Contents lists available at ScienceDirect

International Journal of Heat and Mass Transfer

journal homepage: www.elsevier.com/locate/ijhmt

Development and assessment of the data-informed continuous machine learning approach based on CHF prediction

Meiqi Song [a,b], Fabian Wiltschko [b], Xiaojing Liu [a,*], Aurelian F. Badea [b], Xu Cheng [b,*]

[a] College of Smart Energy (CSE), Shanghai Jiao Tong University (SJTU), 800 Dong Chuan Road, Shanghai 200240, China

[b] Institute of Applied Thermofluidics (IATF), Karlsruhe Institute of Technology (KIT), Kaiserstrasse 12, Karlsruhe 76131, Germany

ARTICLE INFO

Keywords:
Data-informed
Continuous machine learning
Distribution functions
CHF prediction

ABSTRACT

Machine learning (ML) method has attracted more and more interests in engineering applications. Despite extensive efforts in the last decades in the application of the ML-method to thermal- and fluid mechanics, there exist in general some obvious shortcomings. Generally, neither sufficient information about the data base used by the previous researchers nor information about the uncertainty (or error) is included in the ML-model and is not available for the next researchers. This makes the continuous learning process difficult or even impossible.

This paper proposes the new data-informed continuous machine learning (DI-CML) approach, to overcome the above shortcomings. The main feature of the DI-CML approach is to generate a machine learning package, which, in addition to the ML-model, contains the distribution functions of the input variables and the distribution function of the uncertainty (error). With this ML-package, an artificial data base can be produced, which should be as similar as possible to the original data base used for the development of the ML-model. This would make the continuous learning process possible and efficient.

The main idea and the procedure of the DI-CML approach is described. The feasibility of the DI-CML approach is assessed by means of CHF prediction. The large CHF data base provided by the OECD-NEA benchmark working group is used. The accuracy of the CHF prediction by the DI-CML approach is analysed by using different features of data base sets, different methods to derive the distribution functions of the input variables as well as different methods for the generation of the artificial data base. The results confirm the good feasibility of the proposed DI-CML approach. Furthermore, challenges and future research needs are also identified.

1. Introduction

Critical heat flux (CHF) plays a key role in the safe design of many heat transfer systems [1,2]. For many decades, reliable prediction of CHF has attracted strong attention in the research community [5,7]. Due to the complex phenomena involved in two-phase flow and heat transfer, analytical prediction of CHF is impossible. Despite the large number of mechanistic models, their reliability remains questionable and requires further improvement and validation [11]. Recently, many studies have been devoted to the prediction of CHF using the computational fluid dynamics (CFD) approach. However, the sub-models applied in the CFD models for describing various two-phase flow and heat transfer processes are usually not sufficiently validated. The generality of their application is not confirmed and requires also further improvement and validation [13]. Up to now, empirical correlations or look-up tables are the most widely applied CHF prediction methods in engineering applications [5,6]. There exist hundreds of empirical correlations, which were mostly derived based on experimental data bases with limited parameter ranges. It is well agreed that extrapolated application of such correlations to parameters outside their valid range may lead to significant errors and, thus, is not recommended.

Application of Machine Learning (ML) approaches, mainly based on the artificial neural network (ANN), to the CHF prediction started several decades ago [12,16] and attracted more and more attention in the last years [19]. Recently, under the umbrella of OECD/NEA, a working group is established and organizing a benchmark task related to the application of ML-approach to the CHF prediction with >30 institutions. OECD/NEA provides the worldwide largest data bank of CHF obtained in circular tubes with vertical upward water flow and uniform heat flux distribution. This data bank contains >24,500 CHF experimental data points, covering a wide range of parameters.

Despite extensive efforts in the last decades in the application of the ML-approach to engineering problems, there exist in general some

[*] Corresponding authors.
E-mail addresses: xiaojingliu@sjtu.edu.cn (X. Liu), xu.cheng@kit.edu (X. Cheng).

https://doi.org/10.1016/j.ijheatmasstransfer.2025.127474
Received 13 March 2025; Received in revised form 9 May 2025; Accepted 26 June 2025
Available online 3 July 2025

Nomenclature

D	diameter, m
G	mass flux, kg/m^2s
K	data points for comparison
p	pressure, kPa
P	probability distribution function
x	steam quality, -
q	heat flux, W/m^2
CHF	critical heat flux, W/m^2
x	input variables
y	output variable
w	weighting factor in Eq. (5)
ε	error
μ	average value
σ	standard deviation

obvious shortcomings, e.g.

- No sufficient information about the used data base is available. An ML-model was developed based on a large data base used for its training, validation and testing. In general, any empirical models, including ML models, derived with the help of experimental databases, are only valid within the valid range of the input parameters. Furthermore, the reliability of the model in a sub-domain of the valid parameter range is strongly coupled with the number of the data points in this sub-domain. Thus, information of the distribution of the input variables is crucial important for a correct application of the model and interpretation of the reliability. However, an ML-model is usually only a black box and doesn't deliver any (or sufficient) information about the database used such as the distribution of the input variables. It is difficult, for the next researchers to extend the previous ML-model by using their own additional data bases and at the same time to keep its accuracy also for the previous database.
- The accuracy of an ML-model depends on the algorithms applied and is obviously user dependent. With the same data base, different researchers generate different ML-models with different accuracy. Again, no sufficient information about the uncertainty (or error) is included in the ML-model.

The above two shortcomings restrict strongly the continuous learning procedure and, subsequently, its effectiveness in engineering applications.

In the engineering application, continuously improvement and training of models is highly required. Generally, many researchers are working on the same engineering problem at either different institutions or in different time sequences. For example, since more than six decades, extensive experimental studies on CHF were carried out worldwide. A large amount of experimental data bases were generated in different institutions. Due to some practical reasons, such as privacy issue, such data bases are not available for other institutions or researchers. The common situation in the past was that many institutions or individual researchers developed prediction models based on their own data bases. This leads to hundreds of empirical correlations with narrow valid parameter ranges [6].

The same situation also exists for the CHF prediction based on ML-approaches. Various ML-models or neural networks were generated by the individual researchers using their own data bases. Generally, these ML-models don't contain sufficient information about the data base used and the uncertainty (or error) of the models. It is very difficult or impossible for the other researchers to extend these models by adding their own data bases and at the same time to keep high accuracy for both the previous data bases and the additional new data base.

To overcome the above two shortcomings, the present authors propose the so-called data-informed continuous machine learning approach (*DI-CML*). The main feature of the DI-CML approach is to generate an artificial data base, which is as similar as possible to the original data base used for the development of the previous ML-models. The input parameters such as pressure, mass flux, steam quality and diameter should be generated via distribution functions, which were derived according to the previous data bases.

In this paper, the new approach DI-CML is briefly described. The feasibility of the DI-CML approach is assessed with the CHF data base provided by the OECD/NEA benchmark working group.

2. Existing methods for artificial data generation

The continuous learning capability is a crucial issue in the development of ML-models in engineering applications. Conventional ML-approaches, mainly based on artificial neural networks are not, in general, capable of this function. However, in the last years much effort was made to develop algorithms for achieving continuous learning capability. Two approaches are summarized as below.

(A) knowledge distillation

One of the popular approaches for the continual machine learning discussed in the open literature is Knowledge Distillation (KD) based on the generative adversarial network (GAN) methodology [8]. There exist two categories of KD according to their targets. The first category focuses on the generation of a smaller network from a larger network with the purpose of reducing the size and subsequently, the computing effort of the network and to make its application possible to some real-time scenarios [4]. Furthermore, the small network is trained only based on the large network without any data base, the so-called data-free knowledge distillation [9]. In this category there exist again various types of knowledge generators such as feature-based knowledge, relation-based knowledge and response-based knowledge. However, this category doesn't include data-based knowledge generators [4].

The purpose of the second category is the generation of an artificial data base through training a neural network (or data generator). The artificial data base should be similar to the original data base. Various algorithms have been proposed in the last years such as the conditional tabular GAN [15,17]. It was found out that it remains a big challenge for all the existing GAN-based data generators to produce the artificial data base having sufficient similarity to the original data base. Furthermore, it is usually assumed that the variables to be generated are random and independent. This assumption is often not valid in many engineering applications such as in the experimental studies on CHF, the realizable values of various input parameters such as pressure, mass flux and steam quality cannot be adjusted independently. For example, it is very difficult to perform CHF experiments at low pressure and at low mass flux, simultaneously.

(B) Elastic weighting consolidation (EWC)

The main idea of the algorithm, elastic weight consolidation (EWC), is trying to remember the original data through the importance of the weights in the neural network [10]. It protects the knowledge of the original data during training new models by selectively decreasing the plasticity of weights. The importance of the weights and bias in the previous ML-model with respect to the original data base is represented in the Fisher information matrix, which contains the gradient of the lost function to all weights and bias. In extending or training a new ML-model with an additional data base, a penalty term is introduced to modify the effect of various weights and bias on the loss function according to its importance.

$$L_{EWC}(\theta) = L_{new}(\theta) + \lambda \sum_i F_i \left(\theta_i - \theta_i^*\right)^2 \tag{1}$$

Where L is the loss function, is F the Fisher information matrix, θ are the sets of weights and bias. The constant λ controls how strong the regulation should be. This method doesn't directly generate an artificial data base, but nevertheless assumes, that important information of the previous data base is to some degree indirectly included in the Fisher information matrix. However, the direct connection of the Fisher Information Matrix to the previous data base is not proven. For example, the number of the previous data points is not taken into consideration, although the importance of the previous ML-model would be enhanced with the number of the previous data base.

The EWC algorithm requires the generation of the large Fisher information matrix, which contains the sensitivity parameters of all weights and bias of the previous neural network. A complete information package is provided, which, in addition to the black-box ML-model, provides explicitly all values of weights and bias, as well as the Fisher information matrix. This information package could become extremely large.

It is worth mentioning another ML approach, i.e. domain generalization (DG) [18]. The main objective of the DG approach is to train a ML model, which should either provide good prediction in the domain where data are available for training the model or could be applicable to unseen domains. Although the DG approach recently attracted much attention in classifying images and processing natural language [18], there still exist significant challenges and needs in further improvement [14]. Compared to the DG approach, the new approach proposed in the next chapter emphasizes on continuous learning, i.e. to generate the information of the previous databases for further improvement of the previous ML model by including the new databases. Thus, the purpose of the DI-CML approach differs from that of the DG approach.

3. New DI-CML approach

As discussed in Section 2, the existing and widely applied continuous machine learning approaches, such as KD and EWC, are not appropriate for some engineering applications.

Although KD is a well-established and widely applied methodology in ML community and a very powerful method for reducing the complexity of a given ML model, using the original training database. The purpose of the KD method differs from that of the DI-CML approach, i.e. to generate artificial database, which represents as far as possible the previous database used for training the previous ML-model and to realize the continuous learning process.

The approach EWC is a method, indirectly trying to consider the previous database information in the training of the actual ML model through the Fisher information matrix. It requires the storage of a huge amount of information, in addition to the previous ML model. Furthermore, the effectiveness and accuracy of the EWC method cannot easily be identified and needs further investigation. This limits the application of the EWC method for the purpose of continuous learning.

In this paper, a new continuous machine learning approach, the so-called data-informed continuous machine learning (DI-CML) is proposed. The key feature of the DI-CML approach is the generation of an artificial data base via simple distribution functions of the input variables. It requires a minimum amount of information, e.g. a few variables to describe the distribution functions, to be included in the previous ML package. Fig. 1 illustrates the structure of the DI-CML approach. The relationship between the k input variables $(x_1, x_2, \ldots, x_k)$ and the output parameter y is expressed as

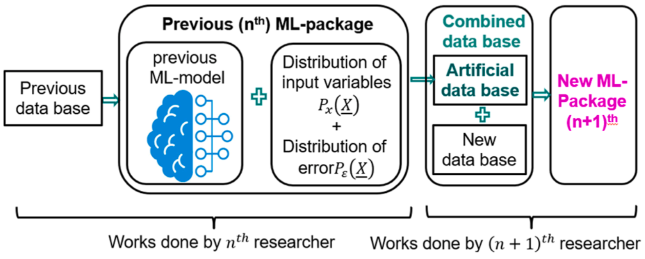

Previous data base
Previous (n^{th}) ML-package
previous ML-model
Distribution of input variables $P_x(\underline{X})$ + Distribution of error$P_\varepsilon(\underline{X})$
Combined data base
Artificial data base
New data base
New ML-Package $(n+1)^{th}$
Works done by n^{th} researcher
Works done by $(n+1)^{th}$ researcher

Fig. 1. Structure of the DI-CML approach.

$$y = f(x_1, \ x_2, \ \ldots, x_k) \tag{2}$$

In this figure, the DI-CML approach is explained with two successive stages of the model development, performed by the previous researcher, identified as the n^{th} researcher, and the present researcher, identified as the $(n+1)^{th}$ researcher. Based on the *previous data base*, the previous researcher generated the previous (n^{th}) ML-package, which includes the *previous ML-model* and the distribution functions for the input variables, $P_x(\underline{X})$, as well as the distribution function for the error $P_\varepsilon(\underline{X})$. One of the key tasks of the DI-CML approach is the generation of both distribution functions. With the previous ML-package, i.e. the previous ML-model and the distribution functions, an *artificial data base* is generated, which has similar features as the previous data base. In this way, the *combined data base* is established, which consists of the artificial data base and the *new data base* of the present researcher. The combined data base is applied to produce the new ($(n+1)^{th}$) ML-package, which includes the new ML-model, new distribution functions for the input variables and the new distribution function for the error, and is then available for the future researchers, who can further extend the $(n+1)^{th}$ ML-package by adding their own experimental data bases. Thus, a continuous and sequential training of the machine learning model becomes possible.

In the following, the procedure to generate the distribution functions of the input variables, and subsequently, the artificial data base is explained.

The previous data base with N data points was applied to develop the previous ML-model and yields

$$y_{m,i} = m(x_{1,i}, \ x_{2,i}, \ \ldots, x_{k,i}), \ i = 1, \ N \tag{3}$$

The error is defined as the difference between the predicted and the measured values of the output parameter and calculated as

$$\varepsilon_i = \left(y_i - y_{m,i}\right) = f(x_{1,i}, \ x_{2,i}, \ \ldots, x_{k,i}) - m(x_{1,i}, \ x_{2,i}, \ \ldots, x_{k,i}), \ i = 1, \ N \tag{4}$$

A thorough analysis of the distribution of the input variables and the error will be carried out, to derive functions, which reasonably represent the real distributions of the input variables $P_x(\underline{X})$ and the error $P_\varepsilon(\underline{X})$. The selection of the method for the derivation of the distribution functions depends on problems considered and, thus, is case-dependent and will be discussed in Section 4. This is also the key challenge of the DI-CML approach and requires continuously improvement in the future.

With the help of the distribution functions derived for the input variables $P_x(\underline{X})$, artificial sets of input variables, $(\widetilde{\underline{X}}_i, \ i = 1, \ N\ )$, will be produced. The number of the artificial data sets N is the same as the number of the data points of the previous data base. For each set of input variables, a value of the output parameter $\widetilde{y}_i$ is then derived based on the previous ML-model, combined with the correction through the error distribution function, i.e.

$$\widetilde{y}_i = m\left(\widetilde{\underline{X}}_i\right) + w_i \cdot \varepsilon\left(\widetilde{\underline{X}}_i\right), \ i = 1, \ N \tag{5}$$

The weighting factor w_i is introduced to consider the strength of the error correction and expected to be dependent on the distribution function of the input variables. For the simplicity, the correction through the error $\varepsilon(\widetilde{\underline{X}}_i)$ is neglected in this paper, i.e. the weighting factor is set $w_i = 0$. Thus, an *artificial data base* with data points N is produced, i.e. $\{(\widetilde{\underline{X}}_i, \ \widetilde{y}_i), \ i = 1, \ N\ \}$, with

$$\widetilde{y}_i = m\left(\widetilde{\underline{X}}_i\right), \ i = 1, \ N \tag{6}$$

The similarity of the artificial data base $\{(\widetilde{\underline{X}}_i, \ \widetilde{y}_i), \ i = 1, \ N\ \}$ to the previous data base $\{(\underline{X_i},\ y_i),\ i=1,\ N\}$ depends on the methods to derive the distribution functions $P_x(\underline{X})$ and determines the accuracy of the DI-CML approach.

The artificial data base is then combined with the new data base (with M data points) of the present researcher, and the *combined data base* is generated, which contains $N+M$ data points. With the combined data base, the new ML-package is produced in the same way as described above and contains the new ML-model, the new distribution functions for the input variables and the new distribution function for the uncertainty (error). The new ML-package is then made available for the future $(n+2)^{th}$ researcher.

4. Assessment of the DI-CML approach

This chapter is devoted to analysing the feasibility of the DI-CML approach. For this purpose, the DI-CML approach is applied to CHF prediction. The experimental data are taken from the OECD/NEA benchmark working group [3]. To describe the CHF in circular tubes with uniform heat flux, four input variables are taken into consideration, i.e. pressure, mass flux, local steam quality and tube diameter. The CHF data base provided by the OECD/NEA benchmark working group contains 24,578 data point and covers the following parameter ranges:

Pressure, p [kPa]:	0.1 – 20.0
Mass flux, G [kg/m^{2}s]:	8.2 – 7964.0
local steam quality, x_{out} [-]:	−0.5 – 1.0
Tube diameter, D [m]:	2.0 – 16.0

As mentioned above, the key target of the DI-CML approach is the generation of the artificial data base. There is no ready-to-use method. The researchers should find out the most suitable way for their specific engineering problems. For the feasibility study purpose in this paper, two different and very simple approaches are applied to generate the artificial data bases and are discussed in two separate sub-chapters, 4.1 and 4.2, respectively. In the first approach described in chapter 4.1, the distribution of *only* one input variable (characterized as reference input variable) is considered, whereas uniform distributions are assumed for all other input variables. In the second approach, all input variables are assumed to be independent, their distributions are considered simultaneously and the results are presented in chapter 4.2.

4.1. *Generation of artificial data base based on one input variable*

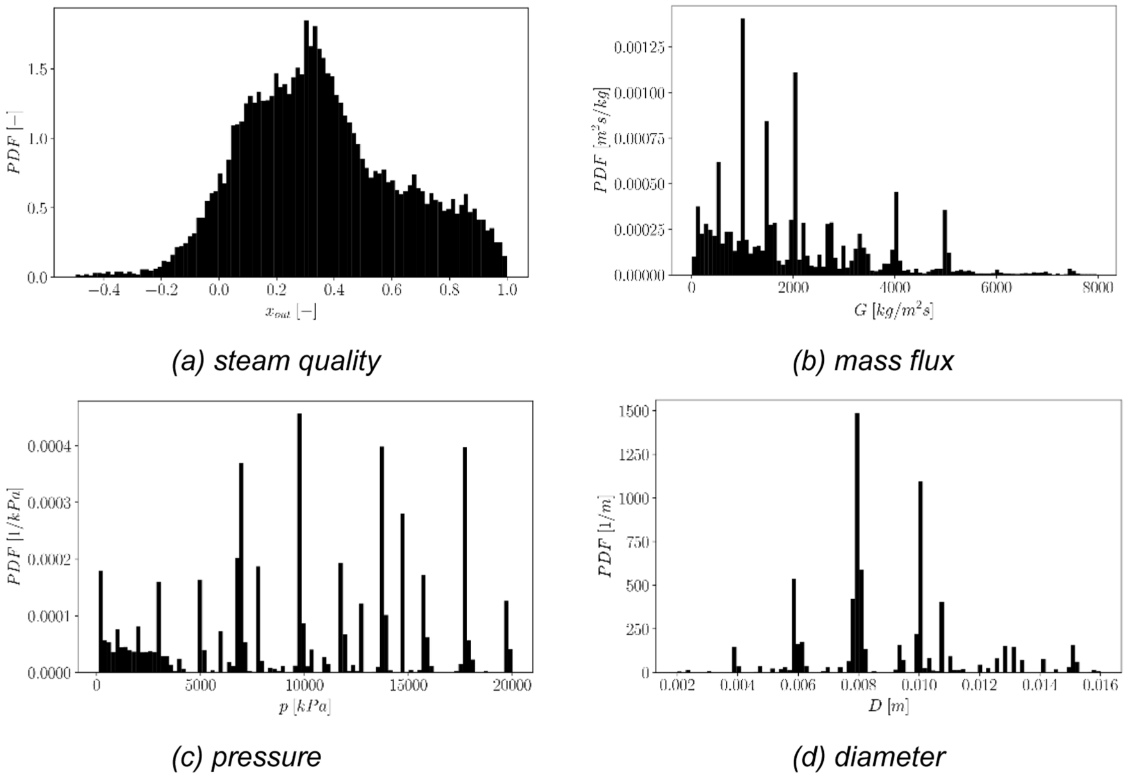
The procedure for the feasibility assessment is divided into 5 steps.

Step 1. ***Generation of the previous data base and the new data base***

The *original data base* provided by the benchmark working group is characterized as DB0. Fig. 2 shows the distribution of the four input variables of the original data base. The distribution of steam quality (Fig. 2a) shows well continuous feature over its entire range, whereas the distributions of both pressure (Fig. 2c) and diameter (Fig. 2d) give strong discrete features. The distribution of mass flux (Fig. 2b) exhibits, in general, continuous behaviour and has, however, strong discrete peaks. In the present study, mass flux is selected as the *reference input variable*, because its distribution includes some features of the distributions of the other three input variables.

The original data base DB0 with 24,578 data points is arranged according to the reference input variable, i.e. mass flux, from large to small values. This arranged data base is divided into two groups with the same number of data points, i.e. 12,289. All data points in the first group have mass flux larger than the mass flux of the data points in the second group. In the entire original database 12,289 data points have their mass flux larger than 1585 kg/m^2, whereas the remaining 12,289 data points have mass flux smaller than 1585 kg/m^2.

Two different options are used to sample the data points from both data groups and to build both the previous data base and the new data base. In the *first sampling option*, 80 % data points of the first data group (i.e. 9831 data points) and 20 % data points of the second data group (i. e. 2457 data points) are randomly sampled to build the *data base 1 (DB1)*. This data base is also referred to "*previous data base*",



(a) steam quality

(b) mass flux

(c) pressure

(d) diameter

Fig. 2. Distribution of the four input variables of the original data base DB0.

corresponding to the terminology used in Fig. 1. The rest data points from both data groups are put together and form the *data base 2 (DB2)* and referred to "*new data base*". In this way, two data bases, i.e. DB1 (previous data base) and DB2 (new data base) are established.

In the *second sampling method*, the entire data points of the data group 1 are used to build the data base 1 (DB1), and the data base 2 (DB2) is also the same as the data group 2. In this case, the mass flux in DB1 is always larger than the mass flux in DB2. According to the second sampling option, a completely different distribution of mass flux of both data bases occurs. The mass flux range in DB2 is totally outside the range in DB1.

Through the two sampling options, together with the original data base, three different structures of data bases are built, and their case-ID names are summarized in Table 1.

In the case identification, the letter *"G"* stands for the case with mass flux as the selected reference input variable. The number represents the sampling option. For example, with the 2nd sampling method, the established two data bases are characterized as DB1-*G2S* and DB2- *G2S*, respectively.

Step 2. ***Derivation of the previous ML-models***

With the data bases DB1, the ML-models (*MLM1)*, corresponding to the *previous model* in Fig. 1, are trained. In this paper, for training of all ML-models the same algorithms and structure of neural network are applied with the following features:

80 % of the data base are used for training und 20 % for testing. The neural structure is DNN with 3 hidden layers and 200 nodes in each layer. In each layer, ReLU activation function is applied. The loss is defined as the mean squared error function. The Adam optimizer is used for the training, and the learning rate is initially set to 10^{-3} and reduced by factor 0.8, if the validation loss does not decrease for 30 training epochs. Finally, training is terminated, when the validation loss does not decrease for 100 epochs. The training data is split into batches of size 50. Each node's weights are randomly initialized, but reproductively, while all biases are initialized by 0. Furthermore, a 5-fold cross validation was carried out, e.g. database DB1 is split into 5 sub-sets, 4 of them were selected for training and one for testing. There are 5 possible combinations of 4 sub-sets used as the training database. The results of the 5-fold cross validation show a negligibly small effect of the sub-sets selection on the results.

It has to be mentioned that the main objective of the present paper is the feasibility study of the proposed DI-CML approach. The conclusions achieved in this paper are at least qualitatively independent of the selection of the neural network architecture. According to the experience gathered in the last years by the present authors, the neural network architecture selected in this paper should be well suitable for the type of problem and the size of the database. Further efforts will be carried out, to study the quantitative effect of the neural network architecture on the results.

The accuracy of the models MLM1 with respect to various data bases is summarized in Table 2. The average value and the standard deviation of this paper are defined as:

$$\varepsilon_i = 100(\widetilde{y}_i - y_i)/y_i \tag{7}$$

$$\mu = \frac{1}{K}\sum_{i=1}^{K}\varepsilon_i \tag{8}$$

Table 1
identification of the cases considered in this study.

	Original data base	1st sampling option	2nd sampling option
Case-ID	*O*	*G1S*	*G2S*

Table 2
Accuracy of the reference ML-models.

Models →	MLM0		MLM1-G1S		MLM1-G2S	
K ↓	μ	σ	μ	σ	μ	σ
	Compared to DB0					
24,578	2.56	19.25	5.66	30.98	−23.97	111.84
	Compared to DB1					
12,289	1.76	15.25	4.44	19.40	−0.06	10.16
	Compared to DB2					
12,289	3.36	22.53	6.88	39.24	−47.92	154.23

$$\sigma = \sqrt{\frac{1}{K-1}\sum_{i=1}^{K}(\varepsilon_i - \mu)^2} \tag{9}$$

K stands for the number of data points used for comparison.

As expected, compared to the original data base DB0, the original ML-model (MLM0) gives the best agreement. The ML-model based on the first sampling option (MLM-G1S) gives much better accuracy than the model MLM-G2S. This is because the mass flux distribution in DB1-G1S has still some similarity with the mass flux distribution in the original data base DB0, whereas the data base DB1-G2S doesn't include the data points with mass flux smaller than 1585 kg/m^{2}s and, thus, has totally different mass flux distribution. The accuracy of the ML-model MLM1-G1S is slightly worse than the original model MLM0. The average deviation increases from 2.6 % to 5.7 %, and the standard deviation from 19.2 % to 31.0 %. Significantly worse agreement was identified between the model MLM1-G2S and the original data base. The average deviation is as large as 24.0 % and the standard deviation is >100 %. These results clearly emphasize that an extrapolated application of a ML-model is not recommended.

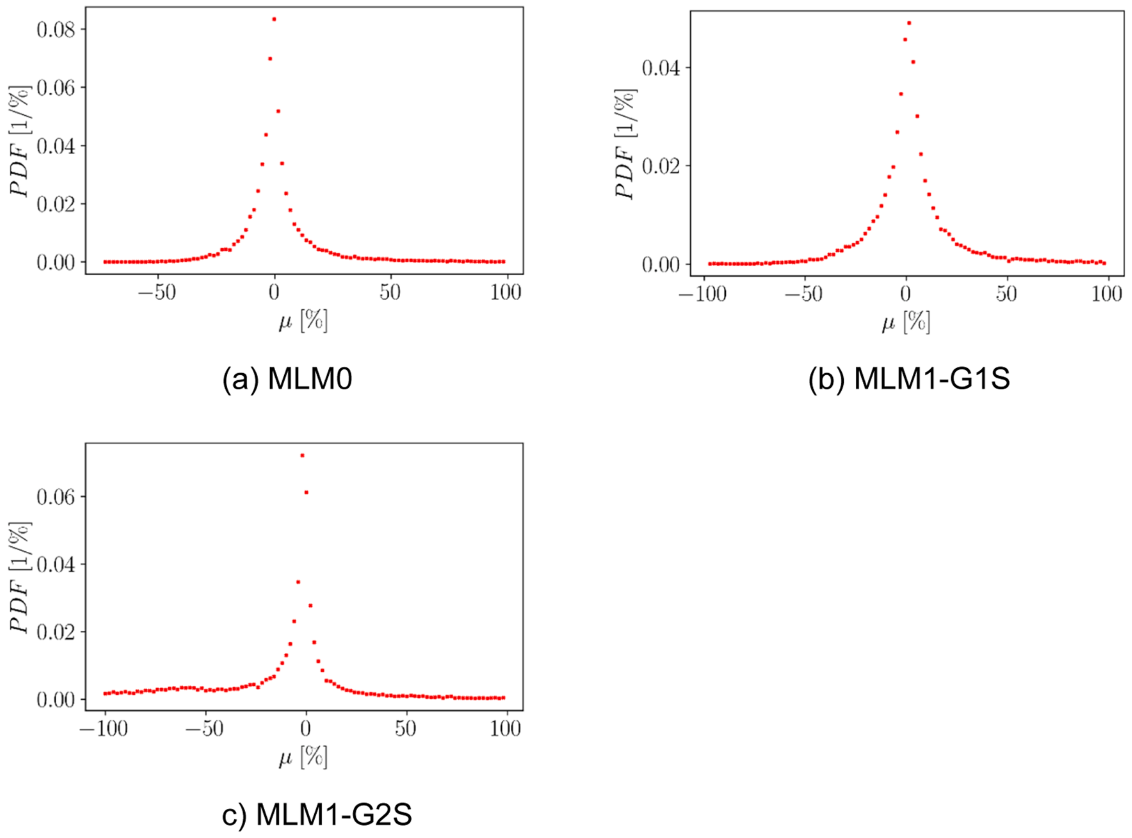
Related to the data base DB1, it is expected that the specifically trained models, i.e. MLM1-G1S and MLM1-G2S, would have higher accuracy compared to the results related to the original based on DB0. This is especially true for the model MLM-G2S, which shows excellent agreement with the data base DB1-G2S. The average deviation is <1.0 % and the standard deviation is 10.2 %.

Comparison with the data base DB2, the accuracy of the model MLM1-G1S is slightly worse than the comparison results with the data base DB1, whereas the model MLM1-G2S gives even larger scattering of the deviation comparison results using DB0. The standard deviation is large than 150 % and the average deviation is close to 50 %. These results emphasize again the unsuitability of extrapolated application of ML-models.

Fig. 3 shows the error distributions of all three ML-models compared to the original data base. In the three figures, data points with deviation larger than 100 % are excluded due to the presentation technique. Related to the three ML models, i.e. MLM0, MLM1-G1S and MLM1-G2S, there are 142, 415 and 4175 data points, respectively, which have deviation larger than 100 %. The larger number of data points with larger deviation is caused by the extrapolated application of the MLM-G1S. In general, it can be recognized that the peak amplitude of PDF near $\mu =$ 0% is highest in case of MLM0. Although the peak amplitude with respect to MLM1-G2S is higher, the width of the peak is smaller. The area beneath the peak curve, which represents the probability of small errors, is then smaller than that with the model MLM1-G1S. Combined with the fact, that much more data points in the artificial data base are out of the valid parameter range of the data base DB1, the standard deviation of the error is much larger by using MLM1-G2S compared to the case with MLM1-G1S.

Step 3. ***Derivation of the distribution functions***

The most challenging task in the DI-CML approach is the derivation of the distribution functions for the input variables. As already mentioned before, there is no ready-to-use method. The researchers



(a) MLM0

(b) MLM1-G1S

c) MLM1-G2S

Fig. 3. Error distribution of THE ML-models compared with the original data base DB0.

should find out the most suitable way for their specific engineering problems. It requires a thorough analysis of the real distributions of the input variables. For the feasibility study purpose, very simple methods are used to characterize the mass flux distribution in this study.

Fig. 4 shows the distribution of the mass flux of both data bases (black pillars), i.e. DB1-G1S and DB1-G2S. Obviously, it is difficult or even impossible, to accurately describe the mass flux distributions with simple functions. As mentioned before, the main purpose of this paper is to assess the feasibility of the DI-CML approach. A more accurate or more reasonable description of the input variable distributions keeps a challenging task for the future continuous studies.

In this paper, two different approaches for the description of the mass flux distribution are considered, to study the effect of the selection of the distribution functions on the accuracy of the DI-CML approach. In the *first approach*, a simple mathematic function, here the Γ-function is taken

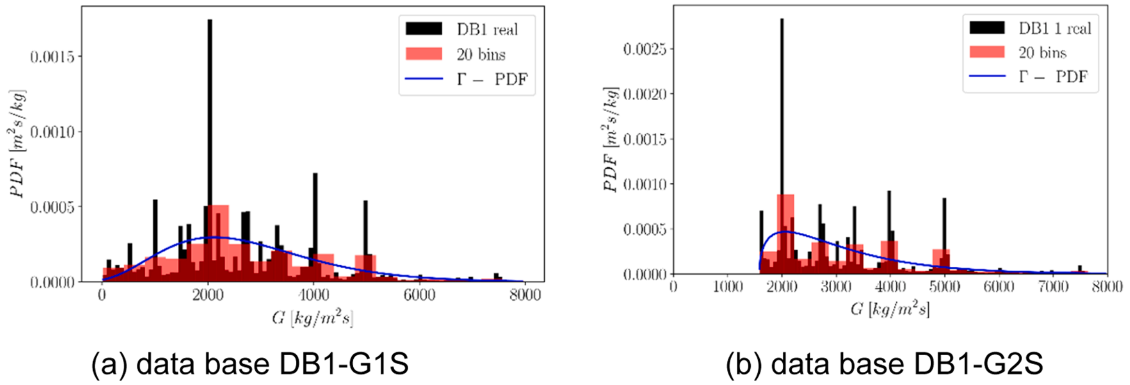
$$PDF(x;\ \alpha,\ \beta,\ \gamma) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}(x-\mu)^{\alpha-1}e^{-(x-\gamma)/\beta} \tag{10}$$

with:

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}e^{-t}dt \tag{11}$$

The three coefficients ($\alpha = shape,\ \beta = scale,\ \gamma = location$) will be optimized, so that the mean error square (MES) has the minimum value. In this way, the following two optimized Γ-functions are derived for the mass flux distribution of both data bases DB-1-G1S and DB1-G2S, i.e.

$$PDF(G) = \frac{1}{\Gamma(5.61)635.5^{5.61}}(G+871.1)^{5.61-1}e^{-(G+871.1)/635.5} \tag{12}$$



(a) data base DB1-G1S

(b) data base DB1-G2S

Fig. 4. Comparison of the mass flux distribution of both data bases.

for the data base DB1-G1S and

$$PDF(G) = \frac{1}{\Gamma(1.18)1330.6^{1.18}}(G - 1584.6)^{1.18-1}e^{-(G-1584.6)/1330.6} \tag{13}$$

for the data base DB1-G2S.

The distributions of both optimized Γ-functions are also presented in Fig. 4 (blue curves). It is clearly seen that the discrete peaks cannot be well captured with the proposed Γ-functions. Thus, future studies are highly required to improve the similarity between the Γ-functions and the real mass flux distribution.

In the *second approach*, the entire mass flux range in DB1 is divided into 20 bins with the same interval

$$\Delta G = \frac{G_{max} - G_{min}}{20} \tag{14}$$

For each bin, the probability density is obtained according to the number of the data points located in the corresponding interval $n_{\Delta G}$:

$$P(G) = \frac{n_{\Delta G}}{N}\frac{1}{\Delta G} \tag{15}$$

where N is the total number of data points, i.e. 12,289. The mass flux distributions according to the 20-bins approach are also presented in Fig. 4 (red pillars) for both data bases DB1-G1S (Fig. 4a) and DB1-G2S (Fig. 4b). It is seen that the mass flux distribution with the 20-bins approach has high similarity with the mass flux distribution of the data bases DB1-G1S and DB1-G2S.

The next task should be the derivation of the distribution function of errors. Assessment of all ML-models based on the corresponding data bases reveals that systematic errors are negligible small, i.e. the error distribution fluctuates around the zero value and is hardly dependent on the values of various input variables. Fig. 5 shows one example of the error distribution versus mass flux for the case *G1S* and indicates, that no clearly systematic errors with respect to the selected input variable (mass flux) occurs. Similar conclusion can be achieved from Table 2 and Fig. 3, where the average errors of ML0 with respect to DB0 and of MLM1 compared to DB1 are very close to zero. In this case, any correction of the predicted output parameter with the error distribution is neither reasonable nor necessary. Thus, in the present study, the correction through the prediction error is neglected. However, it has to be kept in mind that the correction of the prediction values using the error distribution would become important, in case the ML-model shows obvious systematic error distribution, and thus, remains a topic for future studies.

With the steps 2 and 3, the previous ML-package contains the previous ML-model, i.e. MLM1-G1S or MLM1-G2S, the distribution functions of the input variables and the distribution function of error and is now available for the present researcher.

Step 4. ***Generation of the artificial data bases and the new ML-packages***

For the generation of the artificial data base, mass flux is generated according to the proposed distributions, i.e. Γ-functions or 20-bins. Combining the two different sampling options (G1S and G2S) with two different methods for obtaining the mass flux distributions (Γ-functions or 20 bins), totally 4 *artificial data bases* are generated with 12, 289 data points each and identified with the case-ID, G1S-Γ, G2S-Γ, G1S-b and G2S-b, respectively. Here stands "Γ" for Γ-function and "*b*" for bins-approach. It was found that some artificial data points show non-physical meaning, e.g. the generated CHF values are negative. This was possible, because the combination of the input variables, which were generated using the distribution functions, is outside the valid range of the input variable combination used for the training of the previous models MLM1. To avoid such non-physical data points, the minimum value of CHF (25 kW/m^2) was taken. All generated data points with CHF smaller than 25 kW/m^2 will be removed. The value 25 kW/m^2 is selected, which is half of the minimum CHF value (50 kW/m^2) in the original data base provided by OECD/NEA benchmark group.

Combining the artificial data bases with the corresponding data bases DB2 (new data bases), four *combined data bases*, as defined in Fig. 1, are produced with 24, 578 data points each.

With these 4 combined data bases, 4 new ML-models are trained and characterized as MLM2- G1S-Γ, MLM2-G2S-Γ, MLM2-G1S-b and MLM2-G2S-b, respectively. Table 3 summarized the comparison of the four new ML-models with the corresponding combined data bases.

All four new ML-models give similar prediction accuracy, compared with the results of MLM0 on the original data base DB0. This indicates the high feasibility of the DI-CML approach.

With the four combined data bases, four Γ-functions of mass flux are derived accordingly. Herewith, the new ML-packages with the new ML-models and the new distribution functions of input variables are generated and available for the next researchers.

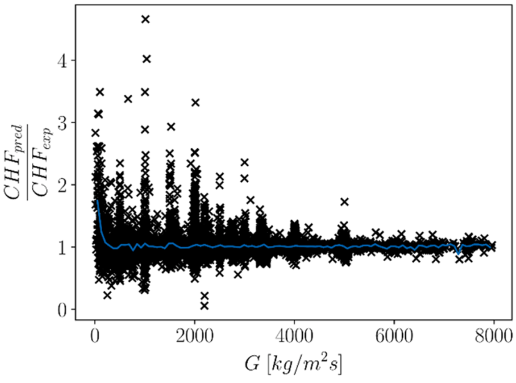


Fig. 5. Error distribution versus mass flux for the case *G1S*.

Step 5. ***Assessment of the new ML-models***

The overall goal of the DI-CML approach is to achieve better results using the new models MLM2 than that using the previous models MLM1 with respect to the total original data base DB0 as well as to the new data bases DB2. Tables 4 summarizes the comparison results with the original data base DB0 as well as with the new data base DB2.

At the first glance, the results are reasonable. With respect to the average error, all models give similar results, and the average deviation is about a few percent, except the model MLM1-G2S, as already explained in Table 2. According to the standard deviation, the original model MLM0 gives the best agreement. This is also expectable. The new models show slightly lower accuracy than the original model, but much improved prediction than the previous models, especially related to the 2nd sampling option, where the distribution of mass flux in the previous data base DB1 and the new data base DB2 is completely different.

The comparison with the new data base DB2 shows a more accurate prediction is achieved using the new models MLM2 compared to the previous models MLM1, especially in case of the 2nd sampling option. Together with the previous section, these results indicate the improvement of the prediction accuracy with respect to the new data base as well as to the entire (original) data base, if the new researchers apply the DI-CML approach, instead of simply use the ML-models provided by the previous researchers. The accuracy improvement becomes more significant, in case the distributions of the input variables of the new data base

Table 3
Summary of the new ML-models with the corresponding combined data bases.

MLM2- G1S-Γ		MLM2- G1S-*b*		MLM2-G2S-Γ		MLM2-G2S-*b*	
μ	*σ*	*μ*	*σ*	*μ*	*σ*	*μ*	*σ*
2.8	23.61	2.34	26.59	3.48	23.79	2.72	23.75

Table 4
Comparison of the new ML-models with the original data base DB0 and the new data base DB2.

Models→	MLM0	MLM1-G1S	MLM2-G1S-Γ	MLM2- G1S-b	MLM1-G2S	MLM2-G2S-Γ	MLM2-G2S-b
Comparison with the original data base DB0							
μ	2.56	5.66	3.19	3.12	−23.97	5.04	3.61
σ	19.25	30.98	21.79	22.84	111.84	26.32	23.38
Comparison with the new data base DB2							
μ		6.88	3.16	3.74	−47.92	6.31	4.79
σ		39.24	23.16	25.14	154.23	30.02	25.69

differ strongly from those of the previous data base.

4.2. Generation of artificial data bases with independent distributions of input variables

In the previous chapter 4.1, it is assumed that all input variables, except mass flux, have uniform distributions. Clearly, this assumption is not consistent with the real world, as shown in Fig. 2. In other side, the distributions of the input variables are dependent on each other, because during the experiments, the four input variables cannot be determined independently, e.g. local steam quality depends on other three input variables and experiments with low pressures and low mass fluxes are difficult to realize. An accurate approach to describe this coupled relationship and to derive distribution functions for the coupled input variables remains a challenging task for the future studies. In this chapter, it is assumed that the distributions of all input variables are independent of each other, to investigate how the selection of the input variable distributions affects the prediction accuracy of the DI-CML approach.

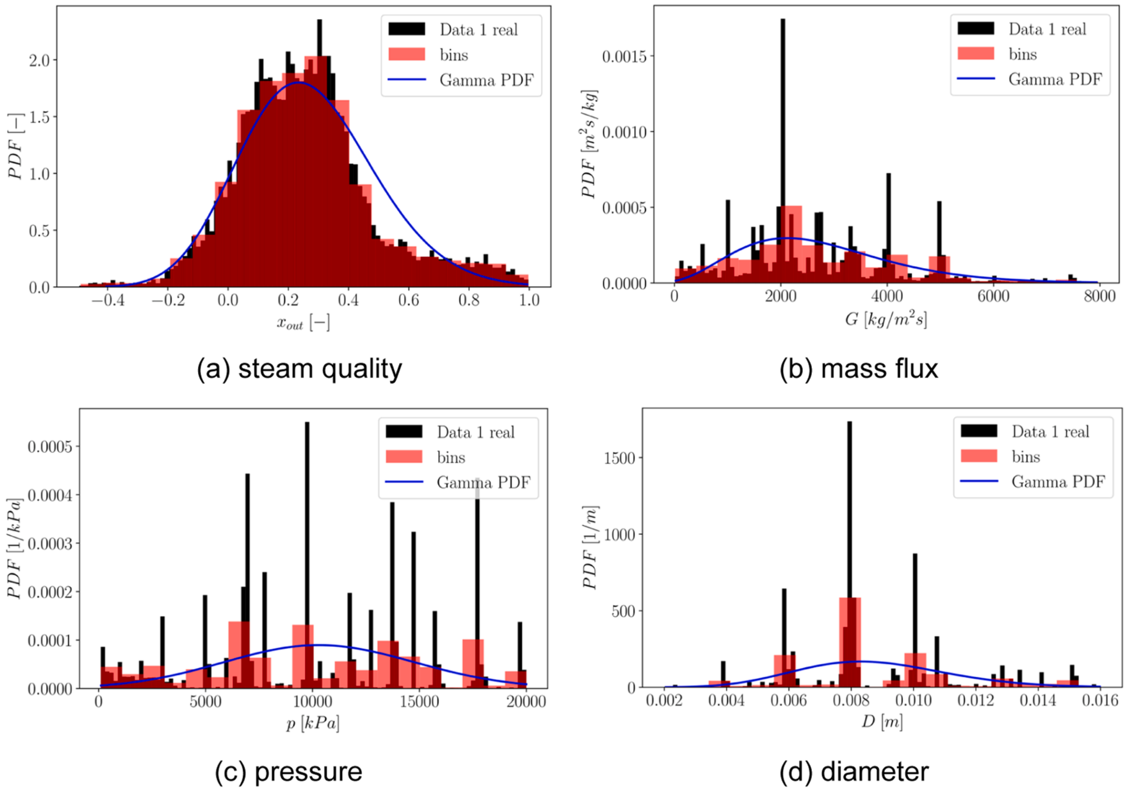
The procedure here is the same as that presented in the previous chapter 4.1. The original data base is arranged according to mass flux as in chapter 4.1. Although the distributions of the input variables are assumed independent, generation of the previous data base (DB1, as indicated in Fig. 1) and new data base (DB2, as indicated in Fig. 1) requires the selection of the reference input variable. To be consistent with the procedure in chapter 4.1, mass flux is also selected here as the reference input variable. As results, two previous data bases (BD1-G1S, DB1-G2S) and two new data bases (DB2-G1S, DB2-G2S) are generated, which are identical as produced in chapter 4.1.

Because the data bases DB1 are the same as in chapter 4.1, the trained ML-models (MLM1, indicated as previous ML-model in Fig. 1) are also the same as derived in chapter 4.1. The main difference between the present procedure from that of the previous chapter is the derivation of the distribution functions of the input variables.

(A) Derivation of the distribution functions

Here the distributions of all four input variables will be derived. Again, both two approaches, using Γ-functions and 20 bins, are considered. The approach using Γ-functions leads to eight optimized Γ-functions for both data bases DB-1-G1S and DB1-G2S and four input variables, respectively. The distributions of the input variables of the data base DB1-G1S according to the optimized Γ-functions (red curve) as well as the 20-bins approach (red pillars), together with the original distributions (black pillars) are presented in Figs. 7.

A good similarity between the real distributions and the Γ-functions is obtained with respect to steam quality. For all other three input variables with discrete features, the discrete peaks cannot be well captured with the proposed Γ-functions. Although the 20-bins approach gives also discrete peaks, the real peaks, especially those with large amplitudes, are flattened significantly. In general, the 20-bins approach shows



(a) steam quality

(b) mass flux

(c) pressure

(d) diameter

Fig. 7. Distributions of the four input variables according to the Γ-functions, 20-bins approach as well as the real original data base DB1-G1S.

improved similarity compared to the Γ-functions, indicating still significant deviation from the real distributions. Thus, future studies are required to improve the similarity between the distribution functions and the real distributions of the input variables.

(B) Generation of the artificial data bases

For the generation of the artificial data base, all four input variables are produced according to their distribution functions independently. Accordingly, four artificial data bases, i.e. DB1-G1S- Γ-*All*, DB1-G2S-Γ-*All*, DB1-G1S- *b-All* and DB1-G2S-*b-All* are created. Here "*All*" indicates that all four input variables are created independently according to their distribution functions. Combining the artificial data bases with the corresponding data bases DB2, four combined data bases are produced.

(C) Assessment of the new ML-models

With these 4 combined data bases, 4 new ML-models are trained and characterized as MLM2- G1S- Γ-*All*, MLM1-G1S- *b-All*, MLM2-G2S- Γ-*All* and MLM2-G2S- *b-All*, respectively. Table 5 compares the four new ML-models with the corresponding combined data bases. In general, good agreement is recognized of all new ML-models with respect to the corresponding combined data bases. The accuracy is comparable to the original ML-model MLM0 on the original data base DB0, where the average value and the standard deviation of the error are 2.56 % and 19.3 %, respectively. Moreover, the accuracy is comparable to the new models derived in chapter 4.1 (see Table 3).

Tables 6 summarizes the comparison of the new ML-models with the original data bases DB0 and the new data bases DB2. With respect to the original data base DB0, all new models give satisfying prediction accuracy, comparable to the original model ML0. They show clearly improvement compared to the previous ML models MLM1. Slightly large deviation (error scattering) of the model MLM2-G2S-*b-all* again is due to the impact of some data points with input parameters outside the valid range on the quality of the artificial data base.

Comparison with the new data base DB2. The results are very similar to the results shown in Table 4. It indicates effect of both selected approaches for the generation of the soft data bases, i.e. uniform distribution for the three input variables or independent distributions of all input variables, is small on the prediction accuracy.

5. Summary and outlook

The data-informed continuous learning (DI-CML) approach is proposed for CHF prediction. The main objective of the DI-CML approach is to generate a combined data base for the model development based on machine learning method. The generated combined data base contains not only the experimental data base of the present researchers, but also the artificial experimental data points of 'all' previous researchers. The artificial experimental data base would have similar features as the original data base of the previous researchers, which is however, not accessible to the present researchers. In this way, a continuous effective learning process becomes possible and enhances the applicability of the ML-approaches in engineering problems.

This paper describes briefly the procedure of the DI-CML approach. With help of the CHF data base provided by the OECE/NEA benchmark working group, the DI-CML was successfully assessed, and its feasibility was well confirmed. The following main conclusions can be achieved:

- Generally, application of a ML-model to new data base leads to a reduction in its prediction accuracy or to failure of prediction, in case the ML-model was trained with data base having completely different valid parameter ranges. Distribution outside the parameter range of their training data base may lead to significant error and is not recommended. Thus, sufficient information of the distribution of the input variables is unavoidable for continuous machine learning.
- With two different sampling options, two data sets with significantly different distributions of the input variables with respect to the previous data base and the new data base were generated. The results showed that the new models based on DI-CML approach give much better prediction than the previous models proposed by the previous researchers, although, only very simplified functions are used to represent the distribution of the input variables. This emphasizes the necessity of the continuous learning and the importance of the proposed DI-CML approach.
- Two different methods were used to generate the artificial data bases. The results show negligibly small effect of the generation methods on the prediction accuracy of the new models. However, it should be kept in mind, that the selection of the generation methods affects the quality of the artificial data base. With the method of independent distribution of the input variables, less data points have their input variable combination outside the valid parameter range, produce no-physical values of CHF and leads to higher quality of the artificial data base.

Table 5
Comparison of the new ML-models with the corresponding combined data base.

MLM2- G1S-Γ-All		MLM2-G2S-Γ-All		MLM2- G1S-*b*-All		MLM2-G2S-*b*-All	
μ	σ	μ	σ	μ	σ	μ	σ
3.93	24.51	0.47	19.53	1.35	17.07	1.65	16.64

Although the necessity and feasibility of the DI-CML approach has been successfully demonstrated, some challenging issues remain, e.g. derivation of more accurate distribution functions of the input variables. As clearly indicated, the similarity of the distribution of the input variables between the database and the distribution functions plays the key role in the accuracy of the DI-CML approach, especially the quality of the artificial database. For many engineering problems such as CHF prediction, the distribution of the input variables has discontinuous features and coupled with each other. Description of such coupled distributions constitutes an extreme challenging task. Although the selection of such distribution functions remains the task of the researchers dealing with their specific problems, establishment of guidelines should belong to the future research work.

Another key feature of the DI-CML approach is the correction of the results by means of the error distribution function, which, however, was not treated in this paper, because no obvious systematic error was observed. However, the error correction needs to be considered, in case a ML-model shows a systematic error distribution. Future studies should provide guidelines for the error correction.

The proposed DI-CML approach is a general approach and applicable also to other engineering problems, where the target parameter is dependent on a group of input variables, independent of the complexity of the problem or the number of the input variables. The key issue is to derive the distribution functions of the (independent or dependent) input variables and the errors. The conclusions achieved in this paper with respect to the feasibility of the DI-CML approach is also applicable to other engineering problems. Further efforts will be made to apply the proposed DI-CML approach to various engineering problems.

CRediT authorship contribution statement

Meiqi Song: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Fabian Wiltschko:** Writing – review & editing, Software, Investigation, Formal analysis. **Xiaojing Liu:** Writing – review & editing, Supervision, Project administration, Methodology. **Aurelian F. Badea:** Writing – review & editing, Methodology, Formal analysis. **Xu Cheng:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

Table 6
Comparison of the ML-models with the original data base DB0 and the new data base DB2.

Models →	MLM0	MLM1-G1S	MLM2-G1S-Γ-All	MLM2- G1S-*b*-All	MLM1-G2S	MLM2-G2S-Γ-All	MLM2-G2S-*b*-All
Comparison with the original data base DB0							
μ	2.56	5.66	4.37	1.68	−23.97	0.79	6.39
σ	19.25	30.98	22.99	24.3	111.84	22.55	63.15
Comparison with the new data base DB2							
μ		6.88	4.24	2.61	−47.92	1.26	3.57
σ		39.24	24.28	25.1	154.23	25.67	24.12

the work reported in this paper.

Acknowledgment

The first author would like to acknowledge the financial support of the program “International Excellent Grants” for her research stay at the Karlsruhe Institute of Technology.

Data availability

The authors do not have permission to share data.

References

[1] M. Bruder, G. Bloch, T. Sattelmayer, Critical heat flux in flow boiling - review of the current understanding and experimental approaches, Heat Transf. Eng. 38 (3) (2017) 347–360, https://doi.org/10.1080/01457632.2016.1189274.

[2] X. Cheng, U. Müller, Review on critical heat flux for water cooled reactors, Forschugszentrum Karlsruhe (2003) 2003. FZKA-6825.

[3] L. Corre, G. Delipei, X. Wu, X. Zhao, Benchmark On Artificial Intelligence and Machine Learning For Scientific Computing in Nuclear Engineering. Phase 1: Critical Heat Flux Exercise Specifications, OECD Publishing, Paris, 2024. NEA Working Papers.

[4] Gou, J.P., Yu., B.S., Maybank, S.J., Tao, A.C., 2021. Knowledge Distillation: a Survey, [cs.LG] 20 May 2021.

[5] D.C. Groeneveld, et al., The 2006 CHF look-up table, Nucl. Eng. Design 237 (2007) (2007) 1909–1922.

[6] D.D. Hall, I. Mudawar, Critical heat flux (CHF) for water flow in tubes - I. Compilation and assessment of world CHF data, Int. J. Heat. Mass Transf. 43 (2000) (2000 a) 2573–2604.

[7] D.D. Hall, I. Mudawar, Critical heat flux (CHF) for water flow in tubes - II. Subcooled CHF correlations David D. Hall, Issam Mudawar, Int. J. Heat. Mass Transf. 43 (2000) (2000 b) 2605–2640.

[8] Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the Knowledge in a Neural Network, [stat.ML] 9 Mar 2015.

[9] M.G. Kang, S.H. Kang, Data-free knowledge distillation in neural networks for regression, Expert. Syst. Appl. 175 (2021) (2021) 114813.

[10] Kirkpatricka, J., et al. 2017. Overcoming catastrophic forgetting in neural networks, [cs.LG] 25 Jan 2017.

[11] Y. Liu, W. Liu, L. Gu, J.Q. Shan, L. Zhang, Existing DNB-type CHF mechanistic models and relations with visualized experiments in forced convective flow boiling: a review, Prog. Nucl. Energy 148 (2022) (2020) 104225.

[12] S.K. Moon, W.P. Baek, S.H. Chang, Parametric trends analysis of the critical heat flux based on artificial neural networks, Nucl. Eng. Design 163 (1–2) (1996) 29–49. IssuesJune 1996.

[13] A. Tentner, et al., Development and Validation Oft Wo-Phase Flow Models and Critical Heat Flux Prediction For the Highly-Scalable CFD Code NEK-2P, Argonne National Laboratory, 2018. ANL-18/34.

[14] Wang, J.D., Lan, C.L. Liu, C., Ouyang, Y.D., Qin, T., Lu, W., Chen, Y.Q.,Zeng, W.J., Yu, P.S., 2022. Generalizing to Unseen Domains: A Survey on Domain Generalization, ,24 May 2022.

[15] Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K., 2019. Modelling Tabular Data using Conditional GAN, [cs.LG] 28 Oct 2019.

[16] Cathey T. Yapo, M.J. Embrechts, R.T. Lahey Jr., et al., Prediction of critical heat fluxes using a hybrid Kohonen-backpropagation neural networks, in: C.H. Dali, et al. (Eds.), Topics in Intelligent Engineering Systems Through Artificial Neural Network, Topics in Intelligent Engineering Systems Through Artificial Neural Network, 2, ASME Press, New York, 1992, p. 1992.

[17] Z.L. Zhao, A. Kunar, R. Birke, H. Van der Scheer, L.Y. Chen, CTAB-GAN+: enhancing tabular data synthesis, Front. Big. Data (2024), https://doi.org/10.3389/fdata.2023.1296508. January 2024.

[18] Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C.C., 2021. “Domain generalization in vision: a survey,” , 2021.

[19] W. Zhou, S. Miwa, H.Y. Wang, K. Okamoto, Assessment of the state-of-the-art AI methods for critical heat flux prediction, Int. Commun. Heat Mass Transf. 158 (2024) (2024) 107844.