buildings

MDPI

Article

Risk-Informed Dual-Threshold Screening for SPT-Based Liquefaction: A Probability-Calibrated Random Forest Approach

Hani S. Alharbi

Civil Engineering Department, College of Engineering, Shaqra University, Dawadmi 11911, Riyadh, Saudi Arabia; h.alharbi@su.edu.sa; Tel.: +966-50-326-0688

Abstract

Soil liquefaction poses a significant risk to foundations during earthquakes, prompting the need for simple, risk-aware screening tools that go beyond single deterministic boundaries. This study creates a probability-calibrated dual-threshold screening rule using a random forest (RF) classifier trained on 208 SPT case histories with quality-based weights (A/B/C = 1.0/0.70/0.40). The model is optimized with random search and calibrated through isotonic regression. Iso-probability contours from 1000 bootstrap samples produce paired thresholds for fines-corrected, overburden-normalized blow count $N_{1,60,CS}$ and normalized cyclic stress ratio $CSR_{7.5,1}$ at target liquefaction probabilities P_{liq} = 5%, 20%, 50%, 80%, and 95%, with 90% confidence intervals. On an independent test set (n = 42), the calibrated model achieves AUC = 0.95, F_1 = 0.92, and a better Brier score than the uncalibrated RF. The screening rule classifies a site as susceptible when $N_{1,60,CS}$ is at or below and $CSR_{7.5,1}$ is at or above the probability-specific thresholds. Designed for level ground, free field, and clean-to-silty sand sites, this tool maintains the familiarity of SPT-based charts while making risk assessment transparent and auditable for different facility importance levels. Sensitivity tests show its robustness to reasonable rescaling of quality weights. The framework offers transparent thresholds with uncertainty bands for routine preliminary assessments and to guide the need for more detailed, site-specific analyses.

Keywords: building earthquake resilience; cyclic stress ratio; dual-threshold framework; machine learning in geotechnics; probabilistic seismic hazard; random forest classifier; risk-informed seismic design; soil liquefaction assessment; standard penetration test

check for updates

Academic Editor: Yong Tan

Received: 22 July 2025
Revised: 14 August 2025
Accepted: 30 August 2025
Published: 5 September 2025

Citation: Alharbi, H.S. Risk-Informed Dual-Threshold Screening for SPT-Based Liquefaction: A Probability-Calibrated Random Forest Approach. *Buildings* **2025**, *15*, 3206. https://doi.org/10.3390/buildings15173206

1. Introduction

Soil liquefaction is a critical limit state in earthquake engineering, characterized by the temporary loss of soil strength and stiffness resulting from the buildup of excess pore water pressure during seismic activity [1,2]. For buildings and their foundations, liquefaction can weaken load transfer, cause differential settlement, and increase lateral spreading demands, threatening structural integrity and serviceability. Liquefaction therefore exerts a significant impact on the stability of foundations, earth structures, and lifelines [3,4]. Recent earthquakes demonstrate this hazard: the 6 February 2023 Turkey–Syria earthquakes (M_w = 7.7 and M_w = 7.6) caused widespread liquefaction and lateral spreading in urban and agricultural areas, damaging critical infrastructure, homes, and levees [5]. Similarly, the 27 July 2022 Northwestern Luzon earthquake in the Philippines (M_w = 7.0) triggered extensive liquefaction, resulting in sand boils and ground fissures that severely impacted local communities [6]. These recent events, along with historically significant ones such as the 2010–2011 Canterbury earthquakes in New Zealand [5] and the 2011 M_w = 9.0

Tōhoku earthquake in Japan [7] underscore the urgent need for reliable criteria to predict liquefaction occurrence and to inform performance-based building design. Although progress has been made in in situ testing, the Standard Penetration Test (SPT) remains one of the most widely used methods for estimating liquefaction resistance [8,9]. As a result, practitioners require more effective assessment tools that can utilize traditional SPT parameters (such as normalized blow count) to evaluate liquefaction risk directly, rather than relying solely on deterministic safety factors.

The seminal SPT-based liquefaction chart by Seed and Idriss [10] introduced a single empirical curve that differentiates between liquefiable and non-liquefiable states based on cyclic stress ratio (CSR) versus SPT blow count. Seed's simplified method was later refined by Seed in 1979 [11] and became the standard; however, it offers a single-risk, fixed CSR–N boundary that fails to account for different failure probabilities. Essentially, the traditional deterministic threshold reflects a fixed safety margin, roughly corresponding to a P_{liq} of about 15% for the design earthquake, without allowing engineers to adjust the acceptable risk level. Post-1990 updates, such as the Idriss and Boulanger refined SPT correlations, which include improved overburden and magnitude scaling corrections [12], continue to produce generic curves without specifying an explicit probability level. As a result, when designing shallow or deep foundation systems for buildings in seismic regions, engineers often must adopt conservative assumptions (e.g., an FS greater than 1.5) or rely on professional judgment that may lack formal risk calibration [4].

In the 2000s, researchers began introducing probabilistic liquefaction models to estimate the likelihood of triggering. Cetin et al. (2004) [13] developed a reliability-based SPT correlation, and Moss et al. (2006) [14] created a complementary CPT-based model, each providing an estimated P_{liq} for specific site parameters. These pioneering studies incorporated risk into the assessment. However, the resulting equations are mathematically complex and cannot be easily simplified into straightforward field curves, limiting their use in routine practice. More recently, machine learning (ML) classifiers have been applied to liquefaction datasets. For example, Rahman et al. (2025) used support vector machines on regional case data and reported excellent predictive performance (area under the receiver operating characteristic (ROC) curve values around 0.98) [15]. Similarly, high accuracy (AUC > 0.95) has been achieved using ensemble methods, such as gradient boosting and RF [16,17]. Nonetheless, these black box models often lack transparent, calibrated threshold values that practitioners can interpret and trust. As noted in a recent review [18], many ML-based liquefaction models are presented with excessive complexity or are not made accessible to engineers, which has hindered their acceptance in practice, especially within the building design community, where prescriptive and auditable criteria are essential.

The limitations discussed above motivate a new approach that blends data-driven rigor with practical simplicity. The present work addresses two key gaps. First, it calls for a dual-parameter, probability-graded liquefaction criterion that directly supports seismic foundation design and is (i) based on field data, (ii) calibrated to an actual P_{liq}, and (iii) simplified for use by engineers. Unlike traditional single-line charts, a probability-explicit criterion allows engineers to choose different safety levels (e.g., 5%, 20%, 50%, 80%, or 95% chance of liquefaction) and determine the corresponding SPT-CSR cutoff values. The second gap pertains to measuring model uncertainty, such as through bootstrap resampling, and accounting for varying case history qualities without adding unnecessary scatter to the final design thresholds. To do this, the work employs a case weighting scheme that emphasizes high-quality SPT data, similar to the expert judgment filters used by Moss et al. [14]. It uses bootstrapped CIs to ensure the thresholds are statistically reliable. Figure 1 shows the study workflow, which includes compiling the SPT database, training the RF model, calibrating probabilities, and converting the results into a straightforward design rule.

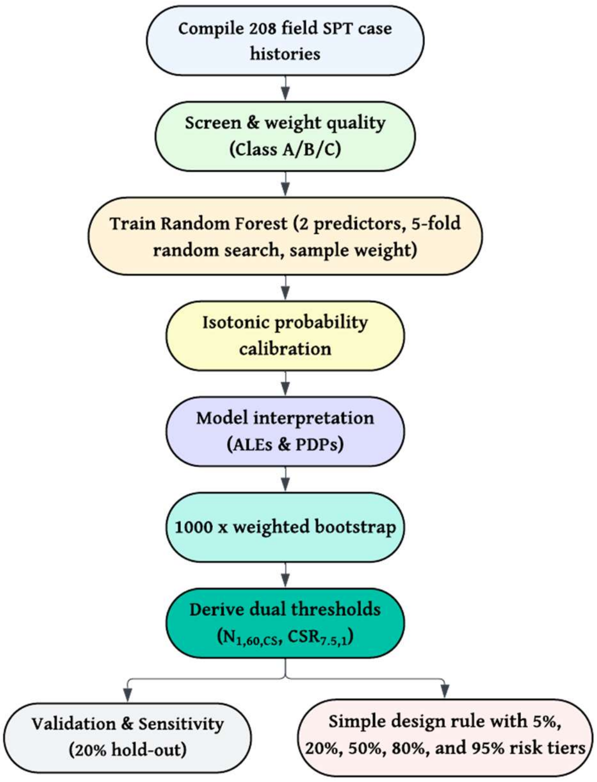
Compile 208 field SPT case histories
Screen & weight quality (Class A/B/C)
Train Random Forest (2 predictors, 5-fold random search, sample weight)
Isotonic probability calibration
Model interpretation (ALEs & PDPs)
1000 x weighted bootstrap
Derive dual thresholds ($N_{1,60,CS}$, $CSR_{7.5,1}$)
Validation & Sensitivity (20% hold-out)
Simple design rule with 5%, 20%, 50%, 80%, and 95% risk tiers

Figure 1. Workflow: SPT data → random forest model → calibration → dual-threshold rule.

The goal is to develop a liquefaction evaluation framework that preserves the familiarity of SPT-based charts while offering rigorous, risk-focused guidance for building projects of different importance levels, distinguishing between ordinary and critical facilities. The contribution is translational: transforming a widely used, openly accessible SPT liquefaction database into a calibrated, probability-based dual-threshold screening rule with 90% bootstrap confidence intervals and an explicit applicability domain through the Coverage Extrapolation Map. The advancement includes four key aspects: (i) a quality-weighted, isotonic-calibrated random forest restricted to $N_{1,60,CS}$ and $CSR_{7.5,1}$ for transparency; (ii) bootstrap extraction of iso-probability contour pairs at selectable Pliq; (iii) delineation of the applicability domain to identify extrapolation; and (iv) a concise decision flow for routine screening. This approach complements ongoing database expansion efforts (e.g., recent global SPT updates), with a focus on operationalizing risk-explicit thresholds for practical use. This study presents a practical, risk-aware tool for preliminary liquefaction assessment, offering insights into the conservatism of traditional charts compared to an optimized, data-driven model that is directly applicable to seismic foundation design in the built environment.

The rest of this paper is organized as follows. Section 2 reviews previous deterministic, probabilistic, and ML methods for SPT-based liquefaction assessment, highlighting remaining gaps. Section 3 describes the field case history database and explains the development of the quality-weighted, probability-calibrated RF model. Section 4 evaluates the model's performance and establishes the probabilistic SPT-CSR dual thresholds. Section 5 compares the proposed criteria with legacy boundaries, assesses the impact of data quality weighting, and discusses limitations and potential future extensions. Section 6 summarizes the key findings and recommendations for implementation.

2. Literature Review

Traditional deterministic, stress-based frameworks are used to assess soil liquefaction potential, relying on empirical correlations between cyclic stress demands and soil resistance. Seed and Idriss (1971) [10] introduced the first simplified stress-based approach, linking the cyclic stress ratio (CSR) required for liquefaction to Standard Penetration Test (SPT) blow counts normalized to a standard overburden pressure and hammer energy ($N_{1,60}$). Their key chart distinguished between liquefiable and non-liquefiable soils based on a binary boundary, calibrated using case histories from the 1964 Alaska and Niigata earthquakes. The simplified liquefaction evaluation method proposed by Seed and Idriss (1971) [10,11] was further developed by Seed (1979) [11] to address cyclic mobility and level ground conditions.

Seed et al. (1985) [19] further refined it by standardizing SPT blow counts to $(N_1)_{60}$ with energy corrections (C_E) and introducing fines content (FC) corrections, building on the overburden normalization factor (C_N) from 1971. These advancements, particularly the $(N_1)_{60}$ standard, have been incorporated into international engineering standards and guidelines, influencing modern building foundation design practice. The Seed and Idriss procedure, as refined, standardizes inconsistent data across case histories using CN, CE, and FC corrections [10,11,19]. Despite these improvements, the deterministic nature of the original method had inherent limitations, including a fixed reference earthquake magnitude of M_w = 7.5 and the inability to evaluate uncertainty explicitly [8,20]. For construction projects characterized by different importance factors, this inflexibility compels designers to either accept an uncertain degree of conservatism or utilize ad hoc safety factors that may not be consistent with performance-based regulatory objectives.

Recognizing the need for risk-focused assessment, probabilistic methods appeared in the late 1980s. Liao and Whitman (1988) employed logistic regression to develop liquefaction probability curves, marking a significant shift from purely deterministic perspectives [21]. Their method provided a quantitative risk framework, although it remained relatively simple due to the limited data and covariates available. Later, Youd and Noble (1997) improved logistic techniques by including more earthquake data, further advancing probabilistic SPT-based criteria [22]. These developments opened the door for explicitly using exceedance probabilities in foundation design decisions.

Cetin et al. (2004) made a significant advance by applying Bayesian statistical updates to liquefaction assessment [13]. Their framework explicitly incorporates uncertainty through a comprehensive reliability-based model that considers site parameters such as vertical effective stress ($\sigma_v{}'$) and fines content (FC). Additionally, Cetin et al. grouped field case histories into three quality categories (A, B, and C). Each field case was rated A–E for data quality; lower-quality cases carry larger variances in the likelihood function, giving high-quality data proportionally greater influence. The resulting contours for P_{liq} = 5–95% provide practical, probability-based triggering thresholds; however, designers must still select an appropriate target probability [23].

Parallel advancements occurred in alternative in situ test-based assessments, exemplified by Moss et al. (2006) [14], who developed a Cone Penetration Test (CPT)-based Bayesian approach, offering probabilistic insights more directly linked to continuous, digitally recorded measurements. Similarly, Kayen et al. (2013) [24] introduced shear-wave velocity (V_s)-based probabilistic models, broadening the options for liquefaction assessment; however, both CPT- and V_s-based methods remain mathematically complex, limiting their routine use by building practitioners, who still rely heavily on SPT data.

In subsequent developments, Idriss and Boulanger (2010, 2014) [12,25] revisited deterministic frameworks by significantly recalibrating the original Seed and Idriss CSR-N curves. Their updated relationships included refined adjustments for fines content and improved magnitude scaling corrections, addressing several empirical limitations of the original model. Boulanger and Idriss (2014) [25] notably introduced a probabilistic interpretation by outlining a CSR-N curve that approximately corresponds to a P_{liq} of 15%. However, their curves were still presented graphically rather than analytically, which limited straightforward integration into automated design workflows or Building Information Modeling (BIM) platforms.

Against this historical backdrop, the past decade has seen an increase in the application of ML techniques to liquefaction research. ML models such as RF, gradient boosting, and Artificial Neural Networks (ANNs) have demonstrated predictive accuracies with AUC values above 0.95 [15–17,26,27]. The primary benefit of these models is their ability to identify complex, nonlinear relationships and interactions in large liquefaction datasets, significantly outperforming traditional regression-based methods in terms of classification accuracy [16]. However, adoption within structural engineering offices has been slow because most ML studies only provide predictive scores and do not supply code-based thresholds that can be directly cited in foundation design reports.

Despite these strengths, ML-based approaches often face criticism for their limited interpretability and practicality. The models, usually referred to as “black boxes,” frequently lack clear, calibrated threshold standards, making it difficult for practitioners to turn high predictive performance into actionable decisions [27,28]. Additionally, ML probabilities often need external calibration to be compatible with reliability-based design frameworks used in building codes.

In response to these concerns, recent research has begun integrating robust classifiers with interpretability tools in liquefaction and soil hazard contexts. For example, Jas et al. expanded Cetin’s database and used XGBoost with Synthetic Minority Over-sampling Technique (SMOTE) resampling to predict liquefaction; Shapley Additive exPlanations (SHAP) analysis showed that equivalent clean sand cone penetration resistance and the permeability coefficient were the primary factors influencing predictions [29]. Hsiao et al. [30] developed XGBoost models to forecast liquefaction-induced lateral spreading following the 2011 Christchurch earthquake; SHAP analysis found PGA to be the most influential global feature, while CPT-derived soil parameters provided valuable local explanatory insights but had lower overall importance. Other groups combined ensemble models and deep learning: a recent Scientific Reports study compared Bi-LSTM, XGBoost, and random forest on an extensive CPT-based database, finding that although recurrent networks achieved the highest accuracy, SHAP applied to the XGBoost model identified normalized tip resistance, peak ground acceleration, and the stress-reduction ratio as the most influential factors [31]. A smaller dataset study on granular soils compared SVM (polynomial and RBF kernels) and random forest classifiers, using model-based feature importance to identify cone tip resistance, overburden stress, peak ground acceleration, earthquake magnitude, liquefied layer thickness, and gravel content as the most influential factors in liquefaction classification [32]. Beyond liquefaction, interpretable frameworks have been used for tunnel convergence and expansive soil swelling; in these cases, SHAP and partial dependence plots quantify how time, rock quality, and the plasticity index influence predictions and reveal threshold behaviors [33,34]. Complementing these efforts on the structural side, a 2025 Structures study on steel diagrid systems trained 21 algorithms on incremental dynamic analysis (IDA)-generated datasets across 4–24-story buildings; tree-based ensembles (Extra Trees, bagging, random forest, Stacking, k-nearest neighbors (KNNs)) achieved $R^2 > 0.95$, and feature-importance/sensitivity analyses highlighted structural weight, fundamental period T_1, and spectral acceleration at the fundamental period ($Sa(T_1)$) as dominant predictors [35]. However, probability calibration and uncertainty quantification were not performed. Overall, these studies demonstrate that integrating advanced ML techniques with SHAP/TreeSHAP, partial-dependence plots (PDPs)/accumulated local effects (ALEs), and similar tools can uncover physically meaningful insights from complex data. Nonetheless, most still report uncalibrated probabilities and stop short of proposing simple, clearly defined decision thresholds.

The quality and completeness of field data remain vital issues in liquefaction modeling. Cetin et al. (2004) [13] pioneered a quality class weighting system (classes A, B, C) that set a standard by measuring data quality uncertainty within probabilistic models. Recent efforts, such as those by Ilgac et al. (2022) [36], have greatly expanded SPT liquefaction databases, compiling an extensive list of more than 400 quality-ranked case histories. Their analyses emphasize how weight assignments heavily influence both regression-based models and ML classifiers, with even minor changes in weighting schemes potentially altering the predictive outcomes. This sensitivity underscores the importance of explicitly integrating data quality considerations into modern probabilistic and ML-based liquefaction evaluations.

Despite significant methodological advances, notable gaps still exist. No existing studies have transformed complex ML classifiers into simple, dual-threshold decision rules that (i) are explicitly calibrated for quantified P_{liq}, (ii) include statistical CI, and (iii) are easy enough to be incorporated into building foundation design checklists. Additionally, the robustness of such thresholds under uniform down-weighting scenarios, which are common when older site investigations are incomplete, has received little attention.

This study addresses these gaps by combining a calibrated RF classifier with probability calibration techniques. PDP and ALE plots enhance interpretability, and bootstrap resampling measures uncertainty, yielding transparent, statistically reliable dual-threshold rules that engineers can easily apply to select risk-appropriate SPT cutoffs for standard, necessary, or essential building facilities. This method blends the predictive power of modern ML with the clarity needed for risk-based building design criteria.

3. Methodology

3.1. Field Case Database and Screening

The dataset used in this study comes from the open-access SPT-based liquefaction case history collection published by Cetin et al. (2018) [37], which expands the foundational catalog originally developed by Cetin et al. (2004) [13]. This compilation gathers high-quality field observations from multiple seismic events, each with geotechnical and seismological parameters relevant for assessing liquefaction potential. To keep the model practical and straightforward, only two predictor variables were retained: (i) $N_{1,60,CS}$ the fines-corrected, overburden-normalized Standard Penetration Test (SPT) blow count, expressed in blows per foot (blows/ft), and (ii) $CSR_{7.5,1}$ the cyclic stress ratio normalized to a reference earthquake magnitude (M_w = 7.5) and vertical effective stress of 1 atm. The binary target variable "Liquefied" indicates field performance observations, where 1 signifies confirmed liquefaction and 0 means no liquefaction. Among the 210 total cases, two records were marked as "Yes/No (marginal)," indicating inconclusive or uncertain liquefaction results. These marginal records were removed to prevent bias during model training, resulting in a final dataset of 208 clear cases: 113 liquefied and 95 non-liquefied. All analyses used the publicly available SPT-based liquefaction case history dataset. No new field data was gathered. The deliberate choice of a single, widely cited source enhances reproducibility and comparability with established curves; the innovation lies in probability calibration, bootstrap threshold extraction, and applicability-domain mapping, rather than expanding the database. In this study, $CSR_{7.5,1}$ denotes site demand normalized to standard conditions. The free field cyclic stress ratio is first obtained from the simplified procedure,

$$CSR = 0.65 \left(\frac{a_{max}}{g}\right)\left(\frac{\sigma_v}{\sigma_v{}'}\right) r_d(z)$$

where a_{max} is peak ground acceleration, g is gravity, σ_v and $\sigma_v{}'$ are total/effective overburden, and $r_d(z)$ is the depth-dependent stress-reduction factor.

Including the depth-dependent stress-reduction factor $\mathrm{r_d}$(z) and overburden effects. CSR is then normalized to $\mathrm{M_w}$ = 7.5 using an MSF and to $\sigma_v{}'$ = 1 atm, resulting in $\mathrm{CSR_{7.5,1}}$. Hence, $\mathrm{CSR_{7.5,1}}$ encompasses standard depth and stress adjustments along with magnitude scaling.

Table 1 presents the definitions, roles, and units for the remaining three columns: $\mathrm{N_{1,60,CS}}$ and $\mathrm{CSR_{7.5,1}}$, which serve as numerical predictors, and the Liquefied column, which is the target variable. Although the original database included additional parameters such as fines content, effective stress, and magnitude, this study's modeling framework deliberately focuses on a simple set of features to improve interpretability and practical use. Restricting predictors to $\mathrm{N_{1,60,CS}}$ and $\mathrm{CSR_{7.5,1}}$ minimizes cross-study heterogeneity (e.g., inconsistent fines content corrections and magnitude scaling corrections), reduces missingness and imputation bias, and aligns with the data most commonly available in geotechnical investigations. The two-parameter specification also supports transparent, auditable criteria favored in practice, addressing the adoption barriers noted for complex ML models. In addition to the predictors and target, each case is assigned a quality label (Class A, B, or C), which is later used to develop a quality-weighted sample scheme during model training. This approach provides a clear and consistent framework for $\mathrm{P_{liq}}$ assessment using only widely available SPT-based parameters.

Table 1. Variables retained for random forest modeling (n = 208 field cases).

Variable	Unit	Role	Description
Liquefaction status	-	Target	Field outcome (Yes = liquefied, No = non-liquefied); encoded 1/0.
$\mathrm{N_{1,60,CS}}$	Blows/ft	Predictor	Fines-corrected, overburden-normalized SPT blow count.
$\mathrm{CSR_{7.5,1}}$	-	Predictor	Cyclic stress ratio normalized to $\sigma_v{}'$ = 1 atm and $\mathrm{M_w}$ = 7.5
Data class	A, B, C	Sample weight	Quality category

Descriptive statistics for the two selected input variables are provided in Table 2. The variable $\mathrm{N_{1,60,CS}}$ ranges from 4.95 to 66.46, with an average of 18.24 and a standard deviation of 11.79. $\mathrm{CSR_{7.5,1}}$ varies from 0.03 to 0.51, with an average of 0.22 and a standard deviation of 0.11. Median values and interquartile ranges are also included, confirming that the dataset covers a wide distribution of soil resistances and seismic demands. This statistical variation suggests that the model can effectively generalize across a diverse range of geotechnical conditions relevant to practical seismic design.

Table 2. Descriptive statistics for key variables (n = 208).

Statistic	$\mathbf{N_{1,60,CS}}$	$\mathbf{CSR_{7.5,1}}$
Count	208	208
Mean	18.24	0.22
Std. dev.	11.79	0.11
Min.	4.95	0.03
25th percentile	9.96	0.14
Median	14.36	0.21
75th percentile	23.15	0.28
Max.	66.46	0.51

The Cetin et al. (2004) [13] SPT-based liquefaction triggering database includes field case histories from various international earthquakes across different tectonic settings such as subduction zones, crustal strike-slip, and typical faulting regimes. It features sites from the United States, Japan, Türkiye, Taiwan, and New Zealand, among others, spanning North America, the Asia-Pacific region, and parts of Europe and the Middle East. The site conditions cover a wide array of geomorphic and depositional environments, including coastal and deltaic plains, estuarine and river alluvium, lake deposits, and reclaimed port fills. The soils mainly consist of clean-to-silty sands, with occasional sandy gravels. The database records a broad range of corrected blow counts ($N_{1,60,CS}$) and equivalent cyclic stress ratios ($CSR_{7.5,1}$), supporting the overall applicability of the probabilistic and deterministic triggering correlations derived from it. This geotechnical and seismic diversity improves the relevance of the dual-threshold screening framework for routine foundation design in coastal and alluvial basin environments.

Figure 2 shows a visual representation of the 208 field cases, plotting $CSR_{7.5,1}$ against $N_{1,60,CS}$, with liquefied cases highlighted in red and non-liquefied cases in blue. The figure clearly illustrates the inverse relationship between cyclic stress demand and in situ resistance, with liquefaction typically occurring in cases of high $CSR_{7.5,1}$ and low $N_{1,60,CS}$. This visual separation confirms the effectiveness of the chosen predictors and further supports the development of a dual-threshold rule based on this two-dimensional space.

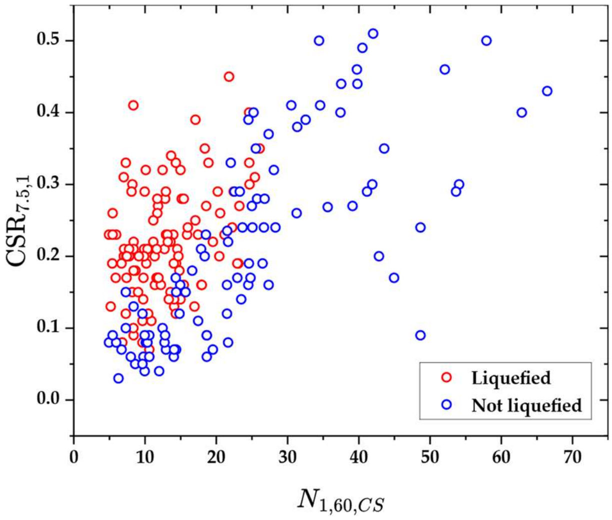

Figure 2. $CSR_{7.5,1}$ vs. $N_{1,60,CS}$ scatter; red = liquefied, blue = not liquefied. Data: Cetin et al. (2018) [37].

To visualize how well the 208 weighted cases populate the ($N_{1,60,CS}$, $CSR_{7.5,1}$) plane, a bivariate kernel density estimate [38] was computed on the standardized predictors (z-scored to a mean of zero and unit variance) using the quality weights w_i = (1.0, 0.70, 0.40). The smallest contours that enclose 50% and 90% of the total weighted probability mass were extracted; these define a Core zone (densest half of the weighted data, where predictions rely on strong interpolation), a Support zone (the next 40% of weighted data, still interpolative but sparser), and an Extrapolation zone (the outer 10%, where little or no data exists and predictions must be used with caution).

Figure 3 shows the resulting Coverage Extrapolation Map (CEM): 60% of the weighted evidence (122 cases) lies in the Core zone, 33% (71 cases) lies in the Support zone, and only 7% (15 cases) falls in the Extrapolation zone, providing a clear basis for judging where data strongly supports subsequent model predictions.

Using data outside the Core/Support zones involves statistical extrapolation; such predictions should only be considered screening-level and must be validated against independent observations. When possible, recent region-specific events (e.g., the 2023 Turkey–Syria earthquakes) should be used as an external check to confirm the transferability of the dual-threshold rule under local geologic and seismotectonic conditions.

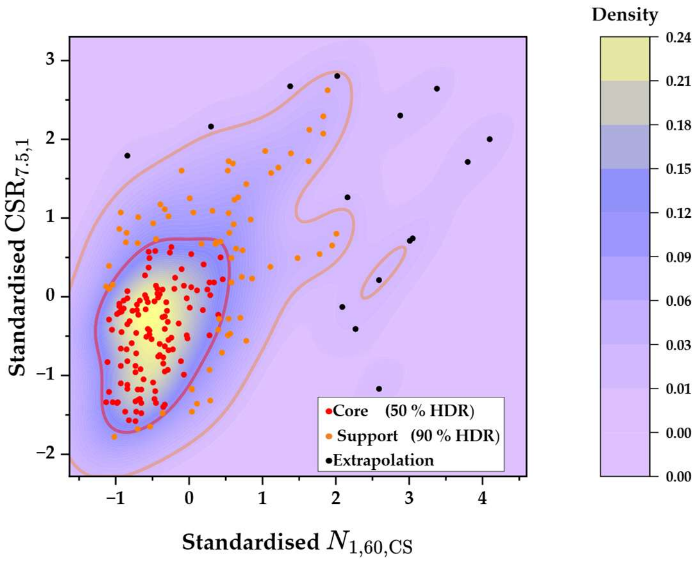

Figure 3. Coverage Extrapolation Map (CEM).

3.2. Pre-Processing and Quality-Weighted Sample Scheme

Cetin et al. [13] classify each field case based on how well the critical stratum is sampled and the statistical spread of its cyclic stress estimate (using quality classes A, B, and C as reported unchanged in Cetin et al. (2018)) [37]. Class A requires at least three corrected SPT blows in the liquefiable layer, well-documented equipment and procedures, and a coefficient of variation for CSR (COV-CSR) no more than 0.20. Class B retains the documentation requirement but allows either $0.20 < \text{COV-CSR} \leq 0.35$ or fewer than three N-values per stratum. Class C still has complete procedural details but permits $0.35 < \text{COV-CSR} \leq 0.50$.

To account for the growing measurement uncertainty while preventing lower-quality data from dominating, the analysis assigns weights of 1.0, 0.70, and 0.40 to Classes A, B, and C, respectively. Quality-based reweighting aligns with reliability-based liquefaction frameworks that lessen the impact of lower-quality cases by assigning larger variances in the likelihood, effectively down-weighting them [13,14,24]. Moving from Class A ($\text{COV-CSR} \leq 0.20$) to Class B ($0.20 < \text{COV-CSR} \leq 0.35$) involves a moderate decrease; the 0.70 factor reflects this initial increase in dispersion. The 0.40 factor for Class C ($0.35 < \text{COV-CSR} \leq 0.50$) indicates significantly greater uncertainty, as documented in the Cetin et al. quality taxonomy and the variance inflation used for lower-confidence cases in CPT- and Vs-based probabilistic models [13,14,24]. Recent meta-analyses of SPT databases also emphasize the sensitivity of model results to case-quality weighting and endorse explicit down-weighting of lower-quality records [36].

3.3. Random Forest Development

The dataset comprises 208 field case histories (113 liquefied, 95 non-liquefied). It was partitioned into training (80%, n = 166) and testing (20%, n = 42) subsets to maintain the original class imbalance, thereby ensuring that both subsets accurately represent the liquefaction proportions and improve evaluation reliability [39]. Sample weights, assigned according to data quality classifications (A = 1.0, B = 0.7, C = 0.4), were utilized to prioritize higher-quality observations during the model training process.

An RF classifier was trained using $N_{1,60,CS}$ and $CSR_{7.5,1}$ as predictors, with the binary liquefaction outcome serving as the target. Hyperparameter tuning was conducted through a randomized search involving 100 parameter combinations within a five-fold stratified cross-validation framework, aiming to optimize the area under the ROC curve [40].

The search space included 60 to 300 trees; max_depth set to None or values between 3 and 15; min_samples_split ranging from 2 to 10; min_samples_leaf ranging from 1 to 10; max_features selected from "sqrt," "log2," or None; and both bootstrap options considered. A fixed random seed (random_state = 42) ensured reproducibility. The best configuration consisted of 97 trees, with max_depth = 5, min_samples_split = 2, min_samples_leaf = 8, max_features = None, and bootstrap enabled. Five-fold stratified learning curves showed an AUC plateau between approximately 80 and 150 trees; choosing 97 estimators at the curve's knee minimized variance without additional cost. A shallow depth ($\leq$5), with at least 8 samples per leaf, provided the best bias–variance trade-off in cross-validated AUC and Brier loss.

RF was selected as the primary classifier because it provides a good balance between predictive accuracy, calibration, and computational efficiency for small tabular datasets [41]. The approach involves building multiple decision trees on bootstrap samples and combining their predictions, which reduces variance and ensures stable performance even with limited data [41–43]. It has a few hyperparameters and is relatively insensitive to their settings, making model tuning easier [42,43]. In contrast, gradient boosting algorithms (e.g., XGBoost, LightGBM) can achieve high accuracy but tend to overfit small datasets and require careful tuning of learning rates, tree depth, and regularization [44–46]. Support vector machines with RBF kernels need cross-validation for probability calibration and are sensitive to the penalty parameter and kernel width [47,48]. Logistic regression offers well-calibrated probabilities but assumes a linear decision boundary and may miss nonlinear interactions present in liquefaction behavior [49]. By combining bagging with random feature selection, RF maintains the ability to model nonlinear relationships while staying stable across bootstrap samples, and isotonic regression can adjust the slight bias in its raw probability estimates.

3.4. Probability Calibration

Predicted probabilities from the RF were recalibrated using isotonic regression [50,51]. A five-fold stratified cross-validation on the training set created an adjusted mapping between raw probabilities and observed liquefaction frequencies, which was then applied to the independent test set. Calibration quality was evaluated using reliability diagrams, which show the average predicted probability versus the observed frequency across ten quantile bins, and by calculating the Brier score loss. Discrimination before and after calibration was compared using the area under the ROC curve to confirm that probability adjustment did not harm overall model performance.

3.5. ALE and PDP Computation

PDPs were created by assessing the model's predicted P_{liq} across the observed range of each predictor, while fixing the other predictor at its median value [52]. ALE profiles were generated by dividing each predictor into 40 quantile-based bins, ensuring each bin contains an equal number of observations for more reliable effect estimates [53]. The local differences in predictions between neighboring bin edges were calculated, and then these differences were cumulatively summed. The curve was centered at a zero mean to emphasize deviations from the average. A two-dimensional ALE surface illustrating the combined effects of $N_{1,60,CS}$ and $CSR_{7.5,1}$ was subsequently produced by integrating local effects along both predictor axes.

3.6. Contour-Based Threshold Extraction and Bootstrap Confidence Intervals

Dual thresholds for $N_{1,60,CS}$ and $CSR_{7.5,1}$ were established using a contour-based extraction method combined with bootstrap resampling to assess uncertainty. The contour-based extraction method involves generating a modeled probability surface over a 2D grid and extracting iso-probability contour lines (level sets) at specified probabilities to identify threshold points [29,54–57]. One thousand bootstrap samples were generated by sampling with replacement from the training set [58,59]. For each replicate, the RF model was retrained and recalibrated, and a complete probability surface over $N_{1,60,CS}$ and $CSR_{7.5,1}$ was generated. Contour lines corresponding to target P_{liq} of 0.05, 0.20, 0.50, 0.80, and 0.95 were then derived from these surfaces to find pairs of threshold values. These threshold pairs were aggregated across all 1000 replicates, and the median threshold for each probability level, along with its 90% CI, was calculated. The procedure captures both model variability and sampling uncertainty, providing robust, probability-based criteria for practical liquefaction evaluation. The number of bootstrap replicates was set to 1000 after a convergence check over 200–2000 replicates; median thresholds and 90% CI endpoints were stabilized by approximately 800 replicates, with only marginal changes afterward relative to the reporting precision; thus, the choice of 1000 balances stability and computational cost.

3.7. Dual-Threshold Rule and Validation Procedure

The probability-based thresholds outlined in Section 3.6 are combined into a dual-threshold screening rule for field use. For a specified P_{liq} level (P = 0.05, 0.20, 0.50, 0.80, or 0.95), a site is classified as susceptible if $N_{1,60,CS}$ does not surpass the median bootstrap threshold and if $CSR_{7.5,1}$ meets or exceeds the paired threshold. The criterion reflects the physical condition under which liquefaction can only occur when soil resistance is low and cyclic loading is sufficiently high [13].

The dual-threshold rule is most suitable for initial risk screening and design checks at sites characterized by level ground, free field conditions where SPT governs the investigation scope and soils are clean-to-silty sands. It is reliably used within the Core/Support coverage of the compiled database (Figure 3), with recommended input ranges of $N_{1,60,CS}$ roughly 5–35 (typically 10–30) and $CSR_{7.5,1}$ approximately 0.08–0.35, assuming the standard corrections applied to form $N_{1,60,CS}$ and $CSR_{7.5,1}$ are appropriate. Situations that often require additional predictors or alternative/auxiliary models include pronounced ground slope or static shear bias; thin interbedded liquefiable layers; excellent soils with high plasticity or fines content outside clean-sand adjustments; gravelly, cemented, or crushable soils; partially saturated or ground-improved deposits; and cases where cyclic mobility is a key concern. In such cases, incorporating variables such as fines content, vertical effective stress, and magnitude scaling factors, or employing CPT- or Vs-based probabilistic procedures alongside this rule, is recommended.

Validation utilized a separate 20% hold-out test set (n = 42) that was entirely independent of all training, calibration, and threshold-setting steps [39]. Observation weights based on data quality classes (A = 1.0, B = 0.70, C = 0.40) were maintained to ensure consistency in penalizing misclassifications. Performance was assessed using a weighted confusion matrix, from which accuracy, precision, recall, and the F_1 score were calculated [60]. A sensitivity analysis then evaluated the robustness of the rule to different weight assumptions by uniformly rescaling the weights for Classes B and C, while keeping the thresholds fixed. This analysis demonstrated how changing the importance of lower-quality records influences predictive performance without altering the current decision criteria.

4. Results

4.1. Model Discrimination and Calibration

The evaluative metrics for the RF models on the test set of 42 cases are summarized in Table 3. The uncalibrated RF model achieved an AUC of 0.96, demonstrating strong discriminative ability, along with an accuracy of 0.92, an F_1 score of 0.93, and a Brier score of 0.10, with an optimal classification threshold of 0.79. In comparison, the isotonic-calibrated version had a slightly lower AUC of 0.95, an accuracy of 0.91, an F_1 score of 0.92, and an improved Brier score of 0.09, indicating better calibration, at an optimal threshold of 0.77. These results highlight the models' strong performance in distinguishing between liquefaction and non-liquefaction cases, with the calibrated model providing marginally better probability estimates, as indicated by the lower Brier score.

Table 3. Performance of RF models on test set (n = 42).

Model	AUC	Accuracy	F_1	Brier	Optimal Threshold
RF (uncalibrated)	0.96	0.92	0.93	0.10	0.79
RF (isotonic-calibrated)	0.95	0.91	0.92	0.09	0.77

Figure 4 shows the receiver operating characteristic (ROC) curves for both models. Panel (a) presents the uncalibrated model's ROC curve, which closely approaches the top-left corner with an AUC of 0.96, greatly surpassing the chance diagonal (AUC = 0.50). Panel (b) displays the isotonic-calibrated model's ROC curve, which maintains high discriminative ability with an AUC of 0.95, again exceeding the baseline, confirming that calibration adjustments do not compromise overall classification performance.

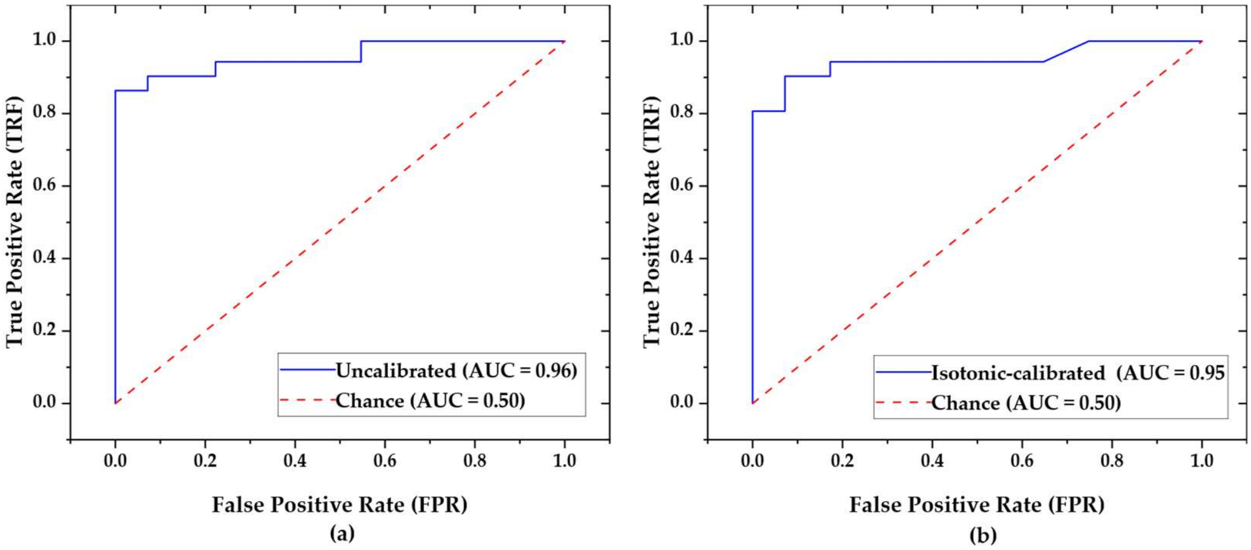

Figure 4. ROC curves: (**a**) uncalibrated, (**b**) isotonic-calibrated.

The calibration quality of the isotonic-calibrated RF model is further examined in Figure 5, which presents a reliability diagram [61] plotting the observed fraction of positive outcomes against the mean predicted probability of a positive outcome. The blue line representing the isotonic-calibrated predictions demonstrates reasonable alignment with the ideal red dashed line (y = x), particularly in the mid-to-high probability ranges. However, minor deviations occur at lower probabilities. This visual assessment confirms the model's well-calibrated probabilistic outputs, where predicted probabilities closely align with empirical frequencies, thereby enhancing its reliability for risk-based decision-making in liquefaction assessments.

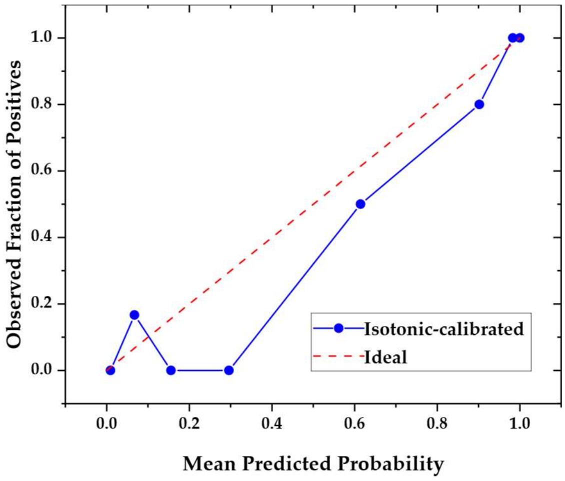

Figure 5. Calibration curve for the isotonic-calibrated RF model.

4.2. Feature Importance

Feature importance in the isotonic-calibrated RF model was assessed using the Gini impurity metric [62], which evaluates each predictor's contribution to the overall decision process by measuring the reduction in node impurity across all trees. As shown in Figure 6, $N_{1,60,CS}$ appears to be the most influential variable, with a Gini importance of 0.515. Meanwhile, $CSR_{7.5,1}$ is close behind at 0.485. This distribution highlights the similar but slightly different roles of in situ soil resistance and seismic demand in distinguishing liquefaction outcomes, consistent with established geotechnical principles, where penetration resistance often indicates susceptibility under specific loading conditions [63].

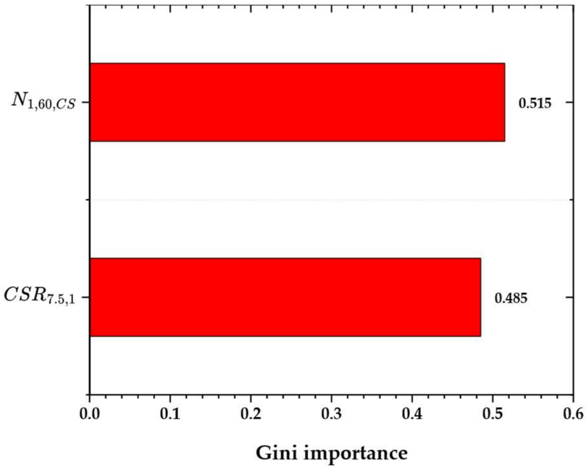

Figure 6. Predictor importances from the calibrated RF.

4.3. ALE and PDP Behavior

To clarify how the calibrated RF model responds to individual predictors, PDPs and ALEs were created, providing insights into the marginal and localized influences of $N_{1,60,CS}$ and $CSR_{7.5,1}$ on P_{liq}. Figure 7 compares the calibrated PDP (in red) and centered ALE (in blue) for each variable. For $N_{1,60,CS}$, both metrics show an apparent nonlinear decrease in predicted probability as blow counts increase, with a steep drop from about 1.0 at low values (below 10) to near zero beyond 30, emphasizing the key role of penetration resistance in reducing liquefaction risk. The ALE curve, being centered, highlights deviations from the average effect, revealing more substantial negative impacts at middle values around 15–25. Conversely, for $CSR_{7.5,1}$, the plots show a sharp rise in probability with increasing stress ratio, transitioning from minimal risk below 0.1 to almost certain liquefaction above 0.3, with the ALE indicating positive deviations in the 0.15–0.25 range, where seismic demands heavily interact with soil properties. The close agreement between the PDPs and ALEs suggests few higher-order interactions, confirming the model's interpretability in this two-variable context [54,57].

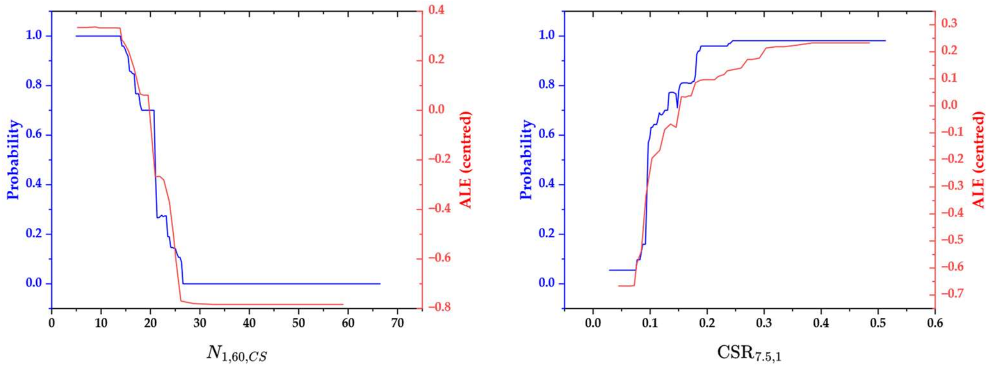

Figure 7. Calibrated partial-dependence and centered ALE for $N_{1,60,CS}$ and $CSR_{7.5,1}$.

Extending the analysis to joint effects, Figure 8 displays the two-dimensional ALE surface, illustrating the combined influence of $N_{1,60,CS}$ and $CSR_{7.5,1}$ on P_{liq} through a color gradient from blue (lower odds) to red (higher odds). Regions with high liquefaction potential are visible where $N_{1,60,CS}$ (<20) is low and $CSR_{7.5,1}$ (>0.15) is high, with contour lines marking transitional zones where small changes in either variable can significantly affect outcomes. Sparse regions, such as those with extremely high $N_{1,60,CS}$ or very low $CSR_{7.5,1}$, should be approached with caution in extrapolation, as the model relies more on interpolation in these areas and may be less accurate. These sparsely populated corners coincide with the Extrapolation zone in the Coverage Extrapolation Map (Figure 3), which explains the visually misleading orange/yellow (high $N_{1,60,CS}$, low $CSR_{7.5,1}$) and green/blue (very high $CSR_{7.5,1}$, very low $N_{1,60,CS}$) patches on the ALE surface.

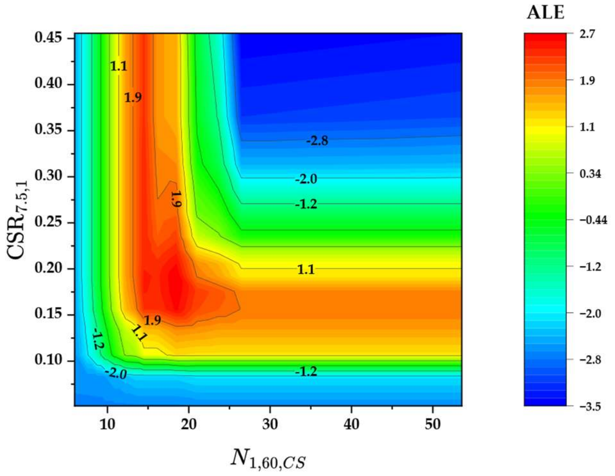

Figure 8. ALE surface for $N_{1,60,CS}$ and $CSR_{75,1}$; warm colors indicate higher liquefaction odds.

4.4. Derived Thresholds

Threshold values for P_{liq} were derived through bootstrapping to quantify uncertainty in the calibrated RF model's predictions, focusing on univariate boundaries for $N_{1,60,CS}$ and $CSR_{7.5,1}$. This method involved resampling the dataset with replacement to generate multiple iterations, from which median thresholds and 90% CIs were calculated for target probabilities ranging from 0.05 to 0.95 [59]. For $N_{1,60,CS}$, the median thresholds decrease steadily as P_{liq} increases, reflecting the inverse relationship between penetration resistance and liquefaction susceptibility: at P_{liq} = 0.05, the median is 26.02 (90% CI: 9.95–47.63), dropping to 15.71 (90% CI: 7.44–20.51) at P_{liq} = 0.95 (see Table 4). The broader intervals at lower probabilities highlight greater variability in defining safe states, where limited data may lead to increased uncertainty in estimation.

Table 4. Bootstrapped dual thresholds (median and 90% CI).

Variable	Target P_{liq}	Median	90% CI Lower	90% CI Upper
$N_{1,60,CS}$	0.05	26.02	9.95	47.63
	0.20	24.22	8.68	26.81
	0.50	21.35	8.68	26.09
	0.80	18.09	7.44	22.98
	0.95	15.71	7.44	20.51
$CSR_{7.5,1}$	0.05	0.21	0.05	0.48
	0.20	0.22	0.08	0.48
	0.50	0.25	0.09	0.49
	0.80	0.26	0.11	0.49
	0.95	0.28	0.14	0.49

In contrast, for $CSR_{7.5,1}$, the median thresholds show a modest increase in P_{liq}, indicating a direct relationship between seismic demand and risk, starting at 0.21 (90% CI: 0.05–0.48) for P_{liq} = 0.05 and rising to 0.28 (90% CI: 0.14–0.49) for P_{liq} = 0.95. The relatively stable upper bound near 0.49 across levels suggests a saturation effect at high demands. Meanwhile, the lower bound expands at higher probabilities, highlighting the need for caution in high-risk assessments. Plotting these thresholds on the CEM (Figure 3) shows that even the extreme-probability points remain within data-supported regions, reinforcing their practical reliability. These dual thresholds provide practical, probability-based decision rules, enabling engineers to adjust conservatism according to acceptable risk levels while considering inherent model and data uncertainties.

4.5. Performance of the Simple Dual Rule

The effectiveness of the proposed dual-threshold rule was assessed on the test set using confusion matrices at selected P_{liq} thresholds (P_{liq} = 0.05, 0.20, 0.50, 0.80, 0.95), measuring the agreement between predicted and observed outcomes. Table 5 shows the average counts for actual non-liquefied cases correctly identified as non-liquefied, false positives (non-liquefied predicted as liquefied), false negatives (liquefied predicted as non-liquefied), and true positives (liquefied correctly detected). At lower thresholds, such as P_{liq} = 0.05, the rule demonstrates a balanced error distribution, with 12.5 true non-liquefied and 9.1 accurate liquefied classifications, although some misclassifications occur (1.4 false positives and 8.5 false negatives). As P_{liq} increases, false positives drop to zero, improving specificity, while false negatives increase, indicating a conservative bias that emphasizes avoiding liquefaction underprediction at the cost of recall. This change highlights the rule's flexibility for risk-averse situations, where higher thresholds favor non-liquefaction predictions in uncertain cases.

Table 5. Confusion matrices for different probability levels.

(P_{liq})	True Non-Liquefied (Pred. Non-Liq.)	True Non-Liquefied (Pred. Liq.)	True Liquefied (Pred. Non-Liq.)	True Liquefied (Pred. Liq.)
0.05	12.5	1.4	8.5	9.1
0.20	13.9	0	9.2	8.4
0.50	13.9	0	12.7	4.9
0.80	13.9	0	12.7	4.9
0.95	13.9	0	13.4	4.2

Complementary performance metrics, summarized in Table 6, further clarify the behavior of the dual rule, including accuracy, precision, recall, and F_1-score. Precision reaches 1 for $P \geq 0.20$, indicating perfect avoidance of false positives and high confidence in optimistic predictions, though at the expense of decreasing recall (from 0.52 at P = 0.05 to 0.24 at P = 0.95), which measures the proportion of actual liquefied cases correctly identified. Accuracy peaks at 0.71 for P_{liq} = 0.20 before declining to 0.57 at higher thresholds. Meanwhile, the F_1-score, which balances precision and recall, is maximized at 0.65 for lower P_{liq} values, suggesting optimal usefulness in scenarios tolerant of moderate conservatism. These metrics confirm the rule's practical value, especially for initial screenings where precision is crucial, while also highlighting the trade-offs involved in threshold selection for detailed geotechnical decision-making.

Table 6. Performance metrics for the simple dual rule.

P	Accuracy	Precision	Recall	F_1-Score
0.05	0.69	0.87	0.52	0.65
0.20	0.71	1.00	0.48	0.65
0.50	0.60	1.00	0.28	0.44
0.80	0.60	1.00	0.28	0.44
0.95	0.57	1.00	0.24	0.39

5. Discussion

5.1. Comparison with Reliability-Based SPT Curves of Cetin et al. (2018)

To put the proposed dual-threshold rule into context, it is compared with the reliability-based SPT correlations of Cetin et al. (2018) [64]. Their curves, derived using maximum-likelihood estimation on the same case history database, provide P_{liq} ranging from 5% to 95% and connect the median contour (P_{liq} = 50%) to an approximate FS of approximately 1.0.

Figure 9 overlays the proposed median thresholds (black markers) on the redrawn contours from Cetin et al. [64]. A visual inspection shows a close agreement in the mid-probability range. For P_{liq} = 50%, the point at $N_{1,60,CS}$ = 21.35 and $CSR_{7.5,1}$ = 0.25 sits close to the median contour ($CSR_{7.5,1} \approx 0.22$). The +14% $CSR_{7.5,1}$ difference yields an implied FS of 0.94, comfortably within the standard design parameters' tolerance.

Table 7 provides a numerical comparison. Implied safety factors were calculated by deriving the cyclic resistance ratio (CRR) from Cetin et al.'s P_{liq} = 50% (FS = 1.0) curve using their Equations (10) and (11) from Cetin et al. (2018) [37] under standard conditions (FC = 5%, M_w = 7.5, $\sigma_v{}'$ = 1 atm) at each median $N_{1,60,CS}$, then dividing that CRR by the corresponding $CSR_{7.5,1}$ median from the current thresholds (FS = CRR/CSR) (an annotated Word document, Supplementary Document S1, details every step of the implied FS calculation). This simple ratio links the probabilistic $CSR_{7.5,1}$ rule to the deterministic FS scale commonly used in design. At the low-risk end (P_{liq} = 5%), the $CSR_{7.5,1}$ value aligns with Cetin's visual estimate, resulting in an implied FS of 1.67 compared to approximately 1.5, making it slightly more conservative. At P_{liq} = 20%, the framework remains mildly conservative (−4% $\Delta CSR_{7.5,1}$; implied FS = 1.37). For higher-risk levels, the trend reverses: the dual-threshold rule requires a larger $CSR_{7.5,1}$ value to indicate liquefaction (e.g., +24% $\Delta CSR_{7.5,1}$ at P_{liq} = 80%), leading to implied FS values between 0.68 and 0.52, slightly below Cetin's estimated range of approximately 0.8–0.7.

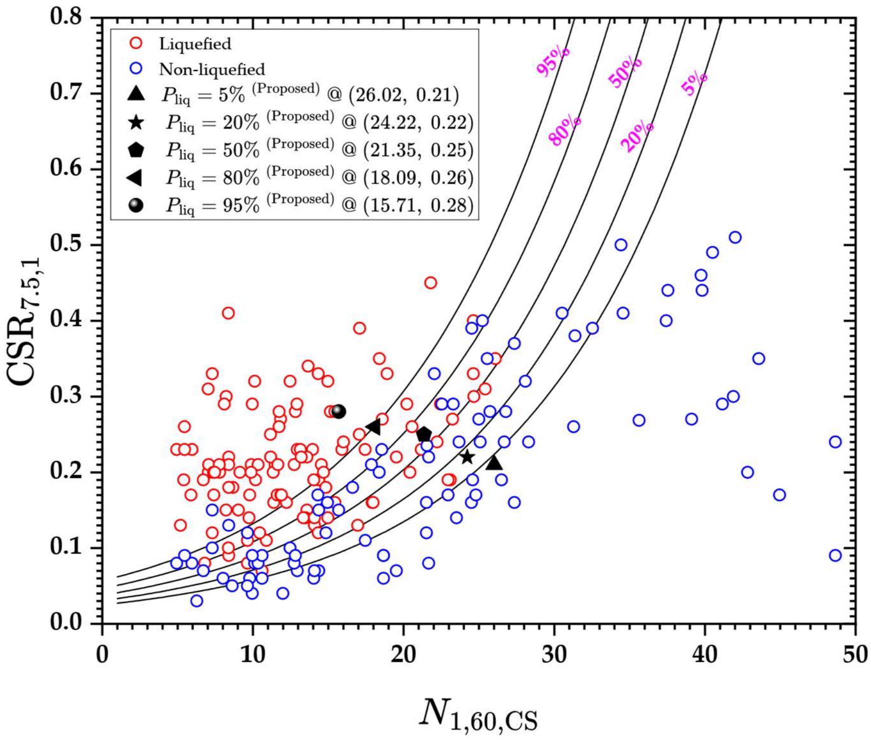

Figure 9. Redrawn probabilistic seismic soil liquefaction triggering curves for clean sands (FC = 5%, Mw = 7.5, σ$_v$′ = 1 atm), illustrating $\text{CSR}_{7.5,1}$ vs. $\text{N}_{1,60,\text{CS}}$ for liquefaction probabilities P_{liq} = 5%, 20%, 50%, 80%, and 95% (adapted from Cetin et al., 2018 [37], Soil Dynamics and Earthquake Engineering, 115, 698–709).

Table 7. Comparison of proposed thresholds with Cetin et al. (2018) [37].

P_{liq}	Proposed $\text{N}_{1,60,\text{CS}}$ Median [1]	Proposed $\text{CSR}_{7.5,1}$ Median [1]	Cetin et al. (2018) [37] $\text{CSR}_{7.5,1}$ at Proposed $\text{N}_{1,60,\text{CS}}$ [2]	Cetin et al. (2018) [37] FS	$\Delta\text{CSR}_{7.5,1}$ (%)	Implied FS (Proposed Model) [3,4]
0.05	26.02	0.21	≈0.21	1.5	0	1.67
0.20	24.22	0.22	≈0.23	1.2	−4	1.37
0.50	21.35	0.25	≈0.22	1.0	+14	0.94
0.80	18.09	0.26	≈0.21	0.8	+24	0.68
0.95	15.71	0.28	≈0.20	0.7	+40	0.52

[1] Proposed threshold (Table 4). [2] Visual estimates from Figure 4 in Cetin et al. (2018) [37]. [3] FS = CRR from Cetin at P_{liq} = 50% (FS = 1) ÷ proposed $\text{CSR}_{7.5,1}$ (using Cetin Equation (11); θ-coefficients from Table 7 Cetin et al. (2018) [37]; FC = 5%, M_w = 7.5, σ$_v$′ = 1 atm). [4] Implied FS (FS > 1 = more conservative, FS < 1 = less conservative). Detailed calculations are provided in Supplementary Document S1.

These differences are modest and arise from methodological choices. Cetin et al. employed a parametric regression with explicit correction factors. In contrast, the present study uses a non-parametric RF that captures nonlinear interactions and provides a bootstrap-based CI.

5.2. *Influence of Quality Weighting*

The quality-weighted sampling scheme, which assigns values of 1.0, 0.70, and 0.40 to Classes A, B, and C, respectively, ensures that higher-quality case histories have greater influence during model training and threshold setting, thereby improving the reliability of the proposed dual-threshold rule. To assess how these weights affect performance, a sensitivity analysis was performed by uniformly rescaling the weights for Classes B and C while maintaining the fixed thresholds from Table 4. This approach simulates typical engineering situations, such as using datasets with varying levels of completeness from older studies. Three weighting scenarios were tested: full uniform weighting (1.0 for all classes), moderate down-weighting (0.7 for B and C), and strict down-weighting (0.4 for B and C). The results were evaluated across five probability levels using accuracy and the F_1-score on the test set.

Figure 10 illustrates how the dual-threshold rule performs with different weight adjustments, showing accuracy (blue bars) and the F_1-score (red bars) for each probability level across various schemes. When the weights are equal (1.0), accuracy reaches a maximum of 0.71 at P_{liq} = 0.20 and decreases to 0.57 at P_{liq} = 0.95, while the F_1 scores exhibit a similar trend, falling from 0.65 to 0.39. This reflects a balanced but diminishing effectiveness as thresholds become less conservative. Slightly lower values occur with moderate down-weighting (0.7), where accuracy ranges from 0.67 at P_{liq} = 0.20 to 0.55 at P_{liq} = 0.95, and F_1 scores from 0.60 to 0.35. This indicates minimal performance loss and emphasizes the scheme's robustness. Strict down-weighting (0.4) results in further declines, with accuracy decreasing to 0.60 at P_{liq} = 0.20 and 0.50 at P_{liq} = 0.95. Meanwhile, F_1 drops from 0.50 to 0.29, as lower-quality cases contribute less and may introduce more noise in sparse areas.

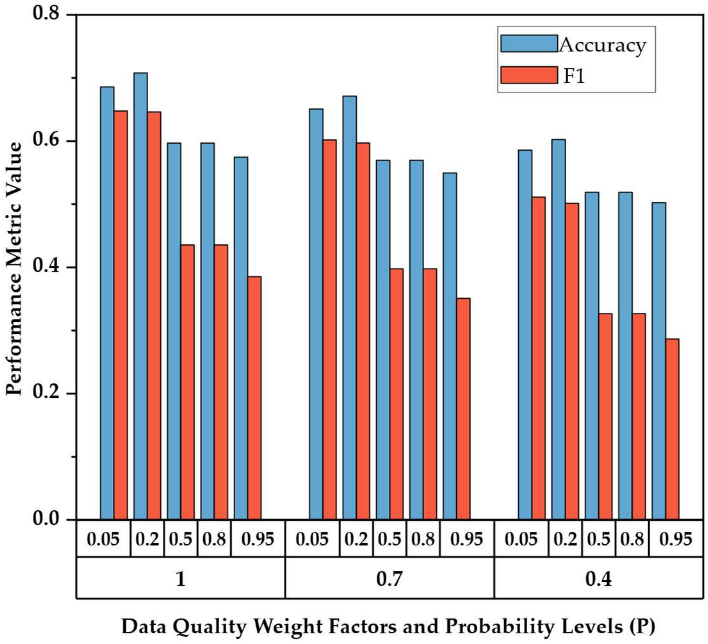

Figure 10. Sensitivity of dual-threshold rule performance to quality weight adjustments.

These variations demonstrate that the quality weighting enhances model stability without significantly compromising performance, as fluctuations remain below 10% across schemes. In practice, this flexibility allows engineers to adjust the rule based on dataset quality, applying stricter weighting for initial evaluations at data-scarce sites while preserving the usefulness of core thresholds in risk-informed seismic foundation design. These findings align with previous probabilistic analyses where lower-quality cases are penalized by increased uncertainty, resulting in slight estimate shifts but greater stability [13,14,36].

5.3. Interpretation of Model Behavior and Key Insights

The calibrated RF model exhibits interpretable behavior consistent with fundamental geotechnical principles, while also revealing subtle patterns through advanced visualization tools. Feature importance analysis using the Gini impurity metric (Figure 6) indicates that $N_{1,60,CS}$ is the most important predictor, with a value of 0.515, slightly exceeding $CSR_{7.5,1}$ at 0.485. Such a distribution highlights the crucial role of in situ soil resistance in mitigating liquefaction risk, as higher penetration resistance indicates denser or stiffer soils that are less prone to pore pressure buildup [65]. Conversely, $\mathrm{CSR}_{7.5,1}$ reflects the seismic demand that can trigger liquefaction under vulnerable conditions [13]. The similar contributions highlight the interconnected relationship between resistance and demand, aligning with empirical observations that liquefaction only occurs when low resistance coincides with sufficient cyclic loading [64,66].

The PDPs and ALEs further clarify these dynamics (Figure 7). For $\mathrm{N}_{1,60,\mathrm{CS}}$, both the PDP and ALE profiles show a nonlinear decrease in $\mathrm{P}_{\mathrm{liq}}$, dropping sharply from nearly 1.0 below 10 to almost 0 above 30, indicating a threshold-like transition around 15–25. It follows that small increases in resistance can result in substantial reductions in risk. The centered ALEs highlight local variations, showing strong adverse effects in this mid-range, which indicates that engineering interventions, such as ground improvement, could be most effective here. In contrast, $\mathrm{CSR}_{7.5,1}$ displays a sharp sigmoid-shaped increase, with the probability rising from minimal levels below 0.1 to nearly certain above 0.3, and ALE positive deviations peaking at 0.15–0.25, pointing to a sensitive zone where site-specific seismic analyses are essential.

Joint effects shown on the two-dimensional ALE surface (Figure 8) reveal high-risk areas (warm colors) at low $\mathrm{N}_{1,60,\mathrm{CS}}$ (<20), and high $\mathrm{CSR}_{7.5,1}$ (>0.15). Contour lines highlight abrupt changes, underscoring the logic of the dual-threshold rule: susceptibility requires both weak resistance and high demand. The data density plot (Figure 3) confirms model reliability in densely sampled regions ($\mathrm{N}_{1,60,\mathrm{CS}}$ = 10–30, $\mathrm{CSR}_{7.5,1}$ = 0.1–0.3), where the analysis is most reliable, but cautions against extrapolating into sparse areas, such as very high $\mathrm{N}_{1,60,\mathrm{CS}}$ or low $\mathrm{CSR}_{7.5,1}$, due to limited data. Those sparse corners align with the Extrapolation zone on the CEM (Figure 3), highlighting why predictions in this area require extra caution.

Key insights include the model's ability to surpass black box limitations through PDPs and ALEs, providing transparent nonlinear insights that foster trust in ML for geotechnics. The bootstrapped thresholds (Table 4) effectively account for uncertainty, with the wider CIs at the extremes reflecting data sparsity. The dual rule's performance (Tables 5 and 6) shows practical trade-offs, such as high precision at conservative levels ($\mathrm{P}_{\mathrm{liq}} \geq 0.20$), which is ideal for avoiding false alarms in foundation design. Overall, these elements position the framework as a link between data-driven prediction and engineering judgment, providing scalable risk assessment for building resilience in seismic regions.

5.4. Practical Implications for Engineering Practice

The proposed dual-threshold framework provides significant practical value for geotechnical and structural engineers involved in seismic building design. It provides a streamlined, risk-calibrated tool that easily integrates into performance-based workflows. By simplifying complex RF predictions into straightforward cutoff pairs for $\mathrm{N}_{1,60,\mathrm{CS}}$ and $\mathrm{CSR}_{7.5,1}$ across five probability levels (Table 4), the rule enables rapid preliminary screenings: a site is considered susceptible if resistance drops below the $\mathrm{N}_{1,60,\mathrm{CS}}$ threshold and demand exceeds the $\mathrm{CSR}_{7.5,1}$ threshold. Such streamlining enables rapid identification of liquefaction hazards without the need for extensive probabilistic calculations. The simplicity addresses a significant obstacle in adopting ML models, as highlighted in the literature, where black box outputs often discourage routine use [67]. Engineers can select thresholds based on project importance; for example, $\mathrm{P}_{\mathrm{liq}}$ = 5% (implied FS $\approx$ 1.67) for critical facilities, such as hospitals, ensuring conservative margins in line with codes, such as that of the American Society of Civil Engineers (ASCE 7), which require FS > 1.2–1.5. Alternatively, $\mathrm{P}_{\mathrm{liq}}$ = 20% (FS $\approx$ 1.37) can be used for ordinary structures to strike a balance between cost and safety. When the simplified assumptions for r_{d}, site amplification, or MSF are questionable (for example, in deep soft basins or where there are significant impedance contrasts), CSR should be obtained from site-specific 1D/2D response or effective-stress dynamic analyses and then normalized to $CSR_{7.5,1}$ before applying the screening rule. Recommended use cases include initial screening and design reviews for standard and critical building facilities where SPT data is the primary source and where site conditions fall within the Core/Support domain described in Section 3.7. Projects requiring higher accuracy or involving complex geology, such as essential facilities, layered or sloped sites, unusual grain sizes or plasticity, or strong static shear, should combine the rule with expanded feature sets or CPT-/Vs-based probabilistic models and, when appropriate, implement site-specific effective-stress analyses. For projects located in basins or depositional settings that are underrepresented in the training catalog, the screening results should be confirmed with an external dataset from a recent regional event (e.g., 2023 Turkey–Syria) before making final design decisions. Regional transfer can be assessed by freezing the learner and recomputing the isotonic probability mapping and bootstrap thresholds with locally sourced cases. If such calibration is not performed, outcomes should be considered screening-level, and reporting the geologic context provides traceability. For daily use, Figure 11 offers a step-by-step flowchart outlining the screening process: P_{liq} selection, input preparation, CEM coverage check, dual-threshold test, confidence-band check, out-of-domain actions, and documentation.

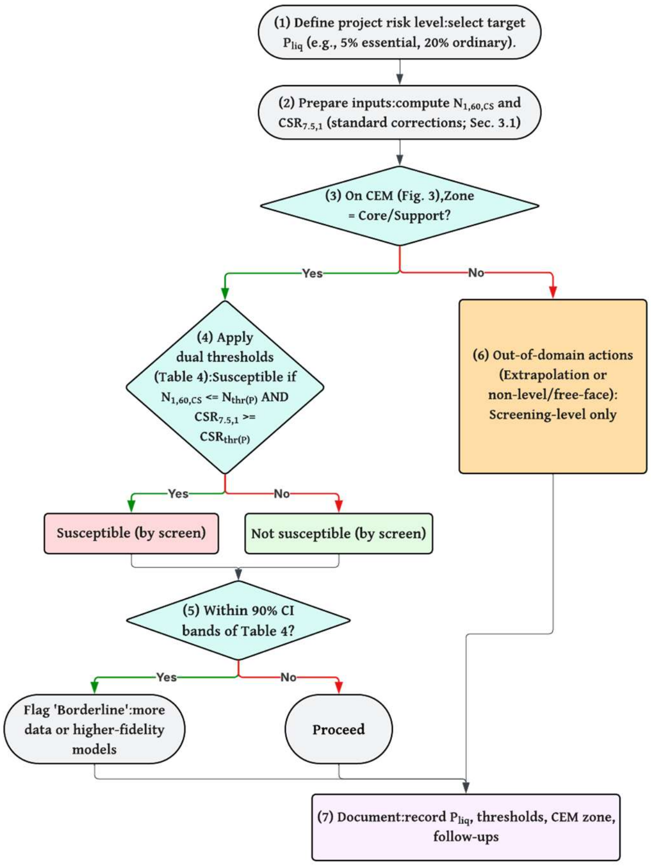
(1) Define project risk level:select target P_{liq} (e.g., 5% essential, 20% ordinary).
(2) Prepare inputs:compute $N_{1,60,CS}$ and $CSR_{7.5,1}$ (standard corrections; Sec. 3.1)
(3) On CEM (Fig. 3),Zone = Core/Support?
Yes
No
(4) Apply dual thresholds (Table 4):Susceptible if $N_{1,60,CS}$ <= $N_{thr(P)}$ AND $CSR_{7.5,1}$ >= $CSR_{thr(P)}$
(6) Out-of-domain actions (Extrapolation or non-level/free-face): Screening-level only
Yes
No
Susceptible (by screen)
Not susceptible (by screen)
(5) Within 90% CI bands of Table 4?
Yes
No
Flag 'Borderline':more data or higher-fidelity models
Proceed
(7) Document:record P_{liq}, thresholds, CEM zone, follow-ups

Figure 11. Dual-threshold decision flow chart for practice (SPT-CSR screening).

The dual thresholds are calibrated for level ground, free field sites; using them on sloping ground or near free faces falls outside the calibration scope and should be considered screening-level only. For such terrains, a recommended extension includes: (i) incorporating a static-shear descriptor (such as $\alpha = \tau_s/\sigma_v'$ or slope angle) and adjusting cyclic demand using a slope or static-bias factor in line with established procedures; (ii) reevaluating triggering with a multivariate or CPT-/V_s-assisted model; and (iii) assessing consequences through lateral spreading or displacement models and, if necessary, site-specific effective-stress dynamic analyses. This approach offers slope-aware checks while maintaining the dual-parameter rule as the initial screening method.

In foundation design, the framework facilitates informed decisions on mitigation strategies, such as ground improvement or deep foundations, by quantifying uncertainty with bootstrapped CIs. For example, wider intervals at low probabilities (e.g., $CSR_{7.5,1}$ CI = 0.05–0.48 at P_{liq} = 0.05) emphasize the need for site-specific investigations in data-sparse areas, encouraging cautious extrapolation. Performance metrics (Table 6) also show their usefulness, with perfect precision (1.0) at $P_{liq} \geq 20$, ensuring a low number of false positives, which is essential for avoiding unnecessary overdesign. Additionally, acceptable F_1 scores (up to 0.65) indicate a balanced recall in risk-averse situations. Comparisons with Cetin et al. (2018) [37] (Figure 9 and Table 7) suggest that the proposed thresholds yield FS margins comparable at low probabilities, thereby increasing conservatism where building resilience is vital.

Broader implications include better integration with BIM platforms, where the rule's parameters could be automated for scenario analyses and adapted to changing datasets. By emphasizing quality weighting and interpretability [29] (e.g., ALE surfaces in Figure 8), the approach builds trust among practitioners, thereby reducing reliance on subjective judgments and promoting fair seismic risk management in vulnerable regions, such as those affected by recent events in Turkey and Syria (2023). Ultimately, this framework enables engineers to tailor liquefaction assessments to specific importance levels, thereby optimizing safety and efficiency in the built environment. To demonstrate practical use, Table 8 features a quick, step-by-step screening example at P_{liq} = 20% using the median thresholds from Table 4. The near-threshold case illustrates how the dual rule can identify susceptibility when a legacy boundary would not, leading to simple confirmatory checks when decisions are sensitive to slight changes.

Table 8. Quick screening example at P_{liq} = 20% (ordinary importance; level ground, free field).

Step	What is Checked	Numbers	Outcome
1	Target risk level	P_{liq} = 20%	-
2	Corrected inputs	$N_{1,60,CS}$ = 24.0 blows/ft; $CSR_{7.5,1}$ = 0.22	-
3	Proposed thresholds (Table 4, medians)	N_{thr} = 24.22; $CSR_{7.5,1,thr}$ = 0.22	-
4	Resistance check	Is $N_{1,60,CS} \leq N_{thr}$? → 24.0 ≤ 24.22	Yes
5	Demand check	Is $CSR_{7.5,1} \geq CSR_{7.5,1,thr}$? → 0.22 ≥ 0.22	Yes
6	Dual-rule result	Both conditions satisfied	Susceptible
7	Legacy boundary at N ≈ 24 (Table 7)	$CSR_{ref} \approx 0.23$; compare 0.22 < 0.23	Not susceptible
8	Takeaway	Near-threshold divergence; prompts confirmatory steps (e.g., CPT/V_s profiling, site-specific response analysis)	-

Notes: Values are illustrative (synthetic), anchored to Table 4 (thresholds) and Table 7 (legacy comparison), and lie within reported 90% confidence bands. Applicability: level ground, free field, clean-to-silty sand sites.

5.5. Limitations and Future Work

While the proposed dual-threshold framework advances practical liquefaction assessment through a calibrated RF model, several limitations should be acknowledged. First, relying on only two predictors: $N_{1,60,CS}$ and $CSR_{7.5,1}$, simplifies the approach for usability but omits the explicit inclusion of additional factors such as varying fines content, vertical effective stress beyond normalization, or site-specific ground motion characteristics. This may reduce accuracy in complex soil profiles. The dataset, derived from Cetin et al. (2018) [37], carries inherent biases, including the overrepresentation of a few well-documented earthquakes and the exclusion of marginal cases, which could limit its applicability to underrepresented regions or recent events (e.g., the 2023 Turkey–Syria earthquakes). Validation on a relatively small test set (n = 42) provides robust metrics but could benefit from larger, independent datasets to verify performance across diverse seismic environments. Although the case histories are geographically broad and geologically diverse, some regions (such as parts of South America and Africa) and very coarse-grained or gravelly deposits are less represented in the catalog. Actionable use is limited to the Coverage Extrapolation Map Core/Support zones (Figure 3) for level ground, free field, and clean-to-silty sand sites. Cases within practical measurement tolerance of a threshold are for screening only and should be treated as provisional. Confidence-interval endpoints that lie outside Core/Support are non-actionable and are reported only to convey uncertainty. A clear validation pathway is recommended: (i) compile an independent set of SPT-based controls from recent events, especially the 2023 Turkey–Syria sequence, for which liquefaction inventories from remote sensing are available; (ii) apply the frozen dual-threshold rule to this set without refitting; (iii) report out-of-sample discrimination (AUC), calibration (Brier score/reliability), and threshold transferability; and (iv) update isotonic calibration and confidence intervals if a distribution shift is detected. This process limits extrapolation error and supports use in regions not fully covered by the current catalog. The lack of new field cases is a limitation; therefore, the released code and workflow are designed to be retrained to update thresholds, and region-specific SPT databases become accessible. Accuracy may decrease if the simplified demand formulation (choice of r_d, MSF, or amplification) does not match site conditions; in such cases, CSR should be computed from site-specific response or dynamic analyses as the input to the normalization step before applying the thresholds.

Beyond scope choices, several practical failure modes can affect screening reliability. Isotonic calibration is non-parametric and tied to the training distribution; under dataset shift, probability reliability may decline even if discrimination remains similar. Near the decision boundary, minor uncertainties in $N_{1,60,CS}$ corrections (e.g., C_E, C_N, fines adjustments) or in mapping site CSR to the reference $CSR_{7.5,1}$ (choice of r_d, duration/MSF, amplification) can alter outcomes, so borderline results should be viewed as provisional. The two-feature model omits covariates such as stratigraphy (thin/interbedded layers), static shear bias, groundwater fluctuations, and effective-stress details; these factors can influence risk in ways not captured by the rule. Thresholds may shift slightly if case-quality labels or weights differ from those assumed. Finally, selecting a P_{liq} level that does not align with facility importance can lead to inconsistent safety margins unless documented in the design record.

Future work could explore these aspects by expanding the model to include multivariate inputs, such as CPT or shear-wave velocity (V_s) data, and utilizing ensemble methods or hybrid physics ML approaches for broader application. A pragmatic two-stage approach is recommended: Stage 1 uses the fixed dual-parameter screening rule; Stage 2 enhances this with additional predictors (such as fines content, σ_v', M_w, q_c, or V_s) using a stacked or ensemble learner with isotonic recalibration. Afterward, thresholds and confidence intervals are re-estimated through bootstrap. Cross-validation should be stratified by event or region to ensure transferability, and CPT- and V_s-based probabilistic methods can act as both parallel validators and supplementary tool inputs. Incorporating recent global case studies would improve the database's diversity, possibly refining thresholds through transfer learning. Extending the framework to handle non-level terrains or coupling it with finite-element simulations could enable comprehensive building analyses, including dynamic response and mitigation strategies. Investigating adaptive weighting through Bayesian updates could further strengthen robustness against data uncertainties. Following these steps and after assessing software packaging and user workflows, the framework will be well-prepared for future integration into digital design environments (e.g., BIM).

6. Conclusions

The research presents a risk-informed dual-threshold framework for SPT-based soil liquefaction assessment, using a calibrated RF model to offer practical, probability-based criteria tailored for seismic building design. By analyzing 208 field case histories from Cetin et al. (2018) [37] with quality-based weighting, the model establishes paired thresholds for $N_{1,60,CS}$ and $CSR_{7.5,1}$ at P_{liq} of 5%, 20%, 50%, 80%, and 95%, including 90% CI derived from bootstrapping (Table 4). The resulting screening rule, which classifies sites as susceptible if resistance drops below the $N_{1,60,CS}$ threshold and demand exceeds the $CSR_{7.5,1}$ threshold, shows strong performance, with precision reaching 1.0 at conservative levels ($P \geq 0.20$) and F_1 scores up to 0.65 (Table 6), indicating reliability for preliminary assessments. Interpretability tools like PDPs and ALEs (Figures 7 and 8) reveal nonlinear behaviors, such as sharp risk transitions around $N_{1,60,CS}$ = 15–25 and $CSR_{7.5,1}$ = 0.15–0.25, providing insights beyond traditional deterministic boundaries.

Comparisons with Cetin et al. (2018) [37] (Figure 9 and Table 7) confirm alignment at moderate probabilities, while highlighting the proposed model's flexibility, with implied FS margins exceeding benchmarks at low P_{liq} (e.g., 1.67 vs. approximately 1.5 at 5%), providing greater conservatism in critical applications. Sensitivity to weighting (Figure 10) underscores robustness, with minimal performance loss under different schemes. These improvements address key gaps in ML adoption and offer an accessible tool for risk-adjusted foundation design that enhances safety and efficiency. Future extensions could include multivariable inputs, further advancing resilient infrastructure in seismic regions.

Applicability is limited to level ground, free field, and clean-to-silty sand sites within the Core/Support region of the Coverage Extrapolation Map. Cases in Extrapolation are screening-level only and require site-specific analyses. The worked screening example and decision flow (Section 5.4, Figure 11; Table 8) demonstrate how risk-graded SPT-CSR thresholds can alter initial decisions compared to legacy criteria and trigger confirmatory checks near thresholds. For regional deployment, the trained model may be frozen while recalculating the isotonic probability map and bootstrap thresholds with local cases. Slope/static-shear and SSI/lateral spreading extensions remain future priorities.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/buildings15173206/s1: The Word file containing the implied-FS calculations is available as Supplementary Document S1 alongside this article.

Funding: This research received no external funding.

Data Availability Statement: The datasets and code supporting the findings of this study are openly available at the following GitHub repository: https://github.com/hssa83/dual-threshold-liquefaction-rf (accessed on 25 August 2025). An archived version of the repository is also available on Zenodo, accessible via the following DOI: https://doi.org/10.5281/zenodo.16329145. The repository contains all cleaned data, Jupyter notebooks, and Excel output files necessary to reproduce the modeling results, bootstrap analysis, and threshold derivation described in the manuscript.

Acknowledgments: The author would like to thank the Deanship of Scientific Research at Shaqra University for supporting this work. Special appreciation is extended to Cetin et al. (2018) [37] for providing the dataset utilized in this study.

Conflicts of Interest: The author declares there are no conflicts of interest.

References

1. Polito, C.P.; Martin, J.R. Plasticity, Flow Liquefaction, and Cyclic Mobility in Liquefiable Soils with Low to Moderate Plasticity. *CivilEng* **2025**, *6*, 31. [CrossRef]
2. Wang, W.; Liu-Zeng, J.; Shao, Y.; Wang, Z.; Han, L.; Shen, X.; Qin, K.; Gao, Y.; Yao, W.; Hu, G.; et al. Mapping of soil liquefaction associated with the 2021 Mw 7.4 Maduo (Madoi) Earthquake based on the UAV photogrammetry technology. *Remote Sens.* **2023**, *15*, 1032. [CrossRef]
3. Mele, L.; Lirer, S.; Flora, A. Induced partial saturation: From mechanical principles to engineering design of an innovative and eco-friendly countermeasure against earthquake-induced soil liquefaction. *Geosciences* **2024**, *14*, 140. [CrossRef]
4. Huang, F.K.; Wang, G.S. A Method for Developing Seismic Hazard-Consistent Fragility Curves for Soil Liquefaction Using Monte Carlo Simulation. *Appl. Appl. Sci.* **2024**, *14*, 9482. [CrossRef]
5. Taftsoglou, M.; Valkaniotis, S.; Papathanassiou, G.; Karantanellis, E. Satellite imagery for rapid detection of liquefaction surface manifestations: The case study of Türkiye–Syria 2023 earthquakes. *Remote Sens.* **2023**, *15*, 4190. [CrossRef]
6. Perez, J.S.; Llamas, D.C.E.; Buhay, D.J.L.; Constantino, R.C.C.; Legaspi, C.J.M.; Lagunsad, K.D.B.; Grutas, R.N.; Quimson, M.M.Y. Impacts of a Moderate-Sized Earthquake: The 2023 Magnitude (Mw) 4.7 Leyte, Leyte Earthquake, Philippines. *Geosciences* **2024**, *14*, 61. [CrossRef]
7. Ko, K.W.; Kayen, R.E.; Kokusho, T.; Ilgac, M.; Nozu, A.; Nweke, C.C. Energy-Based Liquefaction Evaluation: The Port of Kushiro in Hokkaido, Japan, 2003 Tokachi-Oki Earthquake. *J. Geotech. Geoenviron. Eng.* **2024**, *150*, 05024010. [CrossRef]
8. Yang, Y.; Wei, Y. A New Shear Wave Velocity-Based Liquefaction Probability Model Using Logistic Regression: Emphasizing Fines Content Optimization. *Appl. Sci.* **2024**, *14*, 6793. [CrossRef]
9. Alonso-Pandavenes, O.; Torrijo, F.J.; Torres, G. Analysis of the Liquefaction Potential at the Base of the San Marcos Dam (Cayambe, Ecuador)—A Validation in the Use of the Horizontal-to-Vertical Spectral Ratio. *Geosciences* **2024**, *14*, 306. [CrossRef]
10. Seed, H.B.; Idriss, I.M. Simplified procedure for evaluating soil liquefaction potential. *J. Soil Mech. Found. Div.* **1971**, *97*, 1249–1273. [CrossRef]
11. Seed, H.B. Soil liquefaction and cyclic mobility evaluation for level ground during earthquakes. *J. Geotech. Eng. Div.* **1979**, *105*, 201–255. [CrossRef]
12. Idriss, I.M.; Boulanger, R.W. *Soil Liquefaction During Earthquakes*; Earthquake Engineering Research Institute: Oakland, CA, USA, 2008.
13. Cetin, K.O.; Seed, R.B.; Der Kiureghian, A.; Tokimatsu, K.; Harder, L.F., Jr.; Kayen, R.E.; Moss, R.E.S. Standard penetration test-based probabilistic and deterministic assessment of seismic soil liquefaction potential. *J. Geotech. Geoenviron. Eng.* **2004**, *130*, 1314–1340. [CrossRef]
14. Moss, R.E.; Seed, R.B.; Kayen, R.E.; Stewart, J.P.; Der Kiureghian, A.; Cetin, K.O. CPT-based probabilistic and deterministic assessment of in situ seismic soil liquefaction potential. *J. Geotech. Geoenviron. Eng.* **2006**, *132*, 1032–1051. [CrossRef]
15. Rahman, M.M.; Hossain, M.B.; Sayed, A. Prediction of Soil Liquefaction Using Machine Learning Approaches. *Eng. Trans.* **2025**, *73*, 257–277.
16. Liu, C.Y.; Ku, C.Y.; Chiu, Y.J.; Wu, T.Y. Evaluation of liquefaction potential in central Taiwan using random forest method. *Sci Rep.* **2024**, *14*, 27517. [CrossRef] [PubMed]
17. Ghani, S.; Thapa, I.; Kumari, S.; Correia, A.G.; Asteris, P.G. Revealing the nature of soil liquefaction using machine learning. *Earth Sci. Inform.* **2025**, *18*, 198. [CrossRef]
18. Maurer, B.W.; Sanger, M.D. Why "AI" models for predicting soil liquefaction have been ignored, plus some that shouldn't be. *Earthq. Spectra* **2023**, *39*, 1883–1910. [CrossRef]
19. Bolton Seed, H.; Tokimatsu, K.; Harder, L.F.; Chung, R.M. Influence of SPT procedures in soil liquefaction resistance evaluations. *J. Geotech. Eng.* **1985**, *111*, 1425–1445. [CrossRef]
20. Katona, T.J.; Karsa, Z. Probabilistic safety analysis of the liquefaction hazard for a nuclear power plant. *Geosciences* **2022**, *12*, 192. [CrossRef]
21. Liao, S.S.C.; Veneziano, D.; Whitman, R.V. Regression models for evaluating liquefaction probability. *J. Geotech. Eng.* **1988**, *114*, 389–411. [CrossRef]
22. Youd, T.L.; Noble, S.K. Liquefaction criteria based on statistical and probabilistic analyses. In Proceedings of the NCEER Workshop on Evaluation of Liquefaction Resistance of Soils, Salt Lake City, UT, USA, 5–6 January 1996; State University of New York: Buffalo, NY, USA, 1997; pp. 201–205.
23. Cox, B.R.; Griffiths, S.C. *Practical Recommendations for Evaluation and Mitigation of Soil Liquefaction in Arkansas*; University of Arkansas: Fayetteville, AR, USA, 2010.
24. Kayen, R.; Moss, R.E.S.; Thompson, E.M.; Seed, R.B.; Cetin, K.O.; Kiureghian, A.; Tanaka, Y.; Tokimatsu, K. Shear-wave velocity–based probabilistic and deterministic assessment of seismic soil liquefaction potential. *J. Geotech. Geoenviron. Eng.* **2013**, *139*, 407–419. [CrossRef]

25. Boulanger, R.W.; Idriss, I.M. *CPT and SPT Based Liquefaction Triggering Procedures*; Rep. No. UCD/CGM-14/01; University of California: Davis, CA, USA, 2014.
26. Choi, Y.; Kumar, K. A machine learning approach to predicting pore pressure response in liquefiable sands under cyclic loading. In Proceedings of the Geo-Congress, Los Angeles, CA, USA, 26–29 March 2023; pp. 202–210.
27. Şehmusoğlu, E.H.; Kurnaz, T.F.; Erden, C. Estimation of soil liquefaction using artificial intelligence techniques: An extended comparison between machine and deep learning approaches. *Environ. Earth Sci.* **2025**, *84*, 1–22. [CrossRef]
28. Harle, S.M.; Wankhade, R.L. Machine learning techniques for predictive modelling in geotechnical engineering: A succinct review. *Discov. Civ. Eng.* **2025**, *2*, 1–21. [CrossRef]
29. Molnar, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2020. Available online: https://christophm.github.io/interpretable-ml-book/ (accessed on 25 August 2025).
30. Hsiao, C.H.; Kumar, K.; Rathje, E.M. Explainable AI models for predicting liquefaction-induced lateral spreading. *Front. Built Environ.* **2024**, *10*, 1387953. [CrossRef]
31. Ranjan Kumar, D.; Wipulanusat, W. Advancements in predicting soil liquefaction susceptibility: A comprehensive analysis of ensemble and deep learning approaches. *Sci. Rep.* **2025**, *15*, 26453. [CrossRef]
32. Onyelowe, K.C.; Kamchoom, V.; Gnananandarao, T.; Arunachalam, K.P. Developing data driven framework to model earthquake induced liquefaction potential of granular terrain by machine learning classification models. *Sci. Rep.* **2025**, *15*, 21509. [CrossRef]
33. Alharbi, H.S. Efficient Swell Risk Prediction for Building Design Using a Domain-Guided Machine Learning Model. *Buildings* **2025**, *15*, 2530. [CrossRef]
34. Sioutas, K.N.; Benardos, A. Boosting Model Interpretability for Transparent ML in TBM Tunneling. *Appl. Sci.* **2024**, *14*, 11394. [CrossRef]
35. Jahangiri, V.; Akbarzadeh, M.R.; Shahamat, S.A.; Asgari, A.; Naeim, B.; Ranjbar, F. Machine learning-based prediction of seismic response of steel diagrid systems. In *Structures*; Elsevier: Amsterdam, The Netherlands, 2025; p. 109791.
36. Ilgac, M.; Cetin, K.O.; Kayen, R.E. Updated SPT-based seismic soil liquefaction triggering global database. In Proceedings of the Geo-Congress, Charlotte, NC, USA, 20–23 March 2022; pp. 308–317.
37. Cetin, K.O.; Seed, R.B.; Kayen, R.E.; Moss, R.E.; Bilge, H.T.; Ilgac, M.; Chowdhury, K. Dataset on SPT-based seismic soil liquefaction. *Data Brief* **2018**, *20*, 544–548. [CrossRef]
38. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Routledge: Abingdon-on-Thames, UK, 2018.
39. Muraina, I. Ideal dataset splitting ratios in machine learning algorithms: General concerns for data scientists and data analysts. In Proceedings of the 7th International Mardin Artuklu Scientific Research Conference, Mardin, Turkey, 10–12 December 2021; pp. 496–504.
40. Probst, P.; Wright, M.N.; Boulesteix, A. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1301. [CrossRef]
41. Han, S.; Williamson, B.D.; Fong, Y. Improving random forest predictions in small datasets from two-phase sampling designs. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 322. [CrossRef] [PubMed]
42. Weinberge, K. *Bagging and Random Forest (Lecture Notes, Spring 2022)*; Cornell University: Ithaca, NY, USA, 2022.
43. Huang, B.F.F.; Boutros, P.C. The parameter sensitivity of random forests. *BMC Bioinform.* **2016**, *17*, 331. [CrossRef] [PubMed]
44. Boldini, D.; Grisoni, F.; Kuhn, D.; Friedrich, L.; Sieber, S.A. Practical guidelines for the use of gradient boosting for molecular property prediction. *J. Cheminform.* **2023**, *15*, 73. [CrossRef]
45. Niculescu-Mizil, A.; Caruana, R. Obtaining Calibrated Probabilities from Boosting. In Proceedings of the UAI'05: Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence, Arlington, VA, USA, 26–29 July 2005; pp. 413–420.
46. Tarwidi, D.; Pudjaprasetya, S.R.; Adytia, D.; Apri, M. An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. *MethodsX* **2023**, *10*, 102119. [CrossRef]
47. Tsai, C.A.; Chang, Y.J. Efficient selection of Gaussian Kernel SVM parameters for imbalanced data. *Genes* **2023**, *14*, 583. [CrossRef]
48. Huang, Y.; Li, W.; Macheret, F.; Gabriel, R.A.; Ohno-Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 621–633. [CrossRef] [PubMed]
49. Levy, J.J.; O'Malley, A.J. Don't dismiss logistic regression: The case for sensible extraction of interactions in the era of machine learning. *BMC Med. Res. Methodol.* **2020**, *20*, 171. [CrossRef]
50. Zadrozny, B.; Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002; pp. 694–699.
51. Niculescu-Mizil, A.; Caruana, R. Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning, New York, NY, USA, 7–11 August 2005. pp. 625–632.
52. Keller, J. Explainability in Deep Reinforcement Learning. 2024. Available online: https://tud.qucosa.de/en/landing-page/https%3A%2F%2Ftud.qucosa.de%2Fapi%2Fqucosa%253A93844%2Fmets/ (accessed on 25 August 2025).

53. Khoda Bakhshi, A.; Ahmed, M.M. Utilizing black-box visualization tools to interpret non-parametric real-time risk assessment models. *Transp. A Transp. Sci.* **2021**, *17*, 739–765. [CrossRef]
54. Cook, T.R.; Gupton, G.; Modig, Z.; Palmer, N.M. *Explaining Machine Learning by Bootstrapping Partial Dependence Functions and Shapley Values*; Federal Research Bank of Kansas City: Kansas City, MO, USA, 2021.
55. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann Stat.* **2001**, *29*, 1189–1232. [CrossRef]
56. Greenwell, B.M. pdp: An R package for constructing partial dependence plots. *R J.* **2017**, *9*, 421–436. [CrossRef]
57. Apley, D.W.; Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2020**, *82*, 1059–1086. [CrossRef]
58. Davison, A.C.; Hinkley, D.V. *Bootstrap Methods and Their Application*; Cambridge University Press: Cambridge, UK, 1997.
59. Efron, B. Bootstrap methods: Another look at the jackknife. In *Breakthroughs in Statistics: Methodology and Distribution*; Springer: Berlin/Heidelberg, Germany, 1992; pp. 569–593.
60. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef]
61. Silva Filho, T.; Song, H.; Perello-Nieto, M.; Santos-Rodriguez, R.; Kull, M.; Flach, P. Classifier calibration: A survey on how to assess and improve predicted class probabilities. *Mach Learn.* **2023**, *112*, 3211–3260. [CrossRef]
62. Yuan, Y.; Wu, L.; Zhang, X. Gini-impurity index analysis. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 3154–3169. [CrossRef]
63. Duan, W.; Congress, S.S.C.; Cai, G.; Zhao, Z.; Pu, S.; Liu, S.; Dong, X.; Wu, M.; Chen, R. Characterizing the in-situ state of sandy soils for liquefaction analysis using resistivity piezocone penetration test. *Soil Dyn. Earthq. Eng.* **2023**, *164*, 107529. [CrossRef]
64. Cetin, K.O.; Seed, R.B.; Kayen, R.E.; Moss, R.E.S.; Bilge, H.T.; Ilgac, M.; Chowdhury, K. SPT-based probabilistic and deterministic assessment of seismic soil liquefaction triggering hazard. *Soil Dyn. Earthq. Eng.* **2018**, *115*, 698–709. [CrossRef]
65. Muftuoglu, G.M.; Dehghanian, K. Soil Liquefaction Assessment Using Machine Learning. *Artif. Intell. Geosci.* **2025**, *6*, 100122. [CrossRef]
66. Kumar, D.R.; Samui, P.; Burman, A.; Wipulanusat, W.; Keawsawasvong, S. Liquefaction susceptibility using machine learning based on SPT data. *Intell. Syst. Appl.* **2023**, *20*, 200281. [CrossRef]
67. Liu, H.; Su, H.; Sun, L.; Dias-da-Costa, D. State-of-the-art review on the use of AI-enhanced computational mechanics in geotechnical engineering. *Artif. Intell. Rev.* **2024**, *57*, 196. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.