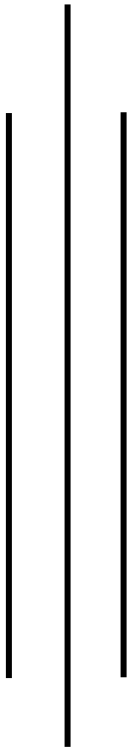
Leveraging Machine Learning to Confirm Invasive Species Reports



Taposh Mollick

Final Project

CEE 609: Environmental Data Science

Master of Science

Division of Environmental Science

State University of New York College of Environmental Science and Forestry

Syracuse, New York

December 2024

Abstract

Invasive species pose significant ecological and economic threats, requiring efficient tools for their detection and management. This study leverages machine learning (ML) techniques, specifically computer vision, to streamline the confirmation process of invasive species reports within iMapInvasives, focusing on the invasive species in New York State. Using 463 Spotted Lanternfly confirmed records selected through stratified random sampling, the combined score was validated as a robust metric for classification confidence through statistical analyses, including ANOVA, Tukey's HSD, and ROC curve analysis. An optimal threshold of 93.61 was identified, achieving an overall classification accuracy of 95.25%. The integration of iNaturalist VisionAPI enabled automated classification into high, medium, or low confidence levels, reducing reliance on manual reviews while ensuring precise identifications. Combined scores proved to be a reliable predictor of classification accuracy, with spatial autocorrelation enhancing confidence by analyzing species proximity to known populations. High-quality images, characterized by clear features, close-up views, and simple backgrounds, significantly improved recognition, while poor-quality images posed challenges. This study demonstrates the transformative potential of ML tools in invasive species management by automating verification processes, reducing manual effort, and enhancing data reliability. While the system effectively bridges automation and expert validation, addressing dataset quality and refining algorithms remain crucial for scalability and broader application. The findings highlight the integration of ML-driven automation as a vital step toward efficient and scalable invasive species monitoring and management solutions.

Keywords:

Invasives Species

Machine Learning (ML)

Computer Vision

iMapInvasives

iNaturalist

1. Introduction

Invasive species carry significant ecological and economic consequences as they can disrupt ecosystem functioning and have adverse effects on agriculture, human health, and transportation (Drake et al., 2003). Invasive species are non-native organisms, either plant or animal, that have successfully established populations and spread in natural or cultivated habitats beyond their original range and cause harm to the environment, the economy, or human health (Colautti and MacIsaac, 2004; Devin and Beisel, 2007). Around 50,000 non-native species have been introduced into the United States. Certain species have proven beneficial and currently contribute to over 98% of the US food system, valued at approximately $800 billion annually. However, some invasive species have led to significant environmental damage, resulting in losses and requiring control costs estimated at nearly $137 billion per year (Pimentel et al., 2000). According to Allen and Bradley (2016) by 2050, regions like the northeastern United States will experience a rise in economic losses from these non-native invasive species as the climate becomes conducive to introducing hundreds of new species.

Simpson et al. (2018) report that around 11,400 non-native species have been established in the United States. However, only a few non-native species display invasive characteristics that lead to significant ecological disruptions and economic costs. The small fraction of the established species is vast, making tracking and managing invasive species difficult. For this reason, automating early detection and rapid response (EDRR) to invasive species through advances in technological applications like computer vision is necessary (Jensen et al., 2020; Martinez et al., 2020; Sun et al., 2017).

The New York State Department of Environmental Conservation (NYS DEC) uses iMapInvasives, an online, mobile-friendly, GIS-based data management system, to track, identify and manage invasive species. It enables natural resource professionals and citizen scientists to report invasive species quickly and easily, facilitating real-time tracking and improving management decisions to protect native ecosystems. Creating verified data within iMapInvasives begins when a user submits a photo of the organism and coordinates and selects a species name from the list of known and anticipated invasive species in the state using online or mobile apps. These records are initially entered as “unconfirmed” presence records. A taxonomic expert receives an email alert or searches for a subset of data in iMap. The expert reviews the submitted photo and its location on the map. If the species name assigned by the observer is correct, the expert changes the record to confirmed; if it is incorrect, the expert

deletes or changes the record. The process from unconfirmed reports to confirmed species takes significant time due to the manual review process. Therefore, automatic and real-time identification using image recognition and computer vision techniques for species identification is necessary to streamline the process and avoid extensive manual reviews.

We aim to develop an efficient computer vision system for species identification to aid in the confirmation of user-submitted photos of invasive species for use by natural and agricultural resource managers and decision-makers. This system will leverage components of existing and open-source machine learning tools to process digital images at a large scale, utilizing finite computation resources and time while also analyzing the geo-coordinates of the locations of invasive species. The target dataset will come from iMapInvasives, a GIS-based invasive species mapping and data-sharing platform across North America that serves as the official database used by New York State for the public and professionals concerned with invasive species.

1.1. Literature review

1.1.1. The Challenges of Invasive Species Management

Invasive species management is a demanding task that requires significant investments of time and financial resources for effective detection and control (Lodge et al., 2006). Invasive species management necessitates a methodical approach, and experts have established transparent review processes to identify invasive species. This not only includes the region-specific impacts of each species but also the projected costs and viability of their management. While shared databases such as iMapInvasives (Jewitt et al., 2021; Van Horn et al., 2021a) offer valuable platforms for collecting invasive species reports from various sources, the sheer volume of data can be overwhelming for manual review and analysis (Terry et al., 2020). However, recent advancements in computer vision and deep learning (DL) present promising opportunities to streamline analysis. By leveraging machine learning (ML) algorithms, it becomes feasible to automate the analysis of user-submitted images and efficiently verify invasive species occurrences (Mohammadi et al., 2018). Integrating ML into invasive species management can significantly enhance the accuracy and efficiency of species identification, enabling more proactive and targeted control strategies. These advancements can transform current management practices by reducing reliance on manual analysis and speeding up decision-making processes.

1.1.2. Computer Vision in Species Recognition

ML, particularly using mobile apps and Convolutional Neural Networks (CNNs), has revolutionized species recognition. These apps allow users to identify plants and animals by uploading a photo, where the app's algorithm analyzes the image, matches it against a known database, and suggests the most likely species (Sun et al., 2017). This technology democratizes species identification, enabling citizen scientists and researchers to contribute to biodiversity monitoring and conservation. The accuracy and efficiency of these algorithms improve as they collect more data. CNNs, in particular, have become the standard for image classification due to their success in the ImageNet Large Scale Visual Recognition Challenges (Krizhevsky et al., 2012; Russakovsky et al., 2015; Willi et al., 2019) and their application in identifying both invasive and native species with high precision (Ashqar and Abu-Naser, 2019; Borges Oliveira et al., 2021; Elias, 2021; Ravoor and Sudarshan, 2020; Sun et al., 2017). Designed to reduce the need for manual preprocessing and excel in computer vision, CNNs represent a significant step forward in the application of ML for environmental and conservation efforts, engaging the public in scientific activities and offering efficient, real-time species identification (Gu et al., 2018; LeCun, 2015).

1.1.3. Existing Machine-Learning and Deep-Learning Platforms

Numerous ML and DL platforms exist, such as iNaturalist (Van Horn et al., 2018), Pl@ntNet (Sun et al., 2017), eBird (Sullivan et al., 2009), PictureThis (Otter et al., 2021), PlantFinder (Dehnen-Schmutz, 2011), to identify invasive and non-invasive plants and animals. Sun et al. (2017) used the geometry and shape of plants such as leaves, flowers, fruits, barks, habits, and other characteristics to identify plant species. Yang et al. (2022) introduced a shape descriptor called the high-level triangle shape descriptor (HTSD) to identify plant species from leaf images. The HTSD combines contour points based on triangle features (CPTFs) and salient point triangle features (SPTFs) using the Fisher vector encoding. Dissimilarities between HTSD characteristics are calculated using the Euclidean distance. Söderkvist (2001) extracted shape characteristics and moment features of the leaves to conduct an analysis on 15 different Swedish tree classes utilizing back propagation (BP) to train a feed-forward neural network. Fu et al. (2004) extracted leaf vein features of living plants for their identification based on an artificial neural network (ANN). Leaf vein extraction is another technique for leaf segmentation that combines the snake technique with cellular neural networks (Li et al., 2005). He and Huang (2008) used the probabilistic neural network as a classifier for plant leaf image identification,

achieving higher accuracy compared to the BP neural network. These diverse approaches demonstrate the versatility and potential of ML techniques for accurate plant species recognition.

CNN is also widely used for animal species recognition (Norouzzadeh et al., 2018). Among the popular convolutional networks (ConvNets) for recognition and classification tasks, notable ones include AlexNet, Network in Network (NiN), VGG16, GoogLeNet, ResNet, and DenseNet (Favorskaya and Pakhirka, 2019). AlexNet introduced the use of Rectified Linear Unit (ReLU) activation functions, enabling nonlinear transformations (Russakovsky et al., 2015). NiN was one of the pioneering architectures that incorporated 1×1 convolutions to enhance the combination of features within the convolutional layers (Lin et al., 2013). VGG16 consisted of sixteen convolutional layers, multiple max-pool layers, and three fully connected layers, showcasing a deep architecture (Simonyan and Zisserman, 2014). GoogLeNet, designed for computational efficiency, utilized significantly fewer parameters than AlexNet while maintaining high accuracy (Szegedy et al., 2015). ResNet introduced the concept of residual learning, where the output of one or more convolutional layers is connected to their original input, facilitating better learning and optimization (He et al., 2016). DenseNet, on the other hand, established connections between every layer and every other layer, resulting in improved classification performance (Huang et al., 2017). Animal biometrics uses quantified methods to represent and recognize animals based on their visual appearances, including shape, structure, color, pattern, and size. It focuses on specific characteristics like the face, coat pattern (for zebra and whales), spot points (for tigers), and muzzle pattern (for cattle). It is used as a pattern recognition system (Kühl and Burghardt, 2013; Ravoor and Sudarshan, 2020). Kumar et al. (2016) applied face recognition and representation approaches such as principal component analysis (PCA), local discriminant analysis (LDA), independent component analysis (ICA), batch-candid covariance-free incremental PCA algorithm (CCIPCA), and independent-candid recognition-based incremental PCA (IND-CCPCA) for representing pixel intensity of facial features from a database of cattle faces for identifying animals. In cattle identification, muzzle point recognition has been accomplished using texture-based approaches like the Speeded Up Robust Feature (SURF) descriptor, yielding 90% identification accuracy with an 8×8 size and the kappa statistics approach (Noviyanto and Arymurthy, 2012). Noviyanto & Arymurthy (2013) proposed a refinement technique using the Scale-Invariant Feature Transform (SIFT) approach for matching muzzle print images, leading to improved cattle identification through refinement of SIFT key point matching.

As image recognition applications become more accessible and user-friendly, the potential for the public to collect data useful to researchers and resource managers greatly increases. iNaturalist is a widely used citizen science platform that leverages machine learning for the recognition of both plant and animal species, providing users with an efficient way to identify and document biodiversity. When users submit images of plants or animals, the iNaturalist system processes these images using ML algorithms, particularly deep learning techniques like CNNs, which analyze the visual features in the images and compare them to a vast database of human-verified species images, allowing the system to suggest potential identifications for the submitted observations (Van Horn et al., 2021b). This ML-based approach not only enhances the accuracy and efficiency of species identification but also enables users with little to no expertise in taxonomy to participate in scientific data collection and contribute to biodiversity research. To display the location of confirmed species on a map, iNaturalist uses the latitude and longitude associated with user-uploaded geotagged images (Van Horn et al., 2018) and the coordinates also help to predict species by using spatial autocorrelation. The platform employs a base map integrated with the geographical information system (GIS) to visualize the spatial distribution of identified species. This feature enhances the usability and practicality of iNaturalist for researchers, conservationists, and managers seeking to understand and monitor species distributions across various geographical regions. Additionally, the geographical coordinates or location of the input images help predict the species by analyzing the geo score or spatial autocorrelation, which is associated with the proximity to the same species in nearby locations. By combining ML-powered species recognition and geospatial visualization, iNaturalist empowers a global community of users to contribute valuable data on biodiversity and support the study and conservation of diverse ecosystems.

1.1.4. Spatial Distribution Mapping of Invasive Species

There are several methods to show the spatial distribution of plant and animal species and predict harmful invasive species, such as field surveys that involve physically visiting an area to record plant species and creating distribution maps (Augustin et al., 1996). Species distribution mapping of invasive species involves collecting accurate location data using GPS, mobile apps for geospatial data collection, and other geospatial tools to predict the spread and impact of these species. Remote sensing uses satellite images to identify and map plant species across large areas (Nininahazwe et al., 2023; Sawaya et al., 2003). GIS combines information about locations and details about those locations to facilitate analysis to show where plants and animals are found and what environmental factors affect them (Guisan and Thuiller, 2005).

Species Distribution Models (SDMs) use data on where plants are found to predict where they might occur based on environmental conditions (Srivastava et al., 2019; Williams et al., 2009). Ecological Niche Modeling (ENM) predicts where species might live based on their ecological needs and can identify possible areas for invasive species to spread (Barve et al., 2011; Kulhanek et al., 2011).

Citizen science involves volunteers gathering data about where plants are found to help monitor invasive species using tools like iNaturalist, eBird, etc. (Dickinson et al., 2012; Johnson et al., 2020). However, while some apps enable location visualization, in others apps, e.g., the PictureThis and Pl@ntNet apps, there is no option to see the spatial distribution on a map of plant species (Sun et al., 2017). iMapInvasives incorporates a map component to show the location and spatial distribution of confirmed and unconfirmed invasive species. Integrating proximity analyses with image recognition enhances confidence in species identification, which is already present in the ML tool VisionAPI of iNaturalist. As we are using VisionAPI for species recognition for iMapInvasive records, the proximity analysis will be applied to our developed tool. For instance, identifying invasive species far from known populations (indicating a low proximity score or geo score for geotagged images) could be critical for early detection using the geographical likelihood model. However, early detection is possible by applying a visually similar model ("Vision") of iNaturalist VisionAPI. By incorporating this early detection approach through visually similar models, we aim to improve the overall management of invasive species.

1.2. Research Gaps and Research Questions

This study addresses the challenge of verifying unconfirmed reports of invasive species within iMapInvasives using machine learning (ML) tools to lessen the time-consuming manual review process. The current process begins when a user submits a photo of the organism and coordinates and selects a species name from the list of known and anticipated invasive species in the state using online or mobile apps. These records are initially entered as "unconfirmed" presence records. A taxonomic expert then receives an email alert or searches for a subset of data in iMap, reviews the submitted photo and its location on the map, and confirms or changes the species identification. This manual review process is time-consuming and resource intensive. Therefore, identification using image recognition and computer vision techniques for species identification is necessary to streamline the process and assist the taxonomic experts with large quantities of records to review.

We aim to develop an efficient computer vision system for species identification to aid the confirmation of user-submitted photos of potentially invasive species. This system will leverage components of existing and open-source machine learning tools to process digital images at a large scale, utilizing finite computational resources and time while also analyzing the locations of invasive species. The target dataset is iMapInvasives, a GIS-based invasive species mapping and data-sharing platform used across North America that serves as the official database used by New York State for the public and professionals concerned with invasive species.

The research questions guiding this study are:

- What insights can be gained from the existing literature on invasive species and image-processing ML methods to inform the automatic identification of invasive species?
- Can we develop a tool that leverages publicly available machine-learning image recognition systems to analyze user-submitted images and confirm invasive species reports accurately?
- Can the developed tool be effectively applied to assist New York State in confirming iMapInvasives invasive species reports?

By addressing these research questions, this study will contribute to the advancement of automated species identification and geospatial mapping, supporting biodiversity research and invasive species management efforts. While we are enthusiastic about the potential of the ML tool, it is essential to note that we do not envision it completely removing the human reviewer. Instead, the tool will provide valuable information to help reviewers make faster and more informed decisions on many records. This collaboration between automated systems and human expertise will ensure accuracy and efficiency in managing invasive species data.

1.3. Objectives and Aims

We aim to develop an efficient computer vision system for species identification. This system will help confirm user-submitted photos of invasive species for use by natural and agricultural resource managers and decision-makers.

This study has three main objectives:

- Explore the literature on invasive species and image-processing ML methods for automatically identifying invasive species.

- Develop a tool to leverage publicly available machine-learning image recognition systems to analyze user-submitted images and confirm invasive species reports.
- Apply the developed tool to assist NYS in confirming iMapInvasives invasive species reports.

2. Materials and Methodology

2.1. Overall Data Process Goal

The overall data processing goal is to enhance the efficiency of confirming invasive species reports within iMapInvasives using machine learning techniques. First, unconfirmed iMapInvasives reports were extracted, including photos, coordinates, and human-assigned species labels. Then we prepared these reports for ML analysis by standardizing and cleaning the data. Finally, we analyzed the prepared reports using the iNaturalist VisionAPI and custom ML code. Fig. 1 illustrates an ML workflow to enhance the confirmation process of invasive species reports. The process begins with input images, which are preprocessed through grayscale conversion, resizing, skeletonization, and segmentation. The dataset is then split into training, validation, and testing sets.

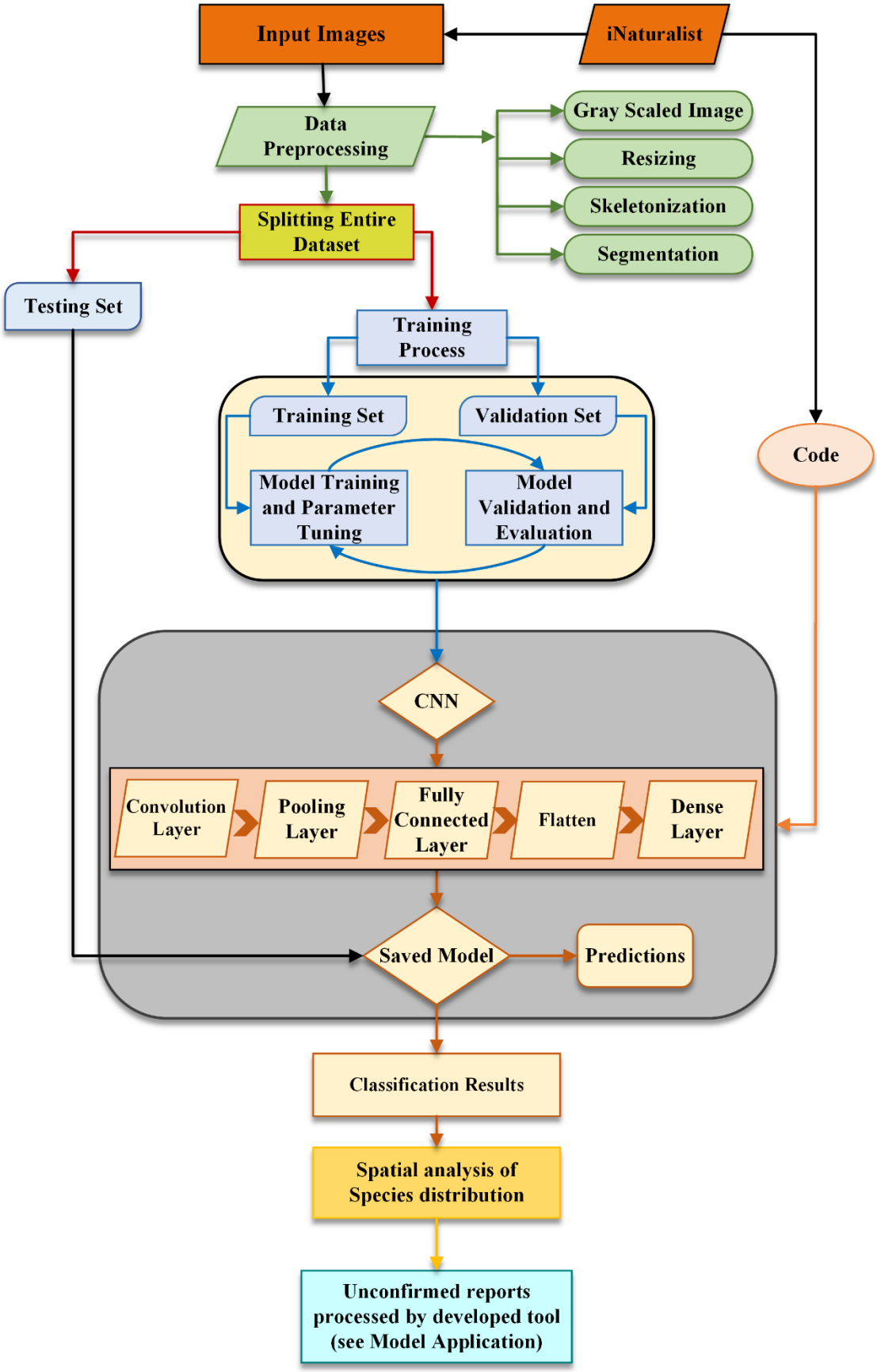


Fig. 1: Methodological flow diagram for CNN model development and training.

The model undergoes parameter tuning using the training set during the training process, while performance is evaluated using the validation set. Feature extraction is performed through a CNN, involving convolution, pooling, fully connected, flattening, and dense layers.

The trained model is used to make predictions. These predictions yield classification results, which are then subjected to spatial analysis to assess species distribution. Finally, the unconfirmed reports are processed using the developed tool, integrating these results into the iMapInvasives platform to improve the management and accuracy of invasive species data. This workflow aims to streamline the confirmation process, reducing reliance on manual review and enhancing data accuracy (see Fig. 1).

2.2. Datasets

iNaturalist is a platform for citizen scientists, enabling them to record and disseminate biodiversity observations worldwide via a dedicated online platform and mobile applications. The CNN model in the iNaturalist VisionAPI was trained, tested, and validated using the iNat2021 dataset. The iNat2021 dataset is freely available and downloadable from the iNaturalist website.

As of 15 June 2024, iNaturalist users had contributed approximately 192,827,502 observations of plants, animals, fungi, and other organisms worldwide, and 369,703 users were active. The users identified a total of 477,585 species worldwide (Fig. 2). Compared to its predecessor, iNat2017, which contained only half as many training images in CNN's algorithm ImageNet, iNat2021 stands out with its expansive collection of 2.7 million training images, 100,000 validation images, and 500,000 test images. These images span across 10,000 distinct species, covering the vast tree of life. Notably, each species within the dataset is represented by a minimum of 152 images in the training set, but more than twice that is available for most species (Van Horn et al., 2021).

Fig. 2: iNaturalist Web Map service for species identification using ML model and citizen science. The rectangular shape and intensity of the orange color represent reports of observations of species from all over the world.

We aim to use ML tools to support the classification of unconfirmed records in iMapInvasives, a North American platform for public sharing of invasive species observations. iMapInvasives is a collaborative, GIS-based database used by professionals and citizen scientists to document invasive species' presence, absence, and treatment (www.imapinvasives.org). In New York State, iMapInvasives features over 250,000 records for approximately 300 invasive species. The platform is essential for tracking invasive species, managing data, and making it accessible for analysis and reporting. Users submit reports, which can include unconfirmed observations that need verification by taxonomic experts (Fig. 3). These records are crucial for the early detection and management of invasive species, but the manual verification process, while thorough, can be time-consuming and resource intensive. By applying machine learning tools, we aim to enhance the efficiency of this verification process, providing taxonomic experts with prioritized and pre-analyzed data to accelerate decision-making and improve the accuracy of species identification.

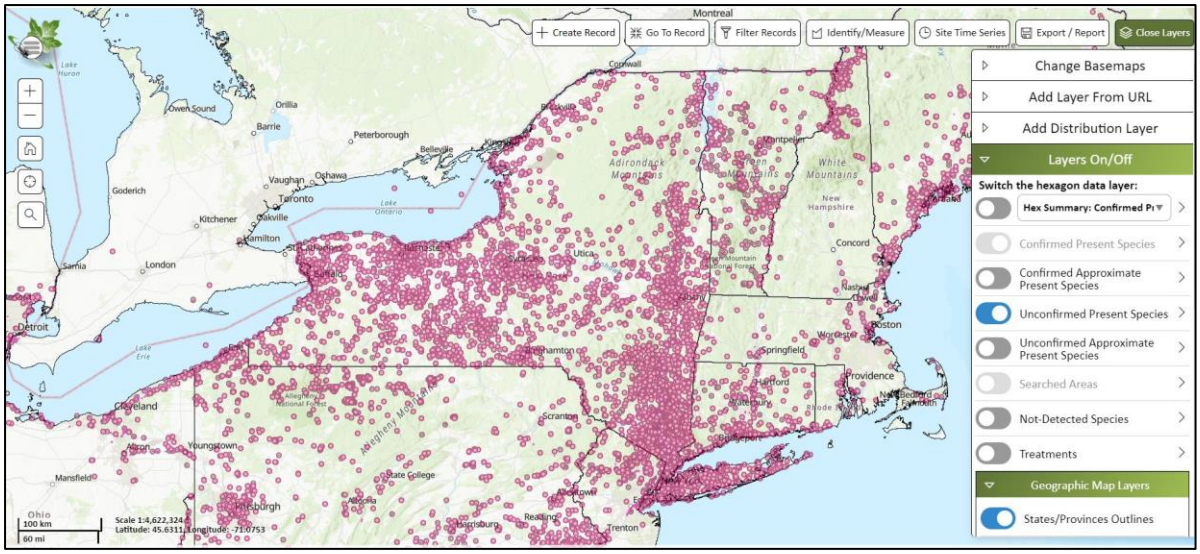


Fig. 3: iMapInvasives tool for invasive species reporting and data management showing unconfirmed present species using pink dots and polygons of the Northeastern United States.

2.3. Study area and sampling techniques

In this research, a total of 463 samples were selected from iMapInvasives confirmed records of the Spotted Lanternfly species across counties in New York State. A stratified sampling method was employed to ensure comprehensive spatial representation across the study area. Initially, the state of New York was divided into strata based on its counties, where each county served as a distinct stratum. This approach allowed for proportional representation of the species across different counties, minimizing geographic bias and ensuring that areas with varying levels of species occurrence were adequately represented.

Within each selected county, a random sampling method was applied to choose specific records. Random sampling ensured that every sample within a given county had an equal probability of being included, which reduced selection bias and improved the representativeness of the data. Combining stratified sampling at the county level with random sampling within counties provided a robust and systematic approach to sample selection, allowing for a balanced and reliable dataset for further analysis of the Spotted Lanternfly distribution and associated characteristics. This methodological framework ensured the integrity of the data while accounting for geographic variation across the state.

2.4. Exploring the creation of a new ML tool

2.4.1. Data Preprocessing

Data preprocessing is a crucial component to enhance the quality and consistency of input images. This multifaceted process includes gray scaling, simplifying images by focusing on structural details rather than color, thereby reducing computational complexity. Various studies employ three image preprocessing techniques to optimize data for analysis by a CNN: resizing, skeletonization, and segmentation (Bradski, 2000; Chandra, 2020; Zhang and Suen, 1984). Resizing adjusts all images to uniform dimensions, ensuring consistency across the dataset, which is crucial for efficient and effective CNN performance. Skeletonization simplifies images to their basic structural outlines, enhancing CNN's ability to recognize species by focusing on essential shapes. Segmentation divides images into distinct regions, isolating specific features for more detailed analysis. These techniques collectively prepare images in a manner that significantly aids CNN in accurately identifying and analyzing different species based on visual characteristics. Together, these steps prepare images for efficient and accurate processing by CNNs, streamlining the path toward effective species identification and beyond.

2.4.2. Splitting Dataset and Dataset Construction

We plan to utilize the iNaturalist VisionAPI, developed from an extensive dataset comprising 2.7 million training images, 100,000 validation images, and 500,000 images for testing. This comprehensive dataset covers 10,000 distinct taxa across 11 classes, ensuring a robust model for species identification. This distribution results in 2,686,843 images for training, 100,000 for validation, and 500,000 for testing. Pre-processing measures, such as resizing images to a maximum of 800px, as recommended by He et al. (2016). The testing set is designated for final validation and evaluation, ensuring the model meets accuracy and performance criteria.

2.4.3. Convolutional Neural Networks (CNNs)

CNNs automate feature extraction in image processing, learning to identify relevant features hierarchically without manual intervention. Initially, CNNs focus on simple, low-level features such as edges and textures, which form the building blocks of images. As a subset of deep neural networks designed for grid-structured data like images, their architecture mimics the human visual system, significantly advancing image recognition tasks (Fig. 4). A typical CNN includes layers such as convolutional, pooling, flattened, and fully connected layers, facilitating a comprehensive analysis of image data (Krizhevsky et al., 2012).

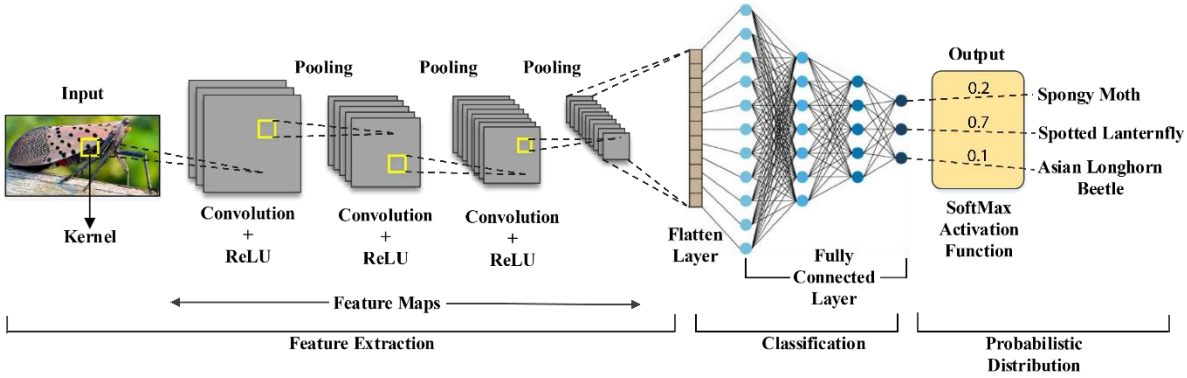


Fig. 4: CNN architecture for species recognition, showcasing layers from input images through convolution, flattening, fully connected layer, to final species classification

2.4.4. Model Predictions and Classification Results

After a Convolutional Neural Network (CNN) model is trained and saved, it is later accessed to identify new images. These images are preprocessed to align with the model's input specifications. During the prediction phase, the model outputs raw scores converted into class labels and associated confidence levels using argmax and SoftMax. The model's performance is evaluated using various metrics, guiding decision-making processes. This workflow, comprising model retrieval, image preprocessing, prediction, and performance evaluation, ensures the model's effectiveness in identifying and categorizing new data (Goodfellow et al., 2016; He et al., 2016; Sokolova and Lapalme, 2009).

2.5. Leveraging an Existing Machine Learning Tool

Fig. 5: Species identification process using iNaturalist VisionAPI and Web Application a) Tree-of-heaven, b) Species recognition result in iNaturalist VisionAPI and c) iNaturalist Web Application.

VisionAPI, developed by iNaturalist, is an advanced image recognition tool designed to identify plants, animals, and other organisms using photographs. It utilizes machine learning and computer vision techniques to identify real-time species, offering a list of potential matches with associated confidence scores. Trained on a vast dataset of labelled images contributed by the iNaturalist community, VisionAPI covers a broad range of taxa and is continuously updated with new data and new species to improve its accuracy over time. Its extensive training data and real-time processing capabilities make it a valuable resource for biodiversity studies, citizen science, and educational applications.

iNaturalist VisionAPI uses two models: visual similarity ("vision") and geographic likelihood ("geo"). For a photo with coordinates, VisionAPI returns scores from each model, then combines them to report what is labelled as the original_combined_score (Fig. 5b). For non-geotagged photos, it uses only the visual similarity model and reports a geo score 0. Visually similar is always true for an included species. Seen nearby is true if the geo score is greater than a threshold used to determine if there are "nearby" observations. The example of tree-of-heaven (Fig. 5a) is used to illustrate the species identification process in the iNaturalist VisionAPI (Fig. 5b) and iNaturalist Web Application (Fig. 5c).

From a technical standpoint, VisionAPI is designed for easy integration into various digital platforms via RESTful API calls. This flexibility allows developers to seamlessly embed sophisticated species identification capabilities within their applications. The API processes requests and returns a structured JSON response, including potential species identifications and their corresponding confidence scores. While the straightforward integration and response format make it an accessible and valuable tool for developers, it is essential to note that the exact methodology used to derive these confidence scores is not publicly disclosed. Nonetheless, leveraging the VisionAPI of iNaturalist can bypass the need to train a new model. We are using VisionAPI to improve the identification process for unconfirmed reports within the iMapInvasives system, enhancing the ability to manage and respond to invasive species observations effectively.

2.6. Set up threshold

The threshold for identifying confirmed Spotted Lanternfly species was established using a statistical approach to ensure accuracy and reliability. First, a one-way ANOVA test was conducted to analyze the differences in the combined score (com_score) across three verification categories: "iMap says correct," "Both iMap and iNat correct," and "Undeterminable." The results revealed significant differences, validating the combined score as a robust indicator for classification accuracy. Tukey's HSD post-hoc test further identified pairwise differences between the groups, supporting the statistical significance of the results. The optimal threshold was then determined using the Receiver Operating Characteristic (ROC) curve and Youden's index, maximizing the balance between true positive and false positive rates. This statistically derived threshold ensures the precise classification of species records, distinguishing between confirmed and unconfirmed identifications.

2.7. Geospatial Component into the Species Identification

Incorporating a geospatial component into species identification can significantly enhance our understanding of species distributions and inform management strategies. After confirming species identification, the focus shifts to using georeferenced observation datasets for spatial analysis. These datasets, excluding images but including geographic coordinates and other metadata like species name and observation date, are stored in databases ranging from simple spreadsheets to complex cloud services. For mapping, GIS platforms (e.g., QGIS, ArcGIS Pro) or web-based tools (e.g., ArcGIS Online, Google Maps API, Mapbox) visualize the data. This georeferencing transforms coordinates into map visuals, facilitating advanced analysis.

In this research, the iMapInvasives web map interface, which already has map visuals for both the confirmed and unconfirmed iMap records. The spatial analysis included steps to identify apparent mistakes, such as aquatic species reported on land, ensuring the accuracy of the data. Additionally, spatial autocorrelation techniques were applied to VisionAPI to identify patterns and distributions of both aquatic and terrestrial species. This approach aims to enhance the understanding of species distribution and support effective management and conservation strategies.

3. Result and Discussion

The classification of Spotted Lanternfly identifications using machine learning algorithms and confidence scores provides a robust framework for assessing the reliability of species

identification. In this study, statistical analyses, including ANOVA and Tukey's HSD tests, were used to evaluate differences in combined score (identification confidence) across three verification categories: "Both iMap and iNat correct," "iMap says correct," and "Undeterminable" which were manually assessed after identification using iNaturalist VisionAPI. Additionally, a threshold of 93.61 was determined using Receiver Operating Characteristic (ROC) curve analysis to separate confirmed from unconfirmed identifications. These analyses aim to validate the effectiveness of combined score as a metric for reliable species classification and ensure the accuracy of identification outcomes. The following sections detail the findings, including statistical validation, threshold determination, and its implications for classification reliability.

3.1. Spatial autocorrelation and identification confidence

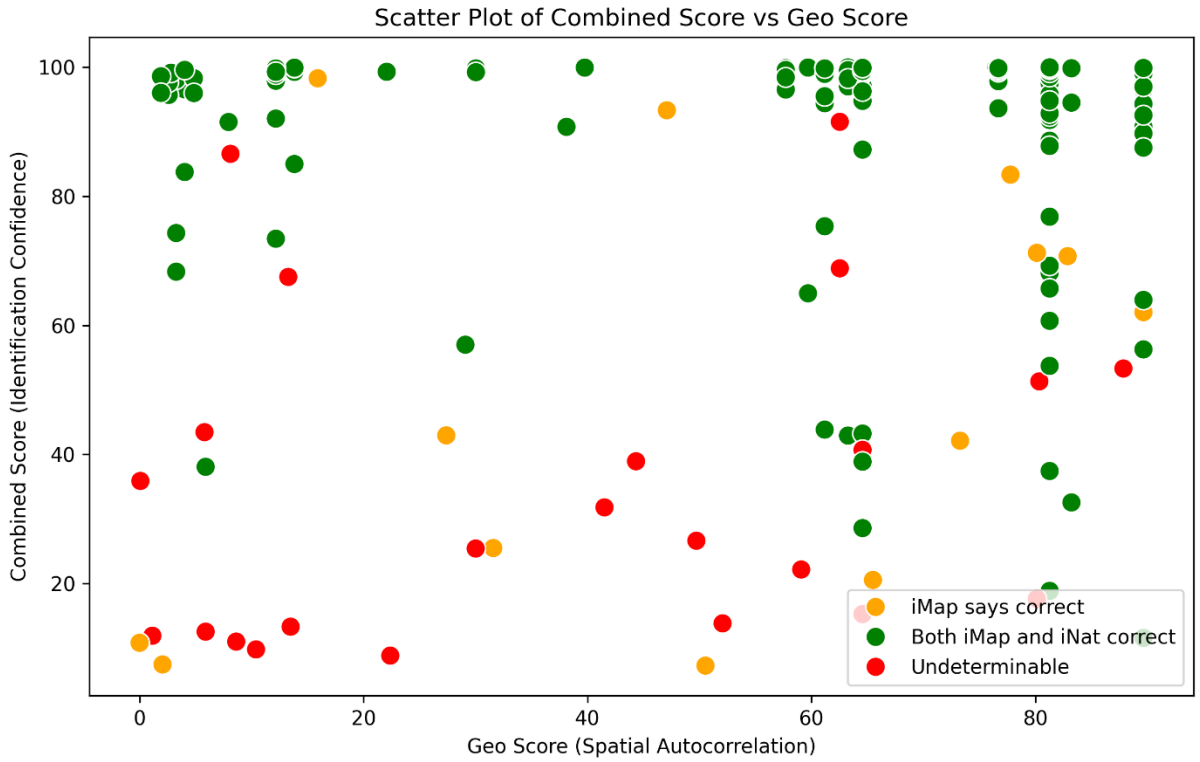


Fig. 6: Relationship between spatial autocorrelation and identification confidence for Spotted Lanternfly records.

The scatter plot (*Fig. 6*) illustrates the relationship between geo score (spatial autocorrelation) and combined score (identification confidence) for species identifications. Correct identifications, represented by green dots, are predominantly found at higher combined scores (above 60) and are distributed across a range of geo scores, particularly at mid-to-high values. This demonstrates the reliability of higher combined scores in ensuring accurate species

identification. In contrast, incorrect identifications, shown as yellow and red dots, are mostly clustered at lower combined scores, although some outliers with high combined scores are present. These anomalies are likely due to mislabeling in the iMapInvasives dataset or complex photo characteristics, such as the presence of multiple species or cluttered backgrounds, which can lead the ML algorithm to focus on non-target species.

The analysis further underscores the combined score as a more reliable predictor of identification accuracy compared to the geo score. Incorrect identifications (both red and yellow dots) are strongly associated with lower combined scores, regardless of their geo scores. This highlights the importance of the combined score in predicting identification accuracy. Moreover, the scatter plot emphasizes the need to address dataset mislabeling and improve image quality to mitigate challenges arising from complex environments or photos containing multiple species. Overall, the combined score proves to be an effective metric for assessing classification reliability.

3.2. Combined score and Geo score

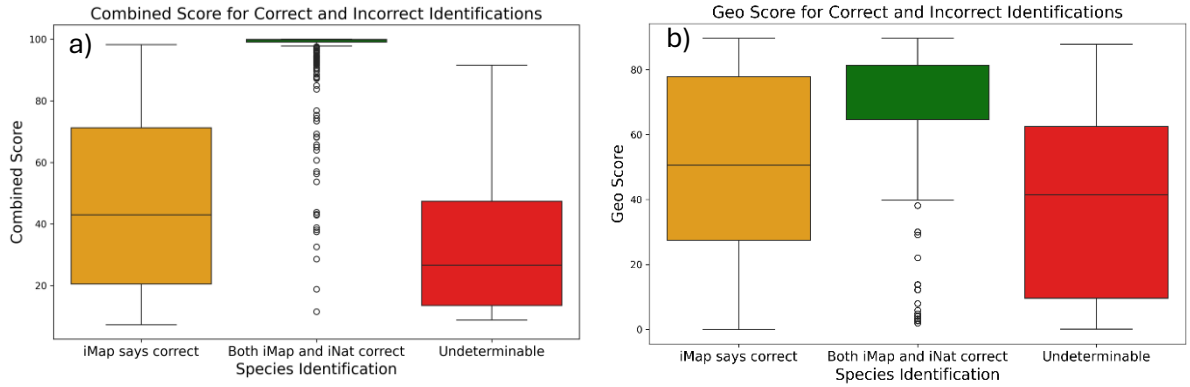


Fig. 7: Combined and geo scores across verification categories for species identification a) Combined score vs verification categories and b) Geo score vs verification categories.

Fig. 7a illustrates the distribution of combined scores and geo scores across three verification categories for species identification. For the combined scores, the "iMap says correct" category (orange) displays a wide range of scores (20–80), with a median of around 40. This indicates moderate confidence in identifications but also reflects some uncertainty. The "Both iMap and iNat correct" category (green) shows consistently high combined scores, close to 100, indicating strong agreement and high confidence in accurate identifications. In contrast, the "Undeterminable" category (red) has low to moderate scores (20–60), with some high outliers. These outliers are likely due to dataset mislabeling or the presence of complex photo characteristics, such as cluttered backgrounds or multiple species in the same image. These

observations emphasize the reliability of high combined scores in confirming correct identifications while highlighting the need for improved data quality to reduce errors in other categories.

The second boxplot (Fig. 7b) focuses on the distribution of geo scores, which reflect spatial autocorrelation. The "iMap says correct" category (orange) has a median geo score of approximately 60, with high variability and some low-score outliers. This suggests moderate reliability but potential issues with dataset errors or false positives. The "Both iMap and iNat correct" category (green) demonstrates a high median geo score of around 80, with minimal variability, confirming strong spatial alignment and reliable identifications. Conversely, the "Undeterminable" category (red) exhibits a low median geo score of approximately 40, with significant variability. This indicates uncertainty and may point to issues such as mislabeling or complex photo characteristics.

3.3. One-way ANOVA test for threshold identification

The one-way ANOVA test assessed differences in combined score across three verification categories: "iMap says correct," "Both iMap and iNat correct," and "Undeterminable." The null hypothesis (H_0) assumed no difference in mean scores, while the alternative hypothesis (H_a) posited significant differences between categories. The test results (F-statistic = 287.41, p-value = 1.03e-81) rejected H_0, confirming significant variations in the mean combined score. "Both iMap and iNat correct" showed the highest scores, reflecting strong classification confidence, while "Undeterminable" had the lowest scores, indicating unreliable identifications. These findings validate combined scores as a reliable metric for distinguishing accurate classifications.

3.3.1. Tukey's HSD comparison test

Table 1: Multiple Comparison of Means - Tukey HSD, FWER=0.05

Group 1	Group 2	Mean difference	p-adj	Lower	Upper	Reject
Both iMap & iNat correct	Undeterminable	-61.5608	0.0	-68.4413	-54.6804	True
Both iMap & iNat correct	iMap says correct	-47.3717	0.0	-56.4212	-38.3221	True

Undeterminable	iMap says correct	14.1892	0.0082	3.0358	25.3425	True

The Tukey's HSD test reveals significant differences in combined score between the three verification groups. The results show that "Both iMap and iNat correct" consistently have the highest combined scores, with a mean difference of -61.56 compared to "Undeterminable" and -47.37 compared to "iMap says correct." This indicates that "Both iMap and iNat correct" is the most reliable category with the highest classification confidence.

On the other hand, "Undeterminable" has the lowest combined scores, with a mean difference of 14.19 below "iMap says correct." These differences are statistically significant, as the p-values for all comparisons are 0.0, and none of the confidence intervals cross zero (Table 1). The rejection of the null hypothesis in all cases confirms that the groups are distinctly different in terms of their combined score.

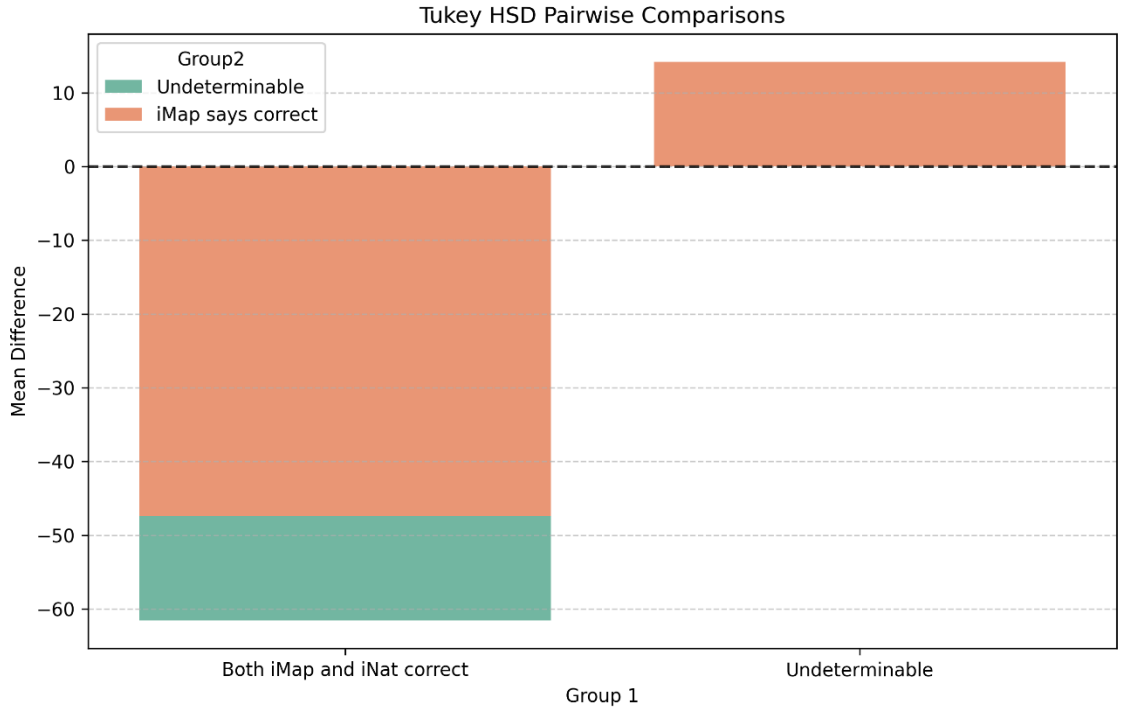


Fig. 8: Tukey's HSD pairwise mean differences in combined scores between the groups.

The Fig. 8 highlights the mean differences in combined scores between verification groups based on Tukey's HSD test. "Both iMap and iNat correct" consistently exhibit significantly higher scores compared to "Undeterminable" (-61.56) and "iMap says correct" (-47.37), confirming its reliability. Additionally, "iMap says correct" scores are moderately higher than

"Undeterminable" (14.19). All comparisons are statistically significant, validating com_score as a robust metric for classification accuracy and distinguishing between reliable and unreliable identifications.

The optimal threshold of 93.61 was determined using the ROC curve analysis, maximizing the Youden Index for distinguishing between correct and incorrect identifications. This threshold ensures a high balance between sensitivity and specificity for reliable classification.

3.3.2. Accuracy assessment and validation of threshold

Table 2: Classification report of accuracy assessment

	Precision	Recall	F1-score	Support
0	0.71	0.67	0.69	36
1	0.97	0.98	0.97	427
Accuracy	-	-	0.95	463
Macro avg	0.84	0.82	0.83	463
Weighted avg	0.95	0.95	0.95	463

The classification results using the threshold of 93.61 achieved an overall accuracy of 95.25%, demonstrating strong performance in identifying "Both iMap and iNat correct" cases (class 1). The identification accuracy can be up to 99% if the input image quality meets the ML requirement (Rzanny et al., 2024). For class 1, the precision, recall, and F1-score were all exceptionally high at 97%, 98%, and 0.97, respectively (Table 2). This indicates the threshold's reliability in distinguishing correct identifications with minimal false positives or negatives.

In contrast, the classification performance for the "Not Both iMap and iNat correct" category (class 0) was comparatively weaker, with a precision of 71%, recall of 67%, and F1-score of 0.69 (Table 2). This is likely due to the smaller sample size and a higher number of misclassifications in this group. The confusion matrix further highlights this imbalance, showing 417 true positives, 24 true negatives, 12 false positives, and 10 false negatives.

3.4. Outcomes: Applying ML tool to iMapInvasives reports

The final product of this research is an ML tool that can assess the human-assigned species label of observations reported to iMapInvasives. The process is visually shown in *Fig. 9*. It begins with an unconfirmed iMapInvasives report, where a user submits a photo with their

labelling. The image is preprocessed as needed and submitted to the recognition machine-learning tool. The tool processes the user-submitted species images and classifies the associated label into high, medium, or low confidence levels of being correct.

A high confidence level indicates that the species in the report is likely correct, which can then prompt data administrators to take different actions depending on the management priority of the species. New York uses a tiered system to classify high-impact species into feasible management strategies (Finley et al., 2023). For common species (Tier 4), it prompts confirmation. It alerts experts for less common species (Tiers 1-3). In species identification, "likely correct" refers to a strong probability that the species has been correctly identified based on the available data or evidence. The term "likely correct" is used because even high-confidence identifications can sometimes be incorrect. Confirmed reports are then updated in iMapInvasives. A medium confidence level indicates uncertainty, necessitating a review by experts. A low confidence level suggests the species is likely incorrect. This prompts alerts to the iMap administrator, who may delete or correct the report and are flagged for further inspection. These reports remain unconfirmed for manual review.

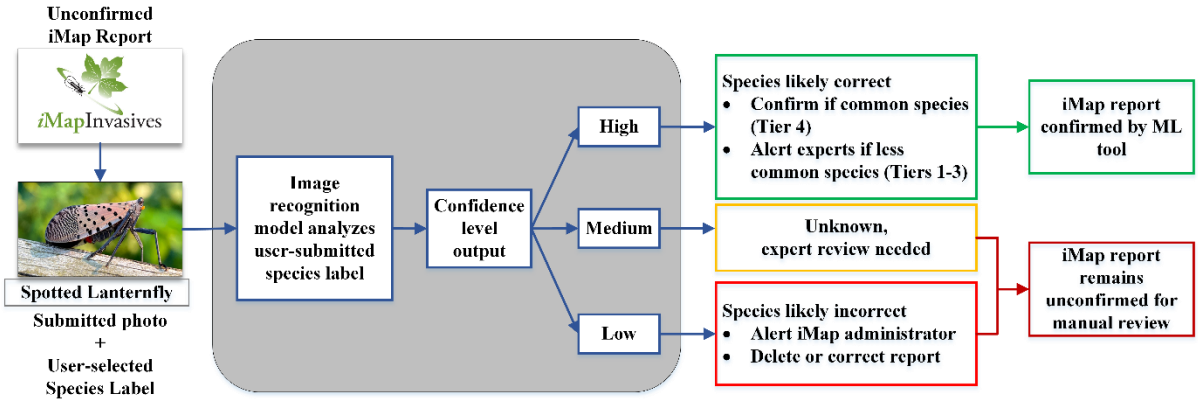


Fig. 9: Identification of unconfirmed species and output confidence level using ML tool

3.5. Recommendations and limitations for image recognition ML species identification

To ensure effective image recognition using machine learning (ML), specific characteristics of good photos play a crucial role in enhancing model performance. High-quality images should prominently feature the target species, captured as close-up shots to maximize visible details (Fig. 10). These photos must be in sharp focus, with the subject clearly discernible. Positioning the target species centrally in the frame is recommended, allowing the ML algorithms to prioritize its features. A solid or simple background is preferable, minimizing distractions and

ensuring the focus remains on the subject. Additionally, including multiple instances of the same species in one image can improve recognition accuracy, as it provides the model with redundant information to validate predictions. Characteristics such as venation details, bark texture, and the structure of leaves further contribute to precise identification, especially in species differentiation tasks.

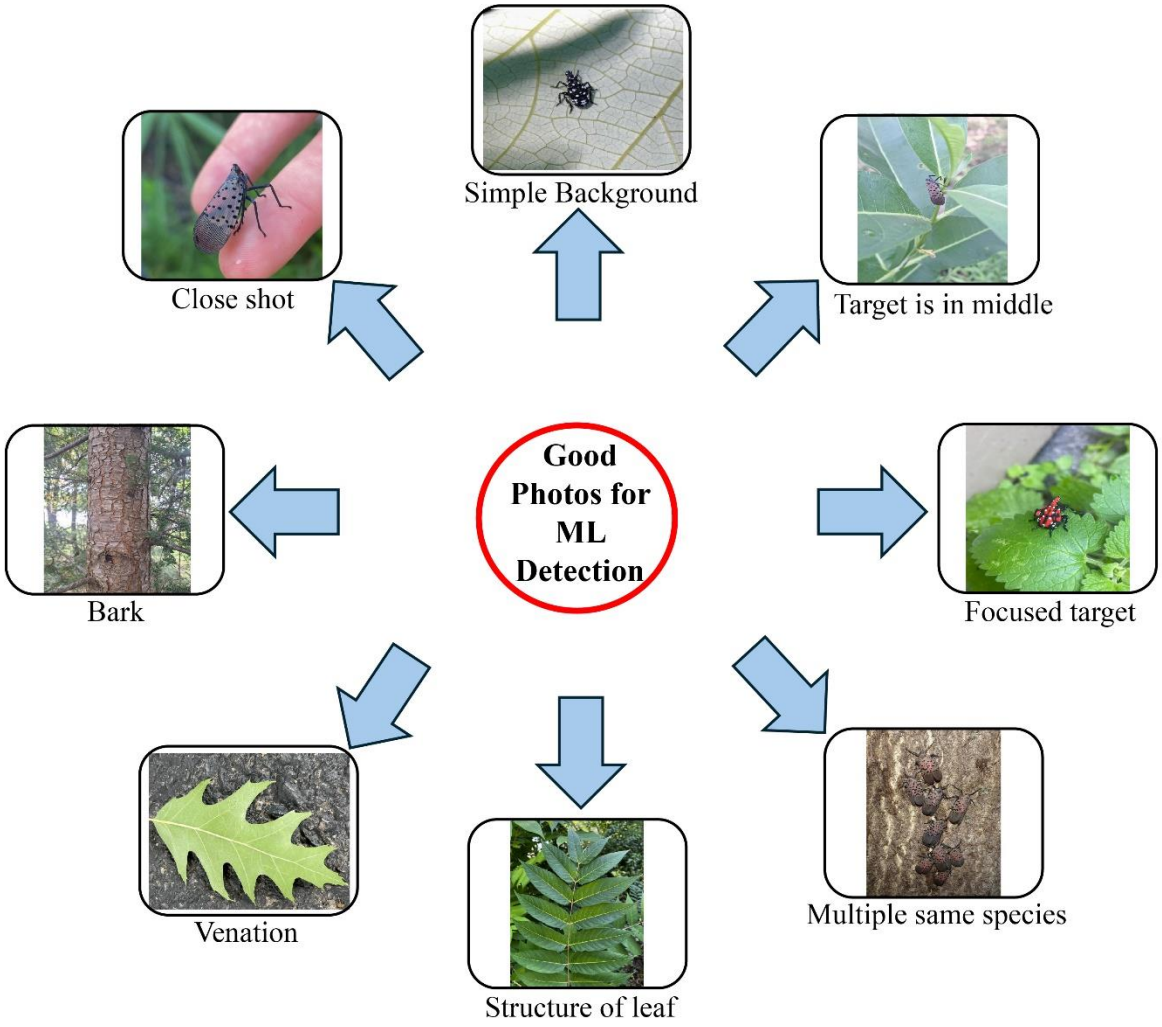


Fig. 10: Characteristics of good photos for image recognition ML tool.

Conversely, poor-quality images can negatively impact recognition accuracy. Common issues include blurry or unfocused targets, distant shots where the subject lacks detail, and poor lighting, particularly overly dark images (Fig. 11). Images featuring multiple species or cluttered environments can confuse the model, making it challenging to identify the intended target. Similarly, photos with ambiguous subjects, such as eggs or fungi that resemble other species, should be avoided. Ensuring proper orientation and focus is equally important to

reduce misclassification. These guidelines underline the importance of providing clear, detailed, and well-composed images for successful ML-based species identification.

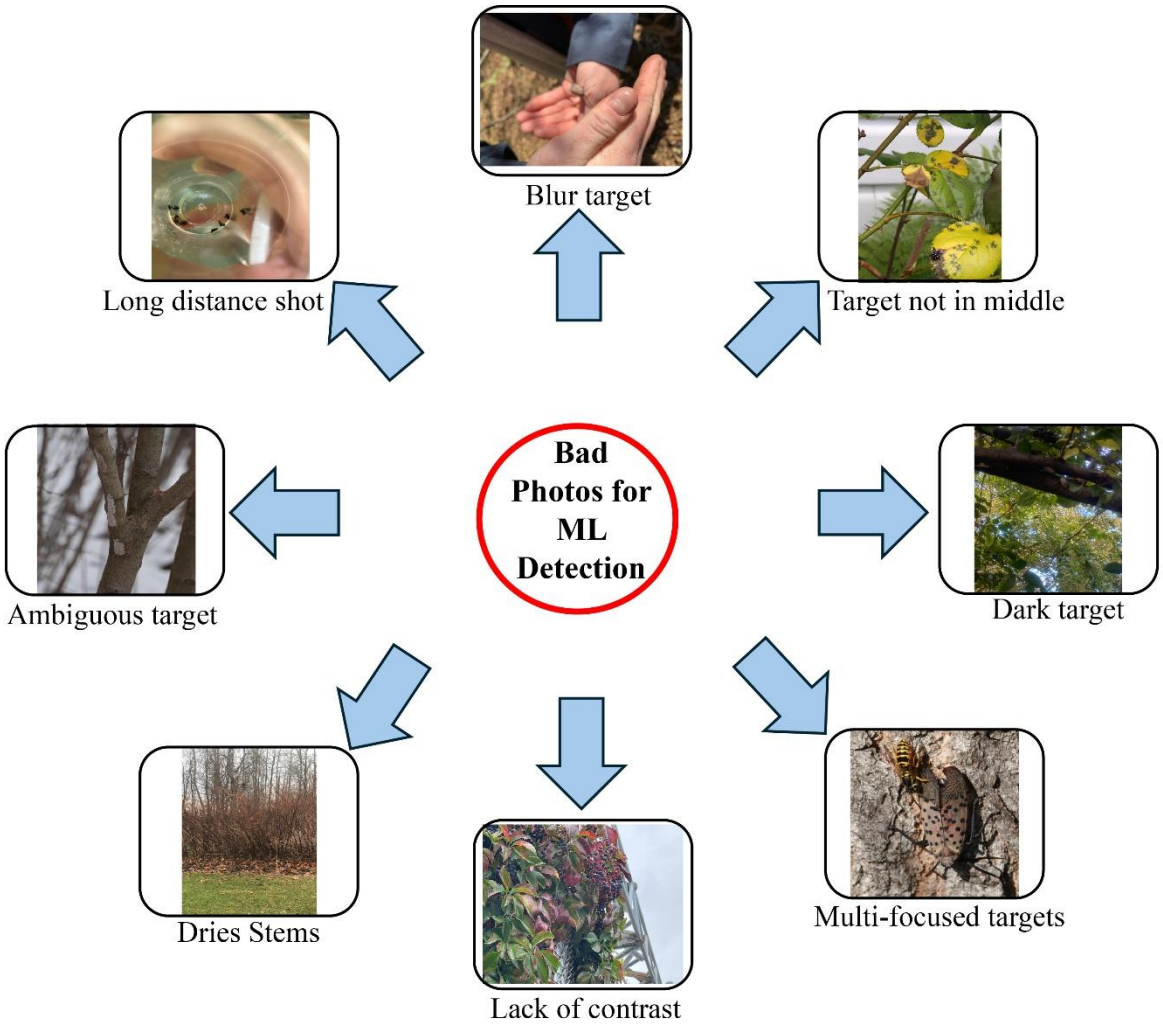


Fig. 11: Characteristics of bad photos or challenges for ML image recognition tool.

4. Conclusion

This study demonstrates the effectiveness of leveraging machine learning (ML) tools, particularly computer vision techniques, to automate the verification process of invasive species reports within iMapInvasives. By integrating statistical methods, machine learning, and spatial analysis, we developed a systematic approach to enhance the accuracy and efficiency of species identification, reducing the time-consuming manual review process. The use of advanced ML tools such as the iNaturalist VisionAPI allowed us to classify user-submitted images into high, medium, or low confidence categories, with an optimal threshold of 93.61 identified through rigorous statistical validation, including ANOVA, Tukey's HSD, and ROC curve analysis. The results highlight that combined scores are a reliable indicator of

classification confidence, with “Both iMap and iNat correct” consistently achieving the highest scores. Spatial components such as geo scores further complemented these results by analyzing species proximity and autocorrelation, reinforcing confidence in accurate identifications. The analysis achieved an overall classification accuracy of 95.25%, confirming the robustness of the developed approach. However, misclassifications, especially in low-confidence categories, emphasize the need for continued improvements in dataset quality, image preprocessing, and ML algorithm refinement. The research also underscores the critical role of high-quality images for effective species identification. Close-up shots, focused targets, simple backgrounds, and detailed features such as leaf structures or bark textures significantly improve recognition accuracy. Conversely, poor-quality images with cluttered environments or ambiguous subjects present challenges to ML algorithms, limiting their performance.

In conclusion, the integration of ML tools into invasive species management represents a transformative step forward, bridging the gap between manual verification and automated decision-making. By streamlining the identification process and ensuring high accuracy, this approach empowers natural resource managers, taxonomic experts, and decision-makers to act more efficiently. Moving forward, addressing data quality issues and enhancing model performance will be essential to further optimize the system's reliability and scalability for real-world applications.

References

Allen, J.M., Bradley, B.A., 2016. Out of the weeds? Reduced plant invasion risk with climate change in the continental United States. Biological Conservation 203, 306–312.

Ashqar, B.A.M., Abu-Naser, S.S., 2019. Identifying Images of Invasive Hydrangea Using Pre-Trained Deep Convolutional Neural Networks. IJCA 12, 15–28. https://doi.org/10.33832/ijca.2019.12.4.02

Augustin, N.H., Mugglestone, M.A., Buckland, S.T., 1996. An Autologistic Model for the Spatial Distribution of Wildlife. Journal of Applied Ecology 33, 339–347. https://doi.org/10.2307/2404755

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., Soberón, J., Villalobos, F., 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. Ecological Modelling 222, 1810–1819. https://doi.org/10.1016/j.ecolmodel.2011.02.011

Borges Oliveira, D.A., Ribeiro Pereira, L.G., Bresolin, T., Pontes Ferreira, R.E., Reboucas Dorea, J.R., 2021. A review of deep learning algorithms for computer vision systems in livestock. Livestock Science 253, 104700. https://doi.org/10.1016/j.livsci.2021.104700

Bradski, G., 2000. The openCV library. Dr. Dobb’s Journal: Software Tools for the Professional Programmer 25, 120–123.

Chandra, S., 2020. Learning to segment images without manually segmented training data [WWW Document]. Amazon Science. URL https://www.amazon.science/blog/learning-to-segment-images-without-manually-segmented-training-data (accessed 9.10.23).

Colautti, R.I., MacIsaac, H.J., 2004. A neutral terminology to define ‘invasive’ species. Diversity and Distributions 10, 135–141. https://doi.org/10.1111/j.1366-9516.2004.00061.x

Dehnen-Schmutz, K., 2011. Determining non-invasiveness in ornamental plants to build green lists. Journal of Applied Ecology 48, 1374–1380. https://doi.org/10.1111/j.1365-2664.2011.02061.x
Devin, S., Beisel, J.-N., 2007. Biological and ecological characteristics of invasive species: a gammarid study. Biol Invasions 9, 13–24. https://doi.org/10.1007/s10530-006-9001-0
Dickinson, J.L., Shirk, J., Bonter, D., Bonney, R., Crain, R.L., Martin, J., Phillips, T., Purcell, K., 2012. The current state of citizen science as a tool for ecological research and public engagement. Frontiers in Ecology and the Environment 10, 291–297. https://doi.org/10.1890/110236
Drake, S.J., Weltzin, J.F., Parr, P.D., 2003. Assessment of Non-Native Invasive Plant Species on the United States Department of Energy Oak Ridge National Environmental Research Park. Castanea 68, 15–30.
Elias, N., 2021. A Novel Method for Automated Identification and Prediction of Invasive Species Using Deep Learning, in: 2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). Presented at the 2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), IEEE, Vancouver, BC, Canada, pp. 1–5. https://doi.org/10.1109/IEMCON53756.2021.9623087
Favorskaya, M., Pakhirka, A., 2019. Animal species recognition in the wildlife based on muzzle and shape features using joint CNN. Procedia Computer Science 159, 933–942. https://doi.org/10.1016/j.procs.2019.09.260
Finley, D., Dovciak, M., Dean, J., 2023. A data driven method for prioritizing invasive species to aid policy and management. Biol Invasions. https://doi.org/10.1007/s10530-023-03041-3
Fu H., Chi Z., Chang J., Fu C.X., 2004. Extraction of leaf vein features based on artificial neural network - studies on the living plant identification I.
Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep learning, volume 1. MIT press Cambridge.
Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T., 2018. Recent advances in convolutional neural networks. Pattern Recognition 77, 354–377. https://doi.org/10.1016/j.patcog.2017.10.013
Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8, 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.
He, P., Huang, L., 2008. Feature extraction and recognition of plant leaf. Journal of Agricultural Mechanization Research 6, 52.
Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.
Jensen, T., Seerup Hass, F., Seam Akbar, M., Holm Petersen, P., Jokar Arsanjani, J., 2020. Employing Machine Learning for Detection of Invasive Species using Sentinel-2 and AVIRIS Data: The Case of Kudzu in the United States. Sustainability 12, 3544. https://doi.org/10.3390/su12093544
Jewitt, A., Antolos, E., Lutz, C., Dean, J., 2021. Targeted Species Projects for Volunteers to Increase Early Detection Capacity: The Water Chestnut Mapping Challenge. Natural Areas Journal 41. https://doi.org/10.3375/043.041.0306
Johnson, B.A., Mader, A.D., Dasgupta, R., Kumar, P., 2020. Citizen science and invasive alien species: An analysis of citizen science initiatives using information and communications technology (ICT) to collect invasive alien species observations. Global Ecology and Conservation 21, e00812. https://doi.org/10.1016/j.gecco.2019.e00812
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25.
Kühl, H.S., Burghardt, T., 2013. Animal biometrics: quantifying and detecting phenotypic appearance. Trends Ecol Evol 28, 432–441. https://doi.org/10.1016/j.tree.2013.02.013

Kulhanek, S.A., Leung, B., Ricciardi, A., 2011. Using ecological niche models to predict the abundance and impact of invasive species: application to the common carp. Ecological Applications 21, 203–213. https://doi.org/10.1890/09-1639.1
Kumar, S., Tiwari, S., Singh, S.K., 2016. Face Recognition of Cattle: Can it be Done? Proc. Natl. Acad. Sci., India, Sect. A Phys. Sci. 86, 137–148. https://doi.org/10.1007/s40010-016-0264-2
LeCun, Y., 2015. LeNet-5, convolutional neural networks. URL: http://yann. lecun. com/exdb/lenet 20, 14.
Li, Y., Zhu, Q., Cao, Y., Wang, C., 2005. A Leaf Vein Extraction Method Based On Snakes Technique, in: 2005 International Conference on Neural Networks and Brain. Presented at the 2005 International Conference on Neural Networks and Brain, pp. 885–888. https://doi.org/10.1109/ICNNB.2005.1614763
Lin, M., Chen, Q., Yan, S., 2013. Network in network. arXiv preprint arXiv:1312.4400.
Lodge, D.M., Williams, S., MacIsaac, H.J., Hayes, K.R., Leung, B., Reichard, S., Mack, R.N., Moyle, P.B., Smith, M., Andow, D.A., 2006. Biological invasions: recommendations for US policy and management. Ecological applications 16, 2035–2054.
Martinez, B., Reaser, J.K., Dehgan, A., Zamft, B., Baisch, D., McCormick, C., Giordano, A.J., Aicher, R., Selbe, S., 2020. Technology innovation: advancing capacities for the early detection of and rapid response to invasive species. Biol Invasions 22, 75–100. https://doi.org/10.1007/s10530-019-02146-y
Mohammadi, M., Al-Fuqaha, A., Sorour, S., Guizani, M., 2018. Deep learning for IoT big data and streaming analytics: A survey. IEEE Communications Surveys & Tutorials 20, 2923–2960.
Nininahazwe, F., Théau, J., Marc Antoine, G., Varin, M., 2023. Mapping invasive alien plant species with very high spatial resolution and multi-date satellite imagery using object-based and machine learning techniques: A comparative study. GIScience & Remote Sensing 60, 2190203. https://doi.org/10.1080/15481603.2023.2190203
Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J., 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences 115, E5716–E5725. https://doi.org/10.1073/pnas.1719367115
Noviyanto, A., Arymurthy, A.M., 2013. Beef cattle identification based on muzzle pattern using a matching refinement technique in the SIFT method. Computers and electronics in agriculture 99, 77–84.
Noviyanto, A., Arymurthy, A.M., 2012. Automatic cattle identification based on muzzle photo using speed-up robust features approach, in: Proceedings of the 3rd European Conference of Computer Science, ECCS. p. 114.
Otter, J., Mayer, S., Tomaszewski, C.A., 2021. Swipe Right: a Comparison of Accuracy of Plant Identification Apps for Toxic Plants. J. Med. Toxicol. 17, 42–47. https://doi.org/10.1007/s13181-020-00803-6
Pimentel, D., Lach, L., Zuniga, R., Morrison, D., 2000. Environmental and Economic Costs of Nonindigenous Species in the United States. bisi 50, 53–65. https://doi.org/10.1641/0006-3568(2000)050[0053:EAECON]2.3.CO;2
Ravoor, P.C., Sudarshan, T.S.B., 2020. Deep Learning Methods for Multi-Species Animal Re-identification and Tracking – a Survey. Computer Science Review 38, 100289. https://doi.org/10.1016/j.cosrev.2020.100289
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y
Rzanny, M., Bebber, A., Wittich, H.C., Fritz, A., Boho, D., Mäder, P., Wäldchen, J., 2024. More than rapid identification—Free plant identification apps can also be highly accurate. People and Nature 6, 2178–2181. https://doi.org/10.1002/pan3.10676
Sawaya, K.E., Olmanson, L.G., Heinert, N.J., Brezonik, P.L., Bauer, M.E., 2003. Extending satellite remote sensing to local scales: land and water resource monitoring using high-resolution imagery. Remote Sensing of Environment, IKONOS Fine Spatial Resolution Land Observation 88, 144–156. https://doi.org/10.1016/j.rse.2003.04.006

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
Simpson, A., Eyler, M.C., Guala, G., Cannister, M.J., Kozlowsky, N., Libby, R., Sellers, E.A., 2018. A comprehensive list of non-native species established in three major regions of the United States: Version 3.0. https://doi.org/10.5066/P9E5K160
Söderkvist, O., 2001. Computer Vision Classification of Leaves from Swedish Trees.
Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information processing & management 45, 427–437.
Srivastava, V., Lafond, V., Griess, V.C., 2019. Species distribution models (SDM): applications, benefits and challenges in invasive species management. CABI Reviews 2019, 1–13. https://doi.org/10.1079/PAVSNNR201914020
Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S., 2009. eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation 142, 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006
Sun, Y., Liu, Y., Wang, G., Zhang, H., 2017. Deep Learning for Plant Identification in Natural Environment. Computational Intelligence and Neuroscience 2017, 1–6. https://doi.org/10.1155/2017/7361042
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9.
Terry, J.C.D., Roy, H.E., August, T.A., 2020. Thinking like a naturalist: Enhancing computer vision of citizen science images by harnessing contextual data. Methods in Ecology and Evolution 11, 303–315.
Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., Mac Aodha, O., 2021a. Benchmarking representation learning for natural world image collections, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12884–12893.
Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., Mac Aodha, O., 2021b. Benchmarking representation learning for natural world image collections, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12884–12893.
Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S., 2018. The INaturalist Species Classification and Detection Dataset. Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8769–8778.
Willi, M., Pitman, R.T., Cardoso, A.W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., Fortson, L., 2019. Identifying animal species in camera trap images using deep learning and citizen science. Methods Ecol Evol 10, 80–91. https://doi.org/10.1111/2041-210X.13099
Williams, J.N., Seo, C., Thorne, J., Nelson, J.K., Erwin, S., O'Brien, J.M., Schwartz, M.W., 2009. Using species distribution models to predict new occurrences for rare plants. Diversity and Distributions 15, 565–576. https://doi.org/10.1111/j.1472-4642.2009.00567.x
Yang, C., Fang, L., Yu, Q., Wei, H., 2022. A Learning Robust and Discriminative Shape Descriptor for Plant Species Identification. IEEE/ACM Trans. Comput. Biol. and Bioinf. 1–1. https://doi.org/10.1109/TCBB.2022.3148463
Zhang, T.Y., Suen, C.Y., 1984. A fast parallel algorithm for thinning digital patterns. Communications of the ACM 27, 236–239.