energies

MDPI

Article

Statistical Safety Factor in Lightning Performance Analysis of Overhead Distribution Lines †

Petar Sarajcev *, Dino Lovric and Tonko Garma

Department of Electrical Power Engineering, FESB, University of Split, R. Boskovica 32, HR21000 Split, Croatia
* Correspondence: petar.sarajcev@fesb.hr; Tel.: +385-21-305-806
† This paper is an extended version of our paper published in Proceedings of the 7th International Conference on Smart and Sustainable Technologies (SpliTech), Bol, Croatia, 5–8 July 2022.

Abstract: This paper introduces a novel machine learning (ML) model for the lightning performance analysis of overhead distribution lines (OHLs), which facilitates a data-centrist and statistical view of the problem. The ML model is a bagging ensemble of support vector machines (SVMs), which introduces two significant features. Firstly, support vectors from the SVMs serve as a scaffolding, and at the same time give rise to the so-called curve of limiting parameters for the line. Secondly, the model itself serves as a foundation for the introduction of the statistical safety factor to the lightning performance analysis of OHLs. Both these aspects bolster an end-to-end statistical approach to the OHL insulation coordination and lightning flashover analysis. Furthermore, the ML paradigm brings the added benefit of learning from a large corpus of data amassed by the lightning location networks and fostering, in the process, a "big data" approach to this important engineering problem. Finally, a relationship between safety factor and risk is elucidated. THe benefits of the proposed approach are demonstrated on a typical medium-voltage OHL.

Keywords: lightning protection; insulation coordination; distribution line; safety factor; machine learning; support vector machine; bagging ensemble

Citation: Sarajcev, P.; Lovric, D.; Garma, T. Statistical Safety Factor in Lightning Performance Analysis of Overhead Distribution Lines. *Energies* **2022**, *15*, 8248. https://doi.org/10.3390/en15218248

Academic Editor: Ayman El-Hag

Received: 17 October 2022
Accepted: 3 November 2022
Published: 4 November 2022

1. Introduction

Lightning performance analysis of medium-voltage (MV) overhead distribution lines (OHLs) constitutes one of the major contributing factors to their secure and reliable operation. Namely, the insulation of MV distribution lines is far more prone to flashover incidents, as a consequence of lightning interactions with the line, than is the respective insulation of the high-voltage (HV) overhead transmission lines. This stems from the two interrelated aspects, notwithstanding the environmental influences, that foster a clear distinction between the lightning performance of distribution and transmission lines: (1) MV insulation has a much lower lightning withstand voltage than the HV insulation, and (2) lightning interaction with the MV lines is more complex, due to the influence of indirect nearby lightning strikes. In other words, lightning performance of MV lines, unlike that of the HV lines, is aggravated by the fact that nearby indirect lightning strikes often have a dominant influence on their operation.

Generally speaking, the interaction of lightning with the overhead electric power lines depends, to a large extent, on the presence or absence of shield wire(s) and can be classified into a total of five different modes, as follows: (1) direct strike to the phase conductor when a shield wire is absent, (2) direct strike to the phase conductor when a shield wire is present (this is known as a shielding failure incident), (3) direct strike to the tower top or to the shield wire along the span (with a consequent so-called backflashover incident), (4) indirect nearby strike when a shield wire is absent, and (5) indirect nearby strike when a shield wire is present. Each mode is associated with an overvoltage that may cause a flashover on the line insulation. As can be seen, there are three modes of direct and two of indirect

interaction. The last two modes of interaction produce an overvoltage that may trigger a flashover incident, through the electromagnetic (EM) coupling of radiated fields from the lightning channel to the line conductors. A lightning channel is essentially behaving like an antenna that radiates strong, high-frequency EM fields far from the location of the strike. These two modes of interaction pose no threat to the HV lines but, at the same time, have a very prominent influence on the MV lines. The shield wire(s), when present on the line, provide to the phase conductors both (a) a shielding effect from the direct strikes, in accordance with the electrogeometric (EGM) theory and (b) a screening effect from the radiated EM fields emanating from the indirect strikes. Both shielding and screening effects depend, primarily, on the number and position of the shield wires on the tower.

All five modes of lightning interaction with overhead power lines have been thoroughly studied, both in case of transmission and distribution lines of different geometries and voltage levels [1,2]. They are only briefly introduced here, while the interested reader is, at this point, advised to consult Refs. [1,3] for additional information. Of the three direct modes of interaction, the backflashover incidents are the most difficult to analyze. This partly stems from the complex nature of the EM wave propagation through the multiconductor, multispan structure which comprises towers, phase conductors, and shield wires, including reflections from the tower's grounding system and adjacent spans. Some additional complicating aspects of the backflashover phenomenon are [1,4]: (1) lightning strokes to the tower tops and along the span length (which initiate different traveling wave patterns), (2) tower height and its grounding impulse impedance, (3) soil ionization, (4) the presence of counterpoise wire, (5) the impact of the nonstandard wave-shape of the backflashover overvoltage on the critical flashover voltage (CFO) of the line insulation, (6) the statistical probability of the time to crest of the lightning current, (7) statistical correlation between amplitudes and time-to-crest values, (8) power-frequency voltage, and (9) the influence of corona on the propagation of traveling waves. Corona attenuates and distorts traveling waves, but also decreases the surge impedance of the shield wire and increases the coupling factor between the shield wire and phase conductors. The EGM theory features prominently in analyzing all three modes of direct interaction, giving in the process rise to the associated notion of the "shielding angle" that is a design feature of the towers. Further complexity in analyzing flashovers on distribution lines (from direct strikes) stems from the possibility of "side strikes" on sloping terrain, the presence of the so-called "rogue" towers, and other exogenous factors (e.g., keraunic levels, orographic factors, the encroachment of nearby structures on the right-of-way of the line, etc.) [1,5].

Two indirect modes of lightning interaction with overhead distribution lines give rise to, probably, the most demanding and complex mathematical models among all five aforementioned modes of interaction. The full-wave EM theory of coupling radiated fields over lossy ground, from the lightning strike channel to the (shielded or exposed) phase conductors, is known to be notoriously complicated; see, for example, Refs. [6–12] for more details and additional information. It is beyond the scope of the present paper to discuss these various numerical approaches to the solution of this complex problem. The associated numerical codes (e.g., FDTD approach in particular) tend to be computationally demanding and expensive to solve, in terms of CPU time and hardware resources. Furthermore, some of the (almost) elusive features of lightning exert an important influence on the overvoltage shape and amplitude that is a consequence of the EM field coupling to the phase conductors [4]. For example, a velocity of the return-stroke current (of the negative downward lightning strike) is one of those elusive but important parameters that features prominently in analyzing indirect lightning strikes to distribution lines.

A secure and reliable operation of an OHL presupposes that the insulation coordination of the line has been properly carried out. Since the OHL has a self-restoring insulation, it is recommended that the statistical method of insulation coordination be applied, as described in the international norm IEC 60071-2:2018 [13]. The statistical method exhibits many advantages over the deterministic method, particularly in that it fully accounts for the stochastic nature of the lightning itself, as well as the statistical characteristics of the

insulation strength. It also brings the notion of flashover probability, risk, statistical safety factor, and others, that are replacing the hard (and often crude) limits of the deterministic (worst-case scenario) approach. Moreover, the advent of lightning location networks (LLNs), which record lightning strike locations and associated amplitudes for strikes over large areas (spanning even whole continents), has ushered in a "big data" paradigm into the lightning analysis domain [14]. Large lightning datasets, coupled with machine learning (ML) techniques, give rise to a new class of models for analyzing the lightning performance of overhead power lines, including their statistical insulation coordination. Using ML techniques has certain advantages over more traditional EM-theory-based methods, particularly in terms of computational speed, reduced model complexity, and reliance on a large corpus of recorded LLN measurements data. ML is able to learn from (real-world) data those (almost intangible) relationships between lightning-current parameters (including strike locations) and OHL flashover probabilities.

One of the most prominent examples of using ML in the analysis of lightning performance of overhead distribution lines was given by Martinez and Gonzalez-Molina in [15,16]. Therein, they applied a feed-forward artificial neural network (ANN) for the analysis of OHL lightning flashovers. The problem was posed as a binary classification, and the ANN was trained on a synthetic dataset generated from the analytical treatment of OHL exposure to lightning. Going forward, two important and interrelated aspects of the problem ought to be emphasized: (1) an insulation flashover is a low-probability event (with all the ramifications that it entails for classification tasks), and (2) any dataset of lightning flashovers on OHLs will, necessarily, be class imbalanced (with important repercussions on the training of ML classifiers). There have been other ML and statistical approaches to analyzing lightning performance of OHLs. For example, Ain et al. in [17] introduced a Gaussian process regression model for the prediction of lightning-induced overvoltages on OHLs. Napolitano et al. in [18] used a stratified-sampling Monte Carlo method for the lightning performance assessment of distribution lines.

The present paper builds on our previous research published in Ref. [19], where the bagging ensemble was first introduced for the lightning assessment of OHL performance. This research is extended here with the introduction of a statistical safety factor. It is argued that the proposed bagging ensemble of support vector machines (SVM) provides not only a robust classifier but brings unique benefits to the statistical treatment of the OHL lightning performance. These emanate primarily from the underlying support vectors, which are unique feature of the SVM. Namely, it is shown how support vectors can be used to construct a curve of limiting parameters (CLP) of the OHL, which features prominently in the statistical methods of insulation coordination; see IEC TR 60071-4:2004 [20] for more information. This is considered to be an original contribution to the state of the art. Furthermore, the proposed ML model provides a foundation for the introduction of a statistical safety factor (SF) to the OHL lightning performance analysis. This is the first time, as far as the authors are informed, that the statistical safety factor is used in the context of the lightning performance analysis of OHLs. Both of these aspects (CLP and SF) fully endorse an end-to-end statistical approach (based on "big data" and ML) to the insulation coordination and flashover performance analysis of OHLs. The interested reader is advised to consult IEC TR 60071-4:2004 [20] for more information related to the use of CLP in insulation coordination and connected studies, which is considered beyond the scope of this paper. The focus of the present paper is on the statistical safety factor and its close relationship with the risk of flashover.

The rest of this paper is organized as follows. Section 2 forms the main body of the paper and presents in Section 2.1 a lightning data generating process, which rests on the Monte Carlo method. It also introduces a dataset on which the subsequent machine learning model is trained and tested. Next, Section 2.2 presents the proposed ensemble learning model, based on SVMs, to study lightning flashovers on overhead distribution lines. It further details the related processes of deriving the curve of limiting parameters for the line, as well as its statistical safety factor. Both stem from the ML model outputs

(i.e., support vectors and model predictions). It also discusses the relationship between the safety factor and a risk of flashover, as well as its use in the pricing of overvoltage protection measures. Section 3 brings a brief discussion of the proposed statistical approach in terms of the international norms IEC 60071 and IEC 62305, along with its limitations and possible future extensions. The paper is concluded in Section 4.

2. Materials and Methods

The materials part introduces a synthetic dataset of lightning flashovers on OHLs, its generation process, and statistical properties. The methods part describes the ML model, its training and testing procedure, and the use of its products in deriving the CLP and the SF for the distribution lines.

2.1. Dataset of OHL Lightning Flashovers

This section briefly introduces a dataset of lightning flashovers for training the machine learning model, which was generated by means of a Monte Carlo simulation; see [19] for more information on the data generating process itself. The main outline of the dataset construction process is depicted in Figure 1. The statistical probability of lightning flashovers on distribution lines, considering all five modes of lightning interaction, is dependent on several parameters. Each of these comes with its own particular statistical distribution, as follows: lightning current amplitudes (I) from a log-normal distribution, lightning return-stroke velocities (v) and lightning strike distances (d) from the uniform distribution, OHL tower's grounding surge impedances (R) from the normal distribution, shield wire's presence/absence on the tower (s) from the *Bernoulli* distribution, and EGM model types (e) from the *categorical* distribution.

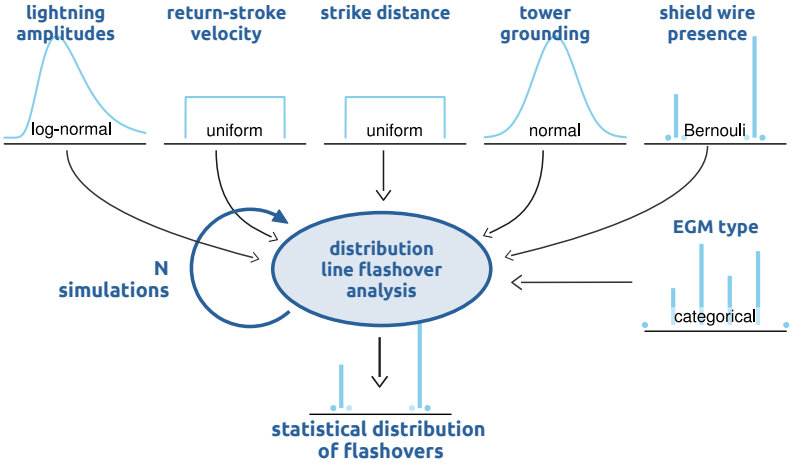

Figure 1. Monte Carlo method of lightning data generating process.

In accordance with the above-stated facts, statistical variables for the Monte Carlo simulation were generated as follows:

$$I \sim \textit{Log-Normal}(31,\, 0.55)\,\text{kA} \tag{1a}$$
$$v \sim \textit{Uniform}(50,\, 500)\,\text{m/s} \tag{1b}$$
$$d \sim \textit{Uniform}(0,\, 500)\,\text{m} \tag{1c}$$
$$R \sim \textit{Normal}(50,\, 12.5)\,\Omega,\; R > 0 \tag{1d}$$
$$s \sim \textit{Bernoulli}(0.5) \tag{1e}$$
$$e \sim \textit{Categorical}(\mathbf{p}) \tag{1f}$$

Since grounding (impulse) impedance cannot possess a negative value (it is a strictly positive real number), the associated normal distribution is cut off on the left-hand side above zero. Furthermore, it was assumed that shield wire(s) were installed in only 50 % of

cases. This fostered data diversity. The EGM could be randomly chosen from six different types (see [1,19] for more information):

$$e \sim f(x|\mathbf{p}) = \prod_{n=1}^{6} p_n^{x_n}, \tag{2}$$

each with its own probability p_n, $n = 1, \ldots, 6$, where $\sum p_n = 1$. Using slightly different EGM variants introduced an additional level of noise into the dataset, which raised the level of difficulty for the model learning.

The simulation started by generating a large number ($N = 10,000$) of samples from each of the statistical distributions. Next, it engaged a lightning flashover analysis, which considered a mode of lightning interaction with the distribution line (direct or indirect). Interaction mode depended on the EGM type and distance of the strike from the line. The mathematical details of the lightning flashover analysis can be found in [1,3,16,19] and are not repeated here. A basic outline of the computational procedure is depicted in Algorithm 1. It can be mentioned that each flashover analysis was carried out in accordance with the EGM theory and a particular mode of interaction. The Rusck's method was used for the analysis of indirect strikes [16]. Each resulting overvoltage that exceeded the CFO of the line accounted for a flashover incident.

Algorithm 1 Lightning flashover analysis on OHL

input OHL geometry (height, s_g, ...).
I, v, d, R, s, $e \leftarrow$ Generate statistical distributions.
flashovers $\leftarrow$ *empty*(list)
for x_0 in d **do** ▷ for each lightning strike
  r_g, $r_c = EGM(I, e)$
  if s = True **then** ▷ shield wire is present
    Compute EGM distances $D_g(r_g, r_c)$ and $D_c(r_g, r_c)$.
    if $x_0 \leq s_g/2 + D_g$ **then** ▷ stroke to shield wire
      V $\leftarrow$ Compute backflashover.
    else if $s_g/2 + D_g < x_0 \leq s_g/2 + D_g + D_c$ **then** ▷ stroke to phase conductor
      V $\leftarrow$ Compute shielding failure.
    else if $x_0 > s_g/2 + D_g + D_c$ **then** ▷ indirect stroke
      V $\leftarrow$ Compute indirect strike with shield.
  else ▷ shield wire is absent
    Compute EGM distance $D_c(r_g, r_c)$.
    if $x_0 \leq s_g/2 + D_c$ **then** ▷ stroke to phase conductor
      V $\leftarrow$ Compute direct strike.
    else if $x_0 > s_g/2 + D_c$ **then** ▷ indirect stroke
      V $\leftarrow$ Compute indirect strike w/o shield.
  if V $\geq$ CFO **then**
    flashover = True
  else
    flashover = False
  flashovers.*append*(flashover)
return flashovers

OHL Dataset Example

The dataset generating process was demonstrated using a typical distribution line, on a flat terrain, with a horizontal arrangement of conductors [16]. The height of the phase conductors was 15 m. The line had double shield wires (when installed), with a separation distance of $s_g = 3$ m and positioned 1.5 m above the phase conductors. The diameter of the phase conductor was 10 mm. The diameter of the shield wire was 5 mm. The CFO of the line insulation equaled 160 kV. The coordinate system was centered on the line itself

and conditions were symmetric in relation to the line. Only downward (negative) lightning strikes were considered, without the possibility of side strikes.

Figure 2 presents a dataset, in terms of the two main attributes: (a) lightning amplitudes and (b) striking distances. It features a scatter plot in the main area of the figure. Flashovers are depicted as red dots, while lightning strikes that do not provoke a flashover are shown as blue dots. The flashover analysis was posited as a binary classification problem [16]. The figure also provides two (independent) marginal distributions, in terms of amplitude and distance of the lightning strokes. The marginal distribution of flashover amplitudes, in particular, featured a fat tail that was not present in the starting *Log-N* distribution. This clearly indicated a direction, statistically speaking, in which lightning amplitudes that triggered flashovers were drifting.

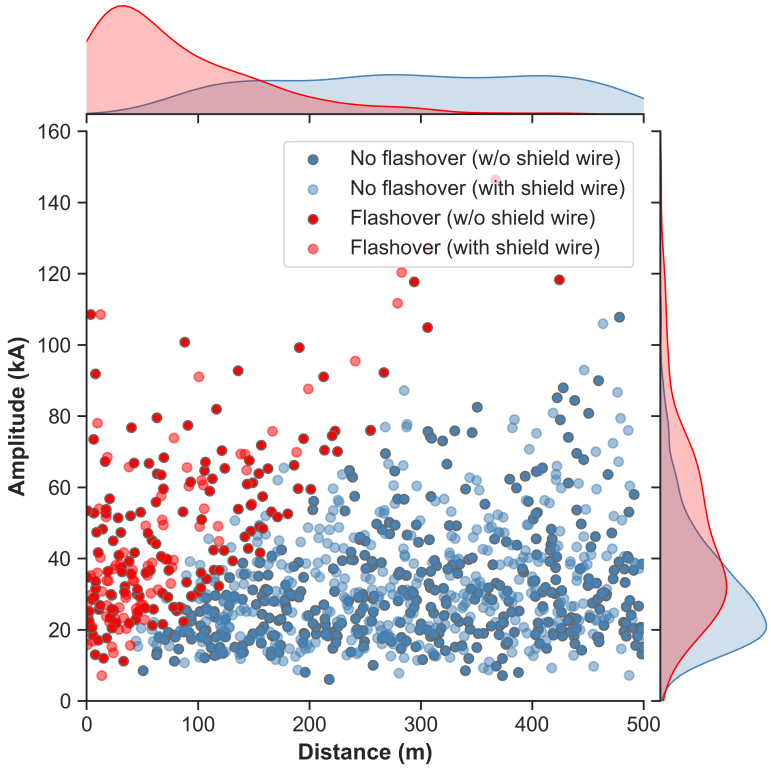

Figure 2. Scatter plot with superimposed marginal distributions of simulated lightning incidents.

Furthermore, Figure 2 indicates that the dataset had a hierarchical structure which distinguished between the presence and absence of a shield wire(s) on the towers, which is depicted by the black edge on the scatter points. The dataset was also class imbalanced (the number of blue points outweighed the number of red ones), which had important repercussions on the subsequent training of the ML models. This imbalance emanated from the fact that a flashover on the distribution line is a low probability event. It can be further deduced from Figure 2 that flashovers were more probable for lightning strikes in the vicinity of the line (red dots clustered to the left-hand side of the figure). These were direct as well as very close nearby indirect lightning strikes. Furthermore, flashovers emanating from indirect strikes were more probable for those associated with larger amplitudes (red dots are predominant in the top portion of the figure). All this was expected and showed that this synthetic dataset emulated reality quite well [16]. Moreover, the screening effect of the shield wire(s) could be discerned by comparing points with and without black edges. This was another notable feature of the dataset that also reflected reality. An instance of the dataset was deposited on Zenodo [21] with a CC BY license.

In order to apply machine learning, the dataset needed to be further processed. First, any extreme outliers in the dataset were removed. These might be particularly associated

with lightning-current amplitudes. Then, the continuous features from the dataset were standardized (i.e., scaled to zero mean and unit variance). Next, the dataset was split, reserving 80 % of the data for training and the remaining 20 % for testing. The training part of the data was then split for the second time, into training and validation sets (with the same 80/20 ratio). Due to the class imbalance in the data, a stratified shuffle split strategy was used during both splittings [22], which preserved the class imbalance rates between training, validation, and test sets.

2.2. Ensemble Learning in OHL Lightning Flashover Analysis

Ensemble learning is an ML paradigm where multiple models, often called base estimators, are trained independently (and even in parallel) and their predictions combined, by some sort of aggregation, to increase the prediction performance [23]. A bagging ensemble is a type of ensemble that is built by means of the bootstrap aggregation of multiple base estimators. The training of each base estimator is performed on a random subset from the training dataset (i.e., bootstrap sample). Aggregation takes predictions from all base estimators and averages them. This kind of ensemble helps reduce overall variance of the final model and helps avoid overfitting at the same time [23]. Here, the proposed bagging ensemble used support vector machines as base estimators and a (weighted or not) "soft voting" strategy for the aggregation. A basic outline of the overall ensemble building process is presented as Algorithm 2. The model was built using the `scikit-learn` and `scipy` Python libraries. The source code was deposited on GitHub [24].

Algorithm 2 Bagging ensemble built from SVM base estimators

input X-features, y-labels
splitter ← *StratifiedShuffleSplit*(splits = 1, test = 20%)
X-data, y-data, X-test, y-test ← splitter.*split*(X-features, y-labels) ▷ 1st
X-train, y-train, X-validate, y-validate ← splitter.*split*(X-data, y-data) ▷ 2nd
estimators ← *empty*(list)
for $m = 1$ to $|\mathcal{M}|$ **do**
    X, y ← Sample random subset from X-train, y-train. ▷ bootstrap sample
    estimator ← *SVM*(C, γ, w) ▷ base estimator
    Pipeline(transformer, estimator)
    Distributions(transformer:[*None*, *StandardScaler*], kernel:[linear, RBF], C, γ, ...)
    model ← *HalvingRandomSearchCV*(Pipeline, Distributions, *StratifiedKFold(k = 3)*, ...)
    model.*fit*(X, y) ▷ fit on sample from train set
    estimators.*append*(model)
if weight = True **then** ▷ weighted ensemble
    weights ← Estimators cross-entropy minimization.
else ▷ equal weights
    weights ← None
ensemble ← *SoftVotingClassifier*(estimators, weights)
ensemble.*fit*(X-validate, y-validate) ▷ fit on validation set
$\hat{y}$ ← ensemble.*predict*(X-test) ▷ predict on test set
score ← metric(y-test, $\hat{y}$)
return $\hat{y}$, score

It can be seen that the training of base estimators (including their hyperparameters optimization) involved a stratified k-fold cross-validation on the random (i.e., bootstrap) sample from the train set. On the other hand, the training of the ensemble as a whole (including weights optimization) used the validation set. Furthermore, the predictions from the ensemble were performed on the test set (never seen before by the model). Each SVM, as a base estimator, was slightly different (see below) and therefore, brought unique qualities to the group (i.e., ensemble), boosting its performance. Furthermore, the individual

predictions of the base estimators from the ensemble were aggregated by averaging their prediction probabilities [23]:

$$f(y|\mathbf{x}) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} w_m \cdot f_m(y|\mathbf{x}), \tag{3}$$

where $\mathcal{M}$ is a set of base models $f_m(y|\mathbf{x})$ in the ensemble, while w_m, $m \in \mathcal{M}$ are model weights. The weights could be determined on the basis of the model's confidence in the predictions, or all models could be assigned equal weights. It was found that equal weighting preserved a higher diversity within the ensemble and produced a slightly better performing final classifier.

For each SVM, the bootstrap training dataset comprised N input vectors $x_1, \ldots, x_N$ with corresponding target (i.e., class) values $t_1, \ldots, t_n$, where $t_n \in \{-1, 1\}$. The SVM solved the following optimization problem [23]:

$$\min_{\zeta, w} C \sum_{n=1}^{N} \zeta_n + \frac{1}{2}||\mathbf{w}||^2 \tag{4a}$$

$$\text{s.t.} \begin{cases} t_n y(x_n) \geq 1 - \zeta_n, \\ \zeta_n \geq 0, \quad n = 1, \ldots, N \end{cases} \tag{4b}$$

where C is the penalty that acts as an inverse regularization parameter, while ζ_n is a slack variable. The dual Lagrangian formulation for the primal in (4) can be written in terms of dual variables $\{a_n\}$, after eliminating slack variables $\{\zeta_n\}$, as follows [25]:

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m) \tag{5a}$$

$$\text{s.t.} \begin{cases} 0 \leq a_n \leq C \\ \sum_{n=1}^{N} a_n t_n = 0 \end{cases} \tag{5b}$$

where $k(\mathbf{x}_n, \mathbf{x}_m)$ is the kernel function. This is a quadratic programming (constrained minimization) problem which can be solved using the standard routines from mathematical programming. The predictions for new points $\mathbf{x}$ are given by [25]:

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b, \tag{6}$$

with

$$b = \frac{1}{|\mathcal{A}|} \sum_{n \in \mathcal{A}} \left(t_n - \sum_{m \in \mathcal{S}} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right) \tag{7}$$

where $\mathcal{A}$ denotes the set of indices of data points having $0 < a_n < C$, while $\mathcal{S}$ represents a set of indices of the support vectors. This set of support vectors, which defines the separation margin between classes, are the only points that contribute to the predictions.

The actual training of SVMs that formed the ensemble (see Algorithm 2) used a pipeline that (1) was fed preprocessed subsamples from the training set, (2) invoked hyperparameter optimization with a stratified k-fold cross-validation, (3) aggregated individual predictions, and (4) returned outputs that included the support vectors and prediction probabilities from the test set. A so-called "hyperband" bandit-based optimization algorithm was used as an optimizer [26]. It is much faster than the more known "random search" (which it extends by adding successive halving and some clever resource management) and has better convergence; see [26] for more information. Each SVM that was part of the ensemble, in addition to hyperparameters, could have different kernel types. Hyperband chose between linear and radial basis function (RBF) kernels and then fine-tuned the RBF kernel coefficient (if it was selected) along with a regularization parameter of the penalty function.

The regularization provided an important safeguard against overfitting of the individual SVMs and was randomly sampled from $C \sim Log\text{-}U(1, 1000)$. Finally, since the dataset was class-imbalanced, each base estimator used a class-weight balancing during training. This step should not be confused with sample weighting, which can be applied in addition to the class weighting.

The so-called Brier score was used as a principal loss metric for training the bagging ensemble, which can be defined as follows:

$$BS = \frac{1}{N}\sum_{n=1}^{N}[y_n - p(x_n)]^2 \tag{8}$$

where y_n is the nth sample's true label and $p(x_n)$ is its positive class probability. As a mean square error, the Brier score is lower with better calibrated predictions, and it remains strictly positive. It is found to be far less sensitive to the class imbalance problem than "accuracy" and other often-reported measures.

Computing individual weights for the base estimators within the ensemble can be achieved by considering their relative scores (on the validation set), as follows:

$$\begin{aligned} &\min_{w}\{-[y \cdot \log L_q + (1-y)\cdot \log(1-L_q)]\} \\ &\text{for } L_q = \sum_{m\in\mathcal{M}} w_m \cdot P|_{y=1}(\mathbf{x}) \\ &\text{s.t.} \begin{cases} 0 \le w_m \le 1 \\ \sum w_m = 0 \end{cases} \end{aligned} \tag{9}$$

where w_m is the model's relative score within the set of $\mathcal{M}$ base models, y is a true class label and $P|_{y=1}$ is a probability estimate of the flashover class. The optimization given by (9) essentially minimizes the cross-entropy between the SVMs within an ensemble. It usually retains only a few best-performing base estimators while discarding others by assigning very small weights to them.

2.2.1. Classifier Performance

The bagging ensemble consisted of three SVMs, which were individually trained with a cross-validation on the bootstrap samples from the training set. The model training resulted in each base estimator having (slightly) different hyperparameters, that could further vary between runs. However, the ensemble as a whole was stable between runs and produced consistent predictions. An example of training results, in terms of model hyperparameters, is presented in Table 1. It can be seen that only two out of three base estimators participated when the ensemble weights were left to be determined by the model training (with linear and RBF kernels), while the third was seen as redundant. The kernel coefficient of type *scale* implemented $1/(n_f \cdot var(X))$, while that of type *auto* used $1/n_f$, where n_f was the number of features and $var(X)$ was the variance of the input features matrix [22].

Table 1. Hyperparameters of base estimators from the bagging ensemble.

Estimator	Transformer	Kernel	Coefficient	Regularization	Weight
SVM-A	StandardScaler	Linear	None	35.8	0.31
SVM-B	None	RBF	Scale	11.2	0.02
SVM-C	StandardScaler	RBF	Auto	2.86	0.67

After training was completed, the bagging ensemble classifier produced a single prediction probability value for each sample in the test set (i.e., probability of positive class). This probability was then converted, based on the classifier's threshold level, to the statement of belonging (or not) to the flashover class. Figure 3 is a testament to the high performance of the classifier. It presents the following measures: (a) the receiver operating

characteristic (ROC) curve, (b) precision–recall (PR) curve, and (c) detection error trade-off (DET) curve of the classifier. All three types of curves were obtained from the test dataset. The area under the ROC curve (i.e., the AUC score) and average precision (i.e., the AP score) are also provided on the figure, both of which confirmed the high performance of the proposed classifier. The presented curves measured the model's performance in terms of different types of errors that it made when predicting class labels [22]. Furthermore, when there is a class imbalance (as is the case here), the PR curve may be superior to the ROC curve in gauging a classifier's performance. Finally, the DET curve can be a valuable aid in the classifier calibration process.

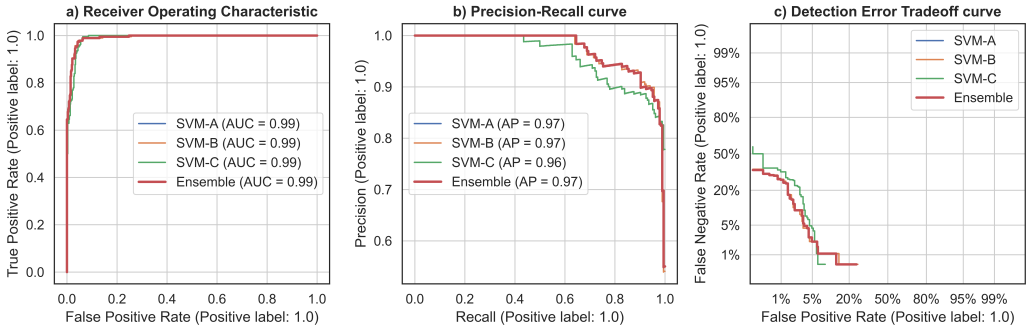

Figure 3. Performance measures: (**a**) receiver operating characteristic, (**b**) precision–recall curve, and (**c**) detection error trade-off curve for the bagging ensemble classifier and individual base estimators.

It can be seen from the figure that even a single SVM could already achieve substantial classification accuracy on this synthetic dataset. However, the ensemble enlarged (and diversified) the pool of support vectors, which helped increase the robustness of the CLP. This is an important feature, particularly if one considers the lightning detection errors and other sources of noise that will pollute a real-world dataset.

2.2.2. Curve of Limiting Parameters

The support vectors from SVMs were considered here as a very important byproduct of the proposed classifier. They supported the decision boundary of the classifier. This boundary in-turn provided a scaffolding for the so-called curve of limiting parameters (CLP). It was found through experimentation and repeated simulations of different lightning datasets (representing different OHL geometries), that a second-degree polynomial fit, based on a least-squares regression, of the support vectors yielded a satisfactory CLP of the OHL which could be used in statistical studies. Hence, Figure 4 presents (in a 2D coordinate space of lightning amplitude and strike distance) the CLP fit of the support vectors and superimposed on the samples from the training set for a better visual reference. The dark shaded region around the CLP curve provides a 95% confidence interval, while the light shaded region depicts a 95% prediction interval. The adjusted R^2 of the regression was around 0.9. The CLP was not a straight line, generally speaking, and its curvature depended on the line height and geometry, the insulation's CFO level, and the local statistical properties of lightning in the area. The support vectors from all underlying SVMs in the ensemble (with any duplicates removed) are highlighted in the figure with orange circles. It can be seen that they "support" (as the name implies) the decision boundary between classes at the same time.

It is important to emphasize that the dataset needs to be sufficiently large in order for the support vectors to cover the region of high-amplitude lightning currents (so that the CLP is well-defined in the broad range of values). Furthermore, since our bagging ensemble employed several SVMs (each slightly different), often between three and at-most ten, their combined decision vectors (without duplicates) were generally robust and insensitive to perturbations and noise in the data. This translated into a robust and stable CLP curve, with tight confidence and prediction intervals. The importance of this stability can be

appreciated by considering the fact that the ML model would typically be applied on data coming from measurements supported by the LLN. These data come with measurement errors related to both lightning amplitudes and strike locations. Namely, the detection accuracy of the LLN strike location (in terms of longitude and latitude coordinates) is defined through an error ellipse that can be wider that 100 m or more.

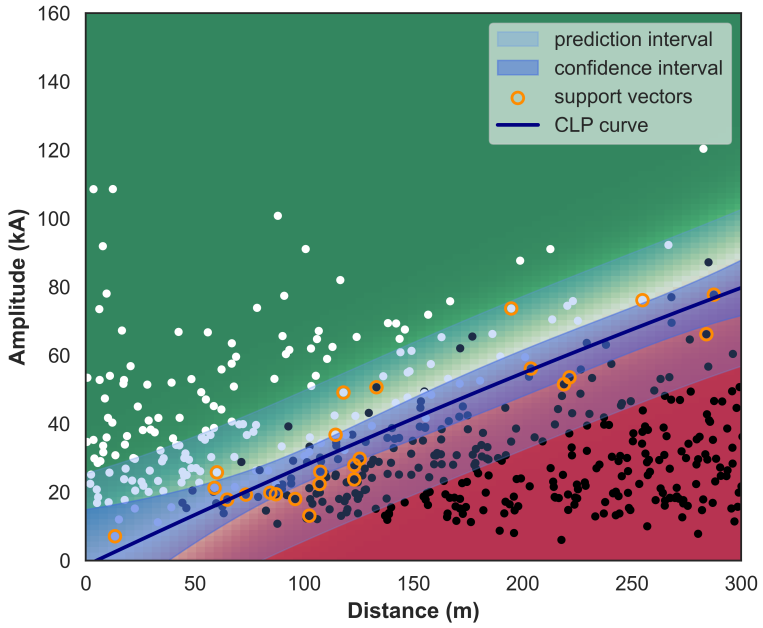

Figure 4. Curve of limiting parameters obtained from the least-squares fit to the support vectors.

2.2.3. Statistical Safety Factor

Furthermore, the ML model's prediction probabilities can be employed in defining a so-called insulation performance function of the OHL, derived in terms of the statistical cumulative distribution function (CDF). Interested reader is at this point advised to consult Ref. [27] for more information on the relationship between a CDF and the insulation's performance function. Namely, a trained classifier returns a prediction probability for each sample from the test set, and these probabilities can be used to construct a CDF of the OHL insulation flashover. Several of these CDFs are presented in Figure 5, considering different strike distances (where scatter points represent class labels from the test set).

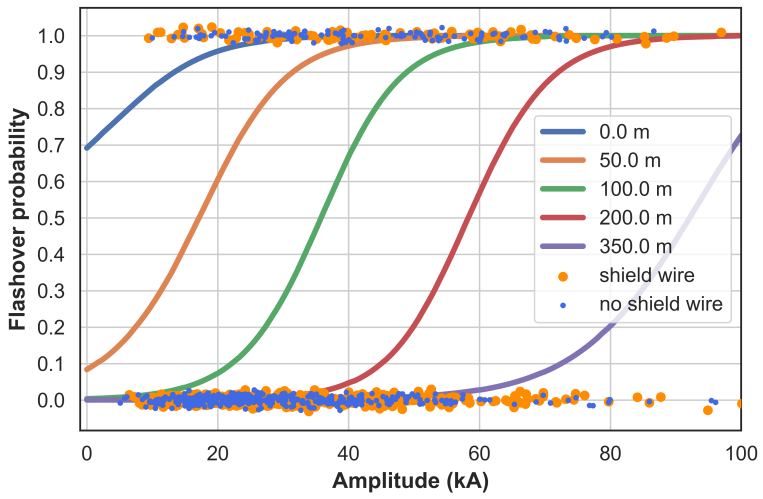

Figure 5. Flashover probabilities in relation to lightning-current amplitudes for several striking distances.

The left-most curve (blue line) represents the flashover probability from direct strikes and is cut off at zero. Other curves represent indirect lightning strikes at various distances from the line. It can be seen that as the strike location moves away from the line, the associated amplitude for attaining the same probability of flashover is increasing. For example, a nearby indirect strike with an amplitude of 30 kA has a 90% probability of evoking a flashover on the line for a strike distance of up to 50 m (orange line), while that probability drops to 30% for a distance of 100 m (green line).

By using the probability density function (PDF) of the lightning-current amplitudes (of negative downward lightning strikes) in combination with the previously obtained CDFs of the line insulation flashover (Figure 5), one can define the statistical safety factor (SF) due to nearby indirect lightning strikes. Namely, the SF is hereafter defined as a quotient between the OHL insulation withstand (taken as the 10% probability of insulation flashover) and the probability of obtaining an amplitude that will be exceeded in no more than 10% of cases. It follows that the SF is a strictly positive number and can be defined for any indirect strike distance d from the line as:

$$SF_d = \frac{L_w(d)}{L_s} \tag{10}$$

where $L_w(d)$ defines a point on the CDF curve of the line's insulation with a 10% probability of flashover, while L_s defines a point on the PDF curve of lightning-current amplitudes with a 10% probability of being exceeded. Here, the PDF was a well-known *Log–N* distribution [28], while the CDF was taken from Figure 5 for any desired distance d from the line. Both of these mentioned points were obtained from the associated quantile (i.e., inverse CDF) functions of the appropriate statistical distributions, as follows:

$$L_w = \hat{F}_d^{-1}(\alpha), \tag{11a}$$

$$L_s = F_s^{-1}(1-\alpha), \tag{11b}$$

with

$$F_s(I) = \int_{-\infty}^{I} f_s(x)dx \tag{12}$$

and

$$f_s(I) = \frac{\exp\left[-\frac{(\ln I - \ln I_\mu)^2}{2\sigma_{\ln I}^2}\right]}{\sqrt{2\pi}\cdot I\cdot \sigma_{\ln I}} \tag{13}$$

where $\alpha = 0.1$ is the threshold, $\hat{F}_d$ is the CDF of lightning flashovers at distance d from the line and $f_s(I)$ is the PDF of the lightning-current amplitudes in which $I_\mu = 31$ kA was the median value and $\sigma_{\ln I} = 0.55$ was the standard deviation [28]. Furthermore, due to the fact that $\hat{F}_d$ was defined by points (Figure 5), a linear interpolation was used in combination with a numerical inversion of this function. On the other hand, the *Log–N* distribution from (12) and (13) had a well-defined quantile function.

The threshold level (α) on both L_s and L_w points was taken at the 10% probability level, as already mentioned. It ought to be emphasized that this is a standard statistical withstand limit of the self-restoring insulation. At the same time, the selected threshold considered the lightning-current amplitudes from the tail of the *Log–N* distribution that had only a 10% chance of being exceeded. In other words, the SF_d, as a single number, tied together the probabilities of two low-probability consecutive events for any strike distance: (1) the probability of obtaining a certain lightning-current amplitude with (2) a probability of insulation being able to withstand the associated overvoltage without flashover. It needs to be stated that these were not independent stochastic events. Moreover, the threshold imposed on the amplitudes could be made more stringent (e.g., at the 5% level) if necessary.

In order to demonstrate the above definition, Figure 6 depicts the graphical construction of the SF for an example of nearby lightning strikes at a distance $d = 100$ m from the distribution line at hand. It ought to be pointed out that points L_s and L_w did not have equal height on the y-axis, and that, actually, two independent y-axes were used in

order to better illustrate the graphical construction of the statistical safety factor. Needless to say, the graphical construction is here provided as a visual aid only, and the SF was computed numerically from the PDF and CDF curves. It should not be forgotten that the CDF curves came directly from the classifier's prediction probabilities. It can be seen from the figure that $Ls = 62$ kA was obtained as a threshold of the PDF distribution (of lightning current amplitudes) with a 10% margin (shaded area in the right-hand tail of the distribution function). At the same time, it can be seen that $L_w = 18$ kA was obtained as a point on the CDF of the line insulation (for $d = 100$ m) flashover characteristic with a 10% probability. Since $L_w < L_s$, the resulting $SF = 0.29 < 1$ did not provide a sufficient safety against flashovers at this particular distance.

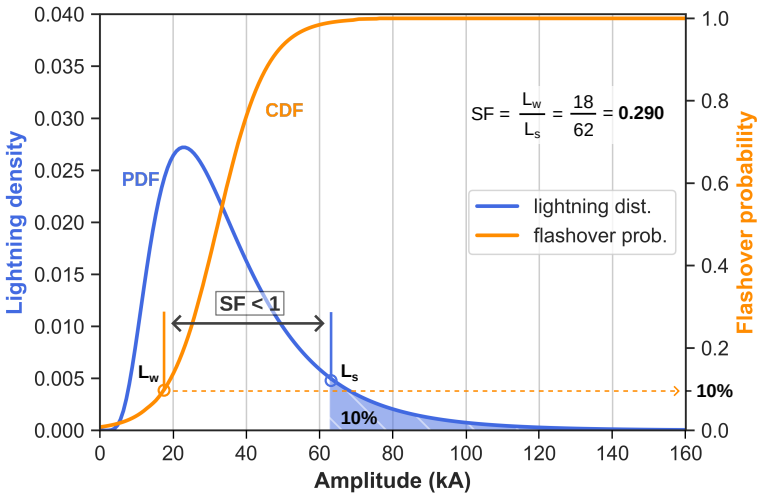

Figure 6. Statistical safety factor for a nearby lightning strike at a distance of 100 m from the line.

Furthermore, at the same time, Figure 7 depicts the same graphical construction of the SF, but for an example of nearby lightning strikes at a distance $d = 350$ m from the distribution line at hand. It can be seen that in this particular case of more distant lightning strikes, although L_s stayed the same (because the lightning ambient conditions did not change), the withstand point increased to $L_w = 88$ kA (for the same 10% probability of withstand). This resulted in $L_w > Ls$, which yielded a much higher statistical safety factor of $SF = 1.42 > 1$. Furthermore, when $L_s = L_w$, it would follow that $SF = 1$.

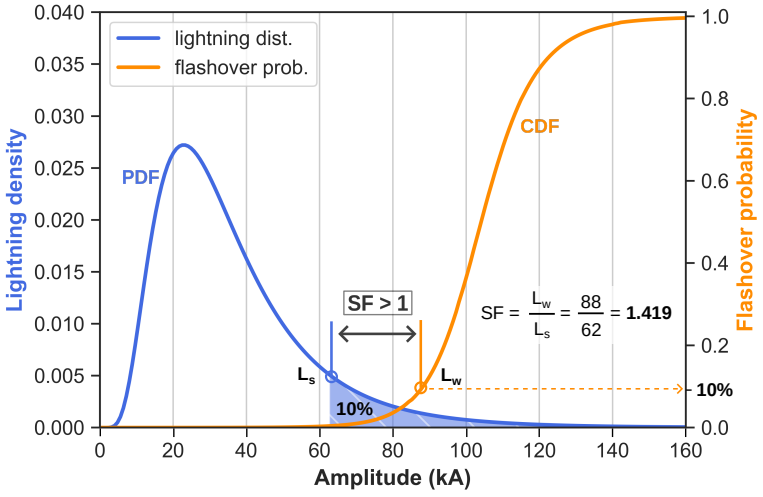

Figure 7. Statistical safety factor for a nearby lightning strike at the distance of 350 m from the line.

2.2.4. Safety Factor vs. Risk

It is important to note that the statistical safety factor is very closely related to the risk of insulation flashover, where the risk can be computed from the following expression [1]:

$$R_d = \int_0^\infty f_s(I)\hat{F}_d(I)dI, \tag{14}$$

where $f_s(I)$ is the PDF of lightning-current amplitudes, while $\hat{F}_d(I)$ is the CDF of the insulation flashover probability for the considered distance of the nearby lightning strikes (from Figure 5). The definite integral in (14) can be computed with sufficient accuracy using the well-known trapezoidal or Simpson's rules.

This relationship between the SF and risk is graphically presented for the OHL at hand in Figure 8, where the SF and risk are given as individual functions of the lightning strike distance from the line in the left-hand part of the figure, while their mutual relationship is depicted in the right-hand part of the figure. It can be seen that this relationship of the safety factor vs. risk was nonlinear. For a safety factor of zero, the risk equaled one, and as the safety factor increased beyond one, the risk dropped to low values, and approached zero asymptotically thereafter. This figure reveals that, for the considered OHL, the risk of flashover at a distance of 100 m was around 50% and it dropped to only 2% at the distance of 350 m.

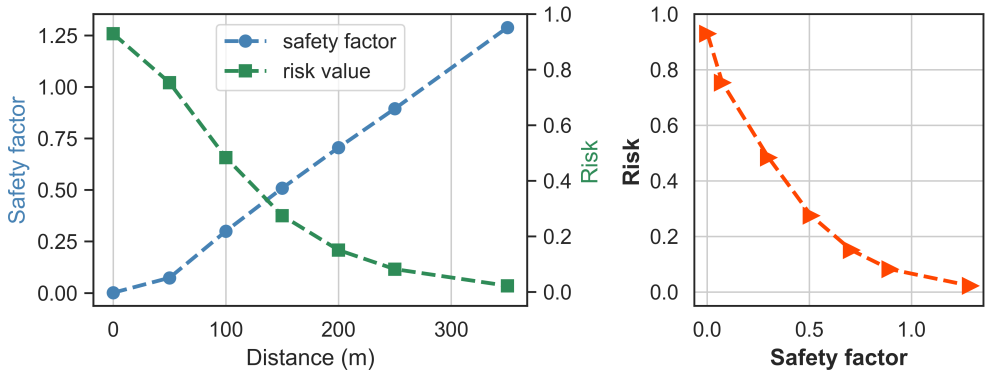

Figure 8. Relationship between the statistical safety factor and the risk of insulation flashover.

The nonlinear relationship between the risk and safety factor could be mathematically described using the following function:

$$R_d = \rho \cdot \exp(-\eta \cdot SF_d) \tag{15}$$

where $\rho = 1$ and $\eta = 3$ were determined from the least-squares fit using the Levenberg–Marquardt algorithm and an exponential weighting of the safety factors by $w = e^{-3x}$. The weighting gave a higher importance to larger SF_d values by decreasing uncertainty. The relationship is graphically presented in Figure 9 using a *semilog* scale for better visual reference. The correlation coefficient for this particular fit equaled $R^2 = 0.98$.

It can be further argued, based on the above presented analysis, that a statistical safety factor above one, i.e.,:

$$SF_d \geq 1 \tag{16}$$

is a sufficient requirement for the purpose of OHL insulation coordination (in terms of the nearby indirect lightning strikes at distance d from the line). In this particular case, it can be seen that the safety factor rose above the threshold of one already at a distance of around 250 m. It needs to be stated that the SF can be increased by translating the CDF curve to higher amplitude levels, which can be accomplished (assuming that the shield wire is already installed) either by (a) increasing the CFO of the line insulation, or by (b) installing the surge arresters. This is exactly what the OHL insulation coordination is all about, where

the statistical safety factor can feature prominently in reconciling the opposing demands between the actual lightning threat levels (i.e., as recorded by the LLN) and the OHL insulation levels (including the possibility of installing protective measures). The statistical approach is reinforcing the safety and reliability aspects of this coordinating process.

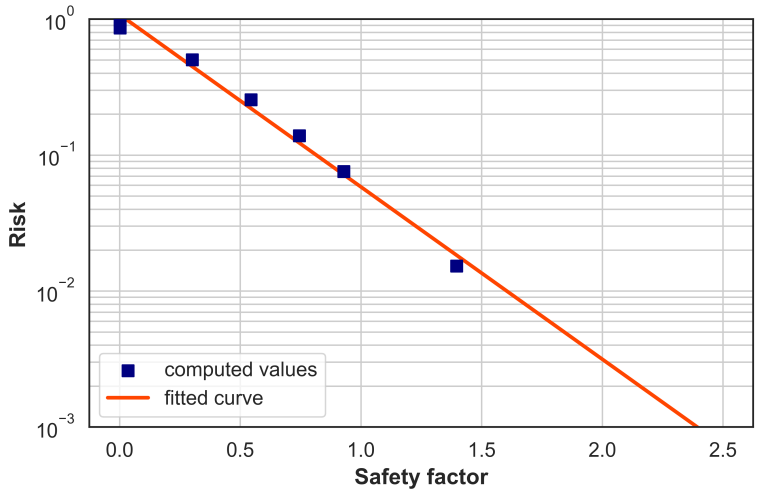

Figure 9. Least -squares fit of the relationship between risk and statistical safety factor.

2.2.5. Profitability of Protection Measures

Investments in lightning and surge protection measures needs to be analyzed in terms of the reductions it brings to the total cost of damages associated with lightning incidents [29]. The projected reductions in costs of damage must come from the decrement of the associated risk of damage (which is congruous to the increment of the statistical safety factor). In other words, the investment in protection measures needs to be recuperated through the savings emanating from a decline in the total cost of damages (which are averted by the installation of protection measures). Hence, based on the IEC 62305-2 standard [30], the profitability of the investment can be analyzed, considering the annual cost of protection measures, by means of the following equation [31]:

$$c_t \cdot \Delta R - c_p \cdot (i + a + m) \geq 0, \quad (17)$$

where c_p is the annual cost associated with protection measures, i is the interest rate (for financing protection measures), a is the amortization rate (calculated as the service life of the protection measures), m is the maintenance rate (which may include inspection and maintenance costs), c_t is the total cost of damages (which includes repair cost, lost revenue due to outage time, and any additional costs inferred from penalties for not serving customers), and $\Delta R = R_0 - R_p$ is the reduction in risk from the initial level (R_0) to the lower level (R_p) associated with the implementation of protection measures. Thus, the procedure assumes that costs can be (roughly) estimated before actually planning lightning and surge protection measures. General information on interest rates, the amortization of protection measures and planning, and maintenance and repair costs must also be available [31]. It can be seen that the investment in protection measures makes economic sense only if the annual saving is expected to be positive. Satisfying inequality (17) can be approached by examining several possibilities and finding that which has the costs of damage as low as possible.

3. Discussion

A statistical treatment of the insulation coordination of high-voltage apparatuses and electrical power stations has been part of the IEC norms for quite some time; see IEC 60071-2:2018 [13] and IEC TR 60071-4:2004 [20]. The probabilistic and risk-based approaches to the lightning protection of electrical installations, and buildings in general,

have long been advocated as part of the IEC 62305-2:2010 [30]. Lightning interaction with wind farms has gone through another revision in the most-recent edition of the IEC 61400-24:2019 [32]. All this points to the ongoing efforts of including the latest research findings into the engineering standards. The same can be said about the associated technical recommendations published by different working groups. However, it needs to be said that it is still habitual among industry experts to consider more traditional approaches, based on field experience and worst-case scenarios, in dealing with these issues. A full probabilistic and risk-based insulation coordination is, unfortunately, still carried-out only is special select cases. The use of the CLP in particular has been underappreciated, although it was given very prominent position in the IEC TR 60071-4:2004. This is unfortunate. The present paper is seen as a contribution in the direction of remedying this situation.

Furthermore, the advent of lightning location networks has completely transformed the way the risk of lightning has been dealt with in the past. For example, the so-called "thunderstorm day", as a measure of lightning activity in an area (criticized for a very long time), has been replaced by much more precise lightning density maps, which are constructed from the LLN's data. Other custom-tailored products of the LLN are often used by insurance companies for determining payments on lightning-related insurance claims. Risk is also being introduced in the process of selection of surge arresters, which is now approached from the point of view of buying insurance [31,33]. However, the introduction of machine learning is still in the nascent phase, particularly when it comes to the lightning analysis of flashovers on overhead power lines. This paper is seen as a contribution to the state-of-the-art and promotes a wider ML adoption for enhancing existing statistical approaches in the fields of insulation coordination and lightning flashover analysis of overhead electric power lines. The proposed ML approach extends the former statistical view of the insulation coordination by learning new relationships directly from the data and applying that knowledge within the existing statistical/engineering framework. That also includes extending the existing framework with the risk-based pricing of protection measures [31].

Model Limitations and Future Extensions

The proposed ML model learned from the synthetic dataset, where Rusck's model featured prominently in the analysis of indirect lightning strikes. It needs to be stated that this is a rather rudimentary model that could not account for some important features, such as the lightning wavefront time duration and earth conductivity. It was retained here for compatibility with Ref. [16]. Better models could be employed, and we implemented two alternatives [24]: (a) the Chowhduri–Gross model and (b) the Liew–Mar model. Both are superior to Rusck's model, but are far more computationally expensive. We also implemented a simplified CIGRE method (see Ref. [1] for more information) of a backflashover analysis as an additional alternative [24]. All these aspects further reinforce synthetic data diversity, increasing the generalization potential of the subsequently trained models. Future research will inspect several of these aspects: a comparison between alternative data generating approaches, the generalization ability of models trained on synthetic data, the treatment of different OHL geometries, testing models with actual lightning data, and others. Future work will also examine in more detail the application of the proposed statistical safety factor in OHL insulation coordination and surge arrester selection, with the emphasis on a pricing of protection measures.

4. Conclusions

This paper presented a novel bagging ensemble classifier, which was built from support vector machines, for the prediction of lightning flashovers on overhead distribution lines. An important benefit that stemmed from the use of an SVM as a base estimator was that it provided support vectors. A set of support vectors from all SVMs that formed the ensemble (with any duplicates removed) served as a basis for fitting the curve of limiting parameters. A least-squares fit with a second-degree polynomial gave rise to a CLP of

substantial precision for subsequent statistical analyses. In addition, the proposed ML model enabled the construction of a CDF of the OHL insulation, which was related to its so-called performance function. On top of this function, we defined a statistical safety factor of the overhead line. The safety factor was closely related to the risk and could be used as its substitute. Both these aspects, the CLP curve and the statistical safety factor, fully supported an end-to-end statistical evaluation of lightning performance of overhead distribution lines and their insulation coordination.

Furthermore, the presented analysis showed that, starting from the ML model's application on the lightning data (e.g., gathered by the LLN), one could derive a statistical safety factor for any OHL, for any foreseeable distance from the line. Carrying out the insulation coordination of the line against nearby indirect lightning strikes, for any particular distance, was a straightforward matter of getting the safety factor to satisfy the inequality $SF_d \geq 1$. This approach had the benefit of fully considering both the random nature of lightning and the stochastic nature of self-restoring insulation's overvoltage withstand strength. The "big data" paradigm and the associated machine learning approach has just started entering this engineering field, and it is argued here that it can bring valuable assistance to the design engineers and decision-makers alike. Specifically, bringing together the statistical safety factor, risk, and profitability of protection measures, bridges the gap between engineering and finance departments, which may streamline the decision-making process by, metaphorically speaking, leveling the playing field.

Author Contributions: Conceptualization, P.S.; methodology, P.S. and D.L.; software, P.S.; validation, D.L. and T.G.; resources, D.L. and T.G.; writing—original draft preparation, P.S.; writing—review and editing, D.L. and T.G.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset was deposited on Zenodo under the Creative Commons Attribution 4.0 International License CC BY 4.0 (https://zenodo.org/record/6406077) (accessed on 1 November 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial neural network
AUC	Area under the ROC curve
AP	Average precision value
CDF	Cumulative distribution function
CLP	Curve of limiting parameters
CFO	Critical flashover voltage
DET	Detection error trade-off curve
EGM	Electrogeometrical model
EM	Electromagnetic
HV	High voltage
IEC	International electrotehnical commission
LLN	Lightning location network
ML	Machine learning
MV	Medium voltage
OHL	Overhead distribution line
PDF	Probability density function
PR	Precision–recall curve
ROC	Receiver operating characteristic
SF	Statistical safety factor
SVM	Support vector machine

References

1. Hileman, A.R. *Insulation Coordination for Power Systems*; CRC Press: Boca Raton, FL, USA, 1999; Chapter 10, pp. 373–460.
2. Furgal, J. Influence of Lightning Current Model on Simulations of Overvoltages in High Voltage Overhead Transmission Systems. *Energies* **2020**, *13*, 296. [CrossRef]
3. Cooray, V. (Ed.) *Lightning Protection*; IET Power and Energy Series 58; The Institution of Engineering and Technology: London, UK, 2010; Chapter 13, pp. 635–675.
4. Chowdhuri, P. *Electromagnetic Transients in Power Systems*; Research Studies Press Ltd.: Taunton, UK, 1996; Chapter 14, pp. 302–334.
5. Sestasombut, P.; Ngaopitakkul, A. Evaluation of a Direct Lightning Strike to the 24 kV Distribution Lines in Thailand. *Energies* **2019**, *12*, 3193. [CrossRef]
6. Agrawal, A.K.; Price, H.J.; Gurbaxani, S.H. Transient Response of Multiconductor Transmission Lines Excited by a Nonuniform Electromagnetic Field. *IEEE Trans. Electromagn. Compat.* **1980**, *EMC-22*, 119–129. [CrossRef]
7. Diendorfer, G. Induced voltage on an overhead line due to nearby lightning. *IEEE Trans. Electromagn. Compat.* **1990**, *32*, 292–299. [CrossRef]
8. Rachidi, F.; Nucci, C.; Ianoz, M.; Mazzetti, C. Response of multiconductor power lines to nearby lightning return stroke electromagnetic fields. In Proceedings of the 1996 Transmission and Distribution Conference and Exposition, Los Angeles, CA, USA, 15–20 September 1996; pp. 294–301. [CrossRef]
9. Pokharel, R.; Ishii, M.; Baba, Y. Numerical electromagnetic analysis of lightning-induced voltage over ground of finite conductivity. *IEEE Trans. Electromagn. Compat.* **2003**, *45*, 651–656. [CrossRef]
10. Ren, H.M.; Zhou, B.H.; Rakov, V.A.; Shi, L.H.; Gao, C.; Yang, J.H. Analysis of Lightning-Induced Voltages on Overhead Lines Using a 2-D FDTD Method and Agrawal Coupling Model. *IEEE Trans. Electromagn. Compat.* **2008**, *50*, 651–659. [CrossRef]
11. Andreotti, A.; Assante, D.; Mottola, F.; Verolino, L. An Exact Closed-Form Solution for Lightning-Induced Overvoltages Calculations. *IEEE Trans. Power Deliv.* **2009**, *24*, 1328–1343. [CrossRef]
12. Sun, J.; Yang, Q.; Xu, W.; Qin, Z.; Wang, K. Lightning-induced Overvoltage of 10 kV Distribution Line Based on Electromagnetic Return-stroke Model Using FDTD. In Proceedings of the 2021 35th International Conference on Lightning Protection (ICLP) and XVI International Symposium on Lightning Protection (SIPDA), Colombo, Sri Lanka, 30 August–4 September 2021; Volume 1, pp. 1–6. [CrossRef]
13. *IEC 60071-2*; Insulation Co-Ordination—Part 2: Application Guidelines. International Standard. IEC: Geneva, Switzerland, 2018.
14. Betz, H.D.; Schmidt, K.; Laroche, P.; Blanchet, P.; Oettinger, W.P.; Defer, E.; Dziewit, Z.; Konarski, J. LINET—An international lightning detection network in Europe. *Atmos. Res.* **2009**, *91*, 564–573. [CrossRef]
15. Martinez, J.; Gonzalez-Molina, F. Statistical evaluation of lightning overvoltages on overhead distribution lines using neural networks. In Proceedings of the 2001 IEEE Power Engineering Society Winter Meeting, Conference Proceedings (Cat. No.01CH37194), Columbus, OH, USA, 28 January–1 February 2001; Volume 3, pp. 1133–1138. [CrossRef]
16. Martinez, J.; Gonzalez-Molina, F. Statistical evaluation of lightning overvoltages on overhead distribution lines using neural networks. *IEEE Trans. Power Deliv.* **2005**, *20*, 2219–2226. [CrossRef]
17. Ain, N.U.; Mahmood, F.; Fayyaz, U.U.; Kasmaei, M.P.; Rizk, M.E.M. A Prediction Model For Lightning-Induced Overvoltages Over Lossy Ground Using Gaussian Process Regression. *IEEE Trans. Power Deliv.* **2022**, *37*, 2757–2765. [CrossRef]
18. Napolitano, F.; Tossani, F.; Borghetti, A.; Nucci, C.A. Lightning Performance Assessment of Power Distribution Lines by Means of Stratified Sampling Monte Carlo Method. *IEEE Trans. Power Deliv.* **2018**, *33*, 2571–2577. [CrossRef]
19. Sarajcev, P. Bagging Ensemble Classifier for Predicting Lightning Flashovers on Distribution Lines. In Proceedings of the 7th International Conference on Smart and Sustainable Technologies, Split, Croatia, 5–8 July 2022; pp. 1–6. [CrossRef]
20. *IEC TR 60071-4*; Insulation Co-Ordination—Part 4: Computational Guide to Insulation Co-Ordination and Modelling of Electrical Networks. International Standard: Technical Recommendation. IEC: Geneva, Switzerland, 2004.
21. Sarajcev, P. Dataset: Lightning Flashover Simulations on Medium Voltage Distribution Lines (Ver. 1.2). Zenodo. 2022. Available online: https://zenodo.org/record/6406077 (accessed on 1 November 2022).
22. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
23. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*, 4th ed.; MIT Press: Cambridge, MA, USA, 2013.
24. Sarajcev, P. Distlines. GitHub. 2022. Available online: https://github.com/sarajcev/distlines (accessed on 1 November 2022).
25. Bishop, C.M. *Pattern Recognition and Machine Learning*; Information Science and Statistics; Springer: New York, NY, USA, 2006; Chapter 7, pp. 325–357.
26. Li, L.; Jamieson, K.; DeSalvo, G.; Rostamizadeh, A.; Talwalkar, A. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *J. Mach. Learn. Res.* **2018**, *18*, 1–52. Available online: https://www.jmlr.org/papers/volume18/16-558/16-558.pdf (accessed on 1 November 2022).
27. Hauschild, W.; Mosch, W. *Statistical Techniques for High–Voltage Engineering*; Peter Peregrinus Ltd.: Herts, UK, 1992. (In English) Originally published in German by VEB Verlag Technik, Berlin, 1984.
28. CIGRE WG. *Lightning Parameters for Engineering Applications*; Brochure 549, CIGRÉ: Paris, France, 2013; Working Group C4.407.
29. Liu, C.H.; Muna, Y.B.; Chen, Y.T.; Kuo, C.C.; Chang, H.Y. Risk Analysis of Lightning and Surge Protection Devices for Power Energy Structures. *Energies* **2018**, *11*, 1999. [CrossRef]
30. *IEC 62305-2*; Protection against Lightning—Part 2: Risk Management. International Standard. IEC: Geneva, Switzerland, 2010.

31. DEHN. *Lightning Protection Guide*, 3rd ed.; Application Guidelines; DEHN: Neumarkt, Germany, 2014.
32. *IEC 61400-24*; Wind Energy Generation Systems—Part 24: Lightning Protection. International Standard. IEC: Geneva, Switzerland, 2019.
33. ABB. *Overvoltage Protection: Metal-Oxide Surge Arresters in Medium-Voltage Systems*, 6th ed.; Application Guidelines; ABB: Wettingen, Switzerland, 2018.