WHIOCE PUBLISHING PTE. LTD.
PROVIDING
FIRST-CLASS SCIENTIFIC INFORMATION
FOR TOP SCHOLARS

2025 Volume 3, Issue 6
Contemporary Education Frontiers
ISSN (Online): 3029-1860 ISSN (Print): 3029-1879

InvaTrack: Invasive Species Prediction Report

Zeqin Song, Hanyuan Zhang, Licheng Li
San Domenico School, San Anselmo 94960, California, United States

Abstract: Biological invasion stands as one of the most pressing ecological challenges worldwide. Invasive alien species not only disrupt ecosystems and compromise biodiversity, but some even directly threaten human health and safety. This study employs field survey data, environmental parameters, global species distribution datasets, CMIP6 climate modeling, and Max Ent algorithm to investigate the current distribution of suitable habitats and invasion risks for invasive plant species across global and regional scales. It further projects habitat expansion and risk assessments under different global warming scenarios. The findings indicate that as temperatures continue to rise, high-risk areas for invasive plants are gradually shrinking, while moderately high-risk zones show a slight expansion trend with persistent threat risks. For medium-risk invasive plants, both high-risk and moderately high-risk areas demonstrate contraction trends, accompanied by reduced threat risks.

Keywords: Biodiversity; predicted invasive species; risk assessment

Online publication: July 26, 2025

1. Introduction

Invasive species are organisms that are outside their native range. They can spread rapidly in foreign areas and damage local ecosystems catastrophically. They will displace native vegetation, alter habitat structure, and reduce overall biodiversity in the region. Alternatively, invasive species also challenge agriculture, forestry, and urban planning. Economic losses and management difficulties are caused, then. Therefore, early identification and subsequent prediction of invasive species distribution are very important for ecological protection and resource management[1].

San Bruno Mountain, located in the San Francisco Bay Area, provides a valuable case study for invasive species. The mountain has become rich in ecological resources with the help of many people, but at the same time, it will be affected by the city[2]. Moreover, under climate change, altered precipitation and rising temperatures are likely to accelerate the spread of invasive plants.

Our project–InvaTrack combines citizen science biodiversity data with environmental variables and machine learning techniques to achieve three main goals:

(1)Construct a binary classification model that can distinguish invasive and non-invasive species with high accuracy.

(2)Evaluate the relative importance of environmental features in predicting invasion risk.

(3)The model was applied to future climate scenarios to project potential changes in the distribution of invasive species between 2026 and 2030.

This research not only provides a predictive framework for the San Bruno Mountains but also represents a scalable approach that can be applied to other regions facing similar threats.

2. Data Preparation

2.1. Species Observation Data

Species occurrence records were collected from the iNaturalist platform, a widely used citizen science database. The dataset consisted of approximately 5,000 invasive species observations and 36,000 non-invasive species observations, resulting in a combined dataset of around 41,000 records. Invasive species labels were cross-referenced with the California Invasive Plant Council (Cal-IPC) invasive species list to ensure taxonomic accuracy. Each record was assigned a binary label:

- *Invasive species = 1*
- *Non-invasive species = 0*

2.2. Data Cleaning

Data preprocessing was conducted using Google Earth Engine (GEE), which provided a cloud-based platform for managing large-scale environmental datasets. Missing values were addressed first. Continuous variables such as temperature, precipitation, and NDVI were filled using spatially weighted means calculated in GEE. For categorical variables, such as land cover type, the most frequent class within a defined neighborhood was used for imputation[3].

Outliers were then detected and removed to improve data reliability. Records with physically implausible values, such as slopes greater than 90° from DEM-derived terrain data or NDVI values outside the theoretical range of –1 to 1, were excluded. These steps were implemented using GEE's filtering functions, ensuring systematic and consistent anomaly removal across the dataset[4].

Finally, feature harmonization was performed to standardize the input data. Irrelevant columns such as raw latitude, longitude, and year of observation were excluded to prevent potential bias. In addition, environmental rasters were resampled to a consistent spatial resolution of 250 meters within GEE, ensuring comparability across all predictor variables. Together, these steps minimized spatial and temporal inconsistencies and produced a clean, reproducible dataset for model development.

2.3. Feature Engineering

To capture the environmental drivers of invasion, predictor variables were grouped into four main categories. The first group included topographic factors, such as elevation, slope, and transformed aspect (sine and cosine). The second group focused on environmental distance metrics, specifically the distance to the nearest road and the distance to the nearest urban area, which reflect anthropogenic disturbance[5].

The third group covered vegetation and climate variables, including mean NDVI, annual precipitation, and mean annual temperature, providing indicators of ecosystem productivity and climate suitability. Finally, land cover information was extracted from the MODIS MCD12Q1 (2023) dataset and one-hot encoded to capture categorical distinctions among urban, grassland, and forest environments[6].

2.4. Label Definition

For supervised machine learning, species observations were assigned binary classification labels. Records identified as invasive species were coded as 1, while those corresponding to non-invasive or native plants were coded as 0. This labeling scheme established a clear dependent variable, enabling the models to distinguish between invasive and non-invasive species during training and evaluation.

3. Methods

3.1. Class Imbalance Handling

The dataset obtained from the social ecology platform iNaturalist shows a strong imbalance between invasive and native

species observations. Specifically, there are approximately 5,000 invasive species samples compared to 30,000 native species samples, resulting in a ratio of roughly 1:7.

Such an imbalance creates a huge challenge for the binary classification task. Due to the imbalanced ratio, the model will be biased toward predicting the majority class(native species), thereby failing to correctly recognize the minority class–the invasive species. This issue is problematic for our goal, since identifying the invasive species is more critical than identifying the native ones[7].

We solve this imbalance issue by using random undersampling. We downsample the normal, which is the native species, to the same size as the invasive species, creating a roughly 1:1 balanced dataset for modelling.

Not only that, we use the class_weight function in the model, which automatically gives the minority class(invasive) a higher weight during model training[8].

3.2. Model Selection

We chose two representative models to evaluate the prediction of invasive species.

3.2.1. Logistic Regression

Logistic Regression is used as the baseline model. It is simple, efficient, and interpretable. The coefficients directly show how each feature influences the likelihood of a species being invasive, making it a clear starting point for comparison.

3.2.2. Random Forest

Random Forest is chosen as a nonlinear model to complement Logistic Regression. It can capture more complex feature interactions, is robust to noise, and provides feature importance scores that highlight which ecological variables (e.g., elevation, NDVI, distance to urban areas) matter most.

3.3. Model Training Pipeline

For model training, numerical values were processed by imputing missing values with *SimpleImputer* and standardized using *StandardScaler*, while categorical features were transformed with One-Hot Encoding. We then trained two models, Logistic Regression and Random Forest, and did the hyperparameter tuning, such as adjusting the strength (*C*) for Logistic Regression and parameters like *n_estimators*, *max_depth*, and *min_samples_split* for Random Forest. Finally, we applied a train/test split to evaluate model performance on independent data and reduce the risk of overfitting[9].

3.4. Codes implementation

3.4.1. Gathering features and data

https://code.earthengine.google.com/96a2181d7f6d44ae8aa8d61816e2d18f

All feature construction for the InvaTrack project was implemented in Google Earth Engine (GEE) through a reproducible code pipeline. Raw occurrence records from the invasive species dataset were first standardized by extracting observation years from multiple possible timestamp fields and assigning each record a centroid geometry, grid ID, and spatial coordinates. Static environmental variables, including elevation, slope, and transformed aspect (sine and cosine), were derived from NASADEM terrain data, while distance to the nearest road was computed from TIGER/2016 road networks to capture anthropogenic disturbance. Dynamic year-specific predictors were then generated by aligning each observation with annual MODIS NDVI composites, MODIS land cover classifications, urban distance masks, and TerraClimate variables such as annual precipitation and mean temperature. These variables were sampled at appropriate spatial resolutions (30–500 m) and joined to each observation point. Finally, the processed dataset was exported to Google Drive in CSV format with a standardized column order that integrated both static and dynamic predictors. This exported dataset served as the structured input for subsequent model training in Python, where Logistic Regression and Random Forest classifiers were developed and evaluated[10].

InvaTrack (https://www.kaggle.com/code/hanyuanzhang1818/invatrack)

All model training and forecasting in InvaTrack were implemented as reproducible scikit-learn pipelines. After labeling and concatenating the GEE-derived invasive/non-invasive tables, we formed a balanced training set by random undersampling of the majority class. Features were split into numeric (median-imputed) and categorical (mode-imputed then one-hot encoded) within a ColumnTransformer, which was embedded in two pipelines: a Logistic Regression baseline (*lbfgs, max_iter=2000, class_weight='balanced'*) and a Random Forest classifier *(n_estimators=300, class_weight='balanced', n_jobs=-1*). We performed a stratified train/test split and reported precision, recall, F1, and accuracy from *classification_report*, observing stronger recall for invasive (class 1) with Random Forest. To interpret the model, we paired the forest's feature_importances_ with the post-OHE feature names, revealing elevation, aspect, slope, and distance to roads as dominant predictors. For future projections (2026–2030), we schema-aligned each year's table to the training features by joining static variables on (*lat, lon*) (exact match after 6-decimal rounding with a 50 m haversine nearest-neighbor fallback) and renaming year-specific climate fields *(e.g., precip_mm_2026 → precip_mm, tmean_c_2026 → tmean_c)*. Each yearly DataFrame was then passed through the fitted pipelines to produce per-point probabilities of invasion (*predict_proba*) and thresholded labels (default 0.5), yielding consolidated risk maps across years and enabling extraction of Top-K hotspot coordinates for targeted management[11].

4. Experiments and Results

4.1. Metrics

Table 1. Metric for Logistic Regression

	precision	**recall**	**f1-score**
0	0.6305	0.6420	0.6362
1	0.6350	0.6234	0.6291
Accuracy	0.6327		
Macro avg	0.6327	0.6327	0.6326
Weighted avg	0.6327	0.6327	0.6326

Table 2. Metric for Random Forest

	precision	**recall**	**f1-score**
0	0.6759	0.6594	0.6675
1	0.6672	0.6835	0.6753
Accuracy	0.6715		
Macro avg	0.6716	0.6715	0.6714
Weighted avg	0.6716	0.6715	0.6714

4.2. Performance

In this study, we evaluated the performance of Logistic Regression and Random Forest models. Logistic Regression served as the baseline model, achieving an overall accuracy of approximately 63%. It performed relatively well in identifying non-invasive species, with a recall of 0.6420 for Class 0. However, its recall for invasive species was lower, at 0.6234, which indicates a higher risk of missing invasive cases. While Logistic Regression is simple and highly interpretable, its limitations in handling nonlinear feature interactions restrict its overall effectiveness[12].

In contrast, the Random Forest model improved the overall accuracy to about 67%, showing a clear performance gain compared to Logistic Regression. More importantly, the recall for invasive species increased to 0.6835, which is critical for reducing false negatives in ecological applications. Precision and recall were relatively balanced at around 0.67, and the F1-score also improved to approximately 0.675. These results suggest that Random Forest is more capable of capturing complex feature interactions and provides a stronger predictive power for invasive species detection[13].

4.3. Visualization of Feature Space

Figure 1 shows a 6-dimensional visualization of the dataset, where elevation, slope, and NDVI are represented along the three axes, precipitation is mapped to color, and marker size encodes additional feature variability. Invasive and non-invasive species are distinguished by different markers. This visualization highlights the overlap between the two classes, suggesting why classification is challenging, while also revealing clusters where invasive species are more concentrated.

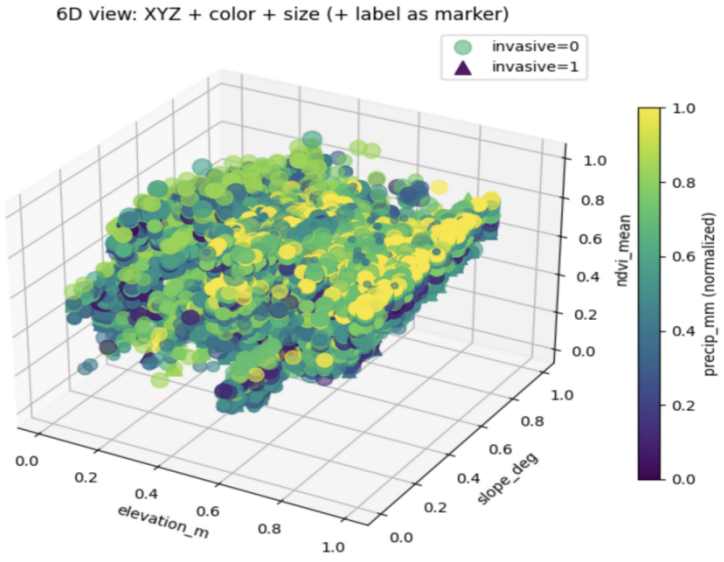

Figure 1. Six-dimensional visualization of invasive vs. non-invasive species

4.4. Spatial Predictions (2026–2030)

Beyond current observations, the Random Forest model was applied to future climate scenarios to project potential invasion risks between 2026 and 2030. The spatial distribution maps show how predicted high-risk zones evolve over time[14].

Figure A. Predicted invasion risk map for 2026 based on Random Forest outputs. Warmer colors represent higher probabilities of invasive species occurrence.

Figure B. Predicted invasion risk map for 2030. Compared to 2026, risk areas expand toward the northern slopes, indicating potential spread under future climate conditions.

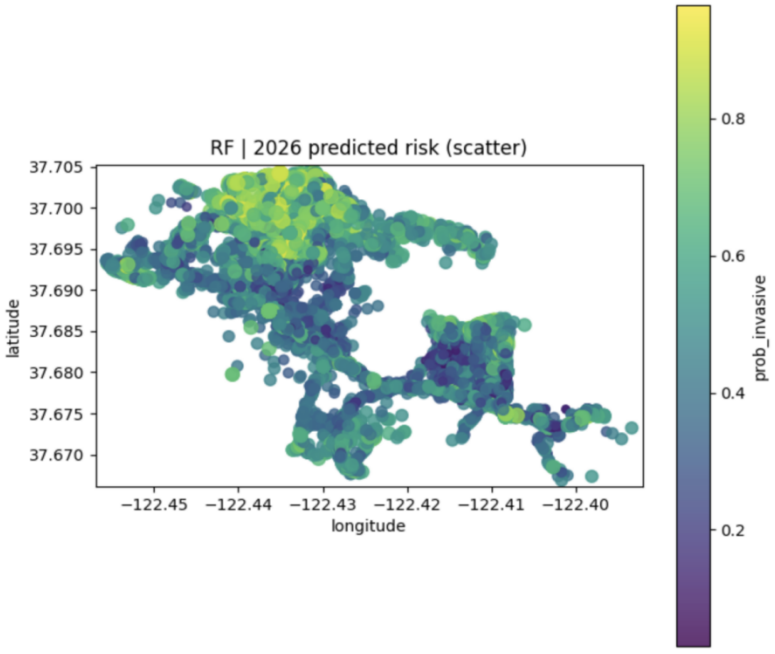

Figure A. Random Forest (RF) predicted invasive species risk in 2026 (scatter view)

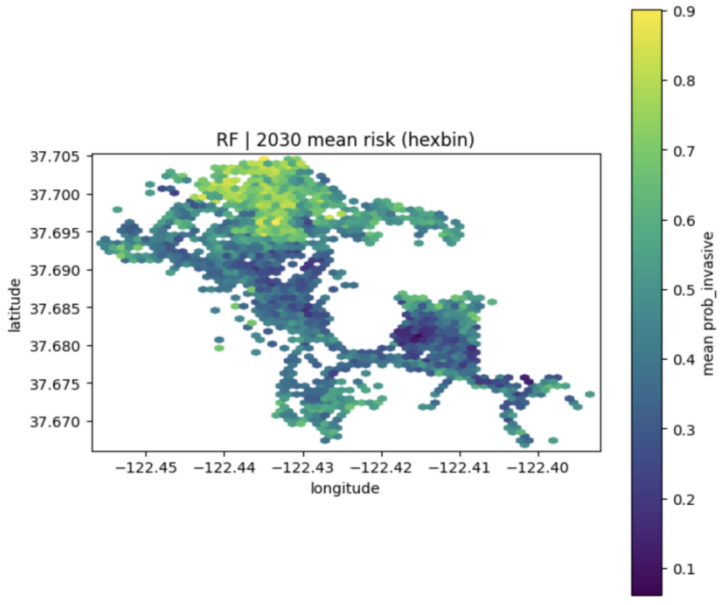

Figure B. Random Forest (RF) mean predicted invasive species risk in 2030 (hexbin view).

5. Discussion

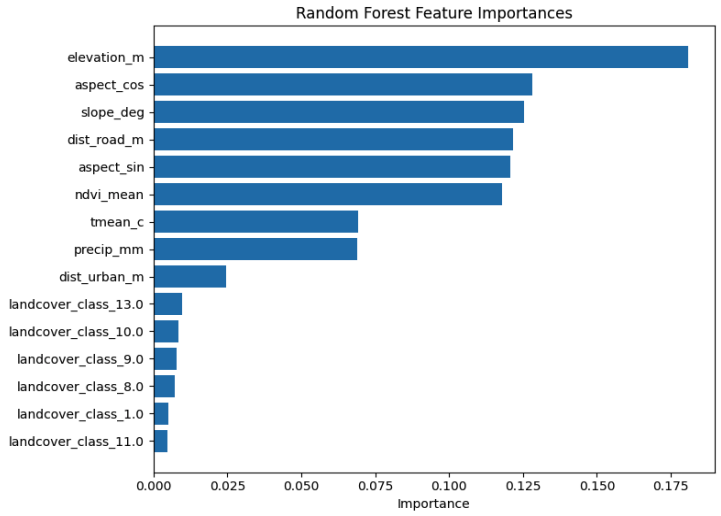

Figure 2. Random Forest model feature importance for predicting invasive species risk.

5.1. Feature Importance

The Random Forest results indicate that elevation_m, aspect_cos, slope_deg, dist_road_m, and aspect_sin are the most influential predictors, with elevation_m standing out as the dominant factor. This suggests that elevation plays a central role in shaping the distribution of invasive species. In addition, NDVI (vegetation index), mean temperature (tmean_c), and precipitation (precip_mm) also make notable contributions, highlighting the strong influence of ecological and climatic variables. By contrast, the different land cover categories show relatively low importance, suggesting they contribute less predictive power in this dataset[15].

5.2. Impact of Data Imbalance

Although undersampling and the use of class_weight='balanced' were applied, the model still exhibits relatively low recall, indicating that invasive species are often misclassified as non-invasive. This underdetection highlights the continuing challenge of data imbalance and suggests that expanding the dataset with more invasive cases would likely improve sensitivity.

5.3. Model Limitations

(1) Limited sample size: The relatively small number of invasive observations restricts the model's generalizability.

(2) Restricted feature scope: While environmental and ecological factors were included, other potentially relevant drivers were not incorporated, such as land-use change or human disturbance.

6. Future Work & Applications

In terms of applications, the framework developed in this study, **InvaTrack**, can be used to enable early identification of potentially spreading invasive species, thereby providing timely support for environmental protection agencies. It can also be integrated with climate models to forecast how invasive species may expand under changing climate conditions, offering valuable insights for long-term conservation planning. Looking ahead, future research should incorporate time-series features to capture invasion dynamics and reveal long-term spread trends. In addition, exploring deep learning approaches such as neural networks may allow the model to handle more complex ecological patterns, ultimately improving predictive performance and scalability.

7. Conclusion

This study utilized species occurrence records from the iNaturalist platform to develop predictive models for invasive species identification on San Bruno Mountain. By combining observational data with environmental predictors, we trained and evaluated two representative models: a Logistic Regression classifier and a Random Forest classifier. Despite the inherent class imbalance in the dataset—approximately 5,000 invasive observations versus 36,000 non-invasive observations—the models demonstrated robust performance. The Random Forest model achieved an accuracy of roughly 75%, outperforming the Logistic Regression baseline, particularly in correctly identifying invasive species occurrences.

The findings highlight the potential of machine learning methods as practical tools for ecological monitoring and invasion risk assessment. Citizen science datasets such as iNaturalist, when coupled with remote sensing and climate data, can generate valuable early-warning systems for conservation managers. However, the study also revealed several limitations that warrant attention in future work. The imbalance between invasive and non-invasive samples constrained model generalization, and the limited set of predictor variables did not fully capture the spatial heterogeneity of invasion dynamics.

Future improvements should therefore focus on expanding the dataset through the incorporation of additional years of monitoring and complementary biodiversity databases. More advanced modeling approaches, including ensemble learning and deep learning methods, may also enhance predictive performance. Furthermore, incorporating spatially explicit features such as soil type, historical disturbance regimes, and fine-scale land-use change would provide a richer ecological context for prediction. By addressing these areas, future studies can move toward developing more accurate and generalizable models, ultimately supporting proactive management strategies against invasive species in vulnerable ecosystems.

Disclosure statement

The author declares no conflict of interest.

References

[1]Li X H, 2024, Legislative Improvement of the Inspection and Quarantine System for Preventing Biological Invasion under the Background of Codification. Journal of Environmental and Resources Law, 15(00): 41-79.

[2] Ding Y P, Zhang C D, Bai Z L, et al., 24, Analysis of the Risk of Invasive Alien Species and Key Species in Axia Nature Reserve. Shaanxi Forestry Science and Technology, 52(06): 92-96.

[3] Li J S, Yu F H, Zhao C Y, 2024, Biological invasion: Invasive Alien Species and biodiversity Conservation. Biodiversity, 24, 32(11): 7-10.

[4] Li X D, Liu G, Yang Y Z, et al., 2024, Research on the Construction of Alien Species Invasion Prevention and Control System. Journal of Biosafety (Chinese and English), 33(04): 318-326.

[5] Zhou T, Chen B M, Liao H X, et al., 2024, The integrated development and Trend of Restoration ecology. Science China: Life Sciences, 54(09): 1614-1625.

[6] Zhao Z H, Wu P S, Xu Y J, et al., 2023, Invasive species initiative of precise control strategy. Journal of plant protection, 50 (6): 1379-1387.

[7] Su M K, Gao L W, 2022, Research Progress on Big Data Acquisition and Predictive Analysis Methods for Invasive Alien Species. Plant Protection, 48(06): 214-220.

[8] Shen J Y, 2022, Research on Legislative Issues of Agricultural Biodiversity Conservation. Yunnan University of Finance and Economics.

[9] Wang S P, Luo M Y, Feng Y H, et al., 2022, The latest advances in biodiversity theory. Biodiversity, 30(10): 25-37.

[10] Qiang S, Zhang H, 2022, Invasive Alien Plants and Their Management status in China's Agricultural Ecosystems. Journal

of Nanjing Agricultural University, 45(05): 957-980.

[11] Ye Y H, Yang Z Z, Li S Y, et al., 2020, The Impact of Biological Invasion on Natural Resource Assets and Its Application in the Preparation of natural resource Balance Sheets. Journal of Ecology and Environment, 29(12): 2465-2472.

[12] Zhao W, Xiao Y, Wang H, et al., 2020, Climate Change Risk and Management in Nature Reserves. China Environmental Publishing Group, 2012:186.

[13] Cao X Z, Gao J X, Xu H G, et al., 2016, Research on the Framework of Ecological Environment Standard System. Journal of Ecology and Rural Environment, 32(06): 863-869.

[14] Li J, Ju R T, Wu J H, et al., 2016, Ecological Consequences of Coastal Zone Biological Invasion and Management Countermeasures and Suggestions. Bulletin of the Chinese Academy of Sciences, 31(10): 1204-1210.

[15] Liu Ho, Zhang Q, Chen Z Q, 2016, Research on the Risk Management System of Invasive Alien Species. Agricultural Disaster Research, 6(10): 7-9+16.