Problem Chosen	2021	Team Control Number
C	MCM/ICM Summary Sheet	2108654

Prediction of Asian Giant Hornets spread and investigation strategies for public reports

Summary

As an increasing number of the Asian Giant Hornets are discovered in Washington State, local species have been negatively affected by these voracious predators. Thus, the government decides to gather public reports to locate these agricultural pests. This article aims to build a prediction model to identify mistaken public reports and help the government prioritize the investigation of the sighting reports.

First, we build a Spread Prediction Model to predict the spread of the Asian Giant Hornets, which consists of the Population Prediction Module and Spatial Distribution Prediction Module. Based on the Logistic function, the Population Prediction Module predicts the population changes of the Asian Giant Hornets over the next few generations. The Spatial Distribution Prediction Module uses the Gaussian kernel to predict the spatial distribution of this pest. As a result, the spread of the pest over time can be predicted approximately.

Then, we construct a fine-grained visual categorization network to solve the image recognition problem, which uses the Destruction and Construction Learning (DCL) method to increase the precision of fine recognition. The network output is then combined with the temporal and spatial information from the report records as the input of some classification models such as Logistic Regression, Support Vector Machine, and Neural Network, which are used to determine whether the report is mistaken or not. The result shows that our model has a 98% accuracy and 80% recall approximately. The image recognition model and classification models constitute our Geo-Season-Vision Classification Model.

Next, the Geo-Season-Vision Classification Model is applied to predict the unprocessed public reports. The result shows that only one report is not mistaken, which can guide the government's further investigation. The future reports can be processed in the same way as well.

Additionally, this article discusses the model updating process when new public reports are available. When the outlier appears, we update the model by putting new reports into the training set and retrain our model. We also explain how we can update the model for different prevention and control strategies.

Eventually, this article proposes some potential criteria to examine if the Asian Giant Hornet has been eradicated and some suggestions to the government.

Keywords: Fine-grained Image Recognition; Neural Network; Machine Learning; Correlation Analysis

Contents

1 Introduction 2

1.1 Problem Background 2

1.2 Restatement of the Problem 2

1.3 Our Approach 2

2 General Assumptions and Notations 3

3 Spread Prediction Model 4

3.1 Population Prediction 5

3.2 Spatial Distribution Prediction 5

3.3 Analysis of the Result 6

4 Geo-Season-Vision Classification Model 6

4.1 Image Recognition Model 6

4.2 Classification Model 11

5 Prediction and Priority 15

6 Updating Model 16

6.1 Basic Ideas 16

6.2 Computer Simulation 16

6.3 Model Update for Different Control Strategies 17

7 Eradicating Evaluation 17

8 Sensitivity and Robust Analysis 18

9 Conclusion 19

9.1 **Strengths and Weaknesses** 19

9.2 Future Works 20

Appendices 23

1 Introduction

1.1 Problem Background

Asian Giant Hornets (Vespa mandarinia) is the largest species of hornet globally, and it can destroy a whole colony of European honeybees in a short time. The life cycle of this hornet is similar to many other wasps. The fertilized queens emerge in the spring and begin a new colony. In the fall, they leave the nest and will spend the winter in the soil.

(a) (b) (c)

Figure 1: **Asian Giant Hornets (Vespa mandarinia)**. Asian Giant Hornets (Vespa mandarinia) is the largest species of hornet globally, and it can destroy a whole colony of European honeybees in a short time. *Image source: Washington State Department of Agriculture*.

Since September 2019, several confirmed sightings of the pest have occurred in Washington State, and if this pest is not controlled, it will cause a severe impact on local honeybee populations. Therefore, The State of Washington has created helplines and a website for people to report these hornets' sightings. Based on these reports, the state should prioritize its limited resources to follow-up with additional investigation.

1.2 Restatement of the Problem

- Address and discuss whether or not the spread of this pest over time can be predicted, and with what level of precision.
- Use only the data set file provided and (possibly) the image files provided to create, analyze, and discuss a model that predicts the likelihood of mistaken classification.
- Use your model to discuss how your classification analyses lead to prioritizing investigation of the reports most likely to be positive sightings.
- Address how you could update your model given additional new reports over time, and how often the updates should occur.
- Address how you could determine that the pest has been eradicated in Washington State.

1.3 Our Approach

The topic requires us to interpret the data provided by the public reports and give government agencies strategies on how to prioritize the public reports for additional investigation. Our work mainly includes the following:

- Based on the data of confirmed sighting reports, a prediction model of the spread of Asian Giant Hornets over time is established;
- Construct the image recognition model;
- Combining the image information with temporal and spatial information, construct the classification model;
- Based on the classification model, the unprocessed reports are predicted, and the priority analysis is given;
- Based on the spread of the pest, This article gives reasonable tactics for model update and discusses the evidence for the eradication of this pest.

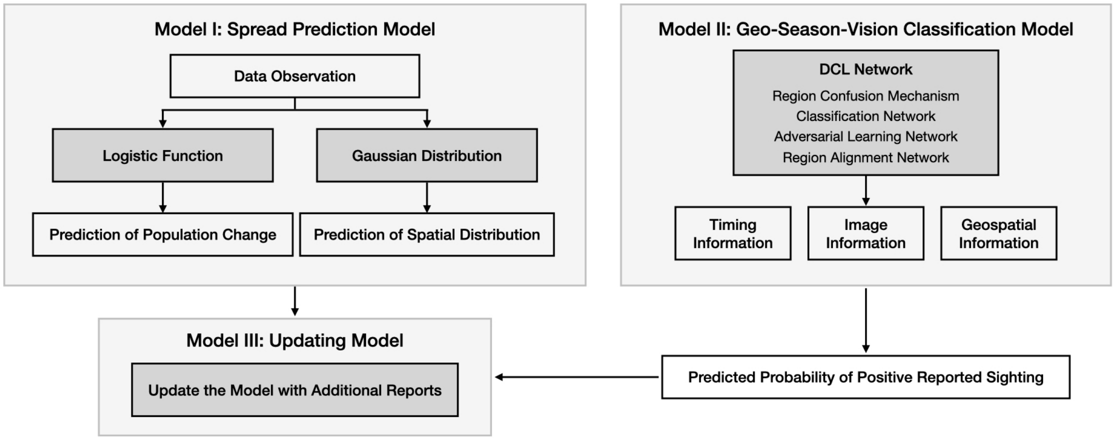

Model I: Spread Prediction Model
Data Observation
Logistic Function
Gaussian Distribution
Prediction of Population Change
Prediction of Spatial Distribution
Model II: Geo-Season-Vision Classification Model
DCL Network
Region Confusion Mechanism
Classification Network
Adversarial Learning Network
Region Alignment Network
Timing Information
Image Information
Geospatial Information
Model III: Updating Model
Update the Model with Additional Reports
Predicted Probability of Positive Reported Sighting

Figure 2: **Model Overview**. Our model consists of three sub-models: the Spread Prediction Model, the Geo-Season-Vision Classification Model, and the Updating Model.

2 General Assumptions and Notations

To simplify the problem, we make the following basic assumptions, each of which is appropriately justified.

- **Assumption 1:** The population changes of Asian Giant Hornets follow the general law of population growth.
  $\hookrightarrow$ **Justification:** Since the life cycle of Asian Giant Hornets is similar to many other wasps, so it is reasonable to assume that it follows the general law of population growth.
- **Assumption 2:** Asian Giant Hornets migrates with equal probability in all directions.
  $\hookrightarrow$ **Justification:** Here, we ignore the actual geography of that place, as we have little data about the trend of migration.

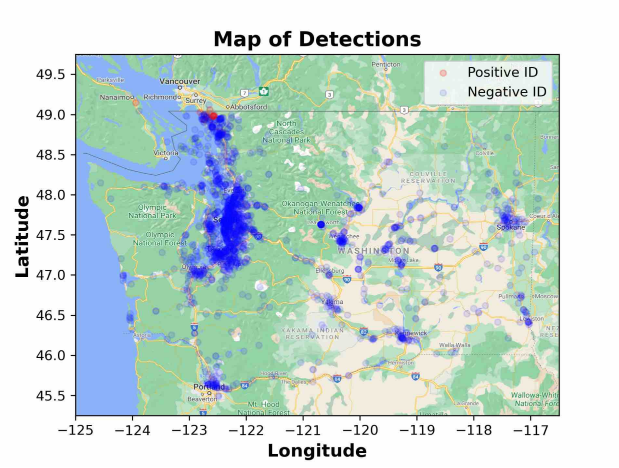


Figure 3: **Positive and Negative Reports Distributions**. According to the scatter map, we have the intuitive impression that positive samples are in one cluster.

- **Assumption 3:** Assume all the positive records are distributed around the beehives.
  ↪ **Justification:** Asian Giant Hornets are social creatures, and the positive reports show a clustering trend in Figure 3.
- **Assumption 4:** Assume the data of the report is accurate.
  ↪ **Justification:** For simplicity but without loss of generality, accurate data can ensure that our models are useful.

Table 1: **Notations**

Symbol	**Description**
N	Population
t	The time from now
r	The growth rate of population
N_m	The maximum amount that the environment can hold
x	The features of a training sample
y	The label of a training sample
θ	The parameter of learning model
$x^{(i)}$	The features of i^{th} training sample
$y^{(i)}$	The label of i^{th} training sample

3 Spread Prediction Model

To predict the spread of Asian Giant Hornets over time, we need to predict population changes and spatial changes. As we have little data, we can not observe the spread trend over time. Therefore we use the Logistic Model to predict the population changes and use Gaussian Kernel to predict the Spatial changes.

3.1 Population Prediction

According to assumption 1, we think the population changes of Asian Giant Hornets follow the general law of population growth. And as this pest is inherited from generation to generation, we use the generation to represent time, e.g. 0^{th} generation for time t_0, 1^{th} generation for time t_1. Then we can construct Logistic Model as

$$\begin{cases} \frac{\partial N}{\partial t} = r(1 - \frac{N}{N_m})N, \\ N(t_0) = N_0, \end{cases} \tag{1}$$

Since this pest is just beginning to invade this region, and we only predict the first five generations, the Logistic Model can be simplified as an exponential growth model. So the population can be expressed as

$$N(t) = e^{st} \tag{2}$$

Using the data of generation 0,1,2, we can calculate the parameter s, so the population can be expressed as

$$N(t) = e^{1.1160t} \tag{3}$$

Then the prediction of the first five generations is shown in Figure 3.

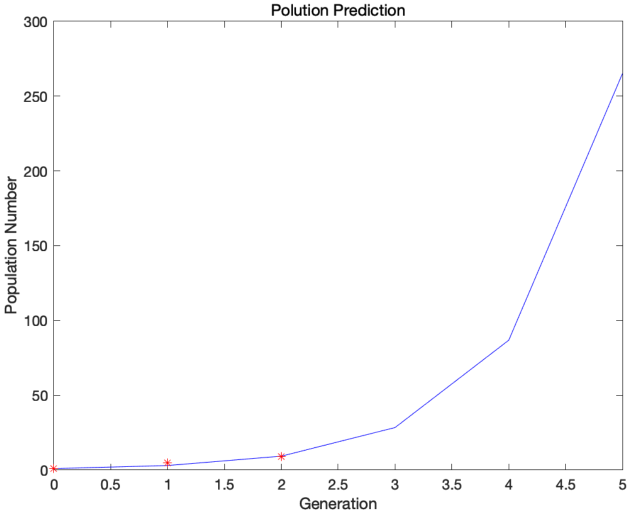


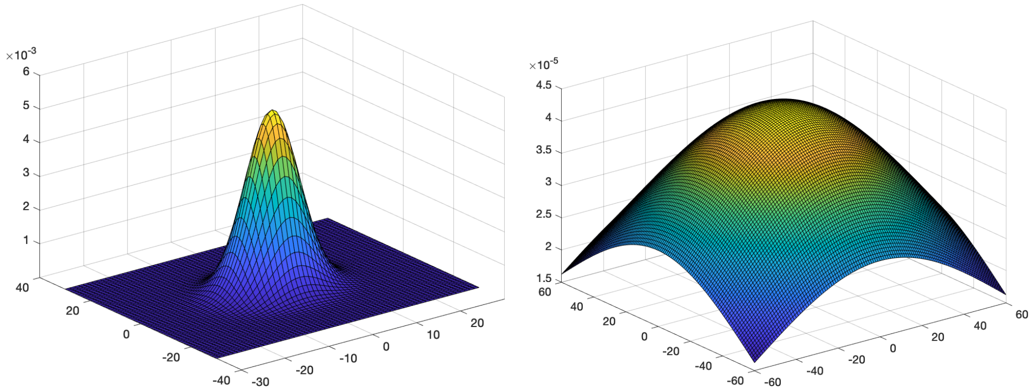
Figure 4: **Prediction of Population**. As generation increases, the population increases exponentially.

3.2 Spatial Distribution Prediction

According to the assumption 2, Asian Giant Hornets are equally likely to move in all directions. From prior experience, the father away from the center, the lower the probability of arrival of this pest. Combining with the information that these pests have a range estimated at 30km for establishing their nest, we can use a Gaussian kernel with a radius of 30km to express their spatial distribution. The probability of distribution $f(a,b)$ at position (a,b) can be expressed as

$$\begin{cases} f(a,b) = \frac{1}{2\pi\sigma^2} e^{\frac{-(a-a_0)^2-(b-b_0)^2}{2\sigma^2}}, \\ \sigma^2 = 30, \end{cases} \tag{4}$$

The spread of the 2^{th} and 3^{th} generation is shown in Figure 4.



(a) 2^{th}generation (b) 3^{th}generation

Figure 5: **Spatial Distribution**. 2^{th} generation with radius of 30km, 3^{th} generation with radius of 60km. As the generation increases, the Gaussian kernel is flatter, which means the probability of staying away from the nest is increasing.

3.3 Analysis of the Result

For the Population changes model, we can only predict changes in the first five generations; the fitting accuracy of historical data is 95.13%. However, the model is limited in the period it can predict. Since the population growth rate is slowing, we cannot make long-term population predictions. For Spatial distribution Prediction, we assume that Asian Giant Hornets migrate with equal probability in all directions. The conclusion is that the farther away from the current nest, the less the pest is distributed. If Assumption 2 is not satisfied, we can not predict the Asian giant hornet's spatial distribution.

4 Geo-Season-Vision Classification Model

In this section, we proposed Geo-Season-Vision Model, based on geographic information, season (timing) information, and vision information.

4.1 Image Recognition Model

The devil is in the details. Distinguishing the Asian Giant Hornets from similar hornets heavily relying on fine-grained object parts, and this process also requires some professional knowledge. This task is called 'fine-grained image recognition' in Computer Vision, opposed to generic object recognition. For fine-grained image recognition, local details are much more deterministic than global structure.

Destruction and Construction Learning (DCL) [1] is a state-of-the-art method for fine-grained image recognition proposed by JDAI in 2019. Based on ResNet-50, 'destruction learning' and 'construction learning' methods are assembled into the model, improves the performance to a higher level. Besides, the DCL method has a clear advantage that the model does not require extra supervision information except category labels in the training process, saving much time for experts to annotate the dataset, which is suitable for emergency cases like biological invasions or infectious disease outbreaks.

4.1.1 Network Structure

The network consists of four subtle parts: Region Confusion Mechanism, Classification Network and Adversarial Learning Network for destruction learning, and Region Alignment Network for construction learning. Figure 6 shows the structure of the whole network as well as a simplified pipeline.

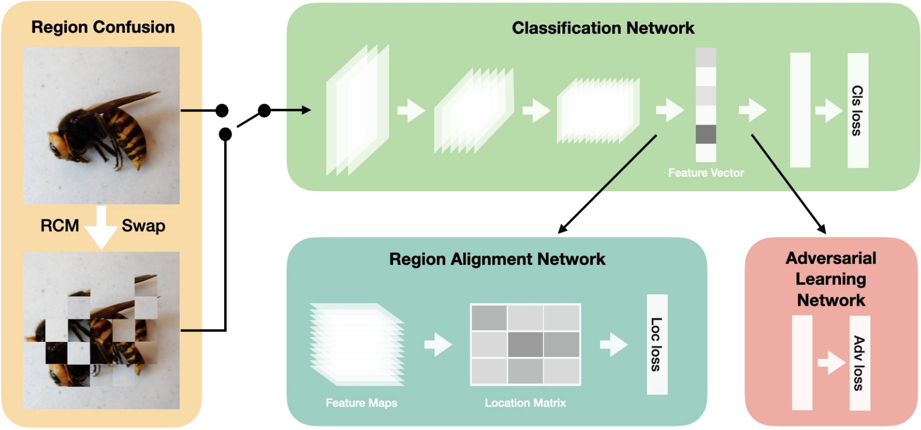


Figure 6: **Network structure**. The network consists of 4 parts: Region Confusion Mechanism, Classification Network and Adversarial Learning Network for destruction learning, and Region Alignment Network for construction learning.

Region Confusion Mechanism Region Confusion Mechanism (RCM) module shuffles the local regions of the input image. As a resemblance to reading, breaking a sentence into out of order words would require readers to concentrate on discriminative words and disregard irrelevant ones. Similarly, if local grid regions in an image are shuffled, the network must learn from discriminative region details for classification.

As shown in Figure 6, RCM aims to destroy the spatial distribution of local regions of the image. Given an input image I, we first uniformly partition the image into $N \times N$ regions $R_{i,j}$, where i and j are the horizontal and vertical index respectively with $1 \leq i, j \leq N$. For the sake of stability, we only swap the partitioned local region $R_{i,j}$ with its neighbors $R_{i-1,j}$, $R_{i+1,j}$, $R_{i,j-1}$, $R_{i,j+1}$. Therefore, the region at (i, j) in original image is swapped to new location:

$$\sigma(i, j) = (\sigma_j^{row}(i), \sigma_i^{col}(j)). \tag{5}$$

While RCM forces the network to focus on discriminative region details through shuffling subregions, it introduces some noise in the area near the borderline between two grids. The network may take these noise as some features and remember them, which is a disadvantage to our learning process.

Classification Network This is the primary classification network that classifies images into fine-grained categories. The original image I, its destructed version $\phi(I)$ and its ground truth one-vs-all label l are joined as $< I, \phi(I), l >$ for training. The Classification Network maps input image I into a probability distribution vector $C(I)$. The loss function of the classification network $\mathcal{L}_{cls}$ can be written as:

$$\mathcal{L}_{cls} = -\sum_{I \in \mathcal{I}} \boldsymbol{l} \cdot \log[C(I)C(\phi(I))], \tag{6}$$

where $\mathcal{I}$ is the training dataset. Since the global structure has been destroyed, the network has to find the discriminative regions and learn the subtle differences among categories to classify these randomly shuffled images.

Adversarial Learning Network To minimize the negative influence of noise introduced by RCM in equation 5, we apply an adversarial loss $\mathcal{L}_{adv}$ to distinguish original images from destructed ones. This module incepts feature vectors from the original image and destructed one separately, comparing the difference between the two feature vectors and calculating the loss. To minimize the adversarial loss, the two feature vectors must be as close as possible; in other words, the Classification Network must concentrate on the real deterministic details and ignore the noise between small grids to output cleaner features.

Inspired by the discriminator in Generative Adversarial Networks [3], the adversarial loss and classification loss work in an adversarial manner to

1. keep destructed-invariant patterns,
2. reject destructed-specific patterns between I and $\phi(I)$.

The loss of the adversarial discriminator network

$$\mathcal{L}_{adv} = -\sum_{I \in \mathcal{I}} \boldsymbol{d} \cdot \log[D(I)] + (\boldsymbol{1} - \boldsymbol{d}) \cdot \log[D(\phi(I))], \tag{7}$$

where $\boldsymbol{d} \in \{0, 1\}^2$ is a one-hot vector label indicating whether the image is destructed or not, $D(I) = \text{softmax}(\theta_{adv} C(I))$ is a discriminator judging whether an image I is destructed or not, and $\theta_{adv} \in \mathbb{R}^{d \times 2}$ is a linear mapping vector.

Region Alignment Network While we pay much attention to the image details, the object's overall spatial structure should also contribute to the classification work. To reserve and make use of the spatial information, we append the Region Alignment Network after the Classification Network to recover local regions' spatial layout. The region alignment loss $\mathcal{L}_{loc}$ is defined as the $L1$ distance between the predicted coordinates and original coordinates, which can be expressed as:

$$\mathcal{L}_{loc} = \sum_{I \in \mathcal{I}} \sum_{i=1}^{N} \sum_{j=1}^{N} \|M_{\sigma(i,j)}(\phi(I)) - [i\ j]^T\|_1 + \|M_{i,j}(I) - [i\ j]^T\|_1 \tag{8}$$

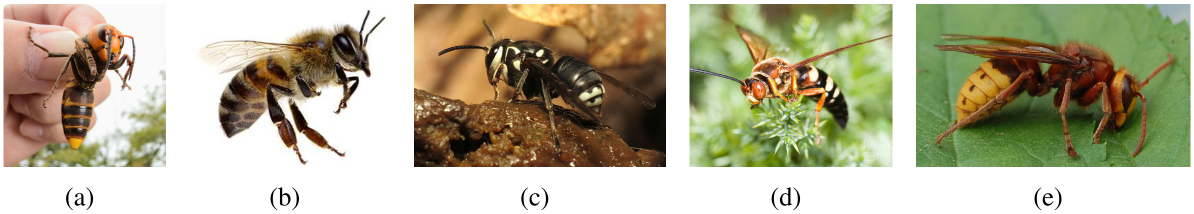
where $M(I)$ is the location matrix derived from the alignment CNN. The region construction loss helps detect the main objects in images and finds the correlation among sub-regions. In a word, the region construction loss can help the network to build a profound understanding of objects and the structure information, such as the shape of objects and semantic correlation among parts of the object.

Destruction and Construction Learning The classification, adversarial and region alignment losses are jointly trained in an end-to-end manner to minimize the total loss

$$\mathcal{L} = \alpha\mathcal{L}_{cls} + \beta\mathcal{L}_{adv} + \gamma\mathcal{L}_{loc}. \tag{9}$$

4.1.2 Experiments and Results

Training and Testing We trained a neural network aiming to find Asian Giant Hornets from these insects resembled with each other. Here we select Asian Giant Hornets and four similar insects (Figure 7) according to $2021MCM\_ProblemC\_Vespamandarinia.pdf$ to build our dataset as Table 2 shows.



(a) (b) (c) (d) (e)

Figure 7: **5 similar insects**. Asian Giant Hornets (a), Bees (b), Baldfaced hornets (c), Eastern cicada killers (d), and European hornets (e).

Table 2: **Dataset**.

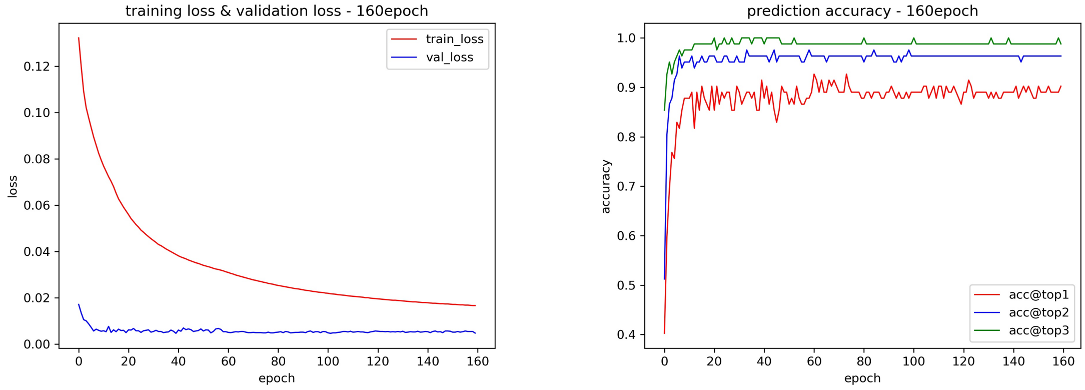
Class ID	Insect	Training Set Num	Testing Set Num
Class 1	Asian Giant Hornets	68	15
Class 2	Bees	70	20
Class 3	Baldfaced hornets	60	12
Class 4	Eastern cicada killers	74	18
Class 5	European hornets	65	17

We trained our model using our own dataset shown in Table 2, and the process is shown in Figure 8.

Table 3: **Classification Result of DCL Network**

Accuracy	Class 1	Class 2	Class 3	Class 4	Class 5
@ top 1	86.67%	91.67%	95.00%	88.89%	88.24%
@ top 3	100.00%	91.67%	95.00%	95.00%	1.0%

The prediction accuracy on the testing set is shown in Table 3. The network outputs five scores for each image, indicating the probability the image is likely to be classified into each category.



(a) Training loss & Validation loss

(b) Prediction Accuracy

Figure 8: **Prediction of Unprocessed samples**. As training goes, the training loss goes down and validation loss converges to a low level, and the prediction accuracy defined by acc@top1, acc@top2, acc@top3 converges to a high level.

Here accuracy@top 1 is the proportion that the ground truth label equals to the highest score label, and accuracy@top 3 is the proportion that the ground truth label is in the top-3 score labels. For Asian Giant Hornets, the network prediction accuracy reaches 86.67% for top 1 case and 100% for top 3 case, and the result is outstanding.

Application Then we can apply the model to real-world cases. Figure 9 shows the prediction results for all Positive and Negative cases in $2021MCMProblemC\_DataSet.xlsx$. Even though many images in the reported data are of poor quality, our model's accuracy reached 92.4%, and 70% positive cases are predicted correctly only use the top-1 score.

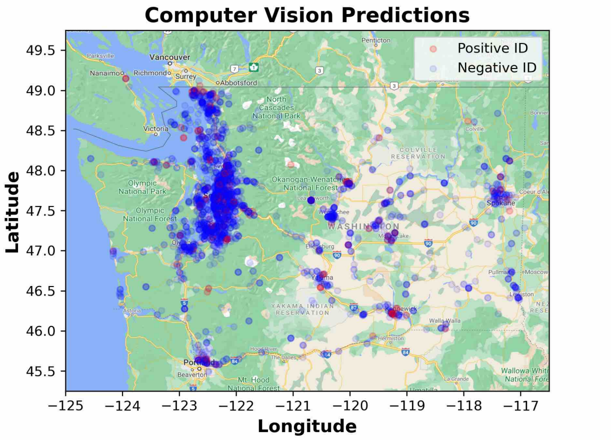


Figure 9: **Computer vision prediction results**. Pure vision based classifier filters about 92.4% negative cases, which is impressive.

4.2 Classification Model

To predict the likelihood of a mistaken report, we apply Logistic Regression, Support Vector Machine model with different kernels, and Neural Network to classify if a new report is a positive sample of Asian Giant Hornets. The detailed description of the models is shown in 4.2.2, 4.2.3, and 4.2.4, respectively. The classification models combine the image information from the Image Recognition Model, temporal information, and spatial information as the input features. The feature extraction process is explained in 4.2.1 in detail.

4.2.1 Data Processing

In the given dataset, we have multiple data categories. To construct a binary classification model, we only use positive reports and negative reports as training samples. After our analysis, we intuitively extract some significative features as our description of a sample. The features and their meanings are shown below.

day Integer, indicating the date difference between the detection date and the base date.
Notes: Choosing 1/1/2019 as the base date. For example, the value of 1/2/2019 is 1, and the value of 12/31/2018 is -1. If the detection date is blank, use 6/30/2020 as an average detection date.

winter Boolean, indicating whether the detection is in winter or not.
Notes: Since the Asian Giant Hornets are in dormancy in winter, a detection in winter might have less validity.

relative_latitude Float, indicating the relative latitude between the latitude of detection and the latitude of known closest positive sample.

relative_longitude Float, indicating the relative longitude between the longitude of detection and the longitude of the known closest positive sample.

s0-s4 Float, indicating the score of the image for five hornet species, respectively.
Notes: s0-Asian Giant Hornet, s1-Bee, s2-Baldfaced Hornet, s3-Eastern Cicada Killer, s4-European Hornet. The scores are results of the Image Recognition Model.

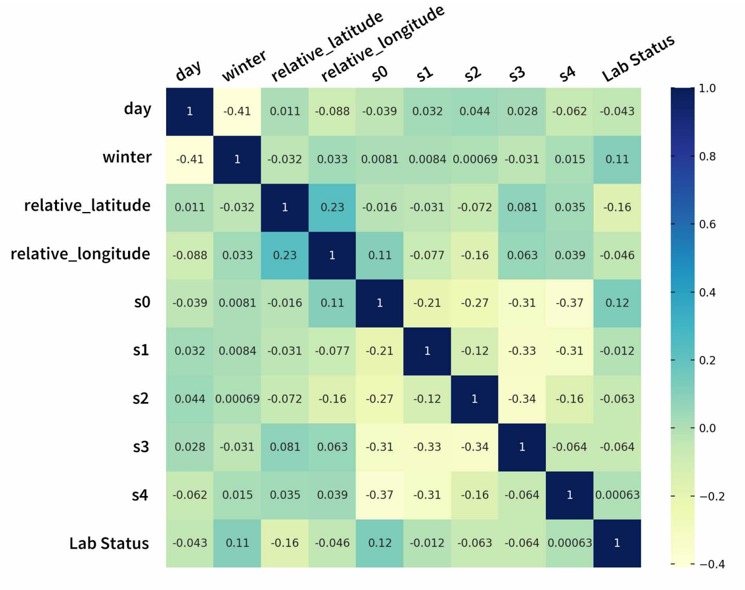
4.2.2 Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function ϕ(z) (shown in Figure 11) to model a binary dependent variable, although many more complex extensions exist. Mathematically, a binary logistic model has a dependent variable with two possible values, which is represented by an indicator variable, where the two values are labeled "0" and "1".

Our binary classification problem fits the setting of binary logistic model, so we choose it as one of our classification models. The logistic regression model is defined as follows.

First, take θ as the parameter of the logistic model, we can define the probability of each sample as:

$$P(y=1|x;\theta) = h_\theta(x) = \phi(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \tag{10}$$



	day	winter	relative_latitude	relative_longitude	s0	s1	s2	s3	s4	Lab Status
day	1	-0.41	0.011	-0.088	-0.039	0.032	0.044	0.028	-0.062	-0.043
winter	-0.41	1	-0.032	0.033	0.0081	0.0084	0.00069	-0.031	0.015	0.11
relative_latitude	0.011	-0.032	1	0.23	-0.016	-0.031	-0.072	0.081	0.035	-0.16
relative_longitude	-0.088	0.033	0.23	1	0.11	-0.077	-0.16	0.063	0.039	-0.046
s0	-0.039	0.0081	-0.016	0.11	1	-0.21	-0.27	-0.31	-0.37	0.12
s1	0.032	0.0084	-0.031	-0.077	-0.21	1	-0.12	-0.33	-0.31	-0.012
s2	0.044	0.00069	-0.072	-0.16	-0.27	-0.12	1	-0.34	-0.16	-0.063
s3	0.028	-0.031	0.081	0.063	-0.31	-0.33	-0.34	1	-0.064	-0.064
s4	-0.062	0.015	0.035	0.039	-0.37	-0.31	-0.16	-0.064	1	0.00063
Lab Status	-0.043	0.11	-0.16	-0.046	0.12	-0.012	-0.063	-0.064	0.00063	1

Figure 10: **Correlation analysis**. The figure illustrates that these features have a weak correlation. Some features like *winter*, *relative_latitude* and *s0* are correlated with the status of the report to some extent.

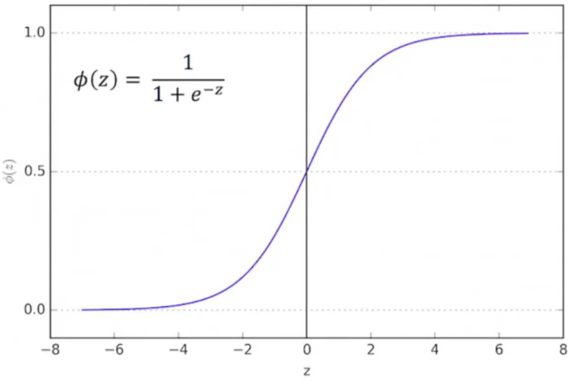


Figure 11: **Logistic Function.** This function is continuously differentiable and continuous, and it can map every value into [0, 1].

$$P(y=0|x;\theta) = 1 - h_\theta(x) = 1 - \phi(\theta^T x) = \frac{e^{-\theta^T x}}{1+e^{-\theta^T x}} \tag{11}$$

Then the log likelihood function can be expressed as:

$$l(\theta) = \sum_{i=1}^{m}(y^{(i)} ln(h(x^{(i)})) + (1-y^{i}) ln(1-h(x^{(i)}))) \tag{12}$$

Finally, We can apply gradient ascent or Newton method to update the parameter θ. The update process of stochastic gradient ascent can be expressed as:

$$\theta_j := \theta_j + \alpha(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)} \tag{13}$$

4.2.3 Support Vector Machine

Support Vector Machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis based on statistical learning frameworks or VC theory. An SVM maps training examples to points in space to maximize the width of the gap between the two categories. More formally, a support-vector machine constructs a hyperplane or set of hyperplanes in a high or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection.

Since our features contain several dimensions, including time, relative location, and image scores, we expect to use SVM to construct a hyperplane to classify the reports.

4.2.4 Neural Network

We construct a fully connected neural network with four hidden layers. The first three layers are activated by the "ReLU" function, and the fourth layer is activated by the "Sigmoid" function. Here are the expressions for the function.

$$\text{Relu function:} \quad f(x) = max(0, x) \tag{14}$$

$$\text{Sigmoid function:} \quad f(x) = \frac{1}{1 + e^{(-x)}} \tag{15}$$

To prevent overfitting, we randomly forgot some neurons at each layer by "dropout." Also, we solve the problem of unbalanced sample proportion by adjusting the weight of the loss function and adding the "bias" layer. The expressions are as follows.

$$min \sum_{i=1}^{n} w_i ||\hat{y}_i - y_i||, \qquad \hat{y}_i = f(x_1, x_2, \cdots, x_9) \tag{16}$$

$$w_i = \frac{\#Pos}{\#Neg}, \qquad y_i = 0, \tag{17}$$

$$w_i = \frac{\#Neg}{\#Pos}, \qquad y_i = 1. \tag{18}$$

In our test example, it is found that this method can effectively solve the problem that all sample points are misjudged as negative by the model. Figure 11 shows the neural network images, and Figure 12 shows the model training curves.

4.2.5 Experiment and Result

In our implementation, we construct five learning models, which are logistic regression, neural network, and SVM, with three different kernels(linear, rbf, and polynomial). Since the portion of positive samples is low, we apply 5-fold cross-validation to evaluate the effectiveness of the models by the average accuracy score, precision score, and recall score.
The results are shown in Table 4.

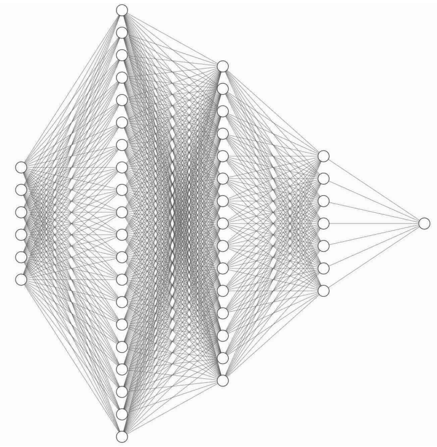
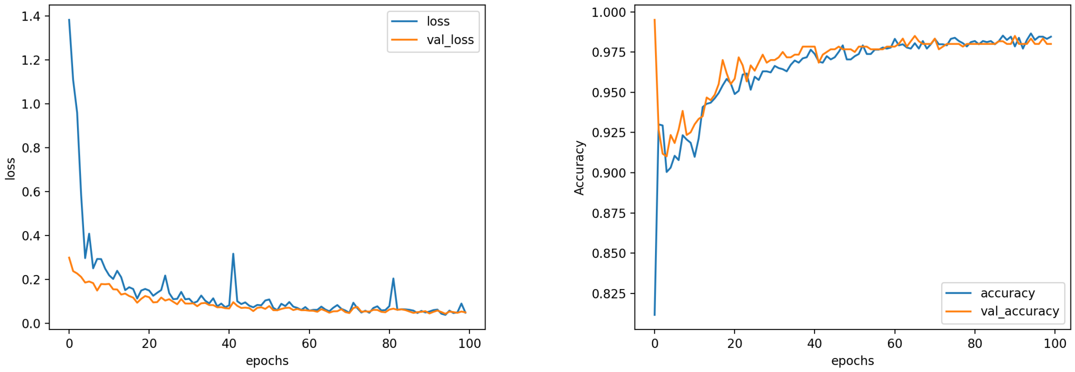


Figure 12: **Neural Network Structure.**



(a) Training loss & Validation loss

(b) Accuracy

Figure 13: **Performance Curves**. (a) shows the training process, as epochs increases, the loss decreases; (b) shows the accuracy of training process, Initially, the accuracy was high because the model predicted all the samples to be negative.

Accuracy The result table shows that the SVM model with RBF kernel or Polynomial kernel and the neural network has higher accuracy than the other three classifiers. It implies that the dataset fits in non-linear models better than linear models. Moreover, all the accuracy scores are higher than 95%, indicating that all classification models can make the correct prediction. However, due to the imbalance of the positive and the negative samples in the dataset, high accuracy might reflect some overfitting problems.

Precision It can be clearly shown that the precision scores of all models are low, which means that the classification models tend to misjudge some negative samples as positive ones. Precision indicates that the positive bound might be too loose to distinguish negative reports from positive ones. Comparatively, the SVM model with RBF kernel and Polynomial kernel has a higher precision score.

Table 4: **Result of Classification Models**

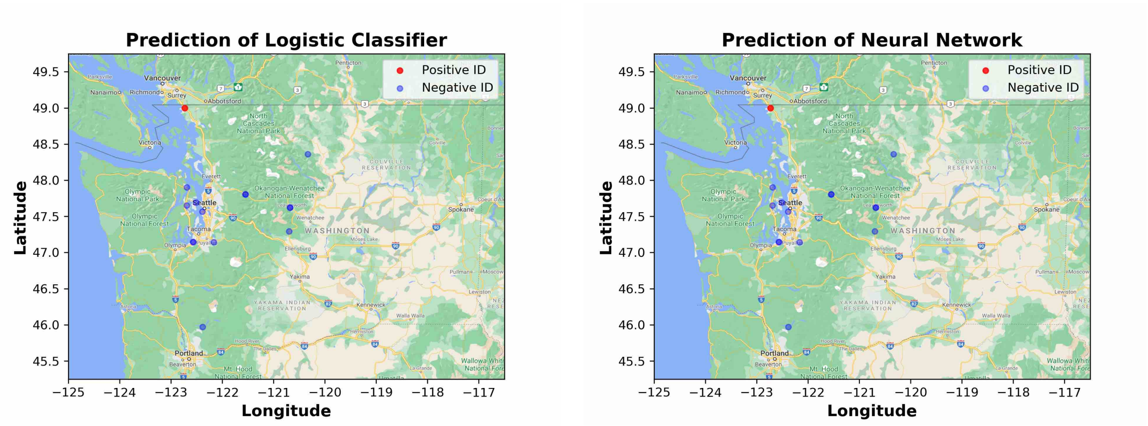
Model	Logistic	SVM-Linear Kernel	SVM-RBF Kernel	SVM-Polynomial Kernel	Neural Network
Accuracy	95.77%	96.59%	98.08%	98.32%	98.53%
Precision	17.17%	24.19%	37.59%	31.76%	21.33%
Recall	80.00%	80.00%	66.66%	66.66%	81.32%

Recall The result table illustrates that all models have high recall scores, especially the logistic regression model, linear SVM model, and neural network model, which means that these models can recognize most of the positive samples in the dataset.

5 Prediction and Priority

In our classification models, we extract the time(day, winter), location(latitude, longitude), and image(s0-s4) information as the features of a sample and fit the model with given positive and negative samples. To prioritize the investigation of the reports most likely to be positive samples, we can use our five classification models to predict the unprocessed reports.

Figure 14 shows the prediction of the logistic model and the neural network. The prediction of SVM models is the same as the results in the figure. Table 5 shows five unprocessed reports which have the highest priority to be investigated according to our model. The result shows that only one report is classified as a positive sample by all five classification models. The map shows that the predicted positive sample's location is close to the cluster of positive samples in the dataset, while other predicted negative samples are far away from northwestern Washington State, where positive samples gather in the dataset. Additionally, the detection date of this predicted positive sample is 8/29/2020, which is in summer. The temporal and spatial information indicates that the prediction is reasonable.



(a) Prediction of Logistic Model

(b) Prediction of Neural Network

Figure 14: **Prediction of Unprocessed samples**. Here we use our classification models to predict the unprocessed reports. The prediction of SVM models are the same as the results above.

Table 5: **Prediction Probability**

Priority	Global ID	Latitude	Longitude	Probability
1	26DDF8E2-DA0C-4F87-A65A-233115BAFCCD	48.9979°	−122.7299°	92.37%
2	C248B633-A567-4C44-BC3B-FC4D7C74EFD1	48.3615°	−120.3334°	19.27%
3	665C417D-A34B-4B58-973A-569EB01C4769	47.8040°	−121.5490°	7.98%
4	72BE3A9B-2F2C-4051-B93E-C2C8875EA63D	47.9046°	−122.6930°	0.53%
5	9BA7BDD9-01A5-4776-99B0-89FCE08CA53B	47.6901°	−122.5098°	0.09%

6 Updating Model

6.1 Basic Ideas

There are two ways of model updating: static updating and dynamic updating. Static updating means that the model is updated after a certain amount of reports has been obtained. Dynamic updating means that the model is updated every time a new report is obtained. Here, we use the method of static updating.

We only update the model with reports that have proven to be an Asian giant hornet. So the frequency of model update depends on the number of positive reports. Here we introduce the definition of concept drift [4]. Conceptual drift in predictive analysis and machine learning represents the phenomenon that feature variables' statistical properties change over time in unpredictable ways. As time goes on, the prediction accuracy of the model will decrease. First, we take five new reports and calculate the average value of each feature variable. Next, we calculate the difference between original feature variables and new feature variables. If it exceeds the threshold, we retrain the model with all the reports that we can obtain now. The threshold value is set by adjusting the value of feature variables and observing the model's robustness.

The algorithm is as follows.

Algorithm 1 Model Updating

```
Input: new reports
Calculate average feature values of the original reports, $d_0 = (\bar{x}_0, \bar{x}_1, \cdots, \bar{x}_9)$
If len(new reports) $\geq 5$    then
    Calculate average feature values of new reports, $d_1 = (\bar{y}_0, \bar{y}_1, \cdots, \bar{y}_9)$
    If $f(d_0 - d_1) \geq threshohold$    then
        Retrain the model
```

6.2 Computer Simulation

We predicted the spatial location of the next generation of Asian giant hornet through computer simulation, modified the value of feature variables according to the spatial information, and predicted through our classification model. The results are shown in Table 6.

The results show that if the Morphological characteristics of the Asian giant hornet do not change, then our model has a high performance of classifying the next generation of Asian giant hornets.

Table 6: **Result of Computer Simulation**

Model	SVM-RBF Kernel	Neural Network
Accuracy	97.73%	98.01%
Precision	35.42%	24.13%
Recall	67.32%	78.87%

6.3 Model Update for Different Control Strategies

Strategy 1: Mass investigation, strict control, and active removal of the hive—suppress Asian Giant Hornets' quantity to a very low level.
Strategy 2: Considering costs and benefits, allowing a certain number of Asian Giant Hornets to survive in the state without damaging the ecology to pursue dynamic balance.

When the number of Asian giant hornets is tiny, we take strategy 1. In this situation, we mainly improve the recall of the model, even if the accuracy and precision are reduced.

When the number of Asian giant hornets is already large, we take the strategy 2. In this situation, we mainly improve the model's accuracy, even if the recall is reduced.

7 Eradicating Evaluation

According to IUCN Red List, the extinction of a taxon is defined as:

A taxon is Extinct when there is no reasonable doubt that the last individual has died. A taxon is presumed Extinct when exhaustive surveys in known and/or expected habitat, at appropriate times (diurnal, seasonal, annual), throughout its historic range have failed to record an individual. Surveys should be over a time frame appropriate to the taxon's life cycle and life form. [2]

So we can check the eradication of the Asian Giant Hornet from temporal dimension and spatial dimension.

- **Temporal Criterion**: The prediction probability of the reports by the classification models are no more than 30% for a year.

  Explanation: Since the Asian Giant Hornet's active period is from spring to autumn, we cannot confirm that it has been eradicated until there are no suspicious positive reports when a new spring is coming. Also, due to the insensitivity of the classification model, we reduce the critical probability to 30% so that the criterion will be tighter.

- **Spatial Criterion**: There is no newly confirmed positive sample near the previous positive samples.

  Explanation: Since the spread of the Asian Giant Hornet must be spatially continuous, we cannot confirm that it has been eradicated until there is no occurrence around their previous habitats.

8 Sensitivity and Robust Analysis

Sensitivity analysis of w

One of our major challenges is to overcome the influence of the unbalanced dataset. Here we introduced a weight factor w:

$$\begin{aligned} w_i &= \frac{\#Pos}{\#Neg}, \qquad y_i = 0, \\ w_i &= \frac{\#Neg}{\#Pos}, \qquad y_i = 1. \end{aligned}$$

And the loss function

$$\mathcal{L} = \sum w_i \|\hat{y}_i - y_i\|$$

. This setting ensures the stability of the model in the following three cases:

- $\#Pos \gg \#Neg$: w_i is large when $y_i = 0$, $\mathcal{L} \approx \mathcal{L}_{Neg}$,
- $\#Pos \approx \#Neg$: $\mathcal{L}_{Pos} \approx \mathcal{L}_{Neg}, \mathcal{L} = \mathcal{L}_{Pos} + \mathcal{L}_{Neg}$,
- $\#Pos \ll \#Neg$: w_i is large when $y_i = 1$, $\mathcal{L} \approx \mathcal{L}_{Pos}$.

In other words, the prediction is not sensitive to the data proportion.

Sensitivity analysis of t

In the Asian Giant Hornet population prediction model of the Problem 1, We take years as the time unit for prediction. However in fact, take years as the unit is a very imprecise approximation. Therefore we study the sensitivity of populations to time. The relationship between the populations and time is approximated by first-order difference,

$$\frac{\partial x}{\partial t} \approx \frac{x(t+\Delta t) - x(t)}{\Delta t} \tag{19}$$

The calculation results are shown in Figure 15.

It is indicated that the populations are susceptible to time, reflecting the rapid growth of the population over time. However, it is essential to note that high sensitivity makes our predictions inaccurate.

Robust Analysis of DCL Model

The prediction accuracy on our testing set and given report attached image dataset is shown in Table 6. Even though many pictures in the report dataset lack accuracy, having problems like blurry, no feature parts of the creatures, and too few pictures to be judged, the DCL model's prediction remains a high accuracy, indicating our model is robust enough.

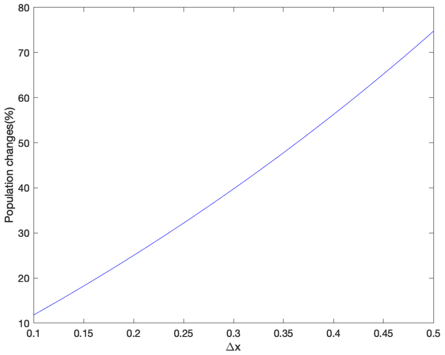


Figure 15: **Sensitivity analysis of Δt.** It shows that when we increase the error with respect to time, the error of population size increases exponentially.

Table 7: **Prediction Results on Different Datasets**.

	Testing set	Report Set
Accuracy	97.33%	92.42%
Recall	86.67%	70.00%

9 Conclusion

9.1 Strengths and Weaknesses

Strengths

- **High prediction accuracy**
  For the Asian Giant Hornet image recognition is very accurate, and the classification model can have more than 98% accuracy.

- **Easy to train, rapid to deploy**
  Our model does not require extra supervision information except category labels, which saves much time for experts to annotate the dataset. This character makes the model is suitable for emergency cases like biological invasions or infectious disease outbreaks.

- **The model can be widely used**
  Image recognition model can be used to identify other animals. Our entire model can be applied to the detection and control of COVID-19.

Weaknesses

- **Low prediction precision**
  The classification models have high probability to mistake some negative samples as positive ones.

- **Notes and comments are ignored**
  The notes and comments information are ignored in feature extraction process. Instead, we only focus on the time, location and image information.

9.2 Future Works

- **Improve prediction precision**
  The precision score can be improved by modulating model parameters. Other models can also be applied to achieve a better performance.

- **Add notes and comments information**
  The words in notes and comments can be extracted as new features of training samples, which may improve the effectiveness of the model.

Memorandum

To: Washington State Department of Agriculture

From: Team #2108654

Subject: Asian Giant Hornets Research

Date: June 22, 2021

Analysis of Public Reports

The public's report content includes time information, spatial information, text description, and attached images or videos. In terms of time, Asian Giant Hornets are mainly active in summer and enter hibernation in winter, which is not easy to be found by the public. In terms of space, the Asian Giant Hornets are centred on the honeycomb and move within a certain radius. Timing (seasonal) and spatial information is useful data for judging whether a sample is positive. However, the report's text description is too vague and lacks an accurate description of the characteristics of hornets, so the text content is unusable. The usability of image information is much higher than that of text information, but many pictures also lack accuracy. Their main problems are blurry images, no feature parts of the creatures, and too few pictures to be judged.

Based on the above data characteristics, we dropped the text description, adopted the time, space, and image characteristics, used computer vision and machine learning related technologies, and tried to judge the samples. Our model achieves an accuracy percentage greater than 95% over the whole report dataset. Generally, the performance of this model is surprising and has practical application value.

Analysis of the Current Situation

At present, the sighting reports of Asian Giant Hornets verified to be positive sightings are concentrated in about 1,000 square kilometers in the northwest corner of Washington State. No confirmed reports have been found in other places in Washington State, indicating that the hornets colony has not been in the whole area, which is a good sign. However, no positive sample report was found does not mean that Asian Giant Hornets were not spread here. Anywhere in the state, there is a possibility that Asian Giant Hornets may exist.

Future Outlook

In the future, the Asian Giant Hornets population's development may have three endings:

1. Human biological invasion defense failed, Asian Giant Hornets raged across the state, proliferated, and finally reached a new ecological balance.

2. Human biological invasion defense and Asian Giant Hornets colony proliferation reach a dynamic balance.
3. Humans defeated the Asian Giant Hornets, and the Asian Giant Hornets became extinct in Washington State.

Take Actions

Prevention and Control Strategies

Consider the trade-offs of different prevention and control strategies:
Strategy 1: Mass investigation, strict control, and active removal of the hive—suppress Asian Giant Hornets' quantity to a very low level.
Strategy 2: Considering costs and benefits, allowing a certain number of Asian Giant Hornets to survive in the state without damaging the ecology to pursue dynamic balance.

Our model can meet the needs of two strategies by adjusting the parameters. For strategy 1, we can increase the model's recall rate of positive samples with the cost of appropriately reduce the overall accuracy. For strategy 2, we can further improve the model's accuracy for all samples, eliminate false positives, and screen out the most likely part of the samples from a large number of potentially positive samples.

Quick Response

When encountering similar crises, quick response is the key. The multiplication speed of the bee colony is extremely fast. Once the emergence of Asian Giant Hornets is observed, corresponding measures should be taken quickly to avoid the large-scale spread of the colony, leading to an exponential increase in the workload of prevention and control in the future. One impressive advantage of our model is that it does not need additional bounding box annotations on objects or parts, which are expensive to collect. This feature enables us to react very fast to those unprepared potential threats.

Public Participation

Guide the public to better participate in the observation report of Asian Giant Hornets. At present, the public lacks knowledge of Asian Giant Hornets critically. A large number of apparent non-Asian Giant Hornets samples are reported as Asian Giant Hornets cases. The messages provided are vague, and the photos are difficult to identify. Dirty data increases the workload of the staff and is not conducive to concentrating resources where needed.

Regional Cooperation

Attach importance to regional cooperation. We noticed that there are also positive reports of Asian Giant Hornets in Canada's bordering area to Washington State. The migration of insects does not require passports and is challenging to detect, so it cannot be controlled through regional blockade strategies. Therefore, while doing a rigorous job in preventing biological invasion in the state, Washington State must also pay attention to cooperation with surrounding areas to prevent Asian Giant Hornets from entering from the border.

References

[1] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[2] International Union for Conservation of Nature, Iucn Species Survival Commission, International Union for Conservation of Nature, and Natural Resources. Species Survival Commission. *IUCN Red List categories and criteria*. IUCN, 2001.

[3] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[4] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363, 2018.

Appendices

Acknowledgement

Tools and Softwares

- **Google earth**. For map data.
- **LaTeX**. For report editing.
- **MATLAB_R2019b**. For analysis.
- **Python**. For coding.
- **Overleaf**. For tex editing.

Libs

- **Tensorflow**. For deep learning.
- **pandas**. For data processing.
- **matplotlib**. For plotting.
- **numpy**. For matrix calculations.
- **PyTorch**. For deep learning.
- **scikit-learn**. For machine learning.

- **OpenCV**. For image processing.