$$O_i^3 = \overline{\omega}_i = \frac{\omega_i}{\sum_{i=1}^{4} \omega_i} \qquad (9)$$

In ANFIS, the function f of the conclusion of a Takagi-Sugeno rule is based on a linear combination of the fuzzy variables [13]:

$$f_i = p_i x_1 + q_i x_2 + r_i \qquad (10)$$

In the penultimate layer, equation 10 is inserted into the nodes [12]:

$$O_i^4 = \overline{\omega}_i f_i = \overline{\omega}_i (p_i x_1 + q_i x_2 + r_i) \qquad (11)$$

At this point, it should be noted that an additional linear combination is required in the respective rule for each additional target vector [15]. Analogous to the premise parameters, the conclusion parameters $\{ p_i , q_i , r_i \}$ are also adapted during the ANFIS learning process using an optimization algorithm. Using equation 6 and equation 11, a possible rule of the system can be defined as follows [12]:

$$IF\ x_1\ is\ \widetilde{A}_1\ \wedge\ x_2\ is\ \widetilde{B}_1\ THEN\ y\ is\ p_1 x_1 + q_1 x_2 + r_1 \qquad (12)$$

In the fifth and final layer, the target vector of ANFIS is determined using a single node. The latter results from the summation of all signals from equation 11 and represents a sharp value instead of a fuzzy value [12, 14]. Equation 13 applies here [12]:

$$O_k^5 = \sum_{i=1}^{4} \overline{\omega}_i f_i \ , k = 1 \qquad (13)$$

Figure 3 shows an example of the ANFIS process or fuzzy reasoning (fuzzy inference) for the first and fourth rule.

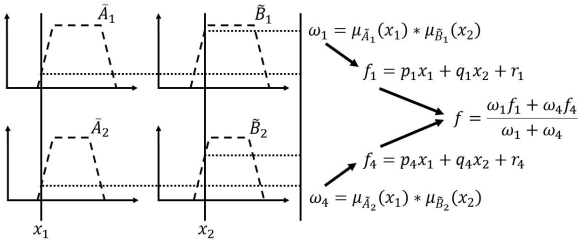

Fig. 3. Fuzzy reasoning of the ANFIS, adapted from [12]

2.3 Linguistic Modifiers

Linguistic modifiers can be used to change the semantics of a linguistic term from a fuzzy set $\widetilde{A}$. An additional operator α serves as the basis, which is integrated into the parameterization of a membership function $\mu_{\widetilde{A}}$ in order to increase flexibility [16]. The powered linguistic modifiers investigated in this paper are mathematically defined as follows [17]:

$$\mu^{\alpha}_{\widetilde{A}}(\cdot) = \begin{cases} \leq \mu_{\widetilde{A}}(\cdot), & \alpha > 1 \\ \geq \mu_{\widetilde{A}}(\cdot), & \alpha < 1 \end{cases} \tag{14}$$

If the condition $\alpha > 1$ applies, $\mu^{\alpha}_{\widetilde{A}}$ is referred to as a concentration or weak modifier. The inverse case is referred to as dilation or a strong modifier [10, 17]. Table 1 lists examples of linguistic modifiers in relation to the value of α.

Table 1. Overview of linguistic modifiers, adapted from [17]

Linguistic modifiers	α-value	Dilation/Concentration
Slightly	0,25	Dilation
More or less	0,5	Dilation
Minus	0,75	Dilation
-	1	-
More	1,5	Concentration
Much more	1,75	Concentration
Very	2	Concentration
Absolutely	4	Concentration

2.4 Grid Partitioning Method

The ANFIS fuzzy inference process shown in Figure 3 is based on a divide-and-conquer method, which divides the n-dimensional input space into specific areas. This corresponds to the premise part of a fuzzy rule. The result of the fuzzy inference is integrated in the specific areas using the conclusion part of the fuzzy rule in the form of the function f described in equation 10 [14, 18]. Each area in the input space therefore represents exactly one fuzzy rule. The structure-oriented grid partitioning method is based on hyper cuboids that are segmented in a grid structure [15]. Equation 15 can be used to calculate the number of specific grids or fuzzy rules [19].

$$n_{rules} = {n_{splits}}^{n_{dimensions}} \tag{15}$$

Where $n_{dimensions}$ stands for the number of dimensions, which is equivalent to the number of input vectors in the ANFIS. The parameter n_{splits} describes the number of splits in the input space [19]. The number of fuzzy sets, membership functions or linguistic terms per fuzzy variable can also be used for this purpose. Figure 4 shows the grid partitioning method for two input vectors with three triangular membership functions each. According to equation 15, nine fuzzy rules or hyper cuboids result from the grid structure.

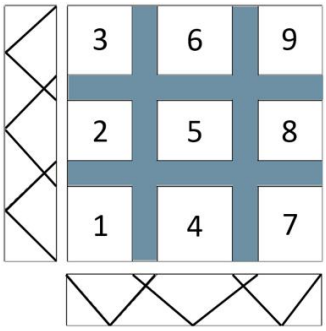

Fig. 4. Grid partitioning method for dividing the input space, adapted from [14]

2.5 XAI metrics and interpretability of fuzzy systems

The problem that neural networks and systems based on artificial intelligence often exhibit black-box behavior demonstrates the need for the application of XAI. According to the Defense Advanced Research Projects Agency (DARPA) of the United States XAI is intended to enable interpretability and increase confidence in AI models in order to achieve effective handling of such models [20]. The causality, transferability and fairness of the systems must also be guaranteed [21]. Despite the lack of a universally valid definition of XAI, the authors of [22] strive to improve transparency, which should result from explanations of a specific decision (local explainability) or the functionalities of the entire model (global explainability). Compared to global explanations, local explanations lead to a higher trustworthiness in AI-based systems due to their

individual and specific explanations [23]. Globally interpretable models are also referred to as ante-hoc systems and can be regarded as a kind of glass-box. Post-hoc systems, on the other hand, focus on local explanations [24]. Here, the interpretation is only made after a model has been trained [25].

Another measure of XAI is the effectiveness of an explanation. According to Gunning and Aha [20], psychological human-in-the-loop experiments are required to measure effectiveness, as automatic determination is not yet possible. The following potential points of reference are suggested, but further research is required at this point. **User satisfaction** is indicated in the form of an assessment of the clarity and usefulness of an explanation. In the **mental model** section, the extent to which local and global explanations contribute to understanding and the strengths and weaknesses of the system are assessed. In addition, the focus is on predictions of the behavior of the model and the intervention of the user. In terms of **performance**, the aim is to answer whether an explanation contributes to an improved decision for a subsequent action by a user and whether the task is fulfilled. The **trust assessment** comprises the current and future confidence in the model in the context of the TAI. **Correctability** tests how easy it is to recognize and correct errors in the model [20].

A suitable matrix for evaluating the interpretability of a fuzzy system is introduced in Table 2. Aspects of complexity-based and semantic-based interpretability are set in relation to the division of the input space and the rule base and presented using a quadrant [26].

Table 2. Quadrant for evaluating the interpretability of a fuzzy system, adapted from [26]

	Segmentation of input space	Rule base
	Q1	**Q2**
Complexity-based interpretability	Number of membership functions Number of input vectors	Number of rules Number of conditions
	Q3	**Q4**
Semantic-based interpretability	Completeness Normalization Distinguishability Complementarity	Consistency of rules Rules fired at the same time

All evaluation metrics that are not self-explanatory by definition are explained in more detail below. The number of conditions depends on the number of conjunctions (AND operations) in the premise of a fuzzy rule. A limit value of five to nine different conditions is set, as a person is not able to process more conceptual units. This parameter correlates strongly with the number of input vectors. Completeness, as listed in the third quadrant, is given if each value of a set can be linguistically represented with at least one fuzzy set of the fuzzy variable. Normalization is achieved with normal membership functions. Distinguishability refers to the fact that each membership function should have a clear linguistic meaning and thus no identical fuzzy sets occur within a fuzzy variable [26]. In order for each membership function of the fuzzy sets to be meaningful for the user, they should also have a convex form [15]. The complementarity parameter of a fuzzy variable evaluates whether the sum of the memberships of each value of a set is equal to 1. If this is true, this will yield the highest semantic interpretability [26]. The consistency of rules contained in the fourth quadrant comprises several aspects. On the one hand, rules with identical premises but different conclusions must be excluded, as these correspond to a complete contradiction [15, 26]. On the other hand, no rule should occur twice in order to avoid redundancies. Partial contradictions caused by overlapping rules or membership functions are permitted and are usually automatically generated by the complementarity of the fuzzy variable aimed for in the third quadrant [15]. The evaluation matrix is completed with the rules fired at the same time. The aim is to reduce the latter for values from a set of a fuzzy variable. The more selective the rules are, the clearer the linguistic interpretations of the individual fuzzy sets remain. Therefore, this parameter strongly dependent on the selected method for partitioning the input space (see section 2.4) [26].

3 Experimental Setup

In this paper, a 2D laser cutting machine in gantry design, serves as the object of investigation for monitoring the condition of the rack and pinion contact. Its cross-section is shown as an example in Figure 5.

As can be seen from the coordinate system, the Tool Center Point (TCP) has three degrees of freedom in the form of three traversing axes. The y and z movements are based on a linear direct drive, while the movement in the x direction is carried out using a rack and pinion drive with two feed axes (X1 and X2 axes). The pinion attached to the output shaft of the single-stage planetary gear engages with the rack and

is rotated by a servo motor. One revolution corresponds to a fixed number of meshes. In case of a worn pinion, increased vibrations or significant changes in the motor current can occur with each meshing. To identify defective pinions, the acceleration of the two feed axes is measured using MEMS sensors while the motor current is recorded by the controller.

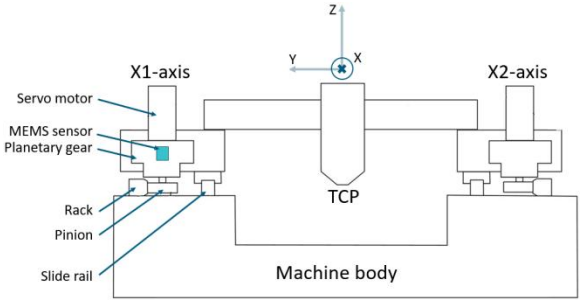

Fig. 5. Machine cross-section of the laser cutting machine

4 Data preprocessing

4.1 Creation of target vectors

The recorded input data as described in section 3 (acceleration of the feed axes and motor current) are based on measurements carried out in the normal state of the machine, i.e. without wear of the pinion and at seven constant traversing speeds per feed axis. In order to be able to draw conclusions about the wear of the pinion, manipulated data is generated synthetically from the original data with the aid of machine-specific knowledge [27].

For this purpose, the data measured in the time domain were converted to a frequency spectrum using a fast Fourier transformation. Due to the fact that an increased amplitude can be observed at the gear mesh frequency, various scaling factors were multiplied to this frequency using a triangular function. Table 3 lists the pinion damage levels generated by the manipulated data as well as their corresponding abbreviations and scaling factors (SF). A linguistic term is assigned to each degree of damage [27].

Table 3. Overview of the pinion damage levels, adapted from [27]

Degree of damage	Abbr.	SF	Linguistic term
No degree of damage (Original data)	Ndod	1	Normal
Slight degree of damage	Sdod	1.5	Slightly increased
Medium degree of damage	Mdod	2	Medium increased
Heavy degree of damage	Hdod	4	Heavily increased

In the next step, bandpass filtering was performed as a function of the traversing speed. The proportional relationship between the gear mesh frequency and the traversing speed of the feed axes meant that both the bandwidth and the center frequency of the bandpass could be adapted to the respective gear mesh frequency. This eliminated irrelevant frequency components from the signals. Afterwards, using an inverse fast Fourier transformation, the signals were transformed back to the time domain, whereupon they were filtered using a moving mean square [27].

In a final step, the two inputs motor current and acceleration of the feed axes are normalized. The latter are integrated into ANFIS as unitless input vectors x_1 (norm. current amplitude) and x_2 (norm. acceleration amplitude) [27]. Figure 6 illustrates the effects of data preprocessing using a scatter plot. The damage levels of the pinion are shown in relation to the input vectors.

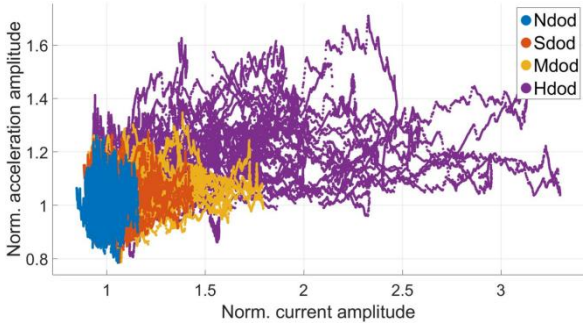

Fig. 6. Scatter plot of the pinion's degree of damage as a function of the input vectors

In contrast to the input vectors, the target vectors SL and SU (see section 2.1) cannot be derived directly from measurements with the real test setup. Instead, the target vectors are based on expert knowledge of the wear of the pinion. Theoretically, as evident from Figure 2, only a forward pass of the data by the ANFIS without least square estimate (LSE) is required to create the target vectors. There is no backward pass and, therefore, no training of the

ANFIS. This approach requires initial membership functions (layer 1) and the conclusion parameters for each fuzzy rule (layer 4) as described in section 2.2. To create the membership functions, the samples of the individual damage levels listed in Figure 6 are first converted into frequency distributions for each input vector. The fuzzy sets are then assigned to each degree of damage on the basis of expert knowledge. The linguistic terms listed in Table 3 are assigned to Gaussian membership functions [27]. The results for both input vectors are shown in Figure 7 and Figure 8.

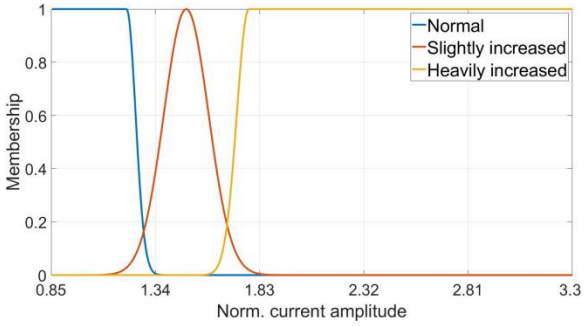

Fig. 7. Initial membership functions of the first input vector, adapted from [27]

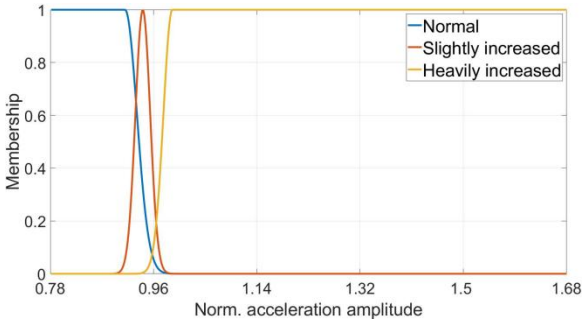

Fig. 8. Initial membership functions of the second input vector, adapted from [27]

As can be seen in Figure 7 and Figure 8, the linguistic term “Medium increased” or the medium degree of damage is neglected at this point. This can be explained by the low differentiability compared to the slight degree of damage [27]. It can also be seen that the other membership functions are initialized to the left or right in a comparatively large value range with the strongest membership. This fulfills the completeness and complementarity of semantic-based interpretability described in Table 2.

In addition to the membership functions, the conclusion parameters were also determined on the basis of expert knowledge. Experts were asked to describe different degrees of damage to the pinion using sliders. These were coupled with the conclusions of Mamdani fuzzy rules. The p- and q-parameters in the Takagi-Sugeno fuzzy rules presented are implemented with 0. Due to the use of the grid partitioning method explained in section 2.4, nine rules result from two input vectors and three membership functions per input vector according to equation 15. For each rule (9 questions), the r-parameters for SL and SU were created and integrated into the fourth layer of ANFIS. At this point and for all further evaluations, assertion A applies [27]:

$$A\text{: "}The\ pinion\ is\ worn\ out\text{"}$$

This will be taken up again in section 6 in order to evaluate the effectiveness of explanations. For each sample, conclusions about the wear of the pinion can already be drawn at this stage of development because, as shown in equation 5, the unsureness U results from a subtraction of the target vectors. This reflects the initial state (M1).

4.2 Computational complexity and data set

The computational complexity of an ANFIS can also be analyzed using one sample. In Table 4, this is expressed by the number of Multiply-Accumulate Operations (MACs) as a function of the number of integrated membership functions (MF). The profilers from Meta Research and Microsoft Deepspeed are used to determine the number of MACs, which show consistent results.

Table 4. Computational complexity of ANFIS depending on the number of MF

Number of MF	Number of MACs
2	32
3	72
4	128
5	200
6	288
7	392
8	512

The formula for computational complexity $CC(n_{MF})$ of ANFIS shown in equation 16 can be derived from the numerical sequence in Table 4.

$$CC(n_{MF}) = 8{n_{MF}}^2 \quad (16)$$

Depending on the architecture and image size, a convolutional neural network requires approx. 435 to 3.09×10^5 MACs to process a pixel [28]. This shows the computing efficiency of ANFIS. It is important to note that the results of the analyses should not be regarded as absolute values. The profilers only serve as an approximate estimate of the computational complexity, as they, for example, neglect the computational costs for an exponential function within a membership function.

The division of the data for training the ANFIS is based on a ratio of 60:20:20 in the sequence: training data set, validation data set and test data set. This is intended to achieve a high degree of generalization of the ANFIS models. The training data set is used to determine the premise and conclusion parameters. The validation dataset is used to tune the hyperparameters and the test dataset as a final evaluation of the performance of the ANFIS. Table 5 provides an overview of the data set.

Table 5. Overview of the data set

	Training set	Validation set	Test set
Total samples	208.050	69.350	69.350
Samples per input vector	104.025	34.675	34.675

5 Proposed Methodology

5.1 Derivation of further models

To derive further models M2 and M4, the target vectors (tv_{init}) from M1 are first manipulated using two different functions. For M2, a continuous S-shaped function is applied to slightly stretch the value range. For M4, an S-shaped function with an embedded discontinuity is employed to evaluate the behavior of ANFIS under a strong manipulation of the value range. In both cases, the value range of the functions is constrained to [0,1], in accordance with the value range of SL and SU (see section 2.1). The S-function is based on equation 17:

$$tv_{man} = \frac{1}{1 + \left(\frac{tv_{init}}{1 - tv_{init}}\right)^{-\beta}}, \beta > 0 \quad (17)$$

The slope of the S-curve in the function sections can be modeled using the factor β. This was set to 1.5 in order to generate a small change in the value range of the target vectors. In contrast, values of 0.1 or 3, for example, lead to a large change in the value range, therefore 1.5 is selected as a robust compromise. The S-function with integrated discontinuity can be expressed using equation 18.

$$tv_{man} = \begin{cases} \frac{0{,}5}{1 + \left(\frac{tv_{init}}{1 - tv_{init}}\right)^{-1.5}}, & tv_{init} \leq 0{,}5 \\ 0{,}5 + \frac{0{,}5}{1 + \left(\frac{tv_{init}}{1 - tv_{init}}\right)^{-1.5}}, & tv_{init} > 0{,}5 \end{cases} \quad (18)$$

Figure 9 shows the function curves of all models including the initial state.

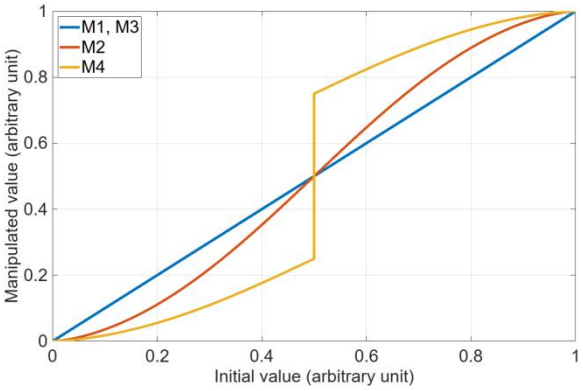

Fig. 9. Various function curves for manipulating the target vectors/deriving the models

With regard to the value range of the distributions of the target vectors, a stretching is achieved by M2. With M4, the discontinuity causes the elimination of all samples between 0.25 and 0.75. Figure 10 illustrates this using the distribution of SL in the training data set. Figure 11 shows similar tendencies in the distribution of SU in the training data set.

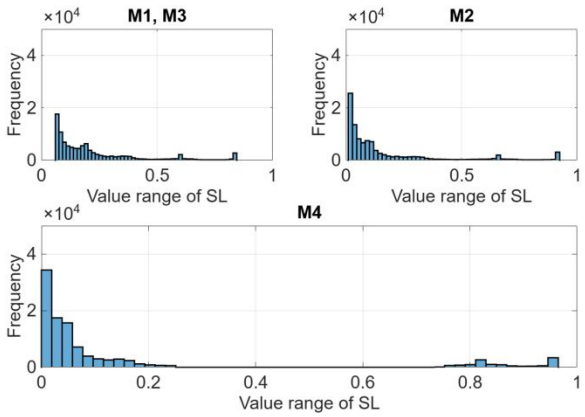

Fig. 10. Comparison of the distributions of SL in the training data set for models M1-M4

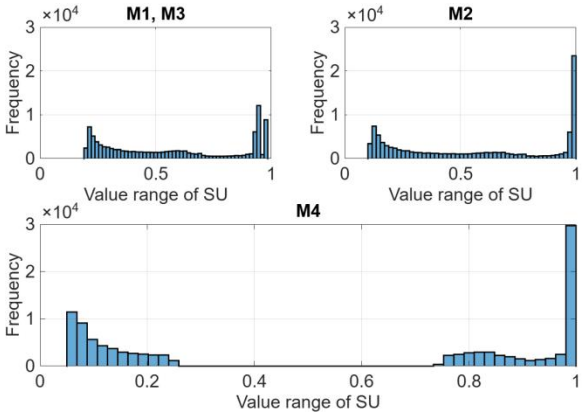

Fig. 11. Comparison of the distributions of SU in the training data set for models M1-M4

As shown in Figure 9, M2 is based on a manipulation of the target vectors, but the same membership functions are used for training this ANFIS as for the initial state M1. M3, on the other hand, does not manipulate the target vectors, as the focus here is solely on the difference between the original data without wear of the pinion and the heavy degree of damage. Accordingly, ANFIS is trained with two membership functions per input vector. For M4, in addition to the manipulation of the target vectors and the integration of the membership function with the linguistic term "Medium increased", an additional synthetic degree of damage with the linguistic term "Moderately increased" is created as a fifth membership function. Due to this fact, the powered linguistic modifiers explained in section 2.3 are used for each fuzzy set. The idea is that the training creates new linguistic terms that shift the boundaries of the damage levels and thus require an interpretation in terms of XAI. An overview of the different models including the initial state is given in Table 6.

5.2 Training of the models

The training of the ANFIS is based on a model in the open-source program library PyTorch. The data of the trained ANFIS are transferred to the MATLAB platform and processed there. A hybrid learning method consisting of an LSE to determine the conclusion parameters and an Adam optimization algorithm to determine the premise parameters is used in this paper. The boundary conditions (BC) for training of M2-M4 are listed in Table 7. A grid search of boundary conditions four and five is used to evaluate the optimal hyperparameters and thus provide the trained ANFIS for evaluating the effectiveness of explanations.

Table 6. Overview of the model features

M	Man. of tv_{init}	Number of MF	Linguistic terms
1	Initial state, no man.	3	Normal, Slightly increased, Heavily increased
2	Man. with equation 17 and $\beta = 1.5$	3	Normal, Slightly increased, Heavily increased
3	No man.	2	Normal, Heavily increased
4	Man. with equation 18	5	Normal, Slightly increased, Medium increased, Moderately increased, Heavily increased

Table 7. Overview of the boundary conditions for ANFIS training

BC	Parameter/Hyper-parameter	Value range/Definition
1	Number of epochs	50
2	Data Shuffling after each epoch	-
3	Hyperparameter of the Adam optimization algorithm	Learning rate: 10^{-3} Betas: (0.9, 0.999) Eps: 10^{-8} Weight decay: 0
4	Iterations per epoch	[1,2,4,8,14,28,56]

5	Learning rate	$[1,10^{-1},10^{-2},10^{-3},10^{-4},10^{-5},10^{-6}]$
6	Target vectors	$0 < SL \leq SU \leq 1$

5.3 Weighted evaluation framework

In a final step, a weighted evaluation framework is presented to evaluate the effectiveness of explanations in relation to XAI and TAI parameters for the initial state (M1) and the trained ANFIS (M2-M4). The evaluation framework contains the parameters described in section 2.5 (model-agnostic) and the parameters belonging to semantic-based interpretability in Table 2 (model-specific). The ranking used here is VDI Guideline 2225 with a score of 0-4. The weights can be freely selected both for each parameter (marked bold) and for each sub-parameter, but must always add up to 100% in each case. Possible strategies for determining the weights include the use of the Analytic Hierarchy Process (AHP) or the Delphi method. Additionally, the weights can be selected based on the specific objectives or the use case of the AI model. After assigning scores to the individual evaluation metrics and deriving the resulting ranking of the AI models, a sensitivity analysis can be applied. This analysis is crucial for identifying the robustness of the model selection process and for understanding which evaluation criteria have the greatest impact on the final decision. An example of this is the one-at-a-time (OAT) approach, which allows for the systematic assessment of how variations in each metric independently influence the overall ranking. Table 8 shows the evaluation framework.

6 Results and Analysis

M2 and M3 show good trainability, expressed by a low Mean Absolute Percentage Error (MAPE) in the test data set of the target vectors SL and SU of 1.95% and 3.18% respectively. The integrated discontinuity in M4 can be mapped less precisely. Here the MAPE is 10.68%. This is primarily due to the heavily manipulated value range when compared to the initial state M1. In addition, only the Gaussian membership function is used in this paper. For a reduction of the model error, triangular or trapezoidal membership functions with more trainable premise parameters would be a promising option. Furthermore, instead of the Adam optimization algorithm, a metaheuristic approach such as particle swarm optimization (PSO) could be used to train the premise parameters. For the optimization of M4 as well as M2 and M3, an extension of the grid search from the parameters mentioned in section 5.2 as well as an extension to the parameters of the optimization algorithm is also conceivable.

6.1 Model-specific evaluation

In terms of local explainability as described in section 2.5, the outputs of the first layer of the ANFIS can be used to examine the model-specific evaluation metrics contained in Table 8 for the segmentation of the input space. Figure 12 and Figure 13 show the membership functions of the initial state and the trained ANFIS for the first and second input vector, respectively.

Table 8. Weighted evaluation framework to determine the effectiveness of explanations, adapted from [20, 26]

Evaluation metric	Weights	Score
Model-agnostic evaluation (MAE)		
User satisfaction	**X%**	
Clarity of the explanation	X%	0-4
Utility of the explanation	X%	0-4
Sum User satisfaction	100%	0-4
Mental model	**X%**	
Understanding individual decisions	X%	0-4
Understanding the overall model	X%	0-4
Ability of the model to make precise predictions	X%	0-4
Sum Mental model	100%	0-4
Performance and trust assessment	**X%**	
User's ability to act through explanations	X%	0-4
Future trust in the model	X%	0-4
Sum Performance and trust assessment	100%	0-4
Correctability	**X%**	
Error detection through explanations	X%	0-4
Troubleshooting through explanations	X%	0-4
Sum Correctability	100%	0-4
Model-specific evaluation (MSE)		
Segmentation of input space	**X%**	
Completeness	X%	0-4
Normalization	X%	0-4
Distinguishability	X%	0-4
Complementarity	X%	0-4
Sum Segmentation of input space	100%	0-4
Rule base	**X%**	
Consistency of rules	X%	0-4

Rules fired at the same time	X%	0-4
Sum Rule base	100%	0-4
Total result (*MAE*+*MSE*)	**100%**	0-4

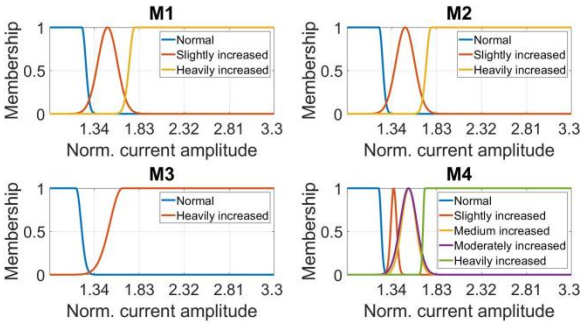

Fig. 12. Comparison of the membership functions of models M1-M4 for the first input vector

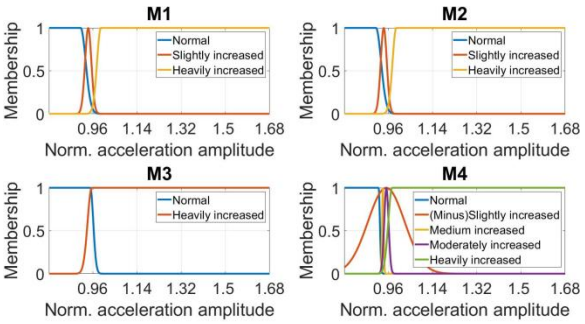

Fig. 13. Comparison of the membership functions of models M1-M4 for the second input vector

As can be seen in Figure 12 and Figure 13, both completeness and normalization are guaranteed for both input vectors and all models. In terms of distinguishability, it can be concluded that the membership functions of the linguistic terms "Medium increased" and "Moderately increased" are almost congruent for the first input vector and M4. This could be due to the fact that originally only four degrees of damage were modelled (see section 4.1) and ANFIS therefore does not recognize any new information in the data. A significant change in a linguistic modifier in M4 can only be seen in the second membership function (marked in red) in the second input vector. Due to an α-value of approx. 0.75, this is a dilation (see Table 1). However, this does not improve the semantic-based interpretability. It is even the case that the midpoint of this membership function with the new linguistic term "(Minus) Slightly increased" lies above the midpoint of the membership function with the linguistic term "Medium increased". This contradiction leads to a low distinguishability of the individual fuzzy sets.

In order to evaluate the complementarity of the models, the accumulated memberships in the input vectors or fuzzy variables are calculated as described in section 2.5. Figure 14 and Figure 15 show the complementarities of the input vectors.

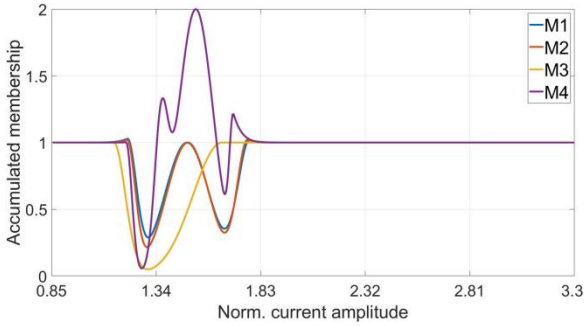

Fig. 14. Comparison of the complementarity of models M1-M4 for the first input vector

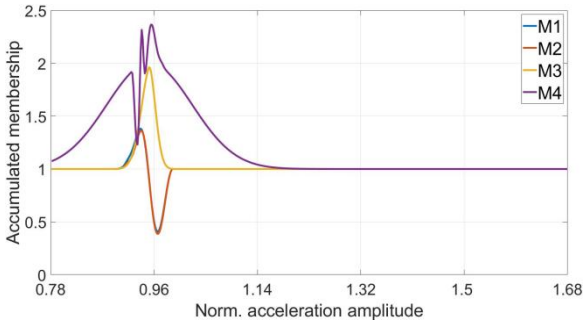

Fig. 15. Comparison of the complementarity of models M1-M4 for the second input vector

As can be seen from the curves of the accumulated memberships, M1 and M2 differ only marginally. Due to the fact that the accumulated memberships of M1 and M2 are in a larger range close to the value 1 compared to the other models, this results in better complementarity. The high accumulated memberships of M4 can be attributed to the low distinguishability of the individual fuzzy sets. This shows the strong correlation of these evaluation metrics. The second input vector shows similar tendencies, but here M1 and M2 differ more strongly from M3. With regard to the model-specific evaluation metrics listed in Table 8 for the segmentation of the input space, it can be concluded that M1 and M2 exhibit slightly better semantic-based interpretability than M3 and significantly better interpretability than M4.

In order to evaluate the rules fired at the same time contained in Table 8, the outputs of the third layer of ANFIS are considered in terms of local explainability. As detailed in section 2.2, the normalized firing strengths are included in this layer. In view of this, the maximum firing strength of the rules in the input space are examined as evaluation metric. A low number of rules fired at the same time is therefore characterized by a range that is close to the value 1. Figure 16 illustrates the results.

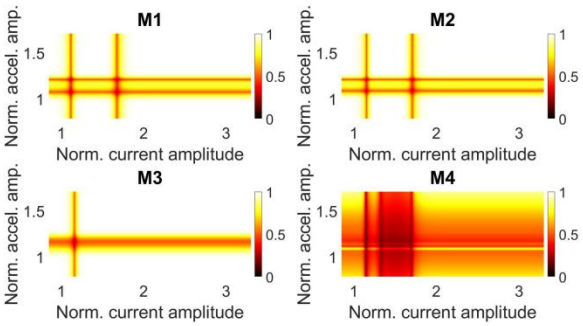

Fig. 16. Comparison of the maximum firing strengths of the rules of models M1-M4 as a function of the input vectors

As can be seen in Figure 16, M3 comprises the largest range in which there is a firing strength of 1. This is due to both the lowest number of membership functions and their small overlap area. Similar results can be seen for M1 and M2, where hardly any differences can be observed, as in the previous evaluations. A clear demarcation in comparison to the other models can be seen in M4. Here there is no area in which a rule has a firing strength of 1. Accordingly, the highest semantic interpretability is given by M3 for this evaluation metric. Figure 16 also shows the method shown in Figure 4 for dividing the input space. According to equation 15, nine hyper cuboids (rules) are mapped for M1 and M2 and four hyper cuboids for M3.

With regard to the consistency of the rules (see Table 8), no identical premises can occur in combination with different conclusions. This is due to the defined linguistic terms for each fuzzy set and the grid partitioning method. Accordingly, complete contradictions are excluded. Furthermore, each rule has a clear semantic meaning, which avoids redundancies. In view of this, the rules for each model are free of contradictions. This completes both the model-specific evaluation metrics and the evaluation of semantic-based interpretability.

6.2 Model-agnostic evaluation

As the TAI plays a central role in this paper alongside the XAI, two model-agnostic evaluation metrics are examined in more detail below. These are the ability of the models to make precise predictions and the future trust in the models. The evaluation is based on the outputs of the fifth layer of ANFIS. There, the target vectors SL and SU as well as the unsureness U to be derived from the support logic explained in section 2.1 are determined.

In order to evaluate the aforementioned model-agnostic parameters, a forward pass of the ANFIS without LSE is performed for both the initial state (M1) and the trained models (M2-M4). Every possible combination of input vectors in the respective value ranges is taken into account as an input. In addition, light blue areas can be seen in some places in Figures 17-19, which indicate that the sixth boundary condition contained in Table 7 is violated after training the models. Furthermore, all samples contained in Figure 6 are displayed transparently in the background for various models. This serves as one way to check the future trust in a trained ANFIS. The heat maps in Figure 17 illustrate the results of the target vector SL.

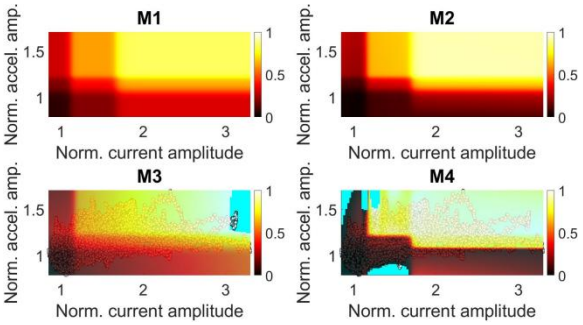

Fig. 17. Comparison of the models M1-M4 with respect to the target vector SL as a function of the input vectors (assertion A is "definitely true")

As shown in Figure 17, M1 and M2 do not violate the sixth boundary condition. The difference between the two models is that M2 predicts higher and lower truth values with regard to wear of the pinion in the boundary areas of the input vectors, i.e. around the minimum and maximum values. This result can be attributed to the S-function shown in Figure 9, as this increases the value range of the target vector, as shown in Figure 10. On the other hand, the minimum differences of the membership functions that depend on this become clear here.

M3 and M4 show the highest truth values for wear of the pinion in certain areas, but in the light

blue colored areas it is not guaranteed that SL remains in the value range of [0,1]. In addition, samples can be seen in these areas that could theoretically occur as inputs to the model in real operation. With regard to the wear of the pinion, its precise condition would therefore not be predictable with these models. This reduces the ability of the models to make predictions about possible states and consequently the trust in the models in terms of TAI.

In a further evaluation, the counter-evidence to assertion A is established. This is equivalent to 1- SU and stands for the truth value that assertion A is "definitely false". Accordingly, there would be no wear of the pinion. Figure 18 illustrates the results.

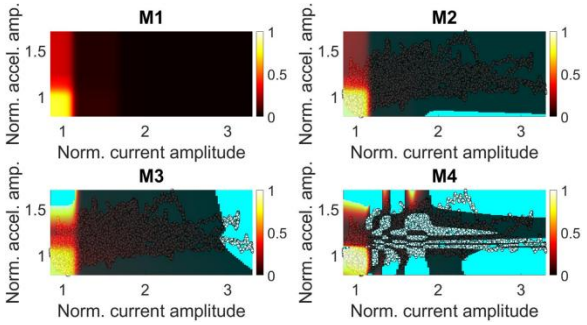

Fig. 18. Comparison of the models M1-M4 with respect to the target vector SU as a function of the input vectors (assertion A is "definitely false")

In contrast to the analysis of the target vector SL, in M2 this time an area occurs in which the sixth boundary condition from Table 7 is violated. However, this is not associated with any loss of trust at this point, as there are no samples contained in the training in this area and there is also a certain distance to them. In comparison to M1, it can be predicted with M2 both in a larger value range as well as with higher truth values that there is definitely no wear of the pinion in low value ranges of the input space. Analogous to Figure 17, M3 in Figure 18 again shows areas in which the sixth boundary condition is violated and also contains possible samples of real operation. This is also the case with M4, although at this point there are hardly any areas in which precise statements can be made about the wear of the pinion. This makes reliable future predictions impossible. This model also clearly shows that the target vector SU has a higher error after training than the target vector SL. This tendency is generally evident in all trained models.

To complete the evaluation of the model-agnostic evaluation metrics, the unsurenesses of the trained ANFIS are shown in Figure 19.

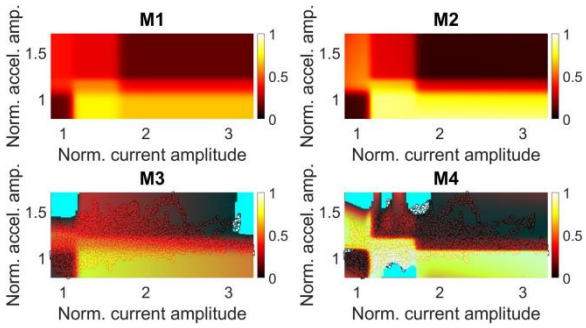

Fig. 19. Comparison of models M1-M4 with regard to the unsureness U as a function of the input vectors

The sixth boundary condition explained in Table 7 also applies to the value range of the unsureness. The results for M3 and M4 are similar to the previous evaluations. For this reason, they will not be discussed in more detail here.

The results of M2 should be emphasized, as the tendencies of the target vectors SL and SU have a positive effect on a low unsureness in the boundary areas of the input space. Compared to M1, conclusions with a higher truth or a lower unsureness with regard to the state of the pinion can be drawn in these areas. The increase of unsureness in all other areas would be justifiable, as no precise statements on the wear of the pinion are required here. The maintenance team should be informed about the condition of the laser cutting machine at this point at the latest. The first two models would therefore guarantee a high level of reliability for future predictions on the wear of the pinion with regard to the XAI and TAI as well as a high level of semantic interpretability in terms of the model-specific evaluation metrics.

6.3 Evaluation by the data scientists

In a final analysis, the models are evaluated by three data scientists familiar with ANFIS functionality using the weighted evaluation framework presented in Table 8. Figure 20 illustrates the average overall results.

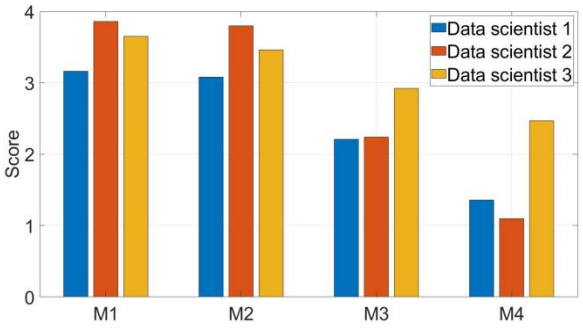

Fig. 20. Scores assigned by the data scientists from the models M1-M4 based on the evaluation framework

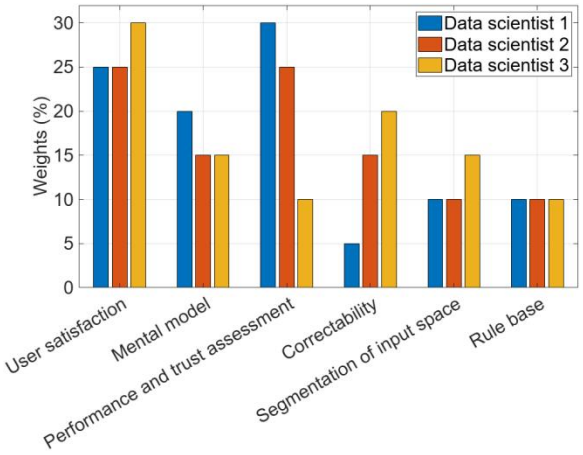

Fig. 21. Assigned weights of the data scientists in relation to the parameters of the evaluation framework

As can be seen from Figure 20, all data scientists rate the effectiveness of the explanations of the individual models in the same order. The differences in the scores are primarily due to the different weights selected as shown in Figure 21. M3 and M4 have a score of 0 in some categories of the model-agnostic evaluation metrics contained in Table 8. This excludes their further use, despite a mean value of 2.46 and 1.64 points respectively. According to VDI Guideline 2225, M1 and M2 have a median score between good and very good.

7 Conclusions

The focus of this paper was the evaluation of ANFIS in the context of XAI and TAI in relation to four different models (M1-M4) for the condition monitoring of a rack and pinion contact. A new methodology in the form of a weighted evaluation framework was developed for this purpose. This offers AI developers and users the opportunity to evaluate and quantify the effectiveness of a model's explanations both model-specifically and model-agnostically.

The approach presented here also shows two ways of predicting the wear of a rack and pinion contact. Firstly, with machine-specific knowledge about a research object and the initial membership functions and r -parameters derived from it. No training of ANFIS is required. Secondly, the manipulated data and the training of the ANFIS were used to show how condition monitoring can be optimized and how changes to the individual parameters of an ANFIS affect explainability and trustworthiness. For both approaches, the developed weighted evaluation framework can be applied and evaluated by experts, for example to assess an industrial implementation. This can be decisive in reducing maintenance and service costs and increasing the availability of a machine The model-agnostic metrics of the evaluation framework are in principle transferable to all AI models. The model-specific metrics are useful for AI models and use cases with integrated fuzzy systems or fuzzy logic.

Since no AI model using a method other than ANFIS is presented in this paper, the model-agnostic evaluation metrics can only be weighed up relatively between the different models. In view of this, no general conclusion can be drawn on the overall explainability of the ANFIS in relation to different training methods such as neural networks. However, due to the comprehensible and step-by-step calculation of the individual layers and their trustworthy representations, the global and local explainability of ANFIS is guaranteed.

Another possible research objective would therefore be to create different AI models for the same use case, evaluate them in relation to the model-agnostic evaluation metrics and compare them with each other. Furthermore, the influence of different forms of membership functions and different hyperparameters of the optimization algorithm on the training of the ANFIS could be analyzed.

Funding

This project (ProKInect N° 02P20A090) is funded by the German Federal Ministry of Education and Research (BMBF) within the “The Future of Value Creation – Research on Production, Services and Work” program and managed by the Project Management Agency Karlsruhe (PTKA). The support

is greatly acknowledged. The authors are responsible for the content of this publication.

References

1. C. Rammer, "Auf Künstliche Intelligenz kommt es an: Beitrag von KI zur Innovationsleistung und Performance der deutschen Wirtschaft.," Bundesministerium für Wirtschaft und Energie (BMWi), 2020.

2. M. Brandt, "Künstliche Intelligenz rechnet sich," [Online]. Available under: https://de.statista.com/infografik/16992/umsatz-der-in-deutschland-durch-ki-anwendungen-beeinflusst-wird/ (Accessed on: 4th August 2024).

3. T. Kraus, L. Ganschow, M. Eisenträger, and S. Wischmann, "Erklärbare KI: Anforderungen, Anwendungsfälle und Lösungen," Technologieprogramm KI-Innovationswettbewerb des Bundesministeriums für Wirtschaft und Energie. Begleitforschung: iit-Institut für Innovation und Technik in der VDI/VDE Innovation + Technik GmbH, 2021.

4. N. Schaaf, S. J. Wiedenroth, and P. Wagner, " Erklärbare KI in der Praxis: Anwendungsorientierte Evaluation von XAI-Verfahren," Fraunhofer IPA, 2021. DOI: 10.24406/publica-fhg-300845

5. C. F. Gethmann et al., "Künstliche Intelligenz in der Forschung: Neue Möglichkeiten und Herausforderungen für die Wissenschaft," 1st ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2022.

6. C. Ehrmann, P. Isabey, and J. Fleischer, "Condition Monitoring of Rack and Pinion Drive Systems: Necessity and Challenges in Production Environments," in Procedia CIRP, vol. 40, pp. 197–201, 2016. DOI: 10.1016/j.procir.2016.01.101.

7. R. Adler et al., "Deutsche Normungsroadmap Künstliche Intelligenz Ausgabe 2," 2022. DOI: 10.13140/RG.2.2.12632.78089.

8. M. Nauta et al.,"From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI," ACM Comput. Surv., vol. 55, no. 13s, pp. 1–40, 2023. DOI: 10.1145/3583558.

9. S. Sithakoul, S. Meftah und C. Feutry, "BEExAI: Benchmark to Evaluate Explainable AI," in Proceedings of the World Conference on Explainable Artificial Intelligence (XAI), Springer, pp. 445–468, 2024.

10. G. J. Klir, "Fuzzy set and fuzzy logic: Theory and applications," New Jersey: Prentice Hall, 1995.

11. M. R. M. Monk, "Support logic programming and its implementation in Prolog," The University of Bristol, 1989.

12. J.-S. Jang, "ANFIS: adaptive-network-based fuzzy inference system," IEEE Trans. Syst. Man Cybern., vol. 23, no. 3, pp. 665–685, 1993. DOI: 10.1109/21.256541.

13. R. Kruse, S. Mostaghim, C. Borgelt, C. Braune, and M. Steinbrecher, "Computational Intelligence: A Methodological Introduction," 3rd ed. Cham: Springer International Publishing; Springer, 2022.

14. C.-U. Yeom and K.-C. Kwak, "Performance Comparison of ANFIS Models by Input Space Partitioning Methods," Symmetry, vol. 10, no. 12, 2018. DOI: 10.3390/sym10120700.

15. C. Borgelt, F. Klawonn, R. Kruse, and D. Nauck, "Neuro-Fuzzy-Systeme: Von den Grundlagen künstlicher Neuronaler Netze zur Kopplung mit Fuzzy-Systemen," 3rd ed. Wiesbaden: Vieweg+Teubner Verlag, 2003.

16. J. Casillas, O. Cordón, F. Herrera, and L. Magdalena, "Accuracy Improvements in Linguistic Fuzzy Modeling," Berlin, Heidelberg: Springer Berlin Heidelberg, 2003.

17. D. P. Rini, S. M. Shamsuddin, and S. S. Yuhaniz, "Particle swarm optimization for ANFIS interpretability and accuracy," Soft Comput, vol. 20, no. 1, pp. 251–262, 2016. DOI: 10.1007/s00500-014-1498-z.

18. J.-S. R. Jang, C. Sun, and E. Mizutani, "Neuro-Fuzzy and Soft Computing-A Computational Approach to Learning and Machine Intelligence," in IEEE Transactions on Automatic Control, vol. 42, no. 10, pp. 1482-1484, Oct. 1997. DOI: 10.1109/TAC.1997.633847.

19. C.-U. Yeom and K.-C. Kwak, "A Performance Comparison of ANFIS models by Scattering Partitioning Methods," in 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, pp. 814–818, 2018. DOI: 10.1109/IEMCON.2018.8614898.

20. D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," AIMag, vol. 40, no. 2, pp. 44–58, 2019. DOI: 10.1609/aimag.v40i2.2850.

21. A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Inf. Fusion, vol. 58, pp. 82-115, 2019.

22. A. Das and P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," ArXiv abs/2006.11371, 2020.

23. A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018. DOI: 10.1109/ACCESS.2018.2870052.

24. A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?," ArXiv abs/1712.09923, 2017.

25. M. Moradi and M. Samwald, "Post-hoc explanation of black-box classifiers using confident itemsets," Expert Systems with Applications, vol. 165, 2021. DOI: 10.1016/j.eswa.2020.113941.

26. M. J. Gacto, R. Alcalá, and F. Herrera, " Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures," Information Sciences, vol. 181, no. 20, pp. 4340–4360, 2011. DOI: 10.1016/j.ins.2011.02.021.

27. W. Zenn, J. Butz and J. Millitzer, "Knowledge-Based AI Model for the Detection of Pinion Wear," 2023 7th International Conference on System Reliability and Safety (ICSRS), pp. 428-433, 2023. DOI: 10.1109/ICSRS59833.2023.10380974.

28. V. Sze, Y. -H. Chen, T. -J. Yang and J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," in Proceedings of the IEEE, vol. 105, no. 12, pp. 2295-2329, 2017. DOI: 10.1109/JPROC.2017.2761740.