ECI Symposium Series, Volume RP5: Proceedings of 7th International Conference on Heat Exchanger Fouling and Cleaning - Challenges and Opportunities,
Editors Hans Müller-Steinhagen, M. Reza Malayeri, and A. Paul Watkinson, Engineering Conferences International, Tomar, Portugal, July 1 - 6, 2007

# INTELLIGENT DISCRIMINATION MODEL TO IDENTIFY INFLUENTIAL PARAMETERS DURING CRYSTALLISATION FOULING

**M.R. Malayeri[1] and H. Müller-Steinhagen[1,2]**

[1] Institute for Thermodynamics and Thermal Engineering (ITW), University of Stuttgart
Pfaffenwaldring 6, D-70550, Stuttgart, Germany, m.malayeri@itw.uni-stuttgart.de
[2] Institute for Technical Thermodynamics, German Aerospace Centre (DLR), Pfaffenwaldring 38-40,
D-70569, Stuttgart, Germany

## ABSTRACT

The introduction of redundant independent variables into any function approximation model, or the neglect of important variables, may result in a correlation with poor prediction and reduced reliability. This paper demonstrates that a novel integrated model of neural networks and genetic algorithms can deal with this problem robustly with good accuracy, while be far less time-consuming compared to lengthy conventional models Furthermore, a redundant variable input was imposed to the model to discern if the approach could identify it among other important variables. Genetic algorithms were exploited as a powerful optimisation tool for the selection of best set of inputs with the help of process "*prior knowledge*" rules.

A comprehensive databank of crystallisation fouling under subcooled flow boiling was used. The resulting model was capable of handling the data and successfully discriminated among several independent inputs if there is any redundant input. The technique may be regarded as a robust method to prevent data over-fitting as well as processes where a large number of inputs are involved such as crude oil fouling.

## INTRODUCTION

Fouling is generally a process with an extremely complicated nature. It involves a considerable number of independent variables with poorly understood interaction between the independent parameters and objective functions. Some of these parameters are:

- Surface temperature
- Bulk temperature
- Bulk composition and chemistry
- Fluid velocity and turbulence
- Physical properties of the working fluid (viscosity, density)
- Surface specifications (material, surface texture, roughness and surface energy)
- Physical properties of the deposit (density, thermal conductivity, stickability)
- Solubility equilibrium
- Chemical kinetics (chemical reaction)

Despite increased attention during the past decades, presently used design procedures involve massive uncertainties, while the recommended correlations and computer models can only be applied to a very limited number of highly idealised deposition processes. These drawbacks may be a result of:

- non-linearity of the fouling process;
- the character of the fouling process which is unsteady-state with potentially high fluctuation;
- the large number of variables and different mechanisms;
- the lack of rigorous understanding of the underlying mechanisms;
- the inherent inadequacy of conventional regression methods to correlate experimental data with an ill-distributed parameter variation.

The utilisation of artificial neural networks, in recent years, has proved to be a pragmatic alternative to address most of these unsatisfactory situations with much better accuracy than conventional parametric regression models. Such exploitation of neural networks as function approximation tool has been undertaken mainly in two distinctive ways: In the first approach, neural networks are merely used to interpolate within the range of experimental results which is usually referred as *black box approach* (Sheikh et al., 1999). The second method is a hybrid approach which includes neural networks in combination with the process "*prior knowledge*" (Malayeri and Müller-Steinhagen, 2001, 2003, 2007). It was found that the results of the second method are more reliable and hence more accurate than the first method.

Although significant progress has been made, the following problems still curtail the optimum use of neural networks:

- For instances where a large number of variables are involved, the selection of inputs may be either a painstaking trial-and-error procedure, arbitrary or merely on a basis of availability. This is not desirable, as introducing redundant inputs or disregarding prominent variables may severely reduce the reliability of any model.

- The proportionality of the relation between independent and objective parameters is not evenly distributed over the domain of the respective deposition process. One such example is crystallisation fouling where it is a diffusion-controlled deposition process at low velocities or reaction-controlled at higher velocities.
- Finally, data bases are often ill-distributed with sparse records in which much weight of the data is concentrated only in a specific domain of data.

The present study is intended to provide a solution for the first problem as stated above. An integrated Genetic Algorithm (GA) and Neural Networks (NN) approach is proposed to facilitate and accelerate the development of an NN regression model for a comprehensive databank of crystallisation fouling under subcooled flow boiling of calcium sulphate solutions (Najibi, 1996). The GA is utilised to identify the most relevant NN input combinations of different independent variables. The resulting model permits the minimisation of a multi-objective criterion that includes NN prediction errors on the training and generalization data sets and, most importantly, a penalty function that satisfies prior knowledge.

The first step in this work is to integrate a suitable neural network with a GA. The second step is to impose a set of process criteria known as "*prior knowledge*" in order to assist the GA to better identify input vectors into the structured neural network. Attempts are subsequently made to utilise the integrated model to discern how the model reacts when it is confronted with redundant parameters.

## PROBLEM STATEMENT

A major problem that arises in all function approximation methods, including regression methods or non-parametric methods such as neural networks, is the ambiguity to select the best combination of variables/ dimensionless groups to be used as inputs to predict a variable of interest. The inclusion of redundant inputs may firstly curtail the training phase and secondly increase the computing time without any improvement in the output accuracy, which ultimately leads to an unreliable network. These consequences may be even worse as the number of influential parameters increases even if defined in dimensionless terms. On the other hand, while choosing the most dominant dimensionless groups, there must be a compromise between the number of dimensionless groups and the accuracy of prediction. Before proposing a methodology for dealing with this problem, it is essential to express the problem in mathematical terms.

Table 1 tabulates a list of N records or data points. In this table M stands for the number of different independent variables, $\mathbf{X_{i,j}}$, $(1<i<N$ and $1<j<M)$ that may influence the objective function of $\mathbf{Y_i}$. Here, **X** may represent dimensionless groups such as Re, Pr, etc.. The ultimate objective is to identify **m** numbers among **M** parameters that are initially being introduced to the model and discard the remaining (M-m) as redundant which would not contribute to better prediction of Y, here fouling resistance. The resulting model is expected to satisfy the following prerequisites:

1. High accuracy of prediction.
2. Phenomenological consistent with the prior knowledge terms that are defined for the model.
3. Minimum complexity by observing the following criteria:
   3.1 The independent input terms, **m**, should be as few as possible.
   3.2 Each input term, **m**, should be highly cross-correlated to the output parameter, **Y**.
   3.3 These input terms should be weakly cross-correlated to each others ($m \neq f(m_i)$).

**Table 1.** Definition of inputs and outputs of the model.

| **M** candidate inputs Dimensionless independent variables | | | | **N** outputs Dependent variable |
|---|---|---|---|---|
| $X_1$ | $X_2$ | … | $X_M$ | $Y_1$ |
| $X_{1,1}$ | $X_{1,2}$ | … | $X_{1,M}$ | $Y_2$ |
| … | … | … | … | … |
| $X_{N,1}$ | $X_{N,2}$ | … | $X_{N,M}$ | $Y_N$ |

## SOLUTION STRATEGY

As stated above, there are two objectives in this study, namely 1) minimum number of independent inputs, **m** and 2) a network with minimum output error for **Y**. The traditional approach to deal with multi-objective problems such as the one here is to optimise primary response function whilst turning into constraints for reliable prediction (Vinnet et al., 1995). In contrast, the GA practice consists of optimising a composite objective function which imposes penalty on any input that does not convince a set of satisfaction indexes that are defined for the model (Tarca et al., 2003).

To simplify the problem, an input selector **S** can be defined which consists of the identification of **m** relevant inputs, out of **M** initial ones. The determination of **S** can be a tedious task with normal optimisation tools such as conjugate gradient or quasi-Newton methods, as the search space of combinations is large and the solution **S** must meet both the process prior knowledge and the accuracy of the resulting NN model. GA technique is thus a more powerful optimisation tool for searching the best input selector **S**.

Genetic algorithms (GAs) have been successfully exploited for problems where the optimum combination among a huge number of input variables is desired within a short period of time (Goldberg, 1989). Based on the mechanisms of natural selection and natural genetics, GAs are capable of highly exploiting the information from already evaluated input combinations, i.e. parent specimens, while ensuring good exploration of the search space. GAs and NNs can be combined in such a way as to ensure a model with excellent predictability while optimising a specifically desired term as the one here as selector input, S (Tarca et al., 2003). The integrated GA-NN procedure used to handle the problem in this work is presented in Fig. 1. More details about genetic algorithms optimisation search can be found in the textbook of Goldberg (1989) and are here only briefly discussed.

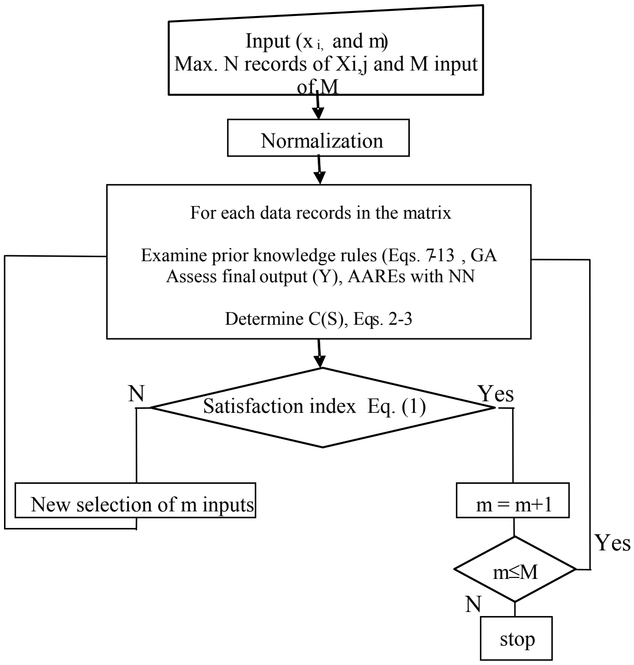
Input (x i, and m)
Max. N records of Xi,j and M input of M
Normalization
For each data records in the matrix
Examine prior knowledge rules (Eqs. 7-13 , GA
Assess final output (Y), AAREs with NN
Determine C(S), Eqs. 2-3
N
Satisfaction index Eq. (1)
Yes
New selection of m inputs
m = m+1
Yes
m≤M
N
stop

**Fig. 1** Flow chart of the integrated GA-NN model.

The GA approach requires a string representation of the *m*-input selector, *S*. Here, **S** is a selection of indices representing some of the candidate input variables of the database listed in Table 1. The genetic algorithm requires a string of one bit representative which is called the encoding modality. In this work, it was chosen as **M**-sized bit strings, permitting merely **m** "one bit" values per string. In this binary representation of solutions, **M** as stated before corresponds to the total number of candidate input variables in the database (Table 1). The "1" at a given rank of the string stands for an input being selected that occupies the same rank in the database. In contrast, "0" stands for an input variable being purged.

To determine the best input selector **S**, the following optimisation criterion, C(S), can be defined (Tarca et al., 2003):

$$C(S) = \min\{C_{opt}(S),\ when\ m_{\min} < m < M\} \qquad (1)$$

$$C_{opt}(S) = \alpha\ AARE\{NN(S)\}_T + \beta\ AARE\{NN(S)\}_G + \lambda\ PPC\{NN(S)\} \qquad (2)$$

In this equation, α, β and λ are coefficients that are picked based on the percentage of data that are used in training, generalization and input selection steps. AARE[NN(*S*)]$_T$ is the average absolute relative error the NN achieves on the training data for a given input combination selector, **S,** AARE[NN(*S*)]$_G$ is the same for the generalisation phase. Finally, PPC[NN(*S*)] corresponds to a penalty for *phenomenological consistency*, ideally guarantees that the model does not breach the expected behaviour of the desired output. This means that all the outputs must satisfy the prior knowledge that is set for the model. In practice, the penalty term must be zero if the resulting network meets all of the process *prior knowledge* rules.

The first generation of "one bit" strings are built in the simplest way. Starting with a null **M**-sized string (all bits are zero), each **m**-input **S** specimen of the first generation is built by turning randomly and with same probability **m** zeros among the **M**'s into ones. The operation is repeated many times till there is no significant improvement in C(S). In this work, for the utilisation of both GA and NN tools, the MATLAB software is employed. As in previous studies, the Radial Basis Function Network (RBF) is utilised as NN tool since it is more accurate for function approximation (Malayeri and Müller-Steinhagen, 2007).

## CASE STUDY AND MODEL IMPLEMENTATION

A comprehensive databank is used in this study which contains 44 sets of experimental fouling resistances as a function of time, reported by Najibi (1996). They were all obtained for $CaSO_4$ scale deposition during subcooled flow boiling in a vertical annulus. The main reasons for using these fouling runs are that i) the experiments were well performed and reproducible, and ii) there is some prior knowledge about the phenomena involved in the deposition process (except during the induction period). The experiments were performed for the following range of operating parameters:

**Table 1** Range of operating parameters.

| V | $T_b$ | $T_s$ | C | I |
|---|---|---|---|---|

| $[m\ s^{-1}]$ | $[^oC]$ | $[^oC]$ | [g/L] | [Mole/L] |
|---|---|---|---|---|
| 0.5-2 | 65-95 | 95-140 | 1.6-2.7 | 0.05-0.3 |

Najibi (1996) showed, as illustrated in Fig. 2, that the deposition rate is controlled by different mechanisms, depending on flow velocity and surface temperature. For the investigated range of flow velocity, fouling rate (a slope of line in the progressive increase of fouling resistance versus time) increases with velocity up to a Reynolds of 30,000 with no noticeable change afterward.

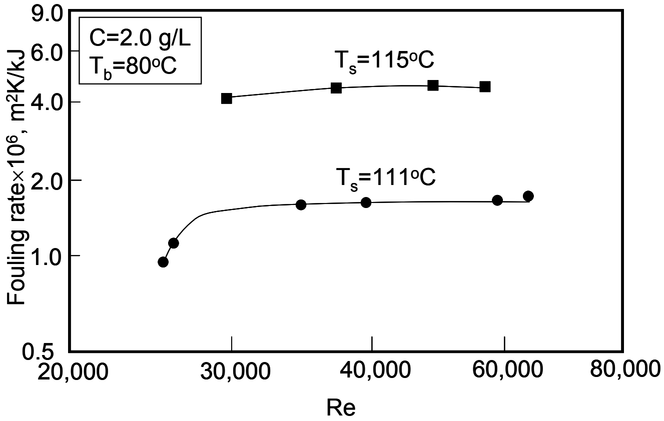

**Fig. 2** Variation of fouling rate as a function of Reynolds number (Najibi, 1996).

The experiments also showed that flow velocity, bulk and surface temperatures, solution concentration and ionic strength influence the fouling process and can be expressed in dimensionless terms as follows:

- Fluid velocity: as stated before at low flow velocities, the mass transfer boundary layer is relatively thick, hence molecular diffusion affects the fouling rate. At high velocities, chemical reaction controls the fouling rate. This effect can be represented in terms of Reynolds number (Re).
- Surface and bulk temperatures: the experimental results show that fouling rates depend mainly on surface temperature, in particular at high fluid velocities. It is also observed that the fouling rate is relatively independent of the bulk temperature over a wide range of velocities. The surface temperature dependence is correlated in terms of ($E_c$=-E/$RT_s$) and the bulk temperature effect in terms of Prandtl number (Pr).
- Solution concentration: the effect of concentration increases as the flow velocity increases, indicating again a change in controlling mechanism from diffusion to reaction. This parameter is represented in form of a ratio of $C_{b,c}=C_b/C^*$, where $C_b$ and $C^*$ are bulk and saturation concentration determined at surface temperature.
- Ionic strength: at constant bulk and surface temperature, an increase of the ionic strength increases the solubility of calcium sulphate and thus decreases the driving force for deposition. Regression analysis done by Najibi (1996) showed that the saturation concentration of calcium sulphate hemihydrate is a function of ionic strength according to:

$$C^* = 10^{a+bz} \tag{3}$$

where parameters a, b and z are:

$$\begin{aligned} a &= 2.047 - 0.01136T \\ b &= -6.5832 + 0.0226T \end{aligned} \tag{4}$$

$$z = \frac{\sqrt{I}}{1 - 1.5\sqrt{I}} \tag{5}$$

- Time is represented in terms of ($\theta=t/t_{max}$) where $t_{max}$ is the longest time an experiment was carried out. It should be pointed out that parameters such as pipe diameter and surface roughness may also influence the fouling process, but are not included in this study because of a lack of experimental evidence.
- There are certainly other important parameters that are not mentioned here. This has been due to the fact that these variables 1) were kept constant throughout the experimentation (such as surface roughness despite its proven effect on the fouling resistance) and 2) were not simply investigated due to the lack of research resources such as surface energy. In order to show and examine the impact of redundant input variables a constant $R_c$ which is virtually considered a representative of surface roughness but with a constant value of 0.1 was deliberately introduced to the model. Therefore only the following dimensionless inputs as M=7 are used:

$$R_f = f\{\mathrm{Re}, E_c, \mathrm{Pr}, \mathrm{C}_{b,c}, \theta, z, R_c\} \tag{6}$$

## MODEL IMPLEMENTATION

The database which is constructed for the integrated model contains **N**=3560 rows and **M**=7 columns of candidate independent inputs given in equation (6). The best **m**-input selector, **S**, must contain a minimum number of elements, **m** which has to be found from all possible combinations of **M**=7 input columns.

To run the GA-NN integrated model, the penalty for phenomenological consistency appearing in the composite criterion equation (1) needs to be formulated. This term must satisfy the *process prior knowledge*. These parameters are described in the previous section that are significantly would affect fouling resistance. In order to mathematically introduce these terms as well as the redundant variable of $R_c$ to the model, they can be re-defined as:

$$\frac{\partial R_f}{\partial T_s} > 0 \quad (7)$$

$$\frac{\partial R_f}{\partial T_b} > 0 \quad (8)$$

$$\frac{\partial R_f}{\partial C} > 0 \quad (9)$$

$$\frac{\partial R_f}{\partial \theta} > 0 \quad (10)$$

$$\frac{\partial R_f}{\partial R_c} > 0 \quad (11)$$

$$\frac{\partial R_f}{\partial z} < 0 \quad (12)$$

$$\frac{\partial \dot{R}_f}{\partial v} \le 0 \quad (13)$$

As it is a tedious and time-consuming procedure to examine all these rules against each input data set, each gradient is evaluated only at two points chosen in the vicinity of the interval edges in the database for each physical variable appearing in equations 7-13. A scale from 0 to 7, measuring the extent of disagreement with physical evidence, is then established to quantify how many rules are violated by the NN model. The PPC term is then expressed simply as the number of rules breached using input selector **S**. If no rules are transgressed by the NN model, PPC[NN(*S*)] equals 0 and the model has no penalty. In contrast, if the model violates all of the rules, the penalty is maximum which means here 7. The coefficients of α and β, the weighting coefficients of the training and generalization AAREs in equation 2 were given the values 0.8 and 0.2, respectively. These values corresponded to the splitting proportions of the initial database into training and generalization sets. Several values for the penalty coefficient λ were tested, and a value of 0.05 was then obtained with respect to minimum computation time.

## RESULTS AND DISCUSSION

Figure 3 presents both the evolutions of the average criterion of the population and of the minimum criterion occurring for the best C(S), as an example of the evolution of the performance of a population through successive generations, The average criterion measures how well the population is doing, as well as how fast it is converging to the optimal solution. The minimum criterion indicates how well the GA has performed in finding a minimum-cost solution in terms of time-consumption. In this figure the sharp decrease occurring at the 25th generation is attributed to the first creation of a fully phenomenologically sought NN model (PPC[NN(*S*)] = 0).

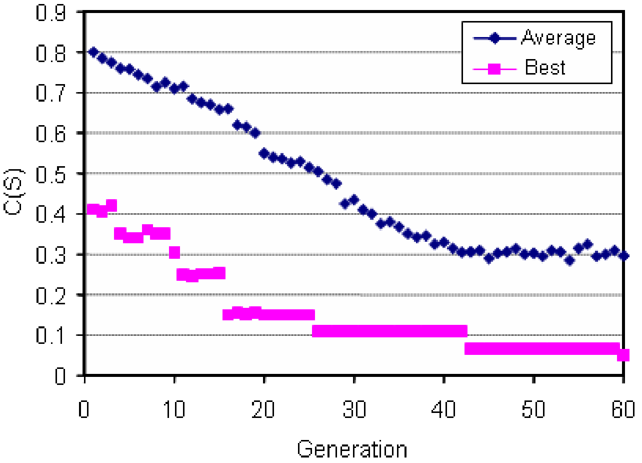

**Fig. 3** Best and population averaged criterion C(S) in a typical integrated GA-NN model.

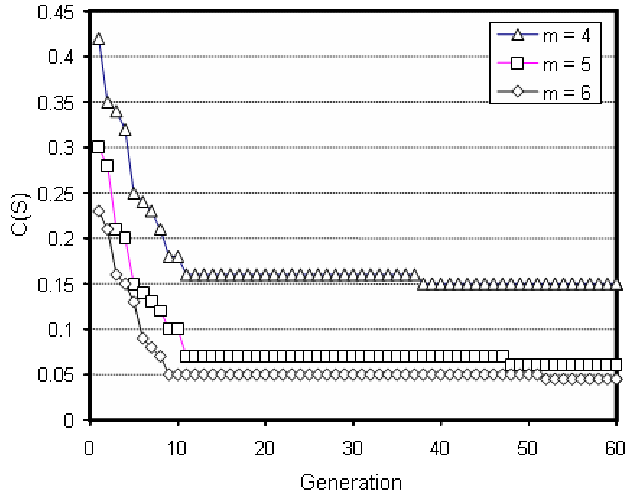

**Fig. 4** Assessment of best input selector, **S** for various number of inputs, **m**.

The integrated model also implies a systematic search with GA of the network for several values of **m**, with the objective of choosing a model with the least complexity, the full phenomenological consistency, and the best accuracy. A search was conducted by launching GA runs for different input numbers such as **m** = 4 – 6 which results are illustrated in Figure 4. It is evident that the first occurrence of having a full phenomenologically consistent model arises, for **m**=4, 5 and 6, after 10 generations. After 22 generations, the penalty term, PPC, becomes zero for the three cases. Interestingly, the introduction of m=6 which represents the Prandtl number does not significantly improve the accuracy of the model. This is consistent with experimental observations where the impact of bulk temperature on fouling resistance is much less than the other important inputs such as velocity and surface temperature. The final AARE is 8.7% for the training phase and 13.5% for the generalisation phase, which are notably better than the results previously reported by Malayeri and Müller-Steinhagen (2001) which were based on the use of all available independent variables without choosing the best selection of inputs as in the present study.

Perhaps the most interesting results come when the model is subjected to $R_c$, as a redundant input, while keeping the other inputs constant. Here, as mentioned before, $R_c$ does not have any physical meaning and was kept deliberately constant. It can be seen in Fig. 5 that C(S) does not change with respect to number of generations with relatively high value of C(S). This implies that this input violates a rule (here most probably defined by equation 13) and thus it may be purged from the list of inputs without the loss of model predictability. One would argue that the introduction of a constant not a parameter to the model and its ability to identify this as a redundant input is a mathematical triviality. It should be pointed out, however, that here the resultant model was able identify the redundant constant systematically without having any prior knowledge of the inputs.

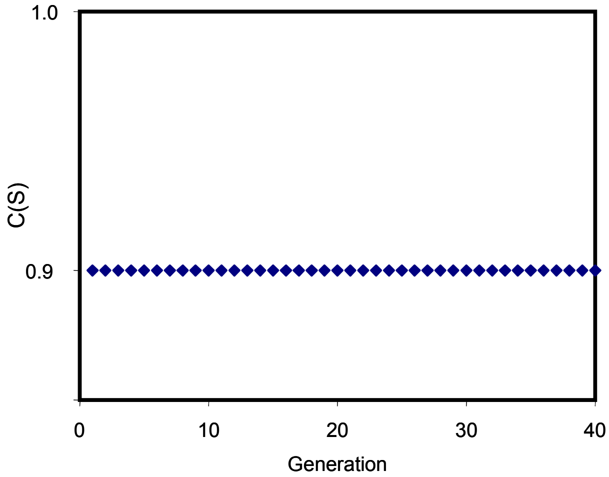

**Fig. 5** Evaluation of the model when it is prone to redundant input, $R_c$ while keeping the other inputs constant.

It should be pointed out that the present work was undertaken for well-conducted fouling runs with only a few input variables and with relatively good in-depth knowledge of underlying fouling mechanisms. However, such an innovative approach may potentially be more helpful for instances where the number of input variables, **m**, is excessively large, and with poorer understanding of basic phenomena governing the respective fouling process. A typical example for this could be crude oil fouling.

## CONCLUDING REMARKS

The integrated GA-NN has proven to be a powerful tool to meet two main objectives: The first objective was to choose the best selection of independent input among a candidate list of both dominant and redundant input variables. The second aim was to predict output, i.e. fouling resistance, with good accuracy. The solution strategy was based on searching among several dimensionless inputs while imposing several rules as *prior knowledge* to help GA as search tool to select the optimum selection of inputs. It should be mentioned, nonetheless, that the optimisation procedure was conducted in the vicinity of only two points in the matrix of data records. This has been due to the laborious and time-consuming procedure of searching the prior knowledge rules for each point within the domain of data points. Finally, with this approach, a satisfactory agreement between predicted and measured fouling resistances has been achieved.

## NOMENCLATURE

| | |
|---|---|
| a,b | defined in equation (4) |
| AARE | Average Absolute Relative Error |
| C | solution concentration, g/L |
| C(S) | composite criterion in equation (1) |
| GA | Genetic Algorithm |
| I | ionic strength, mole/L |
| m | examined number of inputs |
| M | maximum number of candidate inputs |
| N | number of data points |
| NN | Neural Networks |
| PPC | Penalty Phenomenological Consistency |
| $R_f$ | fouling resistance, $m^2$ K $W^{-1}$ |
| $R_c$ | surface roughness representative |
| S | input selector |
| t | time, sec |
| T | temperature, K |
| v | velocity, m/s |
| z | defined in equation (5) |
| RBF | Radial Basis Function |

### Greek Letters

| | |
|---|---|
| $\alpha,\beta,\lambda$ | Coefficients in equation (2) |
| $\theta$ | Time dimensionless term |

### Subscript

| | |
|---|---|
| b | bulk |
| s | surface |

## ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support of the German Research Council (DFG).

## REFERENCES

Goldberg, D.E., 1989, *Genetic Algorithms in Search, optimization, and Machine Learning*, Addison-Wesley, Reading, MA.

Malayeri, M.R., and Müller-Steinhagen, H., 2001, Neural network analysis of heat transfer fouling data, The 4th United Engineering Foundation Conference on *Heat Exchanger Fouling: Fundamental Approaches & Technical Solutions*, Davos, Switzerland, pp. 145-150.

Malayeri, M.R., and Müller-Steinhagen, H., 2003, Analysis of fouling data based on prior knowledge, in "Heat Exchanger Fouling and Cleaning: Fundamentals and Applications", Paul Watkinson, Hans Müller-Steinhagen, and M. Reza Malayeri Eds, ECI Symposium Series, Volume RP1, pp. 145-147.

Malayeri, M.R., and Müller-Steinhagen, H., 2007, Initiation of $CaSO_4$ scale formation on heat transfer surfaces under pool boiling conditions, *Heat Transfer Eng,.* Vol. 28, No. 3, pp. 240-247

Najibi, S.H., 1996, Heat transfer and heat transfer fouling during subcooled boiling of hard water, *PhD thesis,* University of Surrey, UK

Sheikh, A.K., Kamran-Raza, M., Zubair, S.M., and Budair, M.O., 1999, Predicting level of fouling using neural network approach, *UEF conference on Mitigation of Heat Exchanger Fouling and its Economic and Environmental Implications,* 11-16 July, Banff, Canada.

Tarca, L.A., Grandjean, B.P.A. and Larachi, F., 2003, Reinforcing the phenomenological consistency in artificial neural network modeling of multiphase reactors, Chem. Eng. Proc., Vol. 42, pp. 653-662.

Viennet, R.; Fonteix, C.; Marc, I., 1995, New Multicriteria Optimization Method Based on the Use of a Diploid Genetic Algorithm: Example of an Industrial Problem, *Lecture Notes in Computer Science Artificial Evolution;* Alliot, J.-M., Lutton, E., Ronald, E., Schoenauer, M., Snyers, D., Eds.; Springer: Berlin, Vol. 1063, pp 120-127.