Dear authors,

This manuscript proposes a machine learning framework for predicting urban slope instability in Salvador, Bahia, Brazil. The authors integrate a digital terrain model, a database of 350 laboratory test results, probabilistic slope-stability indicators, cumulative rainfall measurements from 40 weather stations, geological structures, vegetation cover, residential density, sewage-service quality, and information about stabilization measures. The target variable is constructed from 13,522 confirmed and unconfirmed emergency calls submitted to the municipality between 2020 and 2025. An XGBoost regression model is trained and evaluated under alternative combinations of radius of influence, time interval, and weighting factor for confirmed calls. The authors report high validation and testing performance, with R^2 values generally exceeding 0.94 and reaching approximately 0.98.

The integration of geotechnical, environmental, and municipal datasets is potentially valuable, and the practical motivation is relevant to urban risk management. However, the current methodological design does not yet demonstrate that the model can predict future slope-instability events. The reported performance may be affected by sample-selection bias, spatial and temporal dependence among derived observations, and information leakage between training and testing subsets. The distinction between predicting physical instability and reproducing patterns in emergency-call activity also requires clarification. These issues must be addressed before the operational claims can be supported.

General questions and essential issues

1. What precisely is the prediction target?
The stated objective is to predict slope instability, but the output variable is a weighted aggregation of confirmed and unconfirmed emergency calls. These calls may reflect physical instability, population density, reporting behaviour, access to municipal services, and public awareness. Please clarify whether the model is intended to forecast: (i) verified mass-movement occurrence, (ii) emergency-call intensity, or (iii) a broader municipal-response indicator. The terminology, evaluation metrics, and practical claims should be aligned with the actual target.

2. How can the model be applied citywide if slopes without emergency calls are excluded?
The manuscript defines an eligible slope as a slope with at least one emergency call within its area of influence. This preselection removes true negative cases. A model trained only on call-positive locations cannot demonstrate its ability to distinguish unstable slopes from stable slopes across the city. Please reconstruct the modelling dataset to include slopes with no calls and time periods without recorded events, or limit the stated objective to ranking call intensity within previously identified risk areas.

3. How have spatial and temporal information leakage been prevented?
Approximately two million derived observations are generated from overlapping slope areas, emergency calls, radii of influence, and time windows. Randomly distributing these observations among training, validation, and testing subsets is likely to place highly related records from the same slope, neighbourhood, rainfall episode, or emergency-call cluster into different subsets. This may substantially inflate the reported R^2 values. Please adopt grouped spatial and temporal validation, such as spatial-block cross-validation and an out-of-time holdout period. Hyperparameter and scenario selection must use only the training and validation subsets; the testing subset should be evaluated once after all decisions are finalized.

4. Is the proposed framework genuinely predictive rather than retrospective?
The rainfall accumulations are calculated relative to the emergency-call date, while the output variable aggregates calls over a time interval associated with the same event. Please specify the forecasting horizon and the exact timestamp at which a prediction would be issued. Every predictor used in an operational forecast must be available before that timestamp. A clear timeline diagram would be useful.

5. Are the evaluation metrics suitable for an early-warning application?
R^2, MAE, MSE, and RMSE are useful for regression assessment but insufficient for evaluating an early-warning system. Please report performance for confirmed mass movements separately and include event-detection metrics appropriate for decision-making, such as sensitivity, specificity, precision, recall, false-alarm rate, calibration, and lead time. Performance should also be compared against simple baseline models, such as rainfall thresholds and logistic regression.

6. How should the geotechnical stability indicators be interpreted?

The calculated Factor of Safety and probability-of-failure values are important model inputs, but they are derived from uniform-slope charts with ru = 0 and formation-level parameter distributions. Please explain whether these quantities are intended as approximate susceptibility proxies or site-specific stability estimates. The distinction is important because the conclusions currently imply a level of physical resolution that may not be supported by the assumptions.

7. How will reproducibility be ensured?
The manuscript states that code and data may be available upon request and subject to institutional agreements. Sensitive emergency-call data understandably require protection. Nevertheless, reproducibility could be improved by releasing anonymized or aggregated records, the feature-construction workflow, the fitted chart coefficients, model hyperparameters, random seeds, and a synthetic demonstration dataset.

Specific comments

Lines 15-42: The literature review should more clearly distinguish among landslide susceptibility mapping, retrospective event classification, nowcasting, and operational forecasting. The novelty of the present study should be framed relative to existing municipal-scale landslide early-warning and susceptibility studies rather than primarily relative to Factor-of-Safety prediction studies.

Lines 27–31: Probabilistic approaches have also been applied to slope-instability assessment under uncertain loading conditions, including Bayesian-network and Newmark-displacement frameworks for earthquake-triggered landslides (Khalaj et al., 2020), as well as reliability-based methods for prioritizing maintenance actions for reinforced slopes under uncertain drainage-related failure conditions (BahooToroody et al., 2021).

Lines 75-97: Please provide the number of drained triaxial and direct-shear tests separately for each soil formation and moisture condition. Explain how results from the two test types were harmonized, what stress ranges were used, and how each slope was assigned a soil formation. The potential sampling bias arising from consultant-requested tests should also be discussed.

Lines 85-86 and Table 1: The sand-and-gravel and clay-and-silt fractions do not consistently sum to 100%. Please explain whether these values were calculated from different subsets of

samples or whether a transcription error is present. Permeability values are presented only as orders of magnitude; indicate the number of measurements and the observed variability.

Lines 100-106: Please justify the 18% slope threshold and the 5 m streamline spacing. Describe how the automated crest, toe, slope height, and slope-angle extraction procedure was validated against field observations or manually interpreted cross-sections.

Lines 107-112: The use of the Michalowski charts should be explained in greater detail. In particular, provide the fitted coefficients used in Equation 1, the interpolation procedure between slope angles, and the treatment of slopes falling outside the chart ranges.

Lines 113-127: Equation 2 and the accompanying notation require revision. The symbol μ is normally used for a mean but is used here for standard deviation. The term COV is described as a coefficient of variation between variables, although the equation appears to require covariance. Please define the covariance calculation explicitly and provide the equations used to calculate probability of failure and the reliability index.

Lines 128-136: The comparison with Slide2 is useful but insufficiently documented. Please state the total number of validation cases, the parameter combinations, the soil-profile assumptions, the groundwater assumptions, and whether identical slip-surface search settings were used. A comparison of error distributions would be more informative than mean error alone.

Lines 137-139: The manuscript states that most emergency calls are concentrated in areas with low saturated Factor of Safety. Please quantify this relationship and report the proportions of confirmed and unconfirmed calls across Factor-of-Safety classes.

Lines 145-152: The representation of geological structures requires additional explanation. Please justify the use of the sine of the angle between the slope section and the nearest structure. Clarify whether structure dip, distance, persistence, and multiple nearby structures were considered or omitted.

Lines 153-158: The statement that residences near the slope toe tend to have a stabilizing effect requires qualification and supporting evidence. Buildings may introduce toe loading, but construction may also involve excavation, drainage modification, and additional water infiltration. Please avoid assigning a universally stabilizing effect without analysis.

Lines 159-162: Please describe rainfall-data completeness, station coverage over time, treatment of missing records, and validation of the inverse-distance-squared interpolation. It would be useful to quantify interpolation uncertainty, especially in areas distant from weather stations.

Lines 163-177: The sewage-service score is potentially informative but the assigned weights require justification. Please explain the basis for assigning values of 0, 0.1, 0.2, 0.3, 0.6, and 1.0 and provide a sensitivity analysis. The variable may also act as a proxy for socioeconomic conditions, reporting behaviour, and infrastructure quality; this should be acknowledged.

Lines 178-181: Please describe how stabilization measures were mapped and dated. Because interventions are likely to target slopes already considered hazardous, this variable may be affected by confounding by indication. Its interpretation should therefore be cautious.

Lines 202-204: Replace “based on threes” with “based on trees.”

Lines 205-209: The discussion of artificial neural networks is not relevant to the adopted XGBoost method and should be removed or rewritten to address gradient-boosting hyperparameter optimization specifically.

Lines 210-213: Please report the RandomizedSearchCV parameter ranges, number of sampled combinations, cross-validation procedure, scoring function, and random seed.

Lines 218-228: A radius of influence of 300 m produces the highest R^2, whereas 150 m is selected as the preferred value based on physical consistency. Please define an objective selection criterion and distinguish between predictive performance and spatial interpretability.

Lines 230-253: Feature-importance interpretation should be strengthened. Several predictors are strongly related or mathematically dependent, including rainfall accumulations over nested durations and the Factor-of-Safety, probability-of-failure, and reliability-index variables. Split weight and gain may therefore be misleading when interpreted individually. Please consider permutation importance or SHAP analysis and avoid causal interpretations.

Lines 255-280: The conclusions substantially overstate the demonstrated capability of the current model. Claims regarding accurate real-time prediction, evacuation warnings, critical thresholds, scalability, and proactive mitigation should be tempered until the model has been evaluated using a leakage-resistant spatial and temporal holdout design and tested prospectively.

Lines 283-288: The code-availability, data-availability, and combined code-and-data-availability statements are repetitive. Please consolidate them into one clear statement and describe which anonymized materials can be released.

Lines 342-343: The DOI provided for Morgenstern and Price (1965) is incorrect. Please replace it with: 10.1680/geot.1965.15.1.79.

Lines 351-352: The Pereira and Barbosa (2023) reference is incorrect and contains a malformed DOI. The reference should identify the censobr R package: "censobr: Download Data from Brazil's Population Census," with DOI: 10.32614/CRAN.package.censobr.

Lines 368-369: This reference conflates ter Braak and Šmilauer with the R Core Team citation. If the authors intend to cite the R environment, the reference should be attributed to the R Core Team. If the Canoco manual is not used in the study, the ter Braak and Šmilauer reference should be removed.

Technical corrections

- **Line 13:** Capitalize "Keywords."
- **Line 21:** Remove the unmatched closing parenthesis after unit weight γ.
- **Line 61:** Add a comma after "occur."
- **Line 75:** Use consistent ASTM standard designations, including the applicable version suffixes.
- **Line 87:** Replace "Atteberg" with "Atterberg."
- **Lines 96-97 and Table 2:** Replace "unity weight" with "unit weight" and correct the units to kN/m^3.
- **Line 102:** Correct the DTM dimensions. The current formatting "5,4299 × 5,3010 pixels" appears erroneous.

- **Line 180:** Replace “emergence call date” with “emergency-call date.”
- **Line 202:** Replace “threes” with “trees.”
- **Figure 6 caption:** Replace panel label “w)” with “e).”
- **Table 3:** Remove the duplicated “20” in the row for Ri = 150 m.
- **Line 243:** Replace “second platoon” with “second group” or “second tier.”
- **Line 247:** Replace “2th place” with “2nd place.”
- **Line 272:** Replace “phisically” with “physically.”
- **Line 290:** The initials “SLM” are repeated in the formal-analysis and validation contribution statement.
- **Competing-interests statement:** Replace “The authors declare have no competing interests” with “The authors declare that they have no competing interests.”
- **Line 364:** Correct the spelling of “Salvador” in the Souza-Oliveira et al. reference.
- Review the manuscript throughout for spacing between numerical values and units, including “150 m,” “4 d,” and “75 m.”

Respectfully,