Electric Power Systems Research 201 (2021) 107537

Contents lists available at ScienceDirect

Electric Power Systems Research

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

A causality based feature selection approach for data-driven dynamic security assessment

Federica Bellizio [a], Jochen L. Cremer [b], Mingyang Sun [*,c], Goran Strbac [a]

[a] Imperial College London, South Kensington, London SW7 2BU, United Kingdom
[b] TU Delft, Mekelweg 5, CD Delft 2628, the Netherlands
[c] Zhejiang University, 38 Zheda Road, Hangzhou City, Zhejiang Province, PR China

ARTICLE INFO

Keywords:
Decision trees
Feature selection
Markov blanket
Power systems operation
Dynamic security assessment

ABSTRACT

The integration of renewable energy sources increases the operational uncertainty of electric power systems and can lead to more frequent dynamic phenomena. The use of classifiers from machine learning is promising to include dynamics in the security assessment of the power system. The training of these classifiers is typically performed offline on synthetically generated operating conditions (OCs) that are similar to real-time operation. However, the uncertainty in the generated OCs and the classifier's inaccuracy is larger the longer the time between offline and real-time operation. Moving the classifier training closer to real-time operation is an important step forward to reduce inaccurate predictions and improve reliability. In this paper, a novel causality-based feature selection approach for an online dynamic security assessment (DSA) framework is proposed. The key novelty is to use the system's physics to learn the causal structure between the features and then select the features based on this causal structure. The proposed approach results in faster computations, is more robust and more interpretable. Moreover, classifiers can be trained closer to real-time operation which enhances the predictive performance. Through a case study using transient stability on the IEEE 68-bus system, the proposed method reduces computational time by 75% in comparison to state of the art feature selection techniques. The proposed workflow showed superior performance in accuracy and robustness against uncertainty compared to conventional machine learning approaches for DSA. The computational benefit was also projected to a dataset of the French transmission system where the approach has the potential to achieve computational savings of up-to two orders of magnitudes.

1. Introduction

The rapid integration of renewable energy sources is increasing the uncertainty surrounding operation of modern power systems, exposing the grid to many kinds of faults [1]. In the past, operators were able to comply N-1 security standards (operation reduced by one equipment) during all of the hours with the classical preventive actions as the power flows were easily predictable. In the near future, the intermittent nature of renewable energy and the demand-side flexibility will make the generation and demand more uncertain, affecting the predictability of the flows. Hence, an efficient operation of the system is close to its limits with smaller safety margins to increase the utilization of the existing assets, and post-fault corrective control actions are important for the reliability [2]. A power system is reliable when it can supply electricity with high enough probability to the end-users at all times (i.e. adequacy) and withstand sudden disturbances without major service interruptions in the real-time (i.e. security) [3].

1.1. Security assessment

The system security differentiates static and dynamic security [3]. The static security refers to the system subjected to a disturbance fulfilling all physical constraints in the post-fault steady-state. It can be assessed in real-time operation by modelling the energy balances for the post-fault static state, or directly for the pre-fault state for the N-1 system in a security-constrained optimal power flow problem, e.g. [4]. However, the assessment of static security does not include whether the system survives the transition from pre-fault to post-fault. This transition

* Corresponding author.
E-mail addresses: f.bellizio18@imperial.ac.uk (F. Bellizio), j.l.cremer@tudelft.nl (J.L. Cremer), mingyangsun@zju.edu.cn (M. Sun), g.strbac@imperial.ac.uk (G. Strbac).

https://doi.org/10.1016/j.epsr.2021.107537
Received 18 January 2021; Received in revised form 25 June 2021; Accepted 16 August 2021
Available online 23 September 2021

is considered in dynamic security. The associated dynamics (e.g., transients) of a contingency may be unacceptable, even if the OC is secure in the steady-state. Assessing the wide range of dynamics involved in the operation is challenging. Hence, the dynamic security is often not considered, instead large static margins are preferred. To relax these static margins and utilize the grid assets in a more efficient manner, a set of typical dynamical phenomena are studied separately, mainly relating to stability in rotor angles (transient stability), in frequency and in voltages [5]. For instance, rotor angle stability refers to the ability of synchronous machines to remain in synchronism after being subjected to a disturbance. Each of these phenomena needs to be analysed and different analytical techniques are used, e.g. transient stability is evaluated by performing an event-type simulation on a large model involving ordinary differential equations. Such a simulation requires numerical integration that is computationally intensive and a separate simulation should be done for each potential OC. Hence, considering dynamic security in real-time operations would require significant online computational resources if the aforementioned numerical approaches are used. To address the challenge of the computational complexity of this task, machine learning has been widely used as it may provide real-time security predictions with almost no computational resources [6].

1.2. The machine learning approach to DSA

Machine learning techniques have attracted attention as they may provide Transmission System Operators (TSOs) with a scalable way of managing reliability [3]. The key concept of the machine learning based approaches is to first generate OCs similar to the ones in real-time and subsequently carry out the training process of machine learning classifiers in an offline manner some days before operation. In this offline stage a vast number of potential OCs are simulated to model the residual uncertainties, e.g. possible network's configurations or contingencies, between the day the classifier is trained and the day the classifier is used. Then, these predictors can be used in real-time operation to instantly infer the post-fault security status of unseen OCs [3,7]. Due to the ability of classifiers to instantly infer the post-fault status, i.e. transient stability in this work, these approaches are promising for predicting various stability phenomena under small and large disturbances, e.g. short-term voltage stability [8]. Various classification models have been investigated for transient stability analysis, such as support vector machines in [9] and feed-forward artificial neural networks (ANN) in [8,10].

Although ANN models perform better in terms of accuracy, Decision Trees (DTs) or DT ensembles have been mostly adopted as they provide a promising trade-off between accuracy performance, computational complexity and interpretability [7,11]. In particular, interpretability is a key requirement of data-driven security assessments as it ensures the ability to understand how a classifier predicts a particular OC with little inspection allowing operators to be still involved in the control loop [12].

Although the offline database is periodically updated, the time distance between the offline and the online stage can compromise the performance of the classifier, irrespective on the type of classifier used. In fact, OCs can change very rapidly over time, resulting in current OCs which are different from those included in the initial knowledge base [1, 10]. The OCs from the online and offline stages can be described as originated from two different probability distributions. Importance sampling is generally a very useful approach to cope with these discrepancies between the training and testing distributions [13]. However, due to the high frequency of probability changes in the OCs, the importance factor cannot be estimated a priori in DSA applications. To track these changes, several efforts have been directed towards a periodic update (e.g. daily or hourly) of both the training database and the classification model. These updates represent a challenging task as they are undertaken very close to real-time operation, i.e. a few hours before operation. In this near real-time stage, more recent information regarding the residual uncertainty deriving from the increasing integration of renewables are included in the training database. The computational intensity of the updates in the near real-time stage is high as the system size requires a non-linear increase of dynamic simulations and this increase (more data) leads to a slower training process. Hence, one strategy for re-training the model is proposed in [14–16], where only the data weights or small portions in an ensemble DT-based model are updated rather than to re-train the full DT. To additionally decrease the computational time in the training of the model, the dimension of the attributes can be reduced by applying Feature Selection (FS) techniques as a pre-processing step [16,17]. The existing FS methods can be broadly classified into filter, wrapper and embedded methods [18,19]. As wrapper and embedded methods require the highest computational times across these three classes, they may not be suitable to online DSA [15]. A few filter methods have been used in the past in DSA, resulting in a moderate trade-off between computations and accuracies [20]. Recently, in machine learning research, causality-based feature selection is investigated as causal features can improve the robustness and the interpretability of the predictive model across different settings [21,22]. These approaches use independence tests to discover the causality between features from data. However, for power system DSA, the physical interconnectivity and the dependency between features are highly related [23]. Considering the system's physical knowledge and not only the data may improve the performance in terms of accuracy, computations and interpretability of causality-based FS. This is what this work investigates, causal feature selection for power system DSA.

1.3. Challenges of designing feature selection to DSA

The first challenge of feature selection for machine learning based DSA refers to their computation efficiency when highly relevant features should be selected [20]. If features are not selected effectively, the classifier results in inaccuracies. Simultaneously, if the feature selection requires a long time, it needs to be done early in the workflow, resulting in discrepancies between offline and real-time so large that the classifier becomes inaccurate. The challenge is to make efficient selection decisions and at the same time keeping the computation time of the selection process as short as possible. Thus, highly relevant features can be selected close to real-time, minimising the discrepancies between offline and real-time.

The second challenge has to do with the robustness of the selected features and the trained classifier against the residual uncertainty between generated and real-time OCs [19]. The real-time OCs will always be (slightly) different from those included in the training database, even if the database generation and classifier training are moved very close to real-time operation. This difference is mainly caused by the high operational uncertainty as the underlying probability distributions of the OCs can change over a few hours [24]. Hence, it is important to identify highly relevant features for improving the robustness against uncertain OCs. The variability of the OCs represents a key challenge in the wider machine learning approaches to reliability assessment [3].

1.4. Proposed approach

A fast approach for FS to address the two aforementioned challenges is proposed. This approach is used in combination with a three-stages learning workflow (Fig. 1) that differs from the traditional two-stages scheme for DSA by the presence of a near real-time stage in between. The proposed approach uses the system's physics and data to discover the causal structure between features. Then, the approach identifies highly relevant features by learning the Markov Blanket (MB) on the derived structure. The key novelty is to inform the Markov Blanket search with the physical interconnectivity of the power system. The proposed approach is beyond using the concept of causality for feature selection [25] and brings a key advantage over other FS approaches that are exclusively data-driven. The correlation structure between the

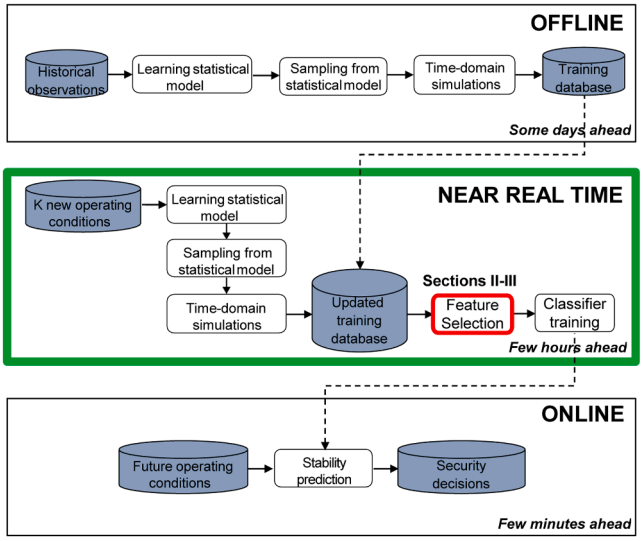

Fig. 1. Data-driven three-stages workflow for classification. In green the proposed workflow and in red the proposed FS approach.

features and dynamic security is highly related to the network topology [23]. Hence, performing the FS based on the system's physics allows identifying features most relevant to security for the given topology. The proposed FS approach results in the highest accuracies and requires less computational times. The proposed workflow has two key benefits:

1. The accuracy of the final classifier is higher as the training is shifted closer to real-time operation, lowering the impact of discrepancies between offline training and real-time operation.
2. The robustness against uncertainties is higher as causal features are selected based on physical knowledge and not purely based on data.

A case study on the IEEE 68-bus system considering transient stability is used to illustrate the performance of the proposed workflow. First, the proposed approach is compared to a two-stage workflow for DSA. Then, the robustness of the classifier is tested on several datasets with varying discrepancies between estimated and actual parameters of the probability distributions of the OCs. Subsequently, the trade-off between prediction performance and computational complexity is investigated. The scalability to a large-scale system is projected with the French Transmission system confirming the potential of the approach.

The rest of the paper is structured as follows. In Section 2, the construction of the graphical model by combining the network knowledge and the availability of a large dataset of measurements is presented in detail. Thereafter, in Section 3, the approach to identify the approximate MB is described. Subsequently, the case study is presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Probabilistic graphical model

To combine the system's physics with a causality-based FS approach, a probabilistic graphical model that captures the statistical correlations according to the physical connections of the network topology is constructed [26]. The power network is defined as a physical graph $G(V,\epsilon)$, where V and ϵ represent the buses and lines, respectively. According to this approach, the voltage measurements obtained by smart meters over time at bus i are associated to a random variable v_i. The following basis justify the focus on voltage measurements: i) These measurements have become more accessible in recent years with usage of high fidelity PMUs. ii) Nodal voltage measurements are characterized by some conditional correlation properties that make them more suitable to an efficient representation of the network topology through graphical models [27].

To describe the probabilistic relationships among the voltage measurements, a joint probability distribution between the voltage variables v_i is computed:

$$p(\mathbf{v}) = p(v_1, v_2, ..., v_n) = p(v_1)p(v_2|v_1)...p(v_n|v_1, ..., v_{n-1}) \tag{1}$$

where v_i represents the voltage measurements at bus i and n is the number of buses. Bus 0 is the reference bus with a unit magnitude and a zero phase angle voltage. If m different voltage measurements are available, calculating this joint probability distribution is computationally expensive as the computational cost would be $\mathcal{O}(m^{n-1})$. To reduce this cost, a simplified distribution $p_a(\mathbf{v})$ can be used to approximate the true distribution $p(\mathbf{v})$ if the minimum information loss is guaranteed. A tree-dependent probabilistic graphical model can be chosen as approximation of $p(\mathbf{v})$:

$$p_a(\mathbf{v}) = \prod_{i=1}^{n} p\left(v_i \middle| v_{pa(i)}\right) \tag{2}$$

where $v_{pa(i)}$ is the direct predecessor, known as parent node, of v_i. In this model, voltages are conditionally independent given their parent nodes' voltage information. This condition holds in a transmission network if the current injections are independent [26]. As the voltages generally remain within the nominal range and the loads can be assumed as being independent, also the current injections can be approximated as independent. The Kullback-Leibler (KL) divergence is used to represent the difference of information contained in $p(\mathbf{v})$ and those contained in $p_a(\mathbf{v})$ about $p(\mathbf{v})$:

$$\begin{aligned} D(p\,\|p_a) &= E_{p(\mathbf{v})}\log\frac{p(\mathbf{v})}{p_a(\mathbf{v})} = \sum p(\mathbf{v})\log\frac{p(\mathbf{v})}{p_a(\mathbf{v})} = \\ &= \sum p(\mathbf{v})\log p(\mathbf{v}) - \sum p(\mathbf{v})\sum_{i=1}^{n}\log p(v_i) \\ &\quad - \sum p(\mathbf{v})\sum_{i=1}^{n}\log\frac{p\left(v_i, v_{pa(i)}\right)}{p(v_i)p\left(v_{pa(i)}\right)} = \\ &= \sum_{i=1}^{n} H(V_i) - H(V_1, .., V_n) - \sum_{i=1}^{n} I\left(V_i; V_{pa(i)}\right) \end{aligned} \tag{3}$$

where I and H indicate the mutual information and entropy, respectively.

Minimizing the KL divergence is equivalent to minimize the information loss when $p(\mathbf{v})$ is approximated with $p_a(\mathbf{v})$. By following the Chow-Liu algorithm [28], the maximum spanning tree algorithm, which is based on mutual information, provides the optimal approximation of $p(\mathbf{v})$ in terms of this minimization. More precisely, the maximum spanning tree is constructed by selecting branches of successively higher values of mutual information and rejecting all branches that involve loops. Then, the undirected graph is transformed to a directed graph by choosing a root variable and setting the direction of all edges to be outgoing from it (the choice of root variable does not change the log-likelihood of the network). The resulting model is a Directed Acyclic Graph (DAG), known as Bayesian Network (BN) [29], where the correlation structure indirectly describes the grid topology. The variables considered are the voltage magnitudes of the buses. Hence, it is assumed that a state-estimation was performed beforehand [30].

2.1. Tree augmented naïve bayes structure

The Chow-Liu algorithm can be extended to learn the maximum likelihood Tree-Augmented Naïve Bayes (TAN) structure instead of the BN. In this model, each feature has as parents the classification target C and at most one other feature. In this study, C corresponds to the post-fault security state and the features represent the bus voltage measurements. The maximum spanning tree is constructed by comparing the

conditional mutual information between each v_i and its parent given C [31].

2.2. Causal dependence in loopy structures

The described Chow-Liu algorithm for learning the TAN model shows low performance when applied to power networks with highly mesh topologies, e.g. transmission network, as it neglects the conditional dependencies of loops. However, taking into account possible loops implies losing the causal dependence between features. The causal dependence between features is crucial to the proposed MB-based approach as the MB can be identified only in a causal model. Two conditions are necessary to learn causal models from data [29]:

- *Markov Condition*: Each attribute $v_i \in \mathbf{v}$ is conditionally independent of its non-effect attributes given its direct cause attributes.
- *Faithfulness*: $\exists$ DAG G that is a perfect map for the probability distribution P of $\mathbf{v}$.

The Markov condition is guaranteed by the independence of current injections. On the other hand, considering loops implies losing the faithfulness assumption as the model would contain cycles. This issue can be solved by introducing an auxiliary variable with fixed value for each loop [29]. An example is provided in Fig. 2, where an auxiliary variable E is introduced between variables A and C in order to describe the conditional dependence between them respecting the causality of the model. Since E is clamped to a fixed value, then it results:

$$p(E) = 1 \Rightarrow p(A, C, E) = p(A, C) \tag{4}$$

Thus, the introduction of E does not change the probabilistic relationship between A and C, and hence the faithfulness assumption is still satisfied.

Similarly, for a general loop l, the probability distribution p_a is:

$$p_a(\mathbf{v}) = p\left(e_l \middle| v_{pa(e_l),1}, v_{pa(e_l),2}\right) \prod_{i=1}^{n} p\left(v_i \middle| v_{pa(i)}\right) \tag{5}$$

where e_l is the auxiliary variable, which is associated to the considered loop, and $\{pa(e_l), 1\}$, $\{pa(e_l), 2\}$ represent the two parent nodes which are linked through l. Following Eq. (4), it results:

$$p(e_l) = 1 \Rightarrow p\left(v_{pa(e_l),1}, v_{pa(e_l),2}, e_l\right) = p\left(v_{pa(e_l),1}, v_{pa(e_l),2}\right) \tag{6}$$

Then, it results:

$$\begin{aligned} D(p \parallel p_a) = & \sum_{i=1}^{n} H(V_i) - H(V_1, .., V_n) - \sum_{i=1}^{n} I\left(V_i; V_{pa(i)}\right) \\ & - \sum p(\mathbf{v}) \log \frac{p\left(e_l, v_{pa(e_l),1}, v_{pa(e_l),2}\right)}{p\left(v_{pa(e_l),1}\right) p\left(v_{pa(e_l),2}\right)} \end{aligned} \tag{7}$$

By adding $p(e_l)$ inside the denominator:

$$\begin{aligned} D(p \parallel p_a) = & \sum_{i=1}^{n} H(V_i) - H(V_1, .., V_n) - \sum_{i=1}^{n} I\left(V_i; V_{pa(i)}\right) \\ & - \sum p(\mathbf{v}) \log \frac{p\left(e_l, v_{pa(e_l),1}, v_{pa(e_l),2}\right)}{p(e_l) p\left(v_{pa(e_l),1}, v_{pa(e_l),2}\right)} - \sum p(\mathbf{v}) \log p(e_l) \end{aligned} \tag{8}$$

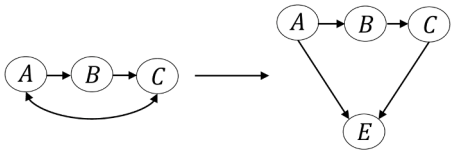

Fig. 2. The auxiliary variable E for describing the loop between A and C respecting the causal dependence.

Since $p(e_l) = 1$, then the following equality holds:

$$\begin{aligned} & D(p \parallel p_a) \\ & = \sum_{i=1}^{n} H(V_i) - H(V_1, .., V_n) - \sum_{i=1}^{n} I\left(V_i; V_{pa(i)}\right) - I\left(e_l; v_{pa(e_l),1}, v_{pa(e_l),2}\right) \end{aligned} \tag{9}$$

The first two terms are both independent of the dependence tree, whereas the last two terms represent the branch weights. The same proof can be done for all loops. As these last terms are both non-negative, minimizing the divergence measure is equivalent to maximize the total branch weight for both directed edges and loops.

2.3. Discretization

The TAN model typically works well on discrete data. In order to deal with continuous variables, the variables are generally discretized and the network is learned on the discretized domain. However, inappropriate discretization intervals may cause strong performance degradation because it may happen that the selected discretization step is able to capture only rough characteristics of the real distribution. The Clustering of $\sqrt{N}$-Interval Discretization (CloNI) is adopted as discretization method because it attempts to minimize the number of intervals while maintaining high accuracy [32]. According to this discretization method, the discretization step is different in the offline and near real-time stage.

3. Approximate Markov Blanket

The TAN model, which includes directed edges and loops, plays a fundamental role in the FS process because of its potential application as a MB technique [25]. The MB of a feature v_i provides a complete picture of the local causal structure around v_i. A number of MB-based FS approaches have been proposed [33]. However, they are slow and inefficient in terms of information found as they are based only on independence tests [34]. A different approach focuses on identifying the MB by first performing a BN learning step [35]. Under the faithfulness assumption, the MB of a variable v_i in a BN is unique and consists of parents, children and spouses of v_i [29].

By following a similar approach, the constructed TAN model is used to identify the MB of the classification class C. Since all of the features are target's children in the TAN model, the MB of C includes all features. In order to select the most relevant features for classification by taking advantage of the causal dependence structure of the TAN model, an approximation of the MB is derived [36].

Definition 1. *For two features v_i and v_j $(i \neq j)$, v_j is an Approximate Markov Blanket (AMB) of v_i if $SU_{j,C} \geq SU_{i,C}$ and $SU_{i,j} \geq SU_{i,C}$, where the symmetrical uncertainty SU measures the correlation between features and between feature and class C.*

Generally, the AMB-based feature selection algorithms discard features which are included in the MB of another feature as redundant to it, hence irrelevant to classification. Since the causal dependencies between features are already known in the derived TAN model, the MB(C) can be directly approximated by performing pairwise comparisons between each parent and children nodes. The mutual information is used as correlation measure. Algorithm 1 shows in detail how the identification of the AMB of C is performed. More specifically, ε is the set of directed edges, for which v_i is the parent of v_j. In contrast, $\widetilde{\varepsilon}$ is the set of loops. AMB(C) and SP(C) are the approximate MB and spouse set of C, respectively. The pairwise comparisons over the directed edges are performed and all features with an higher relevance to C are included in the AMB(C). Their parents are added to the spouse set. From this initial AMB set, some features are deleted by comparing the correlation to C over the loops. Finally, all features $v_{pa(i)}$ whose children v_i have been

Initialization: $\mathrm{AMB}(C) = \{\emptyset\}$, $\mathrm{SP}(C) = \{\emptyset\}$
Output: $\mathrm{AMB}(C)$

for $(i, j) \in \varepsilon$ **do** ▷ *Pairwise comparisons over directed edges*
  if $I(v_j; C) \geq I(v_i; C)$ & $I(v_j; v_i|C) \geq I(v_i; C)$ **then**
    $\mathrm{AMB}(C) = \mathrm{AMB}(C) \cup \{v_j\}$
    $\mathrm{SP}(C) = \mathrm{SP}(C) \cup \{v_i\}$
  else if $I(v_i; C) \geq I(v_j; C)$ & $I(v_j; v_i|C) \geq I(v_j; C)$ **then**
    $\mathrm{AMB}(C) = \mathrm{AMB}(C) \cup \{v_i\}$
  else
    $\mathrm{AMB}(C) = \mathrm{AMB}(C) \cup \{v_i, v_j\}$
  end if
end for
for $(i, j) \in \tilde{\varepsilon}$ **do** ▷ *Pairwise comparisons over loops*
  if $\{v_i, v_j\} \notin \mathrm{SP}(C)$ **then**
    if $I(v_j; C) \geq I(v_i; C)$ **then**
      $\mathrm{AMB}(C) = \mathrm{AMB}(C) - \{v_i\}$
    else if $I(v_i; C) \geq I(v_j; C)$ **then**
      $\mathrm{AMB}(C) = \mathrm{AMB}(C) - \{v_j\}$
    end if
  end if
end for
for $v_{pa(i)} \in \mathrm{SP}(C)$ **do** ▷ *Deletion from spouses*
  if $v_i \notin \mathrm{AMB}(C)$ **then**
    $\mathrm{SP}(C) = \mathrm{SP}(C) - \{v_{pa(i)}\}$
  end if
end for
$\mathrm{AMB}(C) = \mathrm{AMB}(C) \cup \mathrm{SP}(C)$ ▷ *Final* AMB *evaluation*

Algorithm 1. AMB TAN based feature selection.

removed from AMB(C) in the previous step are deleted from SP(C). The final AMB(C) is given by the union of AMB(C) and SP(C).

3.1. Sampling weights

The AMB TAN approach is applied to the updated database in the near real-time stage, as shown in Fig. 1. Due to time restrictions, only few OCs from the near real-time probability distribution are included in the knowledge base. To increase their impact on the FS approach, an importance estimation method is adopted [13]. The sampling weights are calculated through a logistic regression classifier. A selector variable $\delta = 0$ and $\delta = 1$ is assigned to samples from the offline and the near real-time probability distribution, respectively. By applying the Bayes theorem, the weights w can be expressed in terms of δ:

$$w(\mathbf{v}) = \frac{p(\mathbf{v}|\delta=1)}{p(\mathbf{v}|\delta=0)} = \frac{p(\delta=0)}{p(\delta=1)}\frac{p(\delta=1|\mathbf{v})}{p(\delta=0|\mathbf{v})} \tag{10}$$

where the first term is the ratio between the number of samples from the two distributions and the second term is calculated by training a logistic regression classifier with δ as class variable. Then, a weighted form of the mutual information is used to measure the correlation between the features and the class variable [37]. In Algorithm 1, $I(v_i, C)$ is replaced with $wI(v_i, C),\ i = 1...n$, where n is the number of features and $dim(w) = N$, which is the total number of OCs in the updated database.

3.2. Computational complexity

The computational complexity of the proposed FS method is composed of: i) computational time of the CloNI algorithm; ii) computational time of the TAN learning algorithm. Focusing on the discretization method, given n and N the number of variables and samples, the computational time is:

$$\mathscr{O}\left(n\cdot\left(N\log N+\sqrt{N}\right)\right) \approx \mathscr{O}(n\cdot N\log N) \tag{11}$$

By parallelizing the process over n variables, the computational cost is reduced to $\mathscr{O}(N\log N)$. On the other hand, the naïve bayesian model learning generally requires $\mathscr{O}(n^2)$ mutual information tests. It is necessary to check if the number of tests increases when the mutual information tests are replaced by conditional mutual information tests for the TAN learning algorithm. In order to evaluate the conditional mutual information, the training data is partitioned by class values [38]. Then, the mutual information conditioned to each class value is evaluated. Given x, y discrete random variables and z binary variable, it results:

$$\begin{aligned} I(x;y|z) &= \sum_{z\in Z} p(z) \sum_{y\in Y}\sum_{x\in X} \cdot\log\left(\frac{p(x,y|z)}{p(x|z)p(y|z)}\right) \\ &= \frac{k_{z=0}}{N}\cdot I(x;y|z=0) + \frac{k_{z=1}}{N}\cdot I(x;y|z=1) \end{aligned} \tag{12}$$

where $I(x;y|z)$ represents the mutual information between x and y conditioning on z, k_z is the number of samples for which z assumes each class value and N is the total number of samples. By defining D_i the partition for which $Z = z_i$, noticing that the mutual information is symmetrical in D_i and $\sum_i |D_i| = N$, the computational complexity for (12) is:

$$\mathscr{O}\left(\sum_i |D_i|\left(\frac{n}{2}\cdot(n-1)\right)\right) \approx \mathscr{O}\left(N\left(\frac{n}{2}\cdot(n-1)\right)\right) \approx \mathscr{O}\left(\sqrt{N}\cdot\frac{n}{2}\right) \tag{13}$$

where the last equivalence is obtained by parallelizing over n variables and by applying the CloNi algorithm over N samples.

The proposed algorithm has a time complexity which scales linearly in the number of variables. Moreover, the way by which it scales in the number of instances depends only on the chosen discretization algorithm.

4. Case study

Several studies were undertaken to demonstrate the benefits of the AMB TAN FS approach for power system DSA. First the proposed three-stage approach using the AMB TAN FS is compared to the traditional two-stages approach for DSA. Then, the trade-off of accuracy and computations was explored in detail for the proposed FS method and compared to existing techniques. Finally, the computational savings of using the proposed workflow were investigated.

4.1. Test system and assumptions

The IEEE 68-bus system (Fig. 3) was first used for testing the performance of the proposed workflow [39], and later the computations were scaled to a large-scale system (French Transmission System corresponding to 1,955 transmission lines, 1,886 buses and 411 generators). A set of 12000 OCs was generated and each of them represented a pre-fault condition of the system considering the full AC model. These OCs were generated by drawing the active loads from a multivariate Gaussian distribution considering a Pearson's correlation coefficient $c = 0.75$ between all load pairs. Subsequently, by using the method of inverse transformation, the active loads were converted to a marginal Kumaraswamy distribution with the probability density function

$$f(x) = abx^{a-1}(1-x^a)^{b-1} \tag{14}$$

where $a = 1.6$, $b = 2.8$ are shape parameters and $x \in [0,1]$. The benefit of using the Kumaraswamy distribution for modelling the stochastic nature of the loads is that it is highly flexible to adapt to the skewness of the load distributions by appropriately modifying the shape parameters [40]. Finally, the active loads were scaled to be within ±50% of the nominal values. The reactive powers follow the active powers proportionally as constant impedances were assumed. Then, i.i.d power factors were sampled in the range of [0.95,1] for the generators. As the resulting OCs may be infeasible, the full AC model was considered in a mathematical optimization problem to minimize the absolute difference between these power factors. In that way, active and reactive powers corresponding to feasible OCs were obtained for the generators. The optimization problem was implemented in Python 3.5.2 and Pyomo package and solved with IPOPT 3.12.4 [41]. The transients of three-phase faults over 22 different lines were simulated ($k = 1,\cdots 22$). If during 10s simulation time all the differences between each two phase angles of the generators were less than 180°, then the OC i was considered stable $Y_{i,k} = 1$, otherwise unstable $Y_{i,k} = 0$. A fault clearance time of 22 was used. In the resulting 0.1s datasets, the percentage of unstable

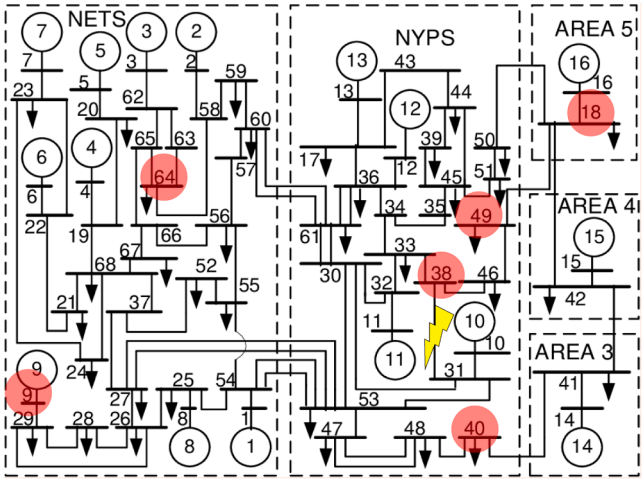

Fig. 3. The IEEE 68-bus system [39]. In red the features selected through the AMB TAN FS approach for the contingency on line [11].

observations was between 1% and 91% with 46% as mean. The simulation was performed in Matlab R2016b Simulink.

In the machine learning part, voltage magnitudes were considered as the relevant features according to Section 2. Consequently, each OC X_i was composed of all 68 buses voltage magnitudes. The data were pre-processed by applying existing FS techniques and the AMB TAN causality-based approach. The Minimum Redundancy Maximum Relevance (MRMR), Correlation-based Feature Selection (CFS) and Joint Mutual Information (JMI) as filter techniques, SVM Recursive Feature Elimination (RFE) as embedded method, and Sequential Forward Selection (SFS) as wrapper method, were used. Then, the CART learning from the *scikit-learn* algorithm was used to train DTs. In Table 1, the mean accuracy performance across all contingencies using DTs was compared against more advanced classification models to show that selecting DTs as models did not impact on the final accuracies. DT with depth equal to 3, SVM with linear kernel [9], AdaBoost and XGBoost with 50 estimators [42,43], and single layer feed-forward ANN with 10 neurons [8], were used. It resulted that all testing accuracies of these approaches were very similar, therefore DTs with maximum depth $D = 3$ were preferred as they are more interpretable. Across the studies, 10 different combinations of training/testing set were computed for each classifier. The training and testing split was 70%/30% in all studies and these two sets were drawn from the same probability distribution, unless indicated otherwise. One DT was learned for each of the 22 datasets. The DT learned for the contingency on line 31-38 using the AMB TAN causality-based approach as pre-processing step is shown in Fig. 4. As part of this study, the size of the training database was varied in the offline and near real-time stage. The F_1 score was used as criterion for the accuracy.

4.2. Inaccuracy of the security rules

In this study, the advantages of the proposed three-stages workflow in terms of the accuracy performance on new OCs were investigated. The offline stage was considered to be several weeks before real-time operation. In the offline stage, it was assumed the operators knew the load distribution parameters to be $a_0 = 1.6, b_0 = 2.8$ and $c_0 = 0.75$. To simulate the changing distribution of OCs, the parameter b was assumed to decrease over time, as shown in Fig. 5. It was assumed operators were aware of this trajectory at the online time. Two training approaches were compared. In the first one the classifier was trained only in the offline stage, whereas in the second one it was re-trained in the near real-time stage. At this stage, a few OCs can be generated and simulated from an updated load distribution that corresponds to the trajectory in Fig. 5. In the two approaches, 12000 OCs were used to perform FS and classifier training. However, in the traditional approach, these 12000 OCs were generated offline from a probability distribution different from the actual one. Then, SFS was used before training the classifier. In the proposed approach, the bulk of data came from the offline stage and only 150 OCs from the near real-time distribution were included in the training database due to limited computational resources. Subsequently, the weight of each OC in the updated database was calculated, AMB TAN approach was used as FS and then the classifier was trained. Subscripts $0, n, t$ were used to indicate the offline, the near real-time and the online stage, respectively.

Table 1
Accuracy performance using different classification models.

	Classification model				
	DT	SVM	AdaBoost	XGBoost	ANN
Accuracy Mean	0.89	0.90	0.89	0.92	0.90

[1] 31–38

Then, the two approaches were tested on 50 new observations for which $a_t = 1.6, b_t = 2.4, c_t = 0.75$ according to Fig. 5. In the proposed approach, the average accuracy across all contingencies improved by roughly 1%. The same study was then repeated for five different linear trajectories in the parameter changes. The results are summarized in Table 2 and demonstrate the improvements in the accuracy across all tested trajectories. The individual contributions of the AMB TAN FS and the classifier training to this improvement were also investigated. By leaving the AMB TAN FS out, the improvement dropped by 0.25%. Finally, the features selected through the proposed approach across all contingencies were analysed to demonstrate the increased interpretability of considering the causal relationships and the physical knowledge in the learning process. The most features included in the MB of each contingency were located very close to the faulted bus, as shown in Fig. 3. At the same time, similar MBs were identified for contingencies with close faulted buses.

4.3. Residual uncertainty

Here, the residual uncertainty in the load distributions that leads to discrepancies between the estimate and the actual probability distribution was investigated. It was assumed the parameter b decreased until one day before the operation. In the near real-time stage, three possible trajectories of b were considered, as shown in Fig. 6. As in the previous study, 150 OCs were generated from the probability distribution of the near real-time stage and used to enrich the offline database. Then, all steps from weighting, applying AMB TAN FS and training the classifier, were performed. Subsequently, the proposed approach and the offline training based approach were tested on three different test sets of 50 OCs corresponding to trajectories shown in Fig. 6 (red dotted lines). The parameters of the three test sets were calculated as follows:

$$b_{t1} = b_n - \Delta\varepsilon,\ b_{t2} = b_n,\ b_{t3} = b_n + \Delta\varepsilon \tag{15}$$

with $\Delta\varepsilon = 0.2$. The result of this test is that the use of the AMB TAN FS in the near real-time stage increased the accuracy mean value over all contingencies by 0.7% for b_{t1}, by 0.9% for b_{t2} and decreased by 0.8% for b_{t3}. Subsequently, the same study was undertaken for five different trajectories as per Table 2. When the change in the load probability distribution in the online stage progresses with the same trend of offline and near real-time stage as for b_{t1} and b_{t2}, the accuracy mean value improved by 0.7% and 0.72%, respectively. If the change in the load probability distribution progresses with a different trend in near real-time and online stage as for b_{t3}, the OCs included in the training database were not informative, and hence the accuracy mean value decreased by 0.5%. However, the decrease in the predictive performance for b_{t3} was lower in mean than the improvements for b_{t1} and b_{t2}.

4.4. Trade-off between accuracy and computational time

The following study focused on the trade-off between accuracy and computational time of the FS approaches for machine learning based DSA. Various FS approaches were used to select the features on which train the DT afterwards by using $|\underline{\Omega}| = 12000$ OCs. Subsequently, the training was repeated 10 times with varying splits of training/testing data to compute an average value for each of the 22 contingencies. In total 220 DTs were trained for each of the 6 FS approaches. The mean values of each contingency for computational time and accuracy are presented in Fig. 7 and summarized in Table 3. The confidence interval (CI) for the accuracies was evaluated considering a confidence level of 95%. Hence, the resulting accuracies do not vary much from the values indicated in Table 3, making then the comparison between the FS approaches in terms of the accuracies more faithful. Three main advantages of the AMB TAN approach can be observed: (1) AMB TAN resulted

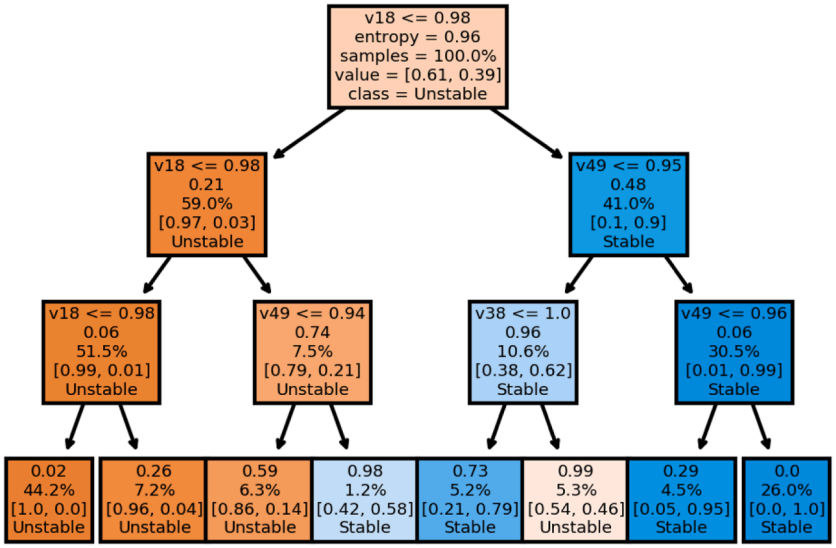

Fig. 4. The DT learned for the contingency on line [21]. The entropy was used as uncertainty measure. The voltages v_{18}, v_{38}, v_{49} were selected as features.

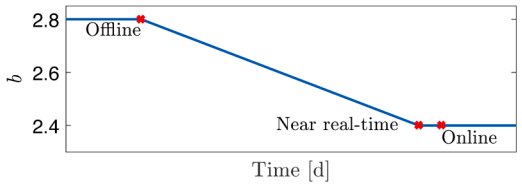

Fig. 5. Trajectory of the parameter b by assuming a constant value between near real-time and online stage.

Table 2
Accuracy improvement by re-training.

Training stage		Testing stage	Accuracy
Offline	Near real-time	Online	improvement
$b_0 = 2.8$	$b_n = 2.2$	$b_t = 2.2$	+ 0.14%
$a_0 = 1.6$	$a_n = 1.2$	$a_t = 1.2$	+ 0.22%
$a_0 = 1.6$	$a_n = 1.4$	$a_t = 1.4$	+ 1.32%
$c_0 = 0.75$	$c_n = 0.25$	$c_t = 0.25$	+ 1.52%
$c_0 = 0.75$	$c_n = 0.50$	$c_t = 0.50$	+ 0.22%

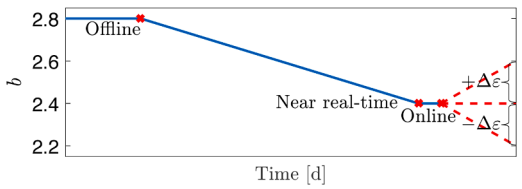

Fig. 6. Trajectory of the parameter b accounting for the residual uncertainty between near real-time and online stage.

in higher accuracies than SVM-RFE, (2) AMB TAN requires 75% less computational time than SFS, resulting in the best trade-off between computational time and accuracy, and (3) the distribution in computational times is narrow. This narrow distribution enables precise estimations of the required computational times, allowing to reliably schedule the FS and classifier training later in the workflow.

The French transmission system was used to illustrate these advantages for larger systems. The computational times were estimated by scaling up from the IEEE 68-bus system and using $\mathcal{O}(n\ /2)$ and $\mathcal{O}(n^2)$ with n number of buses for the AMB TAN and SFS approach, respectively. The resulting computational times are reported in Table 4. There,

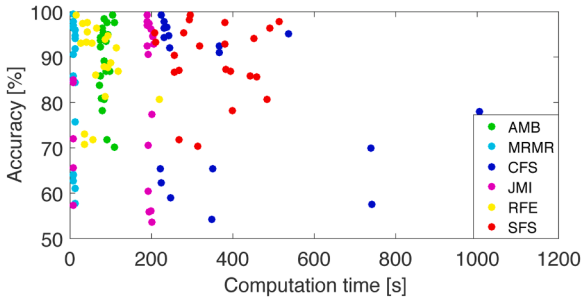

Fig. 7. Accuracy and computation time for [31] contingencies with different FS techniques.

Table 3
Statistical analysis of distributions shown in Fig. 7.

	FS Method					
	AMB	MRMR	CFS	JMI	RFE	SFS
Accuracy Mean	0.89	0.71	0.75	0.77	0.85	0.89
Accuracy Std	0.08	0.24	0.26	0.23	0.13	0.08
Accuracy CI	0.01	0.02	0.02	0.02	0.01	0.01
Time Mean	87*s*	10*s*	353*s*	153*s*	75*s*	352*s*
Time Std	11*s*	3*s*	213*s*	81*s*	47*s*	94*s*

Table 4
Estimate of computational time for FS for a large system.

Approach	Case Study	
	IEEE 68-bus system	French transmission system
AMB TAN	87*s*	∼ 30min
SFS	6min	∼ 75*h*

the SFS approach would require a computational time of 75*h* to identify features, where the AMB TAN approach requires around 30*min*.

4.5. Size of the training database

In this study, the required size of the training database when using various FS approaches was investigated. This was an important analysis to conduct as operators need to decide how to optimally allocate the

computational budget on data generation and training (FS and classifier). The size of the training database $|\underline{\Omega}|$ was varied, as the number of OCs used in FS and classifier training. The four best FS approaches (AMB TAN, MRMR, RFE and SFS) in terms of the accuracy and computations trade-off were first studied with larger database size, i.e. $|\underline{\Omega}| = 40000$, 80000, 120000, to guarantee the scalability of the workflow. For each step, the four different FS were applied and the DT trained. This was repeated 10 times with different training/testing data combinations. The results were averaged and shown in Table 5 for the single contingency with the lowest accuracy across all contingencies. The AMB TAN, RFE and SFS approaches resulted in highest accuracies, however AMB TAN outperformed in terms of computations. Since SFS resulted in similar accuracies to RFE in reduced computational times, only the SFS approach was considered in the following studies. Then, the same analysis was conducted with reducing database size, i.e. $|\underline{\Omega}| = 1000$, 2000,....12000 and applying the best three FS approaches (AMB TAN, MRMR and SFS). The results were averaged and illustrated in Fig. 8 for the same single contingency. The AMB TAN and SFS approach resulted in highest accuracies. However, AMB TAN outperformed SFS in terms of computations. The best database size for this contingency was around $|\underline{\Omega}| = 5000$ as the accuracy did not improve anymore for larger database. At $|\underline{\Omega}| = 5000$, AMB TAN required 70% less computational time than SFS. If in the real-time stage no FS approach was used and the training database was reduced to $|\underline{\Omega}| = 5000$, the accuracy decreased by 5%.

For the same contingency, the required training OCs in the offline stage when introducing the near real-time stage with the AMB TAN approach was investigated. The offline database size was reduced to $|\underline{\Omega}| = 500$, which is the minimum number of OCs to guarantee the feasibility of the SFS approach. It was assumed 4500 OCs were included in the database during the near real-time stage as $|\underline{\Omega}| = 5000$ was shown to be the best size in terms of accuracy performance. The AMB TAN approach was used, the classifier was trained and then tested on 1000 new OCs from the same load distribution. It turned out that the offline database size could be reduced up to 95%, resulting in the same accuracy performance.

4.6. Computational efficiency

In this study, the computational savings obtained by using the proposed workflow for DSA were investigated. For the offline and near real-time stages, the impact of using or not the proposed AMB TAN approach on the computational cost of data generation and training of the machine was analysed. For this comparison, 12000 OCs were available in the offline and near real-time stage. However, the use of the AMB TAN approach allowed to reduce the offline and near real-time training databases to $|\underline{\Omega}| = 500$ and $|\underline{\Omega}| = 5000$, respectively, resulting in same accuracy performance (Section 4.5). Table 6 and Fig. 9 summarize the

Table 5
Accuracy and computation time according to different FS techniques by increasing the size of the training database.

Training data	$\|\underline{\Omega}\| = 40000$		$\|\underline{\Omega}\| = 80000$		$\|\underline{\Omega}\| = 120000$	
	Accuracy	*Time*	*Accuracy*	*Time*	*Accuracy*	*Time*
AMB TAN	0.81	271*s*	0.81	628*s*	0.81	1083*s*
MRMR	0.47	17*s*	0.56	54*s*	0.61	52*s*
RFE	0.81	1116*s*	0.82	4772*s*	0.82	10555*s*
SFS	0.83	1031*s*	0.83	2466*s*	0.83	4296*s*

[3] 22

[3] 22

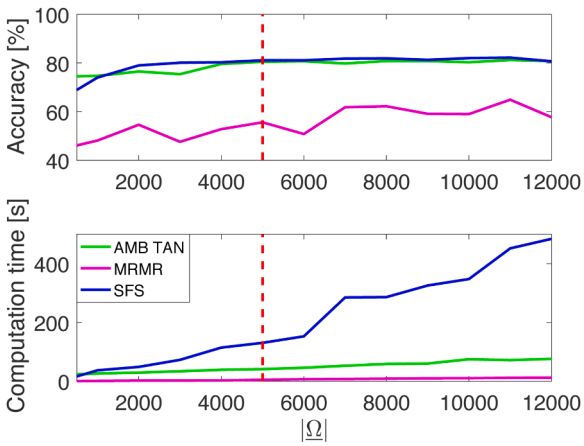

Fig. 8. Accuracy and computation time according to different FS techniques by varying the size of the training database.

computational costs for data generation, feature selection and training of the machine when using the proposed AMB TAN approach and when not using the approach. The simulation time for each observation was 0.2s. The proposed AMB TAN approach enabled a total computational saving in DSA workflows (offline and near real-time stages) up to 75%. This result is a major finding of this work.

4.7. A combination-based approach

To optimize the balance between high accuracy and low computations, a further analysis on a combination-based FS approach was conducted. This is worth investigating because different FS approaches showed the best performance for different contingencies, as shown in Fig. 7. In fact, the performance of MRMR was higher than AMB TAN for some contingencies. For this reason, the performance when combining these two FS approaches was investigated. For all contingencies, MRMR was first applied as its computational times were negligible. Then, only for contingencies where the accuracy was lower than a threshold, AMB TAN was used to select more relevant features. The threshold was selected to be 0.8 and the mean of accuracy and computational time were compared across all 22 contingencies for the AMB TAN versus this combined approach. The result was that the computational time reduced by 35% when using the combined approach, whereas the accuracy drops by only 0.5%. Then, the threshold value was studied in the range of [0,1] with step-size 0.05. The full results are shown in Fig. 10. It turned out that a threshold higher than 0.9 leads to under-performing the AMB TAN approach and 0.8 is the optimal threshold.

4.8. Discussion

The proposed causality-based FS approach in combination with a three-stages workflow showed promising results for online DSA applications, resulting in an accuracy increase of 1%. It resulted in the best trade-off in terms of accuracy and computational performance across

Table 6
Computation times for offline and near real-time stages using and not the proposed AMB TAN approach.

Stage	Offline		Near real-time	
Approach	without FS	with FS	without FS	with FS
N° observations	12000	500	12000	5000
Simulation time	2400*s*	100*s*	2400*s*	1000*s*
FS time	0*s*	23*s*	0*s*	50*s*
DT training time	60*s*	1*s*	60*s*	3*s*
Total	2460*s*	124*s*	2460*s*	1053*s*

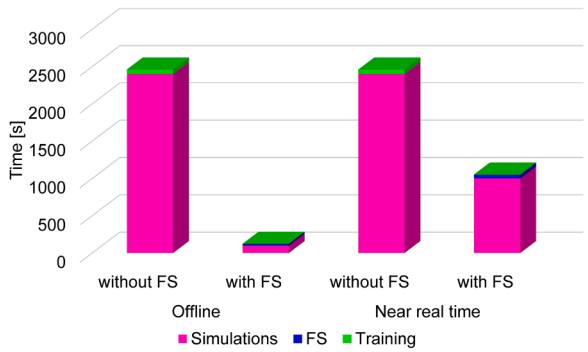

Fig. 9. Computation times for offline and near real-time stages using and not the proposed AMB TAN approach.

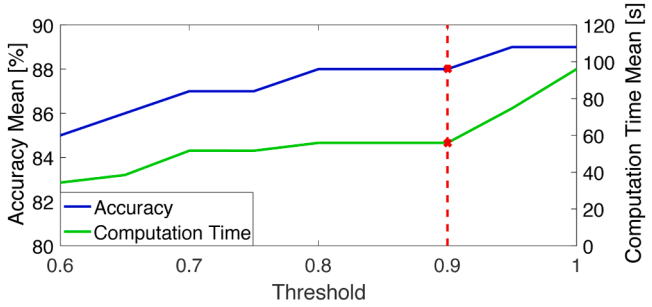

Fig. 10. Mean values of accuracy and computation time by using the combination-based approach and varying the accuracy threshold.

other FS approaches, being four times faster than the SFS approach while obtaining the same high accuracy. The features selected through AMB TAN improved the interpretability of the classifier as they were all located close to the faulted bus and similar features were selected for similar contingencies. Moreover, AMB TAN has a very narrow distribution of the computational times. Consequently, operators can reliably schedule the learning task very close to the real-time operation as the required computational time can be estimated accurately. These improvements in terms of computations are significant higher when moving to a large system, as illustrated on the French system (99% reduction). In terms of robustness against uncertain OCs, the proposed causality-based approach demonstrated to improve the accuracy for various trends in the change of the load probability distributions. The AMB TAN FS approach also reduced the amount of required training data by 60% and by 95% in comparison to the SFS approach for the offline and near real-time stage, respectively. This reduction is significant as a key bottleneck of online DSA approaches is the amount of data needed for large systems, and hence the computational cost for simulating the OCs from this large amount of data. Consequently, the proposed AMB TAN enabled a total computational saving for offline and near real-time stages, including data generation, feature selection and model training, up to 75%. A further reduction in the computational time was obtained by combining AMB TAN approach to another FS approach, resulting in almost the same accuracy (only 0.5% reduction). Overall, all these benefits represent a fundamental step forward to deploy machine learning approaches for DSA.

A few key limitations in designing FS to data-driven DSA approaches still exist. A lot of data are still required and if too little data are available, the FS approaches are not capable of identifying the underlying statistical dependencies, resulting in low prediction accuracy. The machine learning approaches that are used along AMB TAN FS were selected based on their relevance in the literature and their choice does not affect the performance of the proposed workflow, e.g. classification models different from DTs can be used (Table 1). As the MB strongly varies from one contingency to another, single machine learning approaches for each contingency should be trained. In this work, the focus is on FS. However, when designing the entire machine learning workflow, every single step should be investigated and considered when allocating computational budgets. Relying on machine learning based DSA workflow rather than investing in new assets has a risk that should be considered in the decision making process. The proposed combination of the causality-based FS and near-real time stage should be also tested against other stability metrics. Finally, as topology changes may become more frequent, investigating how the proposed FS method performs under topology changes becomes important [44]. The performance of the AMB TAN FS approach should be also tested when the occurrences of severe weather events or topological changes make the assumption on independent currents not valid anymore. In this context, incorporating the physical knowledge into the learning approach of the causal structure between features may also improve the robustness of a single classifier across similar contingencies or network's topologies.

5. Conclusion

The challenge of designing computationally efficient and robust feature selection approaches to machine learning-based DSA was investigated, showing that DSA can suffer from discrepancies between real-time and offline. Not considering these discrepancies along with the time horizon of generating data and training the machine results in inaccuracies. In response, a novel causality-based FS approach in combination with a near real-time stage was proposed to train the classifier closer to real-time operation. By using the system's physics to learn the causal structure and to identify the Markov Blanket, this approach outperformed other FS approaches in robustness, interpretability and significantly improved the computational time while the predictive accuracy is as high as state-of-the-art FS approaches. The IEEE 68-bus system and transient stability were used to illustrate these benefits, showing the required computational time for FS was reduced by 75%. This reduction in the computational time is becoming more significant for large systems as demonstrated for the French transmission system. Moreover, the required training database was reduced by 60%. This reduction is important as often the number of time-domain simulations for a training database is a barrier to using data-driven DSA. The proposed FS approach in combination with a three-stages workflow is a significant step forward to include dynamics in future's security assessment by the support of machine learning, enabling an effective operation of the grid assets closer to their limitations. In the future, the entire workflow should be investigated when allocating computational budgets for other objectives of DSA approaches.

CRediT authorship contribution statement

Federica Bellizio: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – original draft. **Jochen L. Cremer:** Formal analysis, Resources, Supervision, Visualization, Writing – review & editing. **Mingyang Sun:** Resources, Supervision, Visualization, Writing – review & editing. **Goran Strbac:** Funding acquisition, Project administration, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant 62103371, the Engineering and Physical Sciences

Research Council (UK) under the Integrated Development of Low-Carbon Energy Systems programme (EP/R045518/1), and by the TU Delft AI Labs Programme (NL). We are thankful to colleagues from Reseau de Transport d'Electricite who provided expertise that greatly assisted the research. We also thank the reviewers for their insightful thoughts and discussion in the review process.

References

[1] P. Panciatici, G. Bareux, L. Wehenkel, Operating in the fog: security management under uncertainty, IEEE Power Energy Mag. 10 (5) (2012) 40–49, https://doi.org/10.1109/MPE.2012.2205318.
[2] G. Strbac, N. Hatziargyriou, J.P. Lopes, C. Moreira, A. Dimeas, D. Papadaskalopoulos, Microgrids: enhancing the resilience of the European megagrid, IEEE Power Energy Mag. 13 (3) (2015) 35–43.
[3] L. Duchesne, E. Karangelos, L. Wehenkel, Recent developments in machine learning for energy systems reliability management, Proc. IEEE (2020).
[4] F. Capitanescu, J.M. Ramos, P. Panciatici, D. Kirschen, A.M. Marcolini, L. Platbrood, L. Wehenkel, State-of-the-art, challenges, and future trends in security constrained optimal power flow, Electr. Power Syst. Res. 81 (8) (2011) 1731–1741.
[5] P. Kundur, et al., Definition and classification of power system stability IEEE/CIGRE joint task force on stability terms and definitions, IEEE Trans. Power Syst. 19 (3) (2004) 1387–1401.
[6] L.A. Wehenkel, Automatic Learning Techniques in Power Systems, Kluwer Academic Publishers, Norwell, MA, USA, 1998.
[7] I. Konstantelos, G. Jamgotchian, S.H. Tindemans, P. Duchesne, S. Cole, C. Merckx, G. Strbac, P. Panciatici, Implementation of a massively parallel dynamic security assessment platform for large-scale grids, IEEE Trans. Smart Grid 8 (3) (2017).
[8] Y. Zhang, Y. Xu, Z.Y. Dong, R. Zhang, A Hierarchical self-adaptive data-analytics method for real-time power system short-term voltage stability assessment, IEEE Trans. Ind. Inf. 15 (1) (2018) 74–84.
[9] L. Moulin, A.A. Da Silva, M. El-Sharkawi, R.J. Marks, Support vector machines for transient stability analysis of large-scale power systems, IEEE Trans. Power Syst. 19 (2) (2004).
[10] Y. Zhou, Q. Guo, H. Sun, Z. Yu, J. Wu, L. Hao, A novel data-driven approach for transient stability prediction of power systems considering the operational variability, Int. J. Electr. Power Energy Syst. 107 (2019) 379–394.
[11] C. Liu, et al., A systematic approach for dynamic security assessment and the corresponding preventive control scheme based on decision trees, IEEE Trans. Power Syst. 29 (2) (2014) 717–730.
[12] J.L. Cremer, I. Konstantelos, G. Strbac, From optimization-based machine learning to interpretable security rules for operation, IEEE Trans. Power Syst. 34 (5) (2019) 3826–3836.
[13] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, M. Kawanabe, Direct importance estimation for covariate shift adaptation, Ann. Inst. Stat. Math. (2008).
[14] K. Sun, S. Likhate, V. Vittal, V.S. Kolluri, S. Mandal, An online dynamic security assessment scheme using phasor measurements and decision trees, IEEE Trans. Power Syst. 22 (4) (2007) 1935–1943, https://doi.org/10.1109/TPWRS.2007.908476.
[15] Y. Xu, Z.Y. Dong, J.H. Zhao, P. Zhang, K.P. Wong, A reliable intelligent system for real-time dynamic security assessment of power systems, IEEE Trans. Power Syst. 27 (3) (2012) 1253–1263, https://doi.org/10.1109/TPWRS.2012.2183899.
[16] M. He, J. Zhang, V. Vittal, A data mining framework for online dynamic security assessment: decision trees, boosting, and complexity analysis. 2012 IEEE PES Innovative Smart Grid Technologies (ISGT), 2012, pp. 1–8, https://doi.org/10.1109/ISGT.2012.6175766.
[17] M. He, J. Zhang, V. Vittal, Robust online dynamic security assessment using adaptive ensemble decision-tree learning, IEEE Trans. Power Syst. 28 (4) (2013).
[18] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[19] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, H. Liu, Feature selection: a data perspective, ACM Comput. Surv. 50 (2016), https://doi.org/10.1145/3136625.
[20] R. Zhang, Y. Xu, Z.Y. Dong, D.J. Hill, Feature selection for intelligent stability assessment of power systems. 2012 IEEE Power and Energy Society General Meeting, 2012, pp. 1–7, https://doi.org/10.1109/PESGM.2012.6344780.
[21] K. Yu, X. Guo, L. Liu, J. Li, H. Wang, Z. Ling, X. Wu, Causality-based feature selection: methods and evaluations, ACM Comput. Surv. (CSUR) 53 (5) (2020) 1–36.
[22] B. Schölkopf, Causality for machine learning, arXiv preprint arXiv:1911.10500 (2019).
[23] K. Loparo, F. Abdel-Malek, A probabilistic approach to dynamic power system security, IEEE Trans. Circuits Syst. 37 (6) (1990) 787–798.
[24] H. Shimodaira, Improving predictive inference under covariate shift by weighting the log-likelihood function, J. Stat. Plann. Inference 90 (2) (2000) 227–244.
[25] D. Koller, M. Sahami, Toward Optimal Feature Selection. Technical Report, Stanford InfoLab, 1996.
[26] Y. Weng, Y. Liao, R. Rajagopal, Distributed energy resources topology identification via graphical modeling, IEEE Trans. Power Syst. 32 (4) (2017) 2682–2694.
[27] S. Bolognani, Grid topology identification via distributed statistical hypothesis testing. Big Data Application in Power Systems, Elsevier, 2018, pp. 281–301.
[28] C. Chow, C. Liu, Approximating discrete probability distributions with dependence trees, IEEE Trans. Inf. Theory 14 (3) (1968) 462–467, https://doi.org/10.1109/TIT.1968.1054142.
[29] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Elsevier, 2014.
[30] J. Zhao, et al., Roles of dynamic state estimation in power system modeling, monitoring and operation, IEEE Trans. Power Syst. (2020) 1, https://doi.org/10.1109/TPWRS.2020.3028047.
[31] N. Friedman, Building classifiers using Bayesian networks. Proceedings of the Thirteenth National Conference on Artificial Intelligence, AAAI Press, 1996, pp. 1277–1284.
[32] C. Ratanamahatana, CloNI: clustering of N-interval discretization, in: Proceedings of the 4th International Conference on Data Mining Including Building Application for CRM & Competitive Intelligence.
[33] C.F. Aliferis, I. Tsamardinos, A. Statnikov, HITON: a novel Markov blanket algorithm for optimal variable selection. AMIA Annual Symposium Proceedings, 2003.
[34] I. Tsamardinos, C.F. Aliferis, A.R. Statnikov, Algorithms for large scale Markov blanket discovery. FLAIRS Conference, 2003.
[35] J. Shen, L. Li, W.-K. Wong, Markov blanket feature selection for support vector machines. AAAI vol. 8, 2008, pp. 696–701.
[36] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learn. Res. 5 (2004).
[37] E. Schaffernicht, H.-M. Gross, Weighted mutual information for feature selection. International Conference on Artificial Neural Networks, Springer, 2011, pp. 181–188.
[38] N. Friedman, M. Goldszmidt, Discretizing continuous attributes while learning Bayesian networks. Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., 1996, pp. 157–165.
[39] B. Pal, B. Chaudhuri, Robust Control in Power Systems, Springer Science & Business Media, 2006.
[40] R. Singh, B.C. Pal, R.A. Jabr, Statistical representation of distribution system loads using gaussian mixture model, IEEE Trans. Power Syst. 25 (1) (2009) 29–37.
[41] M.B. Cain, R.P. O'neill, A. Castillo, et al., History of optimal power flow and formulations, Fed. Energy Regul. Commission 1 (2012) 1–36.
[42] T. Hastie, S. Rosset, J. Zhu, H. Zou, Multi-class AdaBoost, Stat. Interface 2 (3) (2009) 349–360.
[43] M. Chen, Q. Liu, S. Chen, Y. Liu, C.-H. Zhang, R. Liu, XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system, IEEE Access 7 (2019) 13149–13158.
[44] F. Bellizio, J.L. Cremer, G. Strbac, Machine-learned security assessment for changing system topologies, Int. J. Electr. Power Energy Syst. (2021).