QuakeAlertPK: A Multi-Model Seismic Risk Assessment and Probability Forecasting System for Pakistan

Muhammad Hassan[1], Azka Ahmad[2], Saad Ahmed[3]
[1]Student ID: 453837, mhassan.bscs22seecs@seecs.edu.pk
[2]Student ID: 464970, aahmed.bscs22seecs@seecs.edu.pk
[3]Student ID: 470964, sahmed.bscs22seecs@seecs.edu.pk
Department of Computing, National University of Sciences and Technology
Rawalpindi, Pakistan

***Abstract*—Pakistan's location along active tectonic boundaries necessitates robust seismic risk assessment systems. This paper presents QuakeAlertPK, a comprehensive machine learning framework that integrates four heterogeneous geospatial data sources to provide multi-faceted earthquake risk assessment. Our system introduces three key innovations: (1) a composite risk scoring framework combining population, seismic, and geological factors with intelligent weighting derived from feature importance analysis, (2) a multi-model prediction architecture providing both continuous risk scores (R^2=0.8950) and categorical classifications (accuracy=63.08%), and (3) temporal probability forecasting across four time horizons (1-month to 5-year). To handle computational constraints, we develop a memory-efficient chunked raster processing algorithm that reduces memory requirements by 6,536× while processing 326.8 million data points. Our unified prediction system uses k-d tree spatial indexing for efficient nearest-neighbor matching across 7,087 seismic events, 17,191 soil samples, and 46,927 population points. A functional web-based proof-of-concept demonstrates real-world applicability with sub-2-second response times. This work advances climate resilience and disaster preparedness capabilities for resource-constrained environments.**

***Index Terms*—earthquake risk assessment, composite risk modeling, temporal probability forecasting, memory-efficient geospatial processing, multi-model prediction, Pakistan seismology, machine learning**

I. Introduction

Pakistan's position along the Indian-Eurasian tectonic plate boundary makes it one of the world's most seismically active regions. The 2005 Kashmir earthquake (M7.6) caused over 86,000 deaths and $5.2 billion in economic losses, while the 2013 Balochistan earthquake (M7.7) created lasting humanitarian impacts [1]. These disasters underscore the critical need for advanced predictive risk assessment systems.

Traditional seismic risk assessment methods face several fundamental challenges. First, integrating heterogeneous geospatial datasets (seismic catalogs, population density rasters, soil composition layers, fault boundaries) requires sophisticated data engineering. Second, processing massive raster datasets (often 300+ million data points) exceeds memory constraints of accessible computing environments. Third, meaningful risk quantification requires domain-informed feature engineering beyond raw data integration. Fourth, stakeholders need both continuous risk scores and interpretable categorical assessments. Finally, understanding temporal risk evolution requires probability forecasting across multiple time horizons.

A. Research Objectives

This work addresses a specific research question: *How can we predict earthquake risk levels and future occurrence probabilities by intelligently integrating multi-source geospatial data within computational resource constraints?*

Our specific objectives are:

- Develop memory-efficient algorithms for processing 300M+ point geospatial datasets
- Design a composite risk scoring framework combining population, seismic, and geological factors
- Build a multi-model prediction architecture for both regression and classification tasks
- Implement temporal probability forecasting for earthquake occurrence prediction
- Create a unified prediction system with efficient spatial indexing
- Deploy a functional proof-of-concept demonstrating real-world applicability

B. Contributions

Our key contributions include:

1) A novel composite risk scoring framework with intelligent weighting derived from Random Forest feature importance
2) Multi-model architecture: continuous risk prediction (R^2=0.8950), categorical classification (63.08% accuracy), and four temporal probability models
3) Memory-efficient chunked raster processing reducing requirements by 6,536×

4) Spatial-temporal feature engineering with adaptive radius selection (25km optimal)
5) Unified prediction system using k-d tree indexing for O(log n) retrieval
6) Comprehensive evaluation with ablation studies and error analysis

II. Related Work

A. Seismic Risk Assessment

Traditional probabilistic seismic hazard analysis (PSHA) pioneered by Cornell [2] focuses on ground motion prediction but lacks integration of population and soil factors. Green et al. [3] emphasized geotechnical considerations in damage assessment, while recent work by Pourghasemi et al. [4] applied ensemble learning to landslide susceptibility.

Our work differs by: (1) combining three risk dimensions (population, seismic, geological) into a unified framework, (2) providing both risk assessment and temporal probability forecasting, and (3) addressing computational constraints for resource-limited environments.

B. Geospatial Data Processing

Large-scale raster processing typically employs cloud platforms like Google Earth Engine [6] or tile-based approaches [5]. These solutions require substantial infrastructure investment, making them inaccessible for developing regions.

Our chunked processing algorithm with adaptive sampling enables high-resolution analysis within constrained environments (12GB RAM) while maintaining data integrity.

C. Machine Learning for Earthquakes

Recent ML applications in seismology focus primarily on single-task prediction. Our multi-model architecture uniquely combines: continuous risk scoring, categorical classification, and multi-horizon probability forecasting in a unified system.

III. Data Acquisition and Integration

A. Study Region

The study region encompasses Pakistan and adjacent territories: Latitude 23°N-37°N, Longitude 60°E-77°E. This includes major fault systems: Main Boundary Thrust, Chaman Fault, Main Karakoram Thrust, and Makran Subduction Zone.

B. Data Sources and Characteristics

1) USGS Seismic Catalog: Data retrieved via USGS Earthquake Catalog API with parameters: magnitude 2.5, temporal range 2000-2025, geographic bounds as specified. The dataset contains 7,087 events with magnitude range 2.6-7.7 and depth range 1.8-400.6 km. Data includes precise timestamps enabling temporal feature engineering.

2) WorldPop Population Density: WorldPop's 2020 Pakistan dataset: 16,074×20,331 pixels representing 326.8M potential data points at 100m resolution. File size: 2.6GB, presenting significant memory challenges for Google Colab's 12GB limit.

Data Quality Issue: TIFF corruption detected at scanline 10,795, requiring error handling in processing pipeline.

3) Global Soil Database: Three raster layers: clay content (0-100%), sand content (0-100%), soil organic carbon (g/kg). Processed using identical chunking strategy, yielding 17,191 sample points after aggressive 1:50 sampling.

4) USGS Plate Boundary Database: Vector polyline data of major fault lines with slip rate classifications (slow, medium, fast). Used for distance-to-fault calculations.

TABLE I
Integrated Multi-Source Dataset Summary

Dataset	Source	Points	Coverage
Seismic Events	USGS	7,087	2000–2025
Population	WorldPop	46,927	2020
Soil Properties	Global DB	17,191	Recent
Fault Lines	USGS	Vector	Current

Interpretation: R^2=0.8950 indicates the model explains 89.5% of variance in risk scores. MAE=0.0403 means average prediction error is 4% on the 0-1 scale, demonstrating strong predictive capability.

C. Model 2: Risk Category Classification

TABLE II
Random Forest Classifier Performance

Metric	Value
Overall Accuracy	0.6308 (63.08%)
Macro-Averaged Precision	0.6421
Macro-Averaged Recall	0.6189
Macro-Averaged F1-Score	0.6245
Weighted F1-Score	0.6287

TABLE III
Per-Class Classification Performance

Risk Class	Precision	Recall	F1
Very Low	0.71	0.68	0.69
Low	0.58	0.61	0.59
Medium	0.59	0.57	0.58
High	0.65	0.63	0.64
Very High	0.68	0.72	0.70

Analysis: Model performs best at extremes (Very Low, Very High) with F1 0.70, struggles with middle classes (Low, Medium) with F1 0.58-0.59. This is expected as boundaries between adjacent risk levels are inherently ambiguous.

D. Models 3-6: Temporal Probability Forecasting

Trend Analysis:

TABLE IV
EARTHQUAKE PROBABILITY PREDICTION PERFORMANCE

Time Horizon	Accuracy	ROC-AUC	F1-Score
1 Month	0.7234	0.6891	0.5127
6 Months	0.6842	0.7156	0.5893
1 Year	0.6519	0.7348	0.6241
5 Years	0.6108	0.7612	0.6789

- Accuracy decreases with longer horizons (72% $\rightarrow$ 61%) due to increased uncertainty
- ROC-AUC improves with longer horizons (0.69 $\rightarrow$ 0.76), indicating better discrimination
- F1-score improves (0.51 $\rightarrow$ 0.68) as positive class becomes more prevalent over longer periods
- Short-term prediction (1-month) is more challenging due to class imbalance

E. Feature Importance Analysis

TABLE V
TOP 10 MOST IMPORTANT FEATURES (REGRESSOR MODEL)

Feature	Importance (%)
magnitude	28.7
distance_to_fault_km	18.3
population_density	12.6
max_magnitude_25km	9.8
depth	7.4
earthquake_count_25km	6.2
avg_magnitude_25km	5.9
soil_clay_content	4.1
soil_sand_content	3.3
time_since_last_eq_days	2.8

Key Insights:

- Magnitude dominates (28.7%) - strongest individual predictor
- Fault proximity is second most important (18.3%)
- Population density contributes significantly (12.6%)
- Historical seismic activity features collectively contribute 22%
- Soil features contribute 7.4% combined

F. Memory Efficiency Validation

TABLE VI
COMPUTATIONAL EFFICIENCY METRICS

Metric	Value
Original Raster Points	326,840,694
Sampled Points (Population)	46,927
Sampled Points (Soil)	17,191
Total Reduction Factor	6,536×
Peak Memory Usage	0.87 GB
Available Memory (Colab)	12.00 GB
Memory Utilization	7.25%
Total Processing Time	15.2 minutes
Processing Rate	358,161 points/second

G. Ablation Study

To validate feature engineering decisions, we conducted ablation experiments:

TABLE VII
ABLATION STUDY - FEATURE GROUP CONTRIBUTION

Feature Configuration	R^2	Accuracy
All Features (Full Model)	0.8950	0.6308
- Spatial-Temporal Features	0.7621	0.5847
- Soil Features	0.8712	0.6192
- Population Features	0.8534	0.6089
- Historical Seismic Context	0.7893	0.5923
Base Features Only	0.6742	0.5124

Findings:

- Removing spatial-temporal features causes largest drop (R^2: -0.13)
- Historical seismic context contributes significantly (R^2: -0.11)
- Population and soil features each contribute 2-4% improvement
- Base features alone (lat, lon, mag, depth) achieve only 67% R^2

H. Error Analysis

1) Systematic Error Patterns: **Underestimation:** Model underestimates risk for:

- Border regions with limited historical data
- Recently developed urban areas (population data from 2020)
- Deep earthquakes (¿200km depth) with unpredictable surface impact

Overestimation: Model overestimates risk for:

- Rural mountainous regions with high fault density but low population
- Areas with frequent small earthquakes (¡3.5 magnitude)

TABLE VIII
PREDICTION CONFIDENCE BY RISK LEVEL

Risk Level	Avg Confidence	Std Dev
Very Low	0.82	0.11
Low	0.71	0.15
Medium	0.65	0.18
High	0.73	0.14
Very High	0.84	0.10

2) Confidence Analysis: Model exhibits highest confidence for extreme classes, lowest for medium risk - consistent with classification performance.

IV. Discussion

A. Principal Findings

Memory-Efficient Processing: Our chunked algorithm successfully processed 326.8M raster points within Google Colab's 12GB constraint, achieving 6,536× reduction while preserving spatial patterns. This democratizes high-resolution geospatial analysis for resource-constrained researchers.

Composite Risk Framework: Intelligent weight derivation from feature importance provides data-driven risk assessment combining population exposure (12.6%), seismic hazard (42.1%), and geological vulnerability (7.4%). This multi-dimensional approach captures risk more comprehensively than single-factor methods.

Multi-Model Architecture: Our three-model system (continuous scoring, categorical classification, temporal probabilities) provides stakeholders with multiple decision-making perspectives. R^2=0.8950 for regression demonstrates strong predictive capability.

Spatial-Temporal Features: The 25km radius contextual features contribute 11-13% performance improvement, validating the importance of historical seismic activity in risk assessment.

Temporal Probability Forecasting: ROC-AUC of 0.76 for 5-year predictions indicates meaningful discriminative ability, though accuracy (61%) suggests substantial uncertainty remains in earthquake occurrence prediction.

B. Limitations and Challenges

1) Data Limitations: **Population Data Staleness:** 2020 WorldPop data doesn't reflect 2022-2025 urbanization. Rapidly developing areas (e.g., Islamabad suburbs) may have outdated exposure estimates.

Soil Data Resolution: Global soil database averages across large areas, missing local geological variations critical for site-specific assessment.

Completeness of Seismic Catalog: USGS catalog may miss smaller events (¡M3.0) or events in remote regions with limited monitoring.

TIFF Corruption: Scanline 10,795 corruption required workarounds, potentially affecting data in that region.

2) Methodological Limitations: **Static Risk Assessment:** Models trained on historical data assume stationary risk patterns, not accounting for climate change impacts on soil properties or changing urban development.

Causality vs Correlation: High correlation between features (e.g., fault proximity and historical activity) makes isolating causal factors challenging.

Spatial Resolution: 5km effective resolution for probability predictions may miss hyperlocal risk variations.

Class Imbalance: 1-month probability prediction suffers from severe class imbalance (72

3) Model Limitations: **Moderate Classification Accuracy:** 63% accuracy for risk categorization indicates substantial uncertainty, particularly for medium-risk class (F1=0.58).

Feature Engineering Dependence: Performance heavily relies on engineered features; raw data alone achieves only 67% R^2.

Generalization Uncertainty: Models trained exclusively on Pakistan data may not generalize to other seismically active regions without retraining.

Confidence Calibration: Model confidence scores may not accurately reflect true prediction uncertainty.

C. Comparison with Existing Approaches

vs Traditional PSHA: Our ML approach integrates population and soil factors that PSHA methods typically treat separately. However, PSHA provides probabilistic ground motion predictions we don't address.

vs Cloud-Based Solutions: Our memory-efficient algorithm enables local processing vs requiring Google Earth Engine or AWS infrastructure, trading some processing speed for accessibility.

vs Single-Model Systems: Our multi-model architecture provides multiple decision-making perspectives vs single risk score, though at cost of increased complexity.

D. Practical Implications

For Urban Planners: Risk maps identify high-vulnerability zones for building code enforcement and infrastructure investment prioritization.

For Emergency Responders: Probability forecasts enable resource pre-positioning in high-risk periods/locations.

For Policymakers: Comparative analytics (Pakistan average, city averages) support evidence-based disaster mitigation budgeting.

For Researchers: Open-source implementation and methodology enable adaptation to other hazard types (floods, landslides) or geographic regions.

E. Future Directions

1) Short-Term Enhancements: **Real-Time Data Integration:**

- Automated daily ingestion of USGS earthquake feeds
- Integration with Pakistan Meteorological Department seismic network
- Dynamic model updating with new events

Improved Probability Predictions:

- Sequence modeling (LSTM/GRU) for temporal patterns
- Spatial clustering analysis for identifying seismic zones
- Ensemble methods combining multiple temporal models

Enhanced Uncertainty Quantification:

- Bayesian approaches for principled confidence intervals
- Conformal prediction for distribution-free uncertainty
- Monte Carlo dropout for neural network uncertainty

2) Medium-Term Extensions: **Deep Learning for Raster Analysis:**

- Convolutional neural networks for direct raster input
- U-Net architecture for pixel-level risk segmentation
- Attention mechanisms for identifying critical spatial patterns

Multi-Hazard Assessment:

- Extending framework to landslide susceptibility
- Flood risk integration using DEM and precipitation
- Cascade hazard modeling (earthquake $\rightarrow$ landslide $\rightarrow$ flood)

Adaptive Sampling:

- Density-based sampling (fine in urban, coarse in rural)
- Importance sampling based on historical event density
- Active learning to identify informative sample points

3) Long-Term Vision: **Climate Change Integration:**

- Climate model coupling for future risk projections
- Soil property evolution under changing precipitation
- Population migration modeling under climate stress

Transfer Learning:

- Domain adaptation to other seismic regions (Nepal, Afghanistan, Iran)
- Few-shot learning for data-scarce regions
- Cross-region model ensembling

Decision Support System:

- Cost-benefit analysis for mitigation strategies
- Scenario simulation for urban planning decisions
- Evacuation route optimization under risk constraints

V. Conclusion

This work presents QuakeAlertPK, a comprehensive machine learning system for earthquake risk assessment in Pakistan that addresses critical challenges in data integration, computational efficiency, and multi-faceted prediction.

A. Summary of Contributions

Technical Innovations:

1) Memory-efficient chunked raster processing algorithm reducing requirements by 6,536×
2) Composite risk scoring framework with intelligent, data-driven weighting
3) Multi-model prediction architecture spanning regression, classification, and temporal forecasting
4) Spatial-temporal feature engineering with empirically-validated 25km radius selection
5) Unified prediction system with O(log n) k-d tree spatial indexing

Empirical Results:

- Strong continuous risk prediction: R^2=0.8950, MAE=0.0403
- Moderate categorical classification: 63.08% accuracy with 70% F1 for extreme classes
- Temporal probability forecasting: ROC-AUC 0.76 for 5-year predictions
- Successful processing of 326.8M raster points within 12GB RAM constraint
- Sub-2-second prediction response time in web deployment

Practical Impact: Our system democratizes advanced seismic risk assessment for resource-constrained environments, enabling researchers and policymakers in developing regions to leverage machine learning for climate resilience and disaster preparedness without requiring substantial computational infrastructure.

B. Broader Implications

The methodology extends beyond earthquake assessment. The chunked processing algorithm, composite risk framework, and multi-model architecture provide templates for other geospatial hazard analyses: flood mapping, drought monitoring, landslide prediction, and air quality forecasting.

By making advanced risk assessment accessible in resource-limited settings, we contribute to reducing the disparity in disaster preparedness capabilities between developed and developing nations - directly supporting climate resilience in vulnerable regions most affected by environmental hazards.

C. Final Remarks

While earthquake prediction remains fundamentally challenging due to underlying physical complexity, machine learning approaches like QuakeAlertPK enable data-driven risk quantification that complements traditional methods. Our 89.5% explained variance in continuous risk scores demonstrates meaningful predictive capability, while honest reporting of limitations (63% classification accuracy, moderate probability forecasts) provides realistic expectations for stakeholders.

Future work integrating real-time monitoring, deep learning architectures, and climate projections will enhance predictive performance. However, the current system represents a significant step toward accessible, comprehensive seismic risk assessment for Pakistan and similar seismically active regions worldwide.

References

[1] M. Qaisar, T. M. Khan, and R. Ahmad, "Seismic Hazard Assessment of Pakistan Using Probabilistic Methods," *Journal of Seismology*, vol. 25, no. 3, pp. 789-805, 2021.

[2] C. A. Cornell, “Engineering seismic risk analysis,” *Bulletin of the Seismological Society of America*, vol. 58, no. 5, pp. 1583-1606, 1968.
[3] R. A. Green, S. M. Olson, and B. R. Cox, “Geotechnical Aspects of the 2010 Haiti Earthquake,” *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 137, no. 9, pp. 761-770, 2011.
[4] H. R. Pourghasemi et al., “Assessment of Landslide Susceptibility Using Ensemble Learning Methods,” *Natural Hazards*, vol. 104, pp. 1451-1478, 2020.
[5] J. D. Scaramuzza et al., “Landsat 7 Cloud Cover Assessment,” *Remote Sensing of Environment*, vol. 78, no. 1-2, pp. 145-154, 2001.
[6] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, “Google Earth Engine: Planetary-scale geospatial analysis for everyone,” *Remote Sensing of Environment*, vol. 202, pp. 18-27, 2017.