Generating quality datasets for real-time security assessment
Balancing historically relevant and rare feasible operating conditions

Bugaje, Al Amin B.; Cremer, Jochen L.; Strbac, Goran

DOI
10.1016/j.ijepes.2023.109427

Publication date
2023

Document Version
Final published version

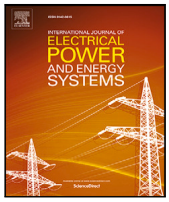
Published in
International Journal of Electrical Power and Energy Systems

Electrical Power and Energy Systems 154 (2023) 109427

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

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



Generating quality datasets for real-time security assessment: Balancing historically relevant and rare feasible operating conditions

Al-Amin B. Bugaje [a], Jochen L. Cremer [b,*], Goran Strbac [a]

[a] Department of Electrical & Electronic Engineering, Imperial College London, London, SW7 2AZ, UK
[b] Department of Electrical Sustainable Energy, TU Delft, Mekelweg 5, 2628 CD Delft, Netherlands

ARTICLE INFO

Keywords:
Data generation
Dynamic security assessment
Machine learning
Power system operation

ABSTRACT

This paper presents a novel, unified approach for generating high-quality datasets for training machine-learned models for real-time security assessment in power systems. Synthetic data generation methods that extrapolate beyond historical data can be inefficient in generating feasible and rare operating conditions (OCs). The proposed approach balances the trade-off between historically relevant OCs and rare but feasible OCs. Unlike conventional methods that rely on historical records or generic sampling, our approach results in datasets that generalise well beyond similar distributions. The proposed approach is validated through experiments on the IEEE 118-bus system, where a decision tree model trained on data generated using our approach achieved 97 % accuracy in predicting the security label of rare OCs, outperforming baseline approaches by 41 % and 20 %. This work is crucial for deploying reliable machine-learned models for real-time security assessment in power systems undergoing decarbonisation and integrating renewable energy sources.

1. Introduction

The power system is undergoing a massive decarbonisation effort by introducing renewable energy sources interfaced with converter-based electronics [1]. This integration ushers a new era of power systems operations, primarily via the introduction of high levels of uncertainty [2] in power generation, together with faster and more complex dynamics in power systems operations [3]. Conventional reliability management designed on legacy power systems relies on assessing the static security for a few pre-select faults. However, this approach to reliability management is unsuitable for tractably analysing the security and adequacy of emerging power grids with low inertia, specifically when assessing dynamic security. An example of this intractability is that a single dynamic security assessment using time-domain simulations can take up to 56 s in large systems [4]. However, to ensure dynamic system security, the system operator (SO) needs to carry out several thousand assessments for multiple operating conditions (OCs) and contingencies.

Machine learning (ML) is promising to improve the situational awareness of SOs without falling into tractability issues [4]. Specifically, ML's ability to infer complex underlying relationships from large and varied datasets and subsequently perform high-speed predictions makes it a competitive alternative to conventional time domain simulations. In recent years, there has been a rise in research outputs in the power system community that investigate how to implement ML in control rooms for power systems operation [5–7]. For security assessment, the ML model learns the security boundary to categorise whether the system in the current OC survives a contingency, labelled as secure or insecure. There, some of the research focuses on efficiently optimising the different stages of the ML workflow, viz: data generation [8], data pre-processing, model training [9] and model evaluation [10]. As ML becomes even more crucial for learning the dynamic security boundary [11] in low inertia systems [12], there is a renewed interest in data generation approaches [13–18] and in more recent works [19–25] to produce representative datasets, especially as low *quality* data leads to training inaccurate models.

The data generation phase of the ML-based security assessment workflow is the first and arguably the most crucial part of the ML workflow, as model performance generally reflects the *quality* of the training data [13]. The availability of data from increased monitoring via PMUs and monitoring tools in control centres [26] suggests that recent works on real-time probabilistic security assessment [10,27] are promising even when there are changes in network topology [28]. However, using only historical records as training data is insufficient [29], and hence simulations are used to generate large synthetic datasets that cover a variety of OCs. The state-of-the-art approaches that move beyond historical records fall under one of three approaches: *historical sampling, generic sampling and importance sampling*, each typically focusing on maximising a specific property of what constitutes *quality* datasets,

* Corresponding author.
E-mail addresses: abb18@imperial.ac.uk (A.-A.B. Bugaje), j.l.cremer@tudelft.nl (J.L. Cremer), g.strbac@imperial.ac.uk (G. Strbac).

https://doi.org/10.1016/j.ijepes.2023.109427
Received 2 April 2023; Received in revised form 19 June 2023; Accepted 4 August 2023
Available online 16 August 2023

Nomenclature

Indices

d	cardinality of set Ω^N
m	index of power system observations
n	index of power system variables

Sets

$\lvert\cdot\rvert$	cardinality of a set
Ω^A	set of generated OCs in proposed phase A
Ω^B	set of generated OCs in proposed phase B
Ω^C	set of clusters in phase B
Ω^g	set of generically generated feasible OCs
Ω^h	set of observed historical data
Ω^J	set of contingencies
Ω^K	subset of power system variables
Ω^N	set of power system variables
Ω^s	set of copula-based generated OCs
$\Omega^{g''}$	set of all generically generated OCs
$\Omega^{g'}$	set of generically generated infeasible OCs

Parameters

λ	Wasserstein distance threshold
S	number of samples to generate
ζ	tolerance parameter
Q	covariance matrix
T_n	target of variable $n \in \Omega^N$
w_k	weight parameter

Variables

$\hat{\boldsymbol{X}}$	random variable approximating true distribution
$\boldsymbol{X}$	random variable representing true distribution
$\mathcal{Y}$	security labels
Z	multivariate gaussian

Others

$\mathcal{E}$	entropy
$\mathcal{V}$	convex hull volume
$\mathcal{W}_p$	pth order Wasserstein distance
$\Phi(.)$	univariate normal CDF
Π^1	share of OCs belonging to class 1
C	copula function
F	cumulative distribution function (CDF)
X^m	vector of a generated OC
X	matrix of power system OCs

respectively, *historical relevance, coverage and discriminative relevance* [4,5]. *Historical relevance* shows how much the generated OCs represent historical power systems operations through variable dependency structure. *Coverage* measures how much of the power system feasible region the generated OCs span. *Discriminative* relevance depicts how much new information the generated OCs add to training an ML model. However, there are pending challenges to curating *quality* datasets for security assessment.

The complexity of the data generation challenge is primarily underscored by three contrasts that define *quality* datasets. An example of these contrasts is between *historical relevance* and *coverage*. Approaches maximising *coverage* (e.g. [14–16,30]) aim to uniformly span all the possible feasible OCs (region A of Fig. 1) and do not consider the

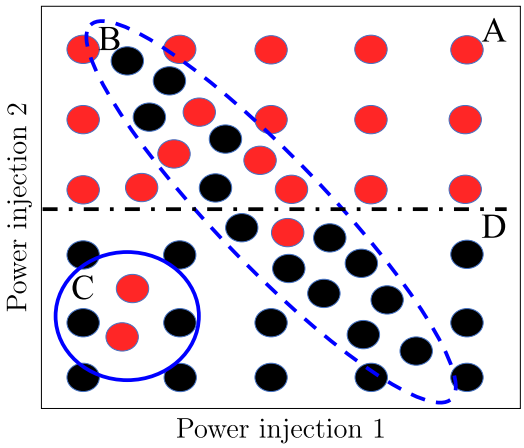


Fig. 1. Quality datasets balance historical relevance (), coverage (), and discriminative relevance (-·-·-·-) so that ML-models can be trained for accurately predicting secure () and insecure () OCs.

dependency structures between variables thereby sacrificing *historical relevance* (region B of Fig. 1). While approaches that consider dependency structures of variables via *historical relevance* (e.g. copula modelling to capture complex non-gaussian marginal distributions and non-linear multivariate dependencies [4,13], autoencoders and conditional variational autoencoders [31]) do not consider other feasible but rare OCs (region C of Fig. 1) thereby sacrificing *coverage*. Another contrast is between *historical relevance* and *discriminative relevance*. Approaches that focus on *historical relevance* (region B of Fig. 1) aim to mimic typical power systems operations and retain variable dependency structures but do not generalise to OCs not in historical records. While approaches that focus on *discriminative relevance* (line D of Fig. 1) aim to target security decision boundaries, so-called high information content regions but do not consider dependency structures between variables that represent typical power system OCs. A further contrast is between *coverage* and *discriminative relevance*. Approaches that focus on maximising *coverage* may miss out on regions that span the security boundary. While approaches that focus on *discriminative relevance* assume a stationary security decision boundary and do not generalise to "rare" OCs (region C of Fig. 1).

The existing gap of all the aforementioned state-of-the-art approaches for generating synthetic datasets for power system security assessment is the lack of a unified approach to efficiently combine all three contrasts of *quality* datasets. The Wasserstein distance [32] addresses this gap and leverages advances in optimal transport research [33] to deal with the comparison of distributions. This ability to compare distributions allows trading-off *historical relevance* and *coverage* of generated OCs thereby preserving relevant dependency structures while generating rare OCs. The Wasserstein distance [32] can be thought of in 1-dimension as the earth mover's distance and calculates how much work it takes to transport the mass of one distribution to another. Besides its intuitiveness, the Wasserstein distance can make meaningful comparisons between distributions with non-overlapping support unlike the popular Kullback–Leibler divergence [34], and by extension the Jensen–Shannon divergence. Other metrics include the total variation and Cramér distances [35], which are generally considered as computationally efficient measures of distance, particularly for low-dimensional data. However, the Wasserstein distance is preferred for distributions with complex structure and high dimensionality [36]. While computing the Wasserstein distance [32] in high dimensions is non-trivial, the Sliced-Wasserstein distance [37] exploits the closed-form of projected one-dimensional distances and has acceptable statistical and asymptotic properties.

1.1. Contributions

The contribution of this paper is the efficient combination of the three properties of *historical relevance, coverage, and discriminative relevance* that were individually considered to generate *quality* datasets in previous research. Our innovative algorithm creates generalised datasets that are more representative, and relevant for many more power system operating settings, therefore making the models trained on our dataset more generic (for example, for higher renewable scenarios where limited historical training data is available). For the first time, this paper proposes a novel unified approach that considers all three properties that define *quality* datasets to generate information-rich and historically relevant datasets while considering rare OCs to train reliable models for power system security. The proposed approach leverages advances in optimal transport research and introduces the Wasserstein distance as a metric. The proposed metric allows to efficiently combine historically relevant OCs modelled with copulas and rare OCs modelled using state-of-the-art split-based generic sampling, which improves the efficiency of data generation. Additionally, the proposed approach uses entropy to redirect sampling to other regions of the feasible space.

The rest of the paper is structured as follows: Section 2 introduces state-of-the-art sampling approaches focusing on maximising single properties and their limitations. Section 3 introduces the proposed unified sampling approach. Section 4 illustrates case studies to compare the performance of ML-models on various datasets. Section 5 concludes the paper.

2. Single criterion-based sampling approaches

The three existing single criterion-based sampling approaches have strengths and limitations for security assessments. Security assessment $X \rightarrow \mathcal{Y}$ takes as an input the power system OCs X and outputs security labels $\mathcal{Y}$ for a set of probable contingencies Ω^J. The security labels $\mathcal{Y}_j^m \in \{0,1\}\ \forall j \in \Omega^J$ represent secure and insecure OCs, respectively, where m is the index of OC X^m. Typically, the input variables are the static pre-fault set-point of all generators and loads. These input variables define the OC X^m and are bounded by the power system's physical limits, such as generator limits, line limits, and complex network constraints.

The first two types of sampling approaches, *historical* and *generic sampling*, focus on generating representative pre-fault OCs X. *Importance sampling* approach focuses on generating X aiming at the inverse $\mathcal{Y} \rightarrow X$ by targeting information-rich regions $\hat{\alpha} \subseteq \alpha$, where α denotes the feasible space containing all possible OCs X. Each approach subsequently focuses on a specific property and has an associated metric to measure sampling quality.

2.1. State-of-the-art historical sampling

Historical sampling focuses on the *historical relevance* property of sampling quality. *Historical sampling* approaches aim to generate similar OCs to observed historical data by learning the underlying distribution and the dependency structures of historical data. The set of historical observations is Ω^h with the data $X_n^m\ \forall\ n \in \Omega^N,\ m \in \Omega^h$, where Ω^N is the set of power systems variables (e.g. loads and injections). For simplicity, we denote X_n as a vector of all observations for the nth variable, X^m as a data vector for the mth observation and $d = |\Omega^N|$ as the cardinality of set Ω^N. From a statistical perspective, the historical data in Ω^h are assumed to be drawn from an unknown true distribution, represented by the continuous random variable $\boldsymbol{X}$. A *historical sampling* approach starts by approximating this true distribution's random variable $\boldsymbol{X}$ by fitting a statistical model $\hat{\boldsymbol{X}}$ to the observed data Ω^h. Then, the approach applies Monte Carlo sampling. Monte Carlo sampling is widely used to randomly sample probability distributions to generate new data. A challenge of *historical sampling* is the separation of marginal distributions from a multivariate distribution with non-linear dependencies, as power system OCs are typically load and generator injection profiles with non-linear dependencies such as renewables. Previous works [4,38,39] consider copula-based sampling models (CSM) in power systems, as copulas can separate the dependency structure of marginal distributions from a multivariate distribution. Another challenge is that OCs can be observed in disjoint clusters that follow distinct statistical characteristics due to unique power system modes, e.g., considering different seasons or times of the day. Previous work in [13] partitions the observed data Ω^h into distinct clusters of similar characteristics profiles. While computing the copula model for stochastic variables like wind farms in high dimensions is not trivial, pair-copula decomposition can mitigate the computational burden. This work approximates $\boldsymbol{X}$ using a combination of copulas and clustering.

A d-dimensional copula, $C : [0,1]^d :\rightarrow [0,1]$ is a cumulative distribution function (CDF) with uniform marginals that provides a suitable way to separate the marginal distributions of X_n from their dependency structure. The multivariate CDF, F_X, with marginal distributions $F_1, \ldots, F_d$ is then

$$F_X(X_1, \ldots, X_d) = C\Big(F_1(X_1), \ldots, F_d(X_d)\Big) \tag{1}$$

which represents Sklar's theorem [40]. The copula, C is unique if all the marginal distributions $F_1, \ldots, F_d$ are continuous. Without loss of generality to other variations of copulas, a single d-dimensional Multivariate Gaussian (MG) copula can model the dependency structure parameterised by the correlation matrix Q as

$$C_Q(\boldsymbol{u}) := \boldsymbol{\Phi}_Q\Big(\boldsymbol{\Phi}^{-1}(u_1), \ldots, \boldsymbol{\Phi}^{-1}(u_d)\Big) \tag{2}$$

where $\boldsymbol{\Phi}(.)$ is the standard univariate normal CDF, $\boldsymbol{\Phi}_Q(.)$ is the joint CDF of a MG variable with mean $\mu = 0$, covariance matrix $\Sigma = Q$, with uniform marginals $u_n = [0,1]$, and $\boldsymbol{u} = (u_1, \ldots, u_d)$. The correlation matrix can be transformed using the Cholesky decomposition $Q = A^T A$, where A is the lower triangular matrix of Q.

The sampling from a MG CSM follows Algorithm 1, where the set of OCs generated with *historical sampling* $\Omega^s = \{\}$ is initially empty. The algorithm then partitions the observed data Ω^h into one of $\mathcal{L} \in \mathbb{N}$ pre-defined clusters. Then, for each independent cluster $l \leq \mathcal{L}$, the algorithm generates a random MG variable $\boldsymbol{Z} \sim \mathrm{MG}_d(\boldsymbol{0}, \mathrm{I}_{\boldsymbol{d}})$, where I_d is the d-dimensional identity matrix. The algorithm then determines $\boldsymbol{\beta} = A^T Z$, and computes $\boldsymbol{U} = (\boldsymbol{\Phi}(\beta_1), \ldots, \boldsymbol{\Phi}(\beta_d))$, whose distribution represents the MG copula from Eq. (2) s.t. $\mathrm{Prob}(U_1 \leq u_1, \ldots, U_d \leq u_d) = \boldsymbol{\Phi}_Q(\boldsymbol{\Phi}^{-1}(u_1), \ldots, \boldsymbol{\Phi}^{-1}(u_d))$. Using the copula property of invariance under monotonic transformations, a resulting random OC m with data vector $\hat{X}^m = (\hat{X}_1^m, \ldots, \hat{X}_d^m)$ is obtained via the standard inverse transform method along each dimension such that $\hat{X}_n^m = F_n^{-1}(U_n)$. The CSM stops after $S^s \in \mathbb{N}$ OCs are generated and added to $\Omega^s \longleftarrow m$. $|\Omega^s|$ is the cardinality of the set. Previous works have explored more complex copula representations, including C and D vine copulas [13,41]. In this work, the choice of the Gaussian copula as a representative method historical sampling is due to its simplicity and scalability in higher dimensions.

An advantage of *historical sampling* is that the set of generated OCs Ω^s retains the dependency structure of observed historical records Ω^h. A limitation of *historical sampling* is that $\hat{\boldsymbol{X}}$ can only generate OCs Ω^s that are statistically similar to OCs in historical records Ω^h.

2.2. State-of-the-art generic sampling approach

Generic sampling focuses on maximising *coverage* of a broad spectrum of varying and physically feasible OCs. *Generic sampling* approaches in power systems typically involve stratified sampling, such as the Latin Hypercube Sampling [14–16] and, recently, the sequential split-based sampling [30].

Algorithm 1 Copula-based Historical Sampling

Require: Ω^h, $\Omega^s = \{\}$, S^s, $l = 1$, $\mathcal{L} \in \mathbb{N}$
Segment Ω^h into $\mathcal{L}$ disjoint clusters
Define arbitrary covariance matrix Σ
Compute Spearman correlation matrix Q
Compute the Cholesky decomposition $Q = A^T A$
for $l \leq \mathcal{L}$ **do**
  while $|\Omega^s| \leq (\lceil \frac{S^s}{\mathcal{L}} \rceil \times l)$ **do**
    Compute $\boldsymbol{Z} = \mathrm{MG}_d(\mathbf{0}, \mathrm{I}_d)$
    Determine $\boldsymbol{\beta} = A^T Z$
    Compute $\boldsymbol{U} = (\Phi(\beta_1), \cdots, \Phi(\beta_d))$
    Compute $\hat{X}_n^m = F_n^{-1}(U_n) \; \forall \, n \in N$
    $\Omega^s \longleftarrow m$
  end while
  $l = l + 1$
end for

Algorithm 2 presents this generic sampling model (GSM) to generate a set of OCs Ω^g that are uniformly distributed across the entire feasible space while ensuring that the physical constraints that represent power systems equality $g(X^{\tilde{m}}) = 0$ and inequality $h(X^{\tilde{m}}) \leq 0$ constraints (e.g., nodal balance, line flow, voltage, phase angle, and generator limits) are met for each generated OC $\tilde{m} \in \Omega^g$ [30]. The generation of OCs follows an exploration of the feasible space similar to a binary tree search that sequentially bisects the input domain to create OCs at specific targets $T_n \; \forall \, n \in \Omega^N$.

$$\begin{aligned} \underset{X^{\tilde{m}}}{\text{minimise}} \quad & \sum_{k \in \Omega^K} w_k (X_k^{\tilde{m}} - T_k)^2 \\ \text{subject to} \quad & g(X^{\tilde{m}}) = 0 \\ & h(X^{\tilde{m}}) \leq 0 \\ & (1-\zeta) T_{\tilde{k}} \leq X_{\tilde{k}}^{\tilde{m}} \leq (1+\zeta) T_{\tilde{k}}, \end{aligned} \tag{3}$$

where w_k and ζ are weight and tolerance parameters. Each target T_n is computed as the mid-way point between consecutively ordered OCs $X_n^{(f)}, X_n^{(f+1)}$, such that $|X_n^{(f)} - X_n^{(f+1)}|$ is the largest Euclidean distance and (f) is the position of the fth largest OC $\forall \, n \in \Omega^N$. $T_{\tilde{k}} \in \mathbb{R}$ is the target computed at the maximum gap called the primary target. $\Omega^K \subset \Omega^N$ is a randomly selected subset of all input variables. $\Omega^{g'}$ is the set of infeasible OCs and $\Omega^{g''} = \Omega^g \cup \Omega^{g'}$. At each iteration, the algorithm updates the set of generated OCs $\Omega^g \longleftarrow \tilde{m}$ with the new OC if $X^{\tilde{m}}$ is physically feasible, otherwise updates the infeasible set $\Omega^{g'} \longleftarrow \tilde{m}$. The algorithm stops with a user-defined criteria $|\Omega^g| \leq S^g$, where $S^g \in \mathbb{N}$.

Algorithm 2 Generic Split-based Sampling

Require: $\Omega^g = \{\}$, $\Omega^{g'} = \{\}$, $\Omega^{g''} = \{\}$, S^g, w_k, ζ
while $|\Omega^g| \leq S^g$ **do**
  Sort $\Omega^{g''} \; \forall \, n \in \Omega^N$
  Compute $T_n \; \forall \, n \in \Omega^N : T_n = \frac{X_n^{(f)} + X_n^{(f+1)}}{2} + X_n$,
  Select $n = \tilde{k} : \max(|X_n^{(f)} - X_n^{(f+1)}|) \; \forall \, n \in \Omega^N$
  Randomly select $\Omega^K \subset \Omega^N$
  Solve optimisation (3)
  if $g(X^{\tilde{m}}) \neq 0$ **then**
    $\Omega^{g'} \longleftarrow \tilde{m}$
  else if $g(X^{\tilde{m}}) == 0$ **then**
    $\Omega^g \longleftarrow \tilde{m}$
  end if
  $\Omega^{g''} = \Omega^g \cup \Omega^{g'}$
end while

An advantage of *generic sampling* is that the set of generated feasible OCs Ω^g covers a much larger volume $\mathcal{V}$ of the feasible space as compared with the set of OCs Ω^s generated by the CSM in Section 2.1, $\mathcal{V}_{\Omega^g} \gg \mathcal{V}_{\Omega^s}$. $\mathcal{V}_\Omega$ is the volume covered by the OCs in Ω. This increase in volume results from the exploration of new OCs that are not presented in historical records. A limitation of *generic sampling* is that the set of generated OCs Ω^g is missing relevant information like dependency structures between variables. As the correlation information from the correlation matrix Q is not considered, many generated OCs may be irrelevant, either as probable OCs or in enhancing the discriminative information for the mapping $X \rightarrow \mathcal{Y}$.

2.3. State-of-the-art importance sampling approach

Importance sampling approaches focus on the *discriminative relevance* property of sampling quality and aim to maximise the information content for the security assessment $X \rightarrow \mathcal{Y}$, as the goal of ML-based security analysis is the correct prediction of security labels for OCs. These approaches assume the existence of an information-rich region (area around line D of Fig. 1) $\hat{\alpha} \subseteq \alpha$ as a subset of the feasible space α that can be explicitly specified [14,29] or obtained from initial OCs such that the entropy $\mathcal{E}$ of the set of OCs (e.g. Ω^s) is maximised [8,42]

$$\mathcal{E} = \sum_{i=1}^{b} -\Pi_{\Omega^s}^i \log_2 \Pi_{\Omega^s}^i \tag{4}$$

b is the number of disparate labels, usually $b = 2$ for secure and insecure labels. $\Pi_{\Omega^s}^i = \frac{|\Omega_i^s|}{|\Omega^s|}$ is the share of OCs that have label i, where $|\Omega_i^s|$ is the number of OCs in Ω^s with label i, i.e. $\Omega^s = \bigcup_{i=1}^b \Omega_i^s$. The maximisation of entropy allows *importance sampling* approaches to generate datasets according to the probability distribution of the security boundary area.

An advantage of *importance sampling* approaches is the generation of balanced datasets by sampling on both sides of the security decision boundary (see line D of Fig. 1) as the result of interpolating between secure and insecure OCs [42]. Otherwise, by fitting a multivariate distribution of feasible OCs within the secure feasible space and generating new OCs [15]. This crucial advantage ensures that the resulting database does not suffer from a class imbalance that can affect ML models' performance. A limitation of *importance sampling* is the exclusion of large regions of the feasible space to focus on a specific region of interest $\hat{\alpha} \subseteq \alpha$ in high-dimensional feasible spaces $\mathbb{R}^d, d \gg 1$. Additionally, *Importance sampling* does not consider the dependency structures of power system variables and can miss on relevant information like different operating modes and seasonality.

3. Proposed unified sampling approach

The proposed approach has two phases: a knowledge discovery phase of generating feasible and diverse pre-fault OCs X and a dataset enrichment phase to generate pre-fault OCs relevant for security assessment $X \rightarrow \mathcal{Y}$. In the knowledge discovery phase (phase A in Fig. 2), the proposed approach trades off copula-based historically relevant OCs and uniformly distributed OCs. This trade-off combines the two properties of maximising *coverage* while retaining *historical relevance*. In the dataset enrichment phase, the proposed approach identifies the most information-rich region of the feasible space to initialise new pre-fault OCs using entropy.

3.1. Knowledge discovery: trading off historical and rare OCs

The trade-off in the knowledge discovery phase between *historical relevance* and *coverage* of OCs minimises the maximum distance between the distribution of the generated OCs and a target probability distribution. Here, the target distribution is the historical distribution which retains the variable dependency structures, for instance, between generation and loads. The proposed Wasserstein distance helps to find a good trade-off by computing the distance between two probability measures.

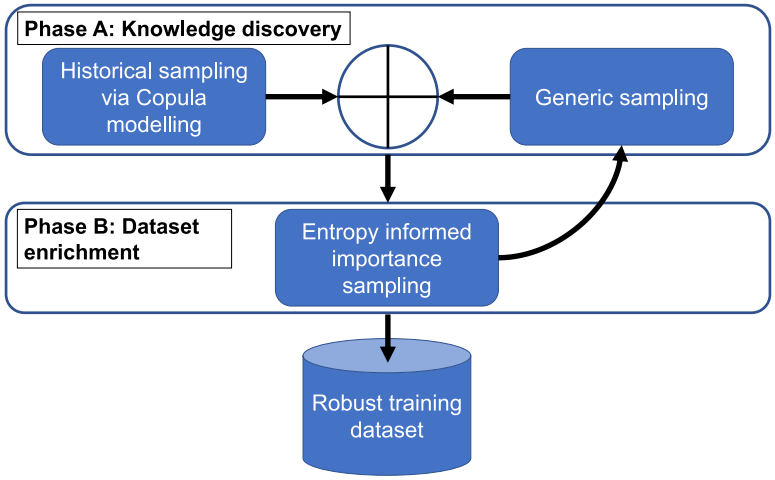


Fig. 2. Proposed unified sampling considering all three properties of *quality* datasets. Phase A combines generic and copula-based historical sampling to generate historically relevant and diverse pre-fault OCs. Phase B directs generating labels to entropy-rich regions.

Definition. We define $\mathcal{P}_p(\boldsymbol{X})\ \forall\ p \geq 1$ as the set of probability distributions $\mathcal{P}_p(\boldsymbol{X}) = \{\eta \in \mathcal{P}(\boldsymbol{X}) : \int_{\boldsymbol{X}} \|\boldsymbol{X}\|^p d\eta(\boldsymbol{X}) < +\infty\}$. The pth order Wasserstein distance $\mathcal{W}_p\ \forall\ \eta, \nu \in \mathcal{P}_p(\boldsymbol{X})$ is

$$\mathcal{W}_p(\eta,\nu) \equiv \inf_{\gamma\in\Gamma(\eta,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\boldsymbol{X}-\hat{\boldsymbol{X}}\|^p\, d\gamma(\boldsymbol{X},\hat{\boldsymbol{X}}) \tag{5}$$

where $\Gamma(\eta,\nu)$ is the set of all joint probability distributions on γ defined on $\mathbb{R}^d \times \mathbb{R}^d$ with respective marginal distributions η and ν. The Wasserstein distance has a closed form of

$$\mathcal{W}_p = \left(\int_0^1 |F_\eta^{-1} - F_\nu^{-1}|^p d\gamma(\eta,\nu)\right)^{1/p} \tag{6}$$

for one-dimensional measures, where F_η and F_ν represent the respective cumulative distributions of η and ν. We assume two sets of OCs, $\underline{\Omega}^h$ and $\overline{\Omega}^h$ with data matrices $\underline{X}$ and $\overline{X}$, where the data $\underline{X}^m$ corresponds to $m \in \underline{\Omega}^h$ (and equivalently for $\overline{\Omega}^h$). The Wasserstein distance between the corresponding data $\underline{X}$ and $\overline{X}$ is

$$\mathcal{W}_2(\underline{\Omega}^h, \overline{\Omega}^h) = \min_{\gamma\in\Gamma(\underline{X},\overline{X})} \sum_{\underline{m}=1}^{|\underline{\Omega}^h|} \sum_{\overline{m}=1}^{|\overline{\Omega}^h|} \gamma_{\underline{m},\overline{m}}\, D(\underline{X}^{\underline{m}}, \overline{X}^{\overline{m}}), \tag{7}$$

where γ is a joint distribution over the two matrices, $\Gamma(\underline{X},\overline{X})$ is the set of all joint distributions, and $D(\underline{X}^{\underline{m}}, \overline{X}^{\overline{m}})$ is the Euclidean distance between the $\underline{m}$-th row of $\underline{X}$ and the $\overline{m}$-th row of $\overline{X}$. Here, the second-order Wasserstein distance is the sum of the distances between each pair of OCs multiplied by the amount of probability mass that must be moved and effectively measures the "distance" between the two matrices of OCs in terms of how much "work" must be done to transform one matrix of OCs into the other. An example in Fig. 3 visualises the Wasserstein distance between two probability distributions, $P(\overline{X})$ and $P(\underline{X})$. The larger the Wasserstein distance between any two probability distributions, the more dissimilar they are. The proposed Wasserstein distance thus allows comparing the probability distributions of synthetically generated OCs Ω^A with historical data Ω^h. This comparison ensures that the distribution of synthetic OCs does not deviate beyond a user-defined threshold distance $\lambda \in \mathbb{N}$ from the historical data distribution, ensuring the dependency between power loads and generators is preserved. This proposal thus allows synthetic OCs to retain the two properties of *historical relevance* and *coverage*.

Algorithm 3 presents the proposed trade-off between *coverage* and *historical relevance* using the proposed Wasserstein distance. The algorithm starts with an empty set of generated OCs $\Omega^A = \{\}$. Subsequently, a copula-based model $\hat{\boldsymbol{X}}$ generates historically relevant OCs Ω^s based on available historical data Ω^h. Only the OCs $\tilde{m} \in \Omega^s$ that satisfy the power system constraints $g(X^{\tilde{m}}) = 0, h(X^{\tilde{m}}) \leq 0$ (e.g., nodal balance, line flow, voltage, phase angle, and generator limits) are considered and serve as a baseline target distribution. The OCs that satisfy the power system constraints populate the set $\Omega^A \longleftarrow \Omega^s$. The algorithm sorts the set Ω^A in ascending order for each variable. Then, for a pre-defined Wasserstein distance threshold $\lambda \in \mathbb{N}$, inserts new "rare" OCs using the CSM in Algorithm 1 such that $\mathcal{W}_2(\Omega^s, \Omega^A) \leq \lambda$, which substitutes the stopping criterion that considers S^g in Algorithm. 2. λ is a parameter where the value is selected by the user based on acceptable trade-offs.

Algorithm 3 Proposed Unified Sampling: phase A

Require: $\Omega^A = \{\}$, $\Omega^s = \{\}$, λ, Ω^h, $\hat{\boldsymbol{X}}$
1: Execute Alg. (1) to generate set of OCs Ω^s using $\hat{\boldsymbol{X}}$
2: $\Omega^A \longleftarrow \tilde{m} : g(X^{\tilde{m}}) = 0, h(X^{\tilde{m}}) \leq 0\ \forall\ \tilde{m} \in \Omega^s$
3: **while** $\mathcal{W}_2(\Omega^s, \Omega^A) \leq \lambda$ **do**
4: Execute Alg. (2) to generate set of OCs Ω^g
5: $\Omega^A \longleftarrow \Omega^g$
6: **end while**

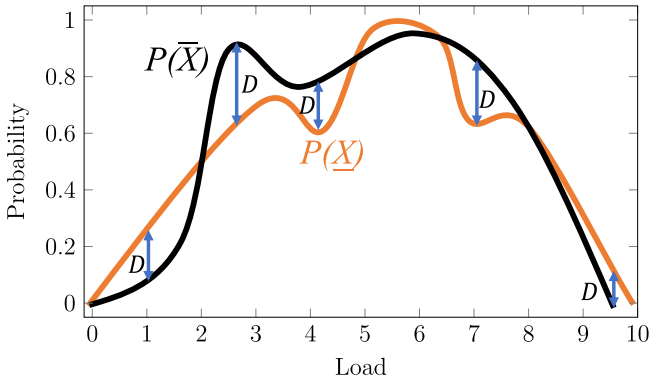


Fig. 3. Wasserstein distance between two probability distributions $P(\overline{X})$ and $P(\underline{X})$ is analogous to the amount of 'work' to transform one distribution into another. 'Work' is the distance moved multiplied by the probability 'mass' at that distance.

3.2. Dataset enrichment and labelling: importance sampling in the context of feasibility

The dataset enrichment and labelling phase of the proposed approach incorporates an entropy-informed re-sampling of pre-fault OCs

Algorithm 4 Proposed Unified Sampling: phase B

Require: Ω^A, $\Omega^B = \{\}$, S^B, K
1: Segment Ω^A into K disjoint clusters
2: Identify $\Omega^A_{\hat{k}}$: $\mathcal{E}_{\hat{k}} = \max(\mathcal{E}_1, \cdots, \mathcal{E}_K)$
3: Compute $[\underline{\rho}_n, \overline{\rho}_n], \forall\, n \in \Omega^N$ for cluster $\Omega^A_{\hat{k}}$ with $\mathcal{E}_{\hat{k}}$
4: $\Omega^B \longleftarrow \Omega^A_{\hat{k}}$
5: **while** $|\Omega^B| \leq S^B$ **do**
6: Execute Alg. (2) : $\underline{\rho}_n \leq X^{\tilde{m}}_n \leq \overline{\rho}_n, \forall\, n \in \Omega^N$
7: $\Omega^B \longleftarrow \Omega^g$
8: **end while**

shown as phase B in Fig. 2. This phase aims to improve the knowledge discovery phase by focusing on information-rich regions of the feasible space.

Algorithm 4 presents phase B of the proposed unified sampling approach. After the initial generation of $|\Omega^A|$ pre-fault OCs Ω^A using Algorithm 3, we perform dynamic simulations $X \rightarrow \mathcal{Y}$ to obtain the security labels

$$\mathcal{Y}^m_j = \{0, 1\},\ \forall\, m \in \Omega^A,\ \forall\, j \in \Omega^J \tag{8}$$

for the set of probable contingencies Ω^J. Subsequently, for each contingency $j \in \Omega^J$, the algorithm segments the OCs in Ω^A using K-means clustering into $K \in \mathbb{N}$ clusters. An example in Fig. 4 illustrates this phase of the proposed approach on three clusters $K = 3$. The set of OCs

$$\Omega^A = \bigcup_{\forall k \in \Omega^C} \Omega^A_k \tag{9}$$

is segregated into K distinct clusters where Ω^A_k is the set of OCs in Ω^A belonging to cluster k and Ω^C is the set of all clusters $|\Omega^C| = K$. In Fig. 4, the circles show the entailing OCs Ω^A_k belonging to each cluster $k = 1, 2, 3$. Each of these clusters k has an entropy $\mathcal{E}_k$ computed as in Eq. (4) substituting $\Omega^s = \Omega^A_k$. The cluster $\hat{k}$ is the cluster that has the highest entropy

$$\mathcal{E}_{\hat{k}} = \max(\mathcal{E}_k |\ \forall\, k \in \Omega^C) \tag{10}$$

with OCs $\Omega^A_{\hat{k}}$. Subsequently, to improve the dataset generated from the knowledge discovery phase Ω^A, we focus on data $X^m_n, m \in \Omega^A_{\hat{k}}$ and compute the bounds of the cluster with the maximum entropy $\Omega^A_{\hat{k}} \subset \Omega^A$ to form a hypercube.

$$\underline{\rho}_n = \min(X^m_n |\ \forall\, m \in \Omega^A_{\hat{k}})\ \forall\, n \in \Omega^N \tag{11}$$

$$\overline{\rho}_n = \max(X^m_n |\ \forall\, m \in \Omega^A_{\hat{k}})\ \forall\, n \in \Omega^N \tag{12}$$

Using the bounds $[\underline{\rho}_n, \overline{\rho}_n], \forall\, n \in \Omega^N$ as additional inequality constraints

$$\underline{\rho}_n \leq X^{\tilde{m}}_n \leq \overline{\rho}_n, \forall\, n \in \Omega^N, \tag{13}$$

in optimisation (3), we generate new pre-fault OCs using Algorithm 2. The sampling stops after generating a user-defined S^B OCs. The final training dataset is then augmented by the pre-fault OCs Ω^B and their respective labels $\mathcal{Y}^m_j = \{0, 1\}, m \in \Omega^B, j \in \Omega^J$ obtained using dynamic simulations for each contingency.

4. Case study

This section examines the effectiveness of the proposed unified sampling approach in generating historically relevant, representative, and balanced datasets for training ML models for security assessment. The first study investigates the trade-off between *historical relevance* and *coverage*. The second and third studies focus on the *historical relevance* and *coverage* metrics, respectively. The fourth and fifth studies analyse the results of security assessments using ML models trained on different databases. The final study presents the results of balancing the label distribution.

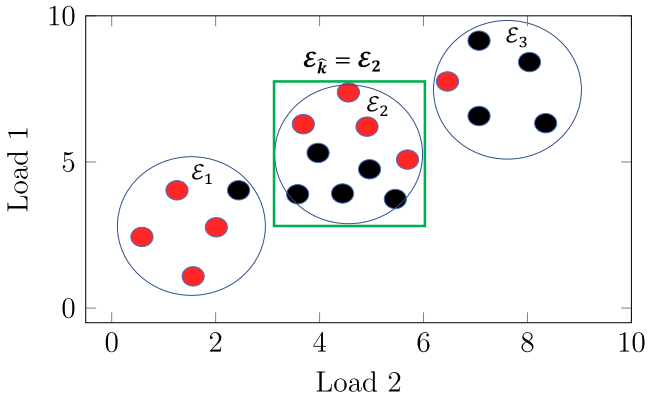


Fig. 4. Proposed phase B computes the entropy of clusters with OCs that are secure (●) and insecure (●). The cluster with the largest entropy initialises the bounds for resampling using Algorithm 2.

4.1. Test system and assumptions

The case studies use the IEEE 118-bus system [43] shown in Fig. 5, where the observed historical data contain 14.250 measurements at 5 min intervals for a period in 2012 provided by the French Transmission SO, RTE. The original dataset spanned over 7,000 load points and 200 wind turbines and was projected relative to the upper limit in the snap-shot onto the IEEE 118-bus test system [44]. We consider a broader and generic definition of uncertainty to represent noise that can be present in historical data. However, noise can also be explicitly considered in our approach. For generating the pre-fault OCs, a DC approximation (of the optimal power flow) is sufficient to demonstrate the challenge and proposed solution. Therefore, we use only active power loads to model and validate the compared approaches. If required, reactive power loads can be similarly modelled. We considered $\mathcal{L} = 10$ disjoint clusters to capture different operation modes. Subsequently, we generate 5,000 OCs (representing loads) using all the approaches in contention. The baselines are the historical model (HM), CSM, and GSM, against the proposed unified sampling model (USM). The HM is data from historical records, the CSM, GSM and the proposed USM (A) and USM (B) approaches are as described in Algorithms 1, 2, 3, 4 respectively. We study USM (A) and USM (B) separately, denoting phases A and B. $K = 10$ clusters were used when studying USM (B) ($K = 10$ was also selected in [13]). The observed historical data were randomly split into training, and testing sets in the ratio of 80:20, and the training set was used to build the CSM.

For the transient studies, a three-phase fault is simulated at bus 12, 15, 49 and 80 for all OCs at time 0.5 s. The fault is cleared by opening the line between buses 12 and 14 after 0.2 s. The transient stability was analysed for 10 s. The post-fault OC was considered as secure ($\mathcal{Y}^m_j = 0$) if the difference between any two generator phase angles is less than 180°, otherwise, the OC is considered insecure ($\mathcal{Y}^m_j = 1$).

For the security assessment, decision trees (DTs) were trained using the CART algorithm [45]. The training settings were set to their default values, except for using Gini impurity to measure the quality of splits instead of entropy and limiting the maximum depth of the trees to 5 (similar as in [46]). The data was split into a training set and a testing set in a 75: 25 ratio, with the feature variable X and the labels $\mathcal{Y}$ serving as the inputs for training the classifier. To address underfitting or overfitting, 10-fold cross-validation was applied and one DT was trained for each contingency $\forall j \in \Omega^J$. We consider the F1-score $= \frac{2Tp}{2Tp+Fp+Fn}$ to measure the test accuracy of the DTs, where Tp, Fp, Fn are the true positives, false positives and false negatives, respectively. Testing data can be used to assess if a trained model performs well. If the performance metric on the testing data is high, we infer that enough data has been used for training. The standard nonparametric two-sample tests from the literature are used to measure *historical relevance*,

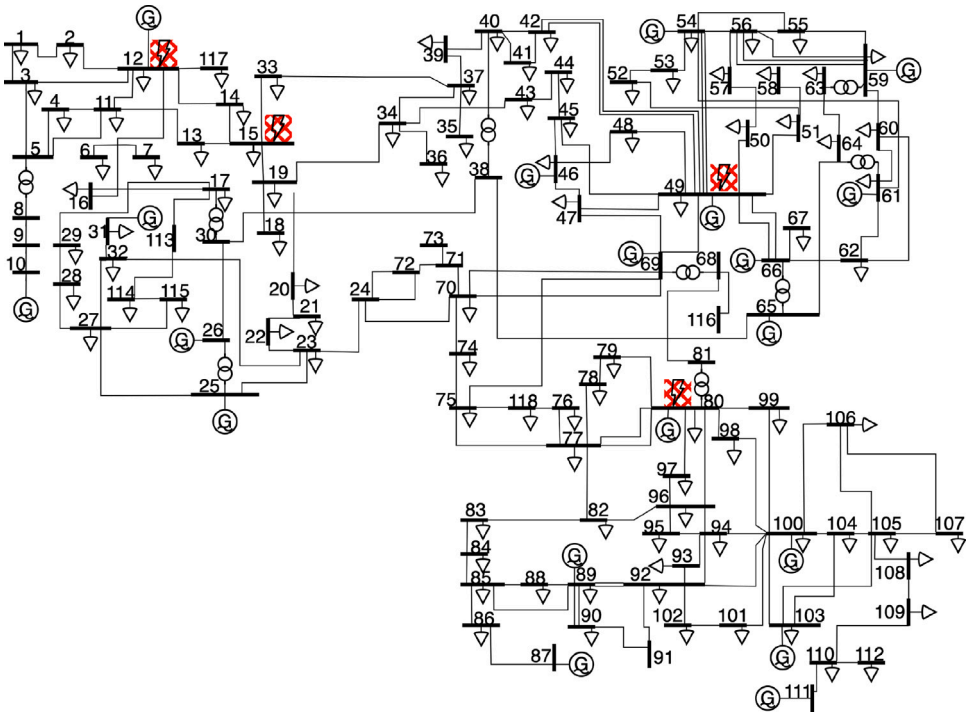


Fig. 5. Single line diagram of the IEEE 118 bus test system showing fault locations at buses 12, 15, 49 and 80.

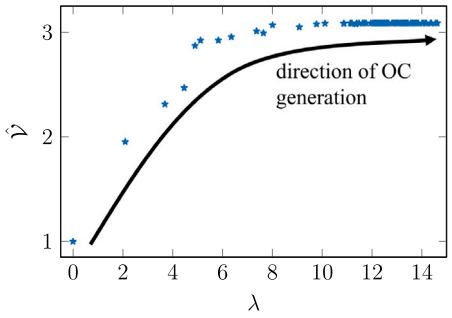


Fig. 6. The relationship between relative volume $\hat{\mathcal{V}}$ and Wasserstein distance threshold λ. $\hat{\mathcal{V}}$ increases with the number of generated OCs and plateaus around $\lambda \geq 10$.

the Kolmogorov–Smirnov (K–S) test and the multivariate energy test. The K–S test investigates whether the generated data from $\hat{\boldsymbol{X}}$ can reconstruct the marginal distributions of the true distribution, while the energy tests show how much variable dependency is maintained in the generated dataset. Finally, The convex hull volume $\mathcal{V}$ occupied by the generated dataset is used to compare the *coverage* of OCs generated by the three models.

All optimisation problems were implemented using the package Pyomo 5.6.8 [47] in Python 3.7.4, and the DC approximation was solved using Gurobi 9.5.0 [48] while using IPOPT 3.13.2 [49] for the AC models of the networks. The DTs were trained with the scikit-learn package version 0.18.1 [50]. The ODEs were solved to simulate the transients using *odeint* in scipy. All studies were conducted on a standard Windows HP desktop running an Intel(R) processor with 64 GB of RAM.

4.2. Trading-off historical relevance and coverage

This section studies the trade-off between *historical relevance* and *coverage* properties of the proposed USM. This trade-off is achieved via the proposed Wasserstein distance $\mathcal{W}_p$. By adjusting the threshold λ of the acceptable Wasserstein distance $\mathcal{W}_p$, the proposed USM can generate new OCs that explore the feasible space.

The result in Fig. 6 shows the coverage metric, $\hat{\mathcal{V}}$, and the threshold λ on the Wasserstein metric, which measures similarity to historical data as new samples are generated. λ plateaus around $\lambda \approx 10$. As a result, subsequent case studies consider the USM approach until $\lambda \leq 10$. However, note that $\lambda \leq 10$ is valid only for this case study and would need to be calculated for other test systems and datasets based on the defined objectives. Fig. 7(a) shows many experiment variations, specifically, the distribution of 1,000 $\mathcal{W}_p$ tests between 250 randomly selected OCs from the observed historical dataset Ω^h and the baseline approaches, the proposed USM, CSM, and GSM. The closer the averages are to 0, the closer the distributions are to the historical distribution. The proposed USM is between CSM and GSM, balancing these two approaches by varying the threshold λ of the acceptable Wasserstein distance $\mathcal{W}_p$ to the target distribution. This trade-off allows the proposed USM to generate historically relevant datasets that sufficiently cover the feasible space simultaneously.

4.3. Historical relevance

This case study investigates the *historical relevance* of the proposed USM (A) and the baseline approaches (HM, CSM, and GSM) using standard statistical two-sample tests. The study randomly generates 1,000 sets of 0.5 % of the generated data and compares them to a set of randomly generated 0.5 % of observed historical test data. The distribution of the p-values from the 1,000 energy tests and 1,28,000 K–S tests for each approach are compared and presented in Figs. 8(a) & 8(b), respectively. For both tests, the null hypothesis assumes data from any two disparate approaches in comparison come from the same model and follow a similar distribution. Therefore, the p-values should be uniformly distributed. As a baseline for comparison, the solid black line in Figs. 8(a) & 8(b) represents the CDF of the p-values that compare data randomly drawn from the observed historical training dataset and the test data. The larger the maximum difference between the CDFs of p-values, the more dissimilar the two datasets are.

The results show that the CSM approach outperforms GSM, as its p-values in the K–S and energy tests come closest to the solid black

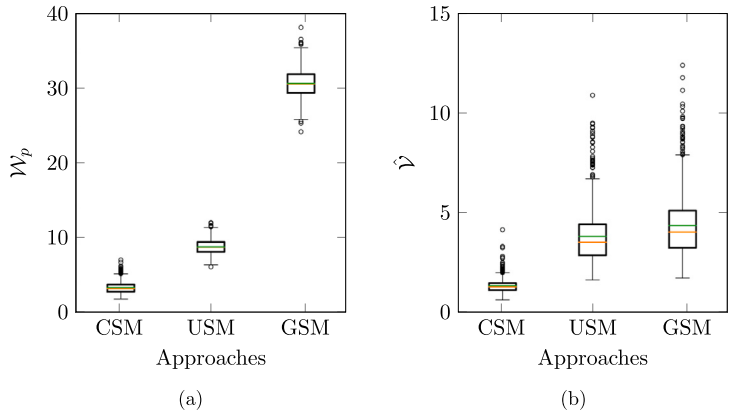


Fig. 7. Boxplots showing mean (—) and median (—) values of 1,000 randomly selected OCs from different approaches corresponding to (a) the proposed $\mathcal{W}_p$ and (b) the normalised $\mathcal{V}$.

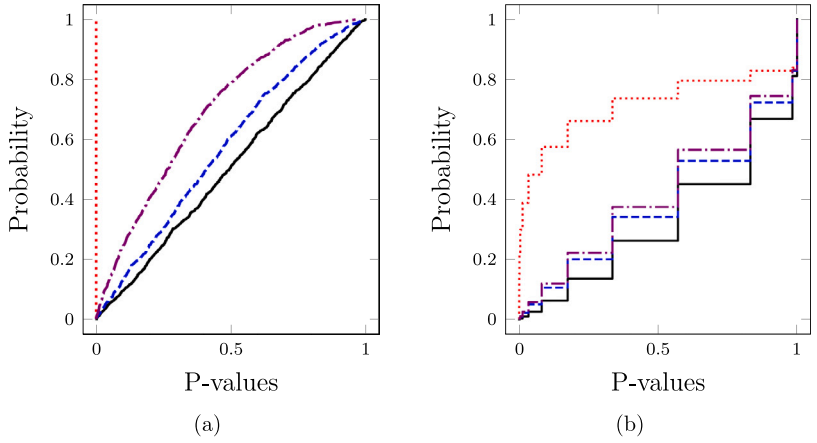


Fig. 8. CDF of p-values corresponding to historical data (—), CSM (- - -), GSM (- - -), and the proposed USM (A) (—) for (a) energy tests and (b) K–S tests.

line of HM. This result is consistent with previous research as CSM can effectively model and capture marginal distributions and dependency structure of variables in higher dimensions. GSM has all 0 p-values for all tests and does not capture the marginal distribution as well, as evidenced by the K–S tests in Fig. 8(b). Notably, in Figs. 8(a) & 8(b), the proposed USM (A) preserves most of the marginal distributions while maintaining a significant variable dependency information, even though the CSM outperforms in that regard.

4.4. Coverage

This case study assesses the *coverage* of generated datasets from the proposed USM (A), CSM and GSM. The results shown in Fig. 7(b) present the volume $\mathcal{V}$ covered by the generated OCs of the different approaches for 1,000 different realisations of variable selection, $\hat{\Omega}^N \subset \Omega^N$, with a cardinality of $|\hat{\Omega}^N| = 3$. The $\mathcal{V}$ values in the figure are normalised by the minimum volume value computed using the historical data, Ω^h, for the selected $\hat{\Omega}^N$ variables, such that $\hat{\mathcal{V}} = \frac{\mathcal{V}}{\min(\mathcal{V}_{\Omega^h})}$. Specifically, the figure shows that the datasets generated by the proposed USM (A) and GSM have similar volume coverage, which on average is significantly larger by as much as 4× more volume than that of CSM-generated datasets. Notably, the proposed USM approach presents the best trade-off among the tested approaches as it covers nearly the same volume as the GSM (90 %), while also providing the additional benefit of high historical relevance, as studied in Section 4.3.

4.5. Security assessment for out-of-distribution OCs

This study compares the performance of ML models trained on datasets from the baseline approaches and tests the models on three types of OCs: historical, generic (for data uniformly covering the feasible space) and rare (for data deviating from typical historical distribution).

Table 1
Results for contingency representing a three-phase fault at bus 12.

Training data	Testing data (F1-score)			Π^1
	historical	generic	rare	
HM	0.89 ± 0.02	0.60 ± 0.21	0.55 ± 0.30	0.14
CSM	0.84 ± 0.04	0.56 ± 0.20	0.41 ± 0.23	0.13
GSM	0.76 ± 0.05	0.99 ± 0.01	0.71 ± 0.09	0.15
USM (A)	0.88 ± 0.03	0.96 ± 0.01	0.98 ± 0.01	0.17

Table 2
Results for contingency representing a three-phase fault at bus 15.

Training data	Testing data (F1-score)			Π^1
	historical	generic	rare	
HM	0.88 ± 0.02	0.49 ± 0.20	0.44 ± 0.15	0.13
CSM	0.82 ± 0.02	0.71 ± 0.16	0.66 ± 0.17	0.14
GSM	0.75 ± 0.07	0.97 ± 0.02	0.69 ± 0.11	0.15
USM (A)	0.86 ± 0.02	0.96 ± 0.02	0.98 ± 0.01	0.18

The results in Tables 1–4 show the average and standard deviation of F1-scores for 100 DT models, where the training and testing data come from the different baseline approaches. The 'rare' OCs in the testing set were randomly selected from data generated using the proposed USM (A). The models trained on the proposed USM (A) data achieved an average F1-score of at least 96 % on these 'rare' OCs, outperforming models trained on data from HM and CSM by 54 % and 57 %, respectively, in contingencies 15 and 12. In contrast, models trained on data from HM and CSM had the worst performance on these

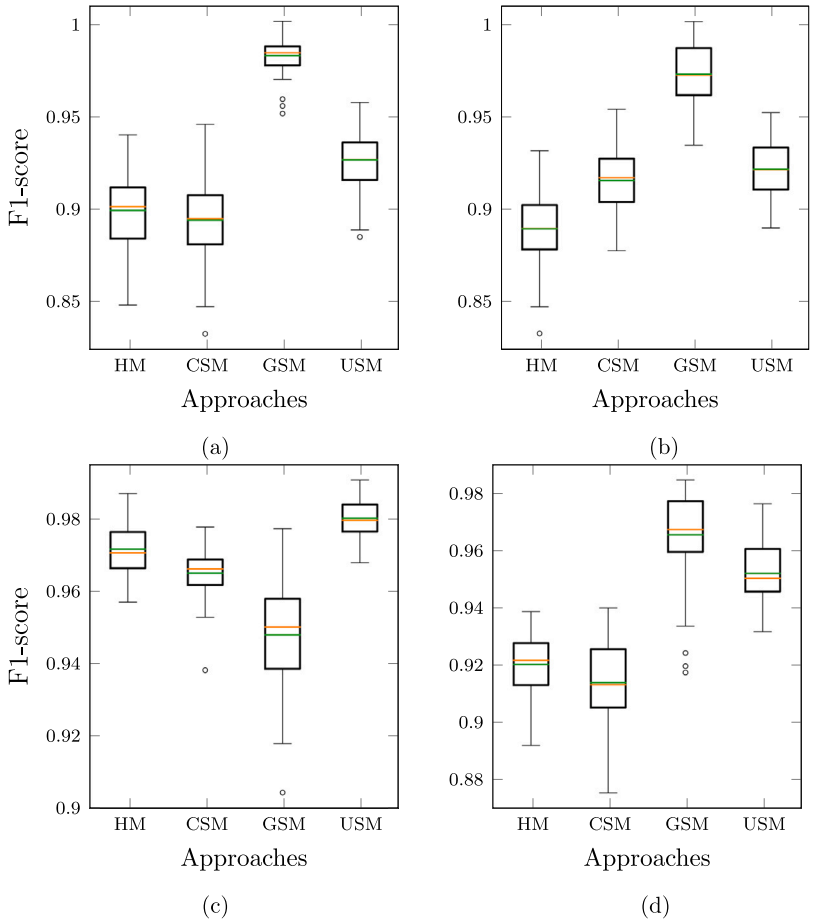


Fig. 9. Boxplots showing mean (—) and median (—) values for 100 models trained with different datasets to predict the post-fault status of a three-phase fault near buses (a) 12 (b) 15 (c) 49 (d) 80. Both training and testing data follow a similar distribution.

'rare' testing OCs as this training data does not consider such rare cases and is biased towards specific data from HM and CSM (e.g. does not generalise well to data from USM (A) that we consider 'rare'). GSM aims to cover the feasible space uniformly, and the DTs trained on this data performs well in some contingencies (e.g., 49 and 80) but still does underperform when using data from the proposed USM (A) for the DTs. These results suggest that the proposed USM (A) can support the development of models that generalise better to OCs from other distributions. For further comparisons, Table 1 presents an example for contingency 12, where the DT models trained on datasets generated by the proposed USM (A) achieved a high accuracy (F1-score of at least 86 %), outperforming models trained on data from CSM and GSM in predicting uniformly distributed and historical OCs, respectively, by as much as 40 % and 12 %. A similar analysis can be made for the other three contingencies in Tables 2–4.

The proposed USM (A) datasets resulted in DT models that are (nearly) as accurate as DTs models with training and testing data from the same distribution. For example, the USM (A) based DT models were within 2 % accuracy of the HM and CSM-based DTs in contingencies 12 and 15. Also, the F1-scores tested on historical HM data on contingencies 49 and 80 showed USM (A)-based DTs are as high as HM-based DTs (Table 4). In comparison, models trained on GSM datasets have an accuracy within 13 % when tested on the same historical data. Additionally, models trained on USM (A) datasets were found to have a maximum deviation of 4 % accuracy from the best model performance in all contingencies when tested on generic sampling data. In contrast, models from HM or CSM can have an accuracy deviation of up to 48 % in all contingencies when tested on the same data. Importantly, these results show that the proposed USM (A) performs with high accuracy across all tested datasets, showing a high level of generalisability to data from other distributions within our test settings.

Table 3
Results for contingency representing a three-phase fault at bus 49.

Training data	Testing data (F1-score)			Π^1
	historical	generic	rare	
HM	0.97 ± 0.01	0.65 ± 0.16	0.60 ± 0.14	0.47
CSM	0.97 ± 0.01	0.64 ± 0.15	0.66 ± 0.17	0.46
GSM	0.93 ± 0.03	0.95 ± 0.01	0.93 ± 0.05	0.30
USM (A)	0.98 ± 0.01	0.91 ± 0.01	0.99 ± 0.01	0.48

Table 4
Results for contingency representing a three-phase fault at bus 80.

Training data	Testing data (F1-score)			Π^1
	historical	generic	rare	
HM	0.92 ± 0.01	0.53 ± 0.21	0.46 ± 0.18	0.27
CSM	0.91 ± 0.01	0.58 ± 0.35	0.56 ± 0.25	0.27
GSM	0.83 ± 0.06	0.97 ± 0.01	0.80 ± 0.09	0.18
USM (A)	0.92 ± 0.01	0.94 ± 0.01	0.96 ± 0.02	0.29

4.6. Security assessment for similar distribution OCs

This study examines the performance of ML models trained and tested on data from similar distributions.

The results in Fig. 9 show the distribution of F1-scores for 100 DT models trained using different datasets to predict the post-fault status of the system following four separate three-phase faults. The results show that all databases can provide input to train accurate models

Table 5
Performance of USM (A) and USM (B) approaches for contingencies representing a three-phase fault at buses 12 and 15.

Approach	Contingency 12		Contingency 15	
	F1-score	Π^1	F1-score	Π^1
USM (A)	0.92 ± 0.02	0.17	0.91 ± 0.02	0.18
USM (B)	0.94 ± 0.01	0.27	0.95 ± 0.01	0.26

for testing data from the same distribution, with an F1-score $\geq 88\,\%$. The results indicate that for contingencies with a more balanced share of labels, such as the fault on bus 49 where the label distribution $\Pi^1_{\Omega^A} = 0.48$, the proposed USM (A) outperforms the other approaches (HM, CSM and GSM). However, for other contingencies, the model trained on GSM datasets has better performance, as the OCs are more uniformly distributed. Notably, the label distribution for GSM datasets for contingency 49 is $\approx 34\,\%$ lower than the other approaches (assuming an ideal distribution $\Pi^1 = 0.5$) which explains the relatively poor performance compared to the other models, albeit with an F1-score $\approx 95\,\%$.

4.7. Balancing the distribution

This section studies the performance of USM (B) as the share of labels Π^1 could impact the performance of models and is thus an important factor to consider when generating training OCs (as also Section 4.5 showed). Therefore, this study focuses on contingencies where the share of labels $\Pi^1 \ll 0.5$, namely contingencies 12 and 14. USM (B) is limited to generate $20\,\%$ of $|\Omega^A|$, e.g., $S^B = 0.2|\Omega^A|$.

The results in Table 5 show the performance of 100 DT models trained with data from the two approaches, USM (A) and USM (B). The results show that in contingency 12, the share of labels improved from $\Pi^1_{\Omega^A} = 0.17$ to $\Pi^1_{\Omega^B} = 0.27$ to and the corresponding F1-score from 0.92 to 0.94. Similarly in contingency 15, the share of labels improved from $\Pi^1_{\Omega^A} = 0.18$ to $\Pi^1_{\Omega^B} = 0.26$. Also, this improved label share improved the F1-score from 0.91 to 0.95.

5. Conclusion

A crucial challenge is generating high-quality datasets for training machine-learned models for real-time security assessment in power systems. Conventional approaches have failed to generate datasets that generalise well beyond similar distributions, resulting in models that are not always accurate. To overcome this challenge, we proposed a novel, unified approach for generating datasets that balance *historical relevance, coverage*, and *discriminative relevance*. Our approach balances historically relevant operating conditions (OCs) with rare but feasible OCs, leading to datasets representing the full range of possible OCs. Experimental results on the IEEE 118-bus system demonstrate the effectiveness of our approach. Our model trained on data generated using our approach achieved an F1-score of 91% for different contingencies and 96% accuracy in predicting the security label of rare OCs, outperforming baseline approaches. We believe our work is an important step towards developing new tools that enable the adoption of machine learning for sensitive tasks such as security assessment in power systems. Future work will investigate generating representative samples for training machine learning security estimators by additionally considering the discrete variable space, and a broader range of fault scenarios and locations.

CRediT authorship contribution statement

Al-Amin B. Bugaje: Conceptualization, Methodology, Data curation, Writing – original draft, Visualization, Investigation, Formal analysis, Software, Validation. **Jochen L. Cremer:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Goran Strbac:** Writing – review & editing, Project administration, Funding acquisition, Resources.

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.

Data availability

Data will be made available on request.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of the abstract of this work, the authors used the tool ChatGPT developed by OpenAI to improve the readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Acknowledgements

This work was supported by a scholarship funded by the Nigerian National Petroleum Corporation, NG, the TU Delft AI Labs Programme, NL, and the research project IDLES, Engineering and Physical Sciences Research Council, UK (EP/R045518/1).

References

[1] Kroposki B, Johnson B, Zhang Y, Gevorgian V, Denholm P, Hodge B-M, et al. Achieving a 100% renewable grid: Operating electric power systems with extremely high levels of variable renewable energy. IEEE Power Energy Mag 2017;15(2):61–73.
[2] Panciatici P, Bareux G, Wehenkel L. Operating in the fog: Security management under uncertainty. IEEE Power Energy Mag 2012;10(5):40–9. http://dx.doi.org/10.1109/MPE.2012.2205318.
[3] Hatziargyriou N, Milanovic J, Rahmann C, Ajjarapu V, Canizares C, Erlich I, et al. Definition and classification of power system stability – Revisited & extended. IEEE Trans Power Syst 2021;36(4):3271–81. http://dx.doi.org/10.1109/TPWRS.2020.3041774.
[4] Konstantelos I, Jamgotchian G, Tindemans SH, Duchesne P, Cole S, Merckx C, et al. Implementation of a massively parallel dynamic security assessment platform for large-scale grids. IEEE Trans Smart Grid 2017;8(3):1417–26. http://dx.doi.org/10.1109/TSG.2016.2606888.
[5] Duchesne L, Karangelos E, Wehenkel L. Recent developments in machine learning for energy systems reliability management. Proc IEEE 2020;108(9):1656–76. http://dx.doi.org/10.1109/JPROC.2020.2988715.
[6] Marot A, Donnot B, Chaouache K, Kelly A, Huang Q, Hossain R-R, et al. Learning to run a power network with trust. Electr Power Syst Res 2022;212:108487.
[7] Bugaje A-AB, Cremer JL, Strbac G. Real-time transmission switching with neural networks. IET Gener Transm Distrib 2023;17(3):696–705.
[8] Krishnan V, McCalley JD, Henry S, Issad S. Efficient database generation for decision tree based power system security assessment. IEEE Trans Power Syst 2011;26(4):2319–27.
[9] Zhang T, Sun M, Cremer JL, Zhang N, Strbac G, Kang C. A confidence-aware machine learning framework for dynamic security assessment. IEEE Trans Power Syst 2021.
[10] Bugaje A-AB, Cremer JL, Sun M, Strbac G. Selecting decision trees for power system security assessment. Energy AI 2021;6:100110.
[11] Liu Y, Shi X-J, Xu Y. A hybrid data-driven method for fast approximation of practical dynamic security region boundary of power systems. Int J Electr Power Energy Syst 2020;117:105658.
[12] Bellizio F, Bugaje A-AB, Cremer JL, Strbac G. Verifying machine learning conclusions for securing low inertia systems. Sustain Energy Grids Netw 2022;30:100656.
[13] Konstantelos I, Sun M, Tindemans SH, Issad S, Panciatici P, Strbac G. Using vine copulas to generate representative system states for machine learning. IEEE Trans Power Syst 2018;34(1):225–35.
[14] Thams F, Venzke A, Eriksson R, Chatzivasileiadis S. Efficient database generation for data-driven security assessment of power systems. IEEE Trans Power Syst 2019;35(1):30–41.
[15] Venzke A, Molzahn DK, Chatzivasileiadis S. Efficient creation of datasets for data-driven power system applications. Electr Power Syst Res 2021;190:106614.
[16] Joswig-Jones T, Baker K, Zamzam AS. OPF-learn: An open-source framework for creating representative AC optimal power flow datasets. In: 2022 IEEE power & energy society innovative smart grid technologies conference. IEEE; 2022, p. 1–5.

[17] Zhu L, Hill DJ, Lu C. Semi-supervised ensemble learning framework for accelerating power system transient stability knowledge base generation. IEEE Trans Power Syst 2021;37(3):2441–54.
[18] Zhu L, Hill DJ. Data/model jointly driven high-quality case generation for power system dynamic stability assessment. IEEE Trans Ind Inf 2021;18(8):5055–66.
[19] Ren C, Xu Y. A fully data-driven method based on generative adversarial networks for power system dynamic security assessment with missing data. IEEE Trans Power Syst 2019;34(6):5044–52.
[20] Pournabi M, Mohammadi M, Afrasiabi S, Setoodeh P. Power system transient security assessment based on deep learning considering partial observability. Electr Power Syst Res 2022;205:107736.
[21] Rossi F, Araujo EP, Mañe MC, Bellmunt OG. Data generation methodology for machine learning-based power system stability studies. In: 2022 IEEE PES innovative smart grid technologies conference Europe. IEEE; 2022, p. 1–5.
[22] Li Y, Zhang M, Chen C. A deep-learning intelligent system incorporating data augmentation for short-term voltage stability assessment of power systems. Appl Energy 2022;308:118347.
[23] Han G, Liu S, Chen K, Yu N, Feng Z, Song M. Imbalanced sample generation and evaluation for power system transient stability using ctgan. In: Intelligent computing & optimization: proceedings of the 4th international conference on intelligent computing and optimization 2021, vol. 3. Springer; 2022, p. 555–65.
[24] Nadal IV, Chevalier S. Scalable bilevel optimization for generating maximally representative OPF datasets. 2023, arXiv preprint arXiv:2304.10912.
[25] Mollaiee A, Ameli MT, Azad S, Nazari-Heris M, Asadi S. Data-driven power system security assessment using high content database during the COVID-19 pandemic. Int J Electr Power Energy Syst 2023;150:109077.
[26] Sevilla FRS, Liu Y, Barocio E, Korba P, Andrade M, Bellizio F, et al. State-of-the-art of data collection, analytics, and future needs of transmission utilities worldwide to account for the continuous growth of sensing data. Int J Electr Power Energy Syst 2022;137:107772.
[27] Liu Y, Wang J, Yue Z. Improved multi-point estimation method based probabilistic transient stability assessment for power system with wind power. Int J Electr Power Energy Syst 2022;142:108283.
[28] Papadopoulos PN, Milanović JV. Probabilistic framework for transient stability assessment of power systems with high penetration of renewable generation. IEEE Trans Power Syst 2016;32(4):3078–88.
[29] Yan R, Geng G, Jiang Q. Data-driven transient stability boundary generation for online security monitoring. IEEE Trans Power Syst 2020;36(4):3042–52.
[30] Bugaje A-AB, Cremer JL, Strbac G. Split-based sequential sampling for realtime security assessment. Int J Electr Power Energy Syst 2023;146:108790.
[31] Wang C, Sharifnia E, Gao Z, Tindemans SH, Palensky P. Generating multivariate load states using a conditional variational autoencoder. Electr Power Syst Res 2022;213:108603.
[32] Bernton E, Jacob PE, Gerber M, Robert CP. Approximate Bayesian computation with the wasserstein distance. J R Stat Soc Ser B Stat Methodol 2019;81(2):235–69.
[33] Villani C. Optimal transport: old and new, vol. 338. Springer; 2009.
[34] Jiang B. Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In: International conference on artificial intelligence and statistics. PMLR; 2018, p. 1711–21.
[35] Bellemare MG, Danihelka I, Dabney W, Mohamed S, Lakshminarayanan B, Hoyer S, Munos R. The cramer distance as a solution to biased wasserstein gradients. 2017, arXiv preprint arXiv:1705.10743.
[36] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: International conference on machine learning. PMLR; 2017, p. 214–23.
[37] Nadjahi K, De Bortoli V, Durmus A, Badeau R, Şimşekli U. Approximate Bayesian computation with the sliced-wasserstein distance. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing. IEEE; 2020, p. 5470–4.
[38] Hagspiel S, Papaemannouil A, Schmid M, Andersson G. Copula-based modeling of stochastic wind power in europe and implications for the Swiss power grid. Appl Energy 2012;96:33–44.
[39] Zhang N, Kang C, Singh C, Xia Q. Copula based dependent discrete convolution for power system uncertainty analysis. IEEE Trans Power Syst 2016;31(6):5204–5.
[40] Sklar M. Fonctions de repartition an dimensions et leurs marges. Publ Inst Stat Univ Paris 1959;8:229–31.
[41] Sun M, Konstantelos I, Strbac G. A deep learning-based feature extraction framework for system security assessment. IEEE Trans Smart Grid 2018;10(5):5007–20.
[42] Genc I, Diao R, Vittal V, Kolluri S, Mandal S. Decision tree-based preventive and corrective control applications for dynamic security enhancement in power systems. IEEE Trans Power Syst 2010;25(3):1611–9.
[43] Illinois Institute of Technology (IIT), IEEE 118-bus System Data, URL http://motor.ece.iit.edu/Data/.
[44] Sun M, Cremer J, Strbac G. A novel data-driven scenario generation framework for transmission expansion planning with high renewable energy penetration. Appl Energy 2018;228:546–55. http://dx.doi.org/10.1016/j.apenergy.2018.06.095.
[45] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. International Group 1984;432:151–66.
[46] Cremer JL, Konstantelos I, Strbac G. From optimization-based machine learning to interpretable security rules for operation. IEEE Trans Power Syst 2019;34(5):3826–36.
[47] Hart WE, Laird CD, Watson J-P, Woodruff DL, Hackebeil GA, Nicholson BL, et al. Pyomo-optimization modeling in python, vol. 67. Springer; 2017.
[48] Gurobi Optimization. Gurobi optimizer reference manual. 2018.
[49] Wächter A, Biegler LT. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Math Program 2006;106(1):25–57.
[50] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12:2825–30.