Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2115921

A Hornet Relative Density Distribution model Based on Kernel Density Estimation and Xgboost Classifier

Summary

To address the hornet-confirming challenge, we propose a species relative density distribution model based on **Kernel Density Estimation(KDE) Method**. KDE method can calculate the contribution of all samples to the density function at each space position (the decay mode of density contribution with distance is determined by and kernel function), so as to calculate the density distribution function of species in the whole space. We estimate the uncertainty level of the prediction at location X by the density of report beside X since the denser the samples in the vicinity of X, the more accurate the prediction of the model. This results of this method is proved to be intuitive and robust by experiments.

To predict the species distribution over time, we devide the whole timespan into several time windows and propose a updated strategy synthesizing the historical distribution and new report information in the current time window. The length of the time window is just how often we update the model. Minimum appropriate time window length is determined by whether enough reports is received within the time window.

We then built a **Xgboost classifier** for unverified reports prediction. Feature like the month the report occurred, description about the bee is extracted. What's more, **image classification algorithm based on EfficientNet** and **Sentiment Classification algorithm based on BERT** are used to dig information from images dataset, witness' description and laboratory comments. In order to solve the imbalance of image samples, data enhancement and **Focal Loss** functions were used to effectively adjust the process of image preprocessing and CNN training. However, suffered from severe data imbalance, Xgboost classifier performs poorly. Therefore, we make the choice to use the species distribution function which is independent of unverified report prediction.

Based on model established above, we propose a report priorities model and set up conditions to determine whether hornet species has been eradicated.

To assess report priorities, both the **uncertainty level of report's location** and the **predicted credibility** of report itself are taken into account.

The **eradication** of hornets can be defined as the maximum value of its relative density function is less than a certain threshold.

Our framework shows a strong robustness. The species relative density estimation model doesn't depend on meshing the space and provides a continuous prediction. Metrics of our species relative density distribution model is stable with regard to manually given parameter.

Keywords: KDE-method, Species Density Estimation, EfficientNet, Xgboost classifier

Contents

1 Introduction

1.1 Problem Background

The nest of the scientific name Vespa mandarinia was discovered on Vancouver Island in British Columbia, Canada, in September 2019. The nest was quickly destroyed, but the news of the incident spread quickly across the area. Since then, there have been several confirmed sightings of vermin in the neighboring state of Washington, as well as many false sightings. Researchers need to effectively identify the most likely species identification among reports so that they can effectively document and predict the path of species spread.

1.2 Problems to be solved

Here are the problems to be solved:

1. We should build a model to predict the **geographic distribution** of hornets; Explain **the method and how often this model is updated** when new reports are received.
2. We should build a binary classifier to predict **how likely a report is to be positive**;
3. We should build a model to evaluate the **Priority Level** of unverified sightings.To guide the government agencies about which reports should be investigated first.
4. We should gave **conditions** under which we could prove that hornet species had **disappeared for studied area**.

2 Preparation of the Models

2.1 Assumptions

To simplify our model and eliminate the complexity, we make the following main assumptions in this literature. All assumptions will be re-emphasized once they are used in the construction of our model:

1. The spread of this pest over time can be described by a ground truth probability function $f^*(x, y, t)$, it is the proportion of established hornets in all bees at position (x, y) and time t. However, with limited information we can only get a estimation $\mu(x, y, t)$ with variance $\sigma(x, y, t)$.
2. $f(x, y, t)$ may lacks special structure but it is continuous, that would make it easy to optimize. f is derivative-free(evaluations do not give gradient information).
3. Government agencies has limited resource and verify a report can be costly. Therefore, evaluating f is expensive and the number of times we can evaluate it is severely limited. The purpose of carrying out a investigation is to make the estimation of f as accurate as possible.
4. Unverified report can also povide information about the probability of hornet activity. The estimated probability $p(info)$ can be extracted by pictures, descriptions and other information.

2.2 Notations

In this work, we use the nomenclature in **Table 1** in the model construction. Other nonefrequent-used symbols will be introduced once they are used.

Table 1: Notations

Symbol	Definition
$X = (x, y), t$	The longitude, latitude and time
$f^*(X, t)$	The ground truth probability function that hornets are present
$\hat{\mu}_i(X, t)$	The estimation of $f^*(X, t)$ based on new report in ith time window
$\mu_i(X, t)$	The estimation of $f^*(X, t)$ based on new report in ith time window
$\sigma_i(X, t)$	The Uncertainty Level of relative density prediction at position X
p_i	The Unverified Report Credibility

2.3 Data Filtration and Abnormal Value Processing

Since the major subject of our study is the spread of hornets, time is a key variable. Therefore, report without valid date is worthless. We drop reports whose 'Detection Date' is '<Null>' or before 2000 year. Moreover, report before 2010 is so sparse that not every year there's a new report. So we drop reports before 2010. Then we plot the density function for the rest of the data as is shown in Figure 1.

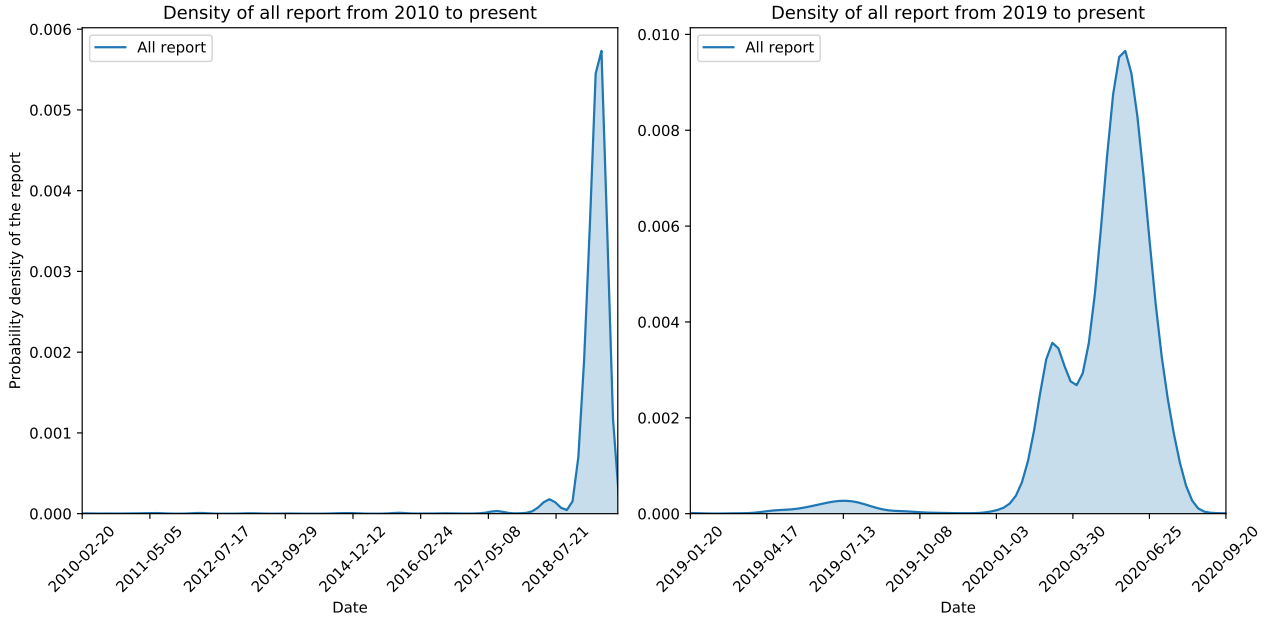

Figure 1: Distribution of all report over time

According to Figure 1, proportion of reports before May 2019 is negligible. Therefore our research

focus on reports from 2019-5 to present. Figure 2 shows the number of reports in each Lab Status and in each month. Obviously this is a highly unbalanced data set for more than 96% reports is negative.

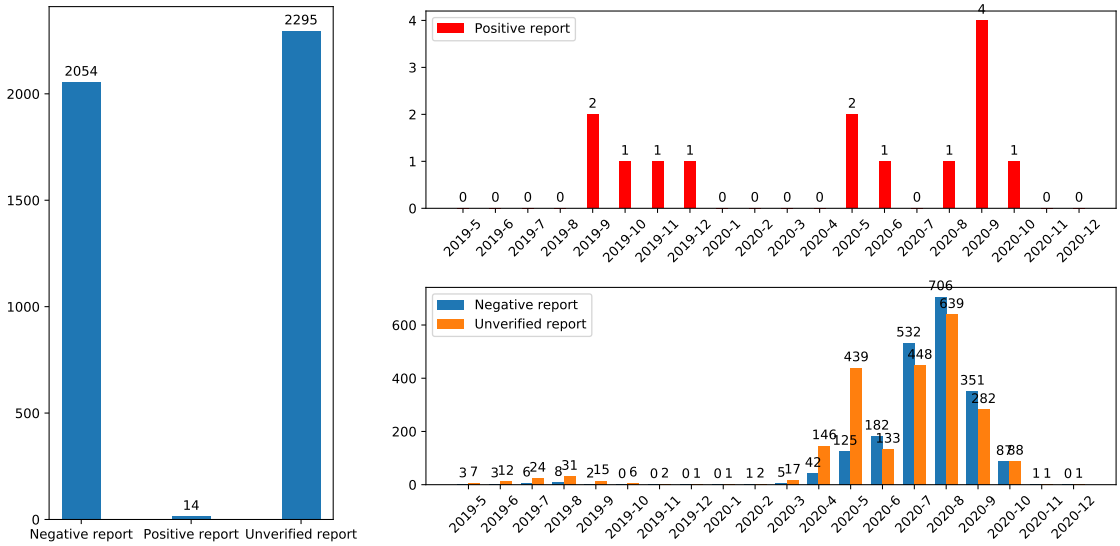

Figure 2: Distribution of all report over Lab Status and month

3 Model Construction

3.1 Quantitative description of hornets distribution

3.1.1 Model building

Predicting the spread of hornets can be seen as a presence-only modeling of species distributions. Many methods have been used for such purpose. Generalized linear models (GLMs) and generalized additive models (GAMs) is widely used from species distribution to environmental management[2]. A Bayesian approach [3] proposed modeling presence versus a random sample. ENFA, Hirzel et al proposed Environmental-Niche Factor Analysis using presence localities together with environmental data for the study area[4].

Our method modeling species geographic distributions is **Kernel Density Estimation Method** (**KDE-Method**)[1]. In our data set, additional environmental data is not provided and available geographic information of reports is just longitude and latitude. KDE-Method is suitable for modeling spacial disctance and can output a distribution function estimating species density.

Unlike histograms, kernel estimators are smooth and does not depend on the width of the bins and the end points of the bins. Kernel estimators smooth out the contribution of each data point over a local neighbourhood of that data point. The contribution of data point X_i to the estimate at some point X depends on the between X_i and X are.The extent of this contribution is dependent upon the shape of the kernel function adopted and the width (bandwidth) accorded to it. If we denote the kernel function as K and its bandwidth by h, the estimated density at any point X is

$$\hat{f}(X) = \frac{1}{n}\sum_{i=1}^{n} K\left(X, X_i\right), \tag{1}$$

where $\int K(t)dt = 1$ to ensure that the estimates integrates to 1. X is a position vector. The kernel function K is usually chosen to be a smooth unimodal function with a peak at 0. RBF kernel is one of the most widely used kernels due to its similarity to the Gaussian distribution. This kernel can be mathematically represented as

$$K\left(X, X'\right) = \exp(-\gamma\|X - X'\|^2), \tag{2}$$

where γ is a positive parameter.

We use Python API **sklearn.svm.OneClassSVM** in to solve the kernel density. **Figure 3** shows the kernel density of all reports since 2019-5 and the parameter $\gamma = 3$.

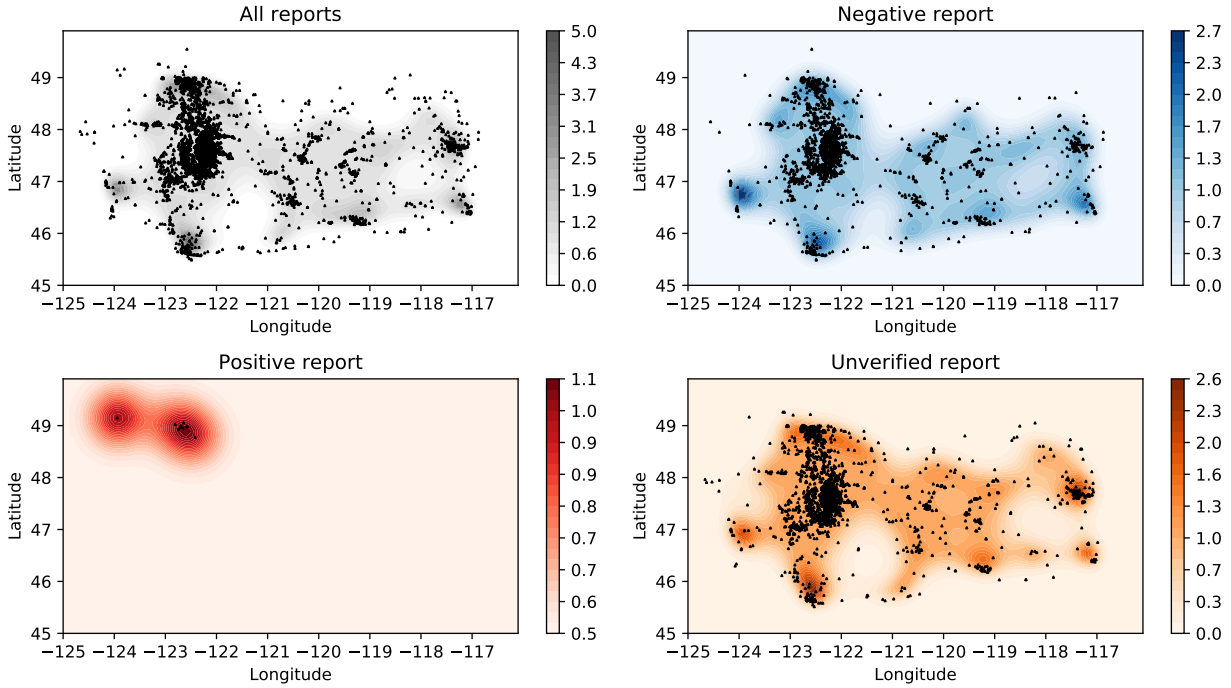

Figure 3: The Kernel Density of All Reports since 2019-5

To describe the spread process of hornets, we split all the reports by date in time window of 30 days. Then we estimate the kernel density of positive and negative reports represented as:

$$\hat{f}(X, t, \Delta t, label) = \frac{1}{n}\sum_{X_i \in S(t, label)} K\left(X, X_i\right), \tag{3}$$

where $label = 0, 1, 2$ represents **negative, positive and unverified** report respectively. $S(t, \Delta t, label)$ represents set of all reports with time window of length Δt after time t and has **Lab Comments** corresponding to $label$. n is the size of $S(t, \Delta t, label)$.

Intuitively, the proportion of established hornets in all bees is the ratio of density of positive and negative reports.

$$\hat{\mu}(X,t,\Delta t) = \frac{\hat{f}(X,t,\Delta t,1)}{\hat{f}(X,t,\Delta t,0) + \hat{f}(X,t,\Delta t,1)} \tag{4}$$

According to **Equation 4** unverified report does not have any contribution to μ. But in fact, we can build a model to infer how likely an unverified report is to be a positive example. This model take image file, notes and lab comments as input, probability as output denoted as

$$p_i = p(Image_i, Note_i, LabComment_i). \tag{5}$$

If p_i is available, $\hat{\mu}$ can be corrected to be

$$\hat{\mu}(X,t,\Delta t) = \frac{\hat{f}(X,t,\Delta t,1) + \frac{1}{n}\sum_{X_i \in S(t,label)} p_i K(X,X_i)}{\hat{f}(X,t,\Delta t,0) + \hat{f}(X,t,\Delta t,1) + \hat{f}(X,t,\Delta t,2)} \tag{6}$$

The 'hat' sign on μ indicates that $\hat{\mu}$ only depends on reports within time window $(t, t+\Delta t)$ and does not take former reports into consideration. However, it is reasonable to assume that a part of hornets which have been recorded before still inhabit in the original area. Therefore, final estimation μ take both former estimation and new reports into consideration. $\mu_i(X)$ represents the proportion of established hornets at position X in ith time window.

$$\mu_i(X) = \frac{w \cdot \mu_{i-1}(X) + \hat{\mu}_i(X)}{w+1}, i = 1,2,...,n \tag{7}$$

where the n is the number of time windows the whole timespan is devided into. w is the weight of the historical information. $\mu_i(X)$ measures the distribution of hornets over a period of time. **Figure 4** shows the sequence $\mu_i(X)$ estimated by **Equation 4** with $w = 0.5$. It can visually show the spread of hornets over time.

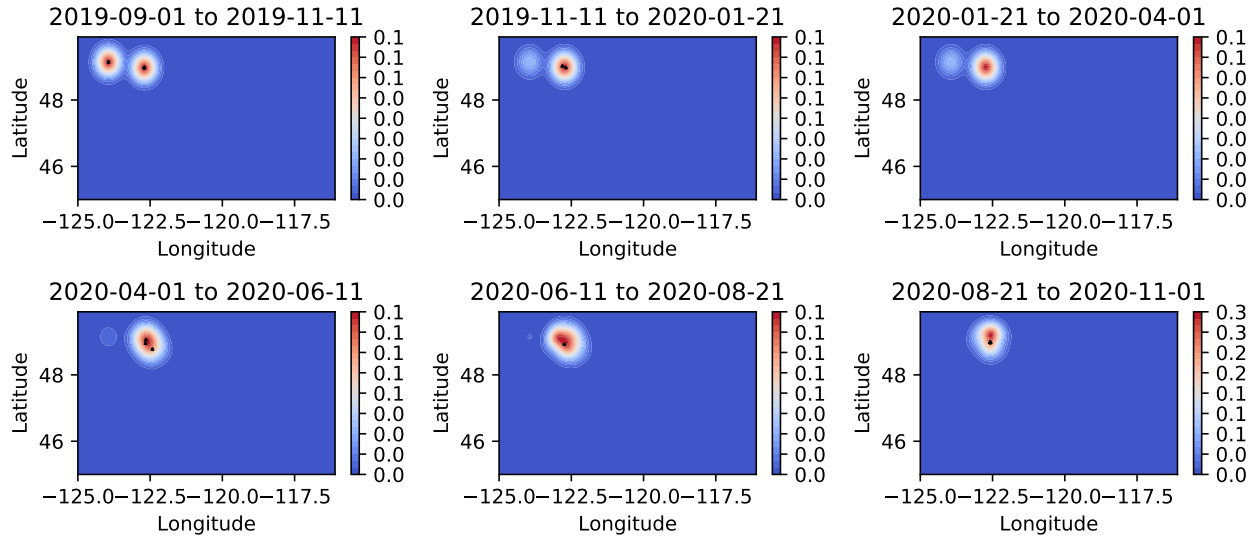

Figure 4: $\mu_i(X)$ since 2019-9 with $\gamma = 7, w = 1$

Since the report received is only the partial sampling of the actual species distribution, $\mu(X,t)$ has a certain amount of uncertainty $\sigma(X,t)$. The more spatially dense the samples are, the more accurate the estimation of the density function will be. The uncertainty level of $\mu(X,t)$ can be defined as

$$\sigma(X,t,\Delta t) = \exp(-\hat{f}(X,t,\Delta t,0) - \hat{f}(X,t,\Delta t,1)). \tag{8}$$

$\sigma_i(X) \approx 1$ means that the samples near X in ith time window is so sparse that we are completely unsure of the density estimate.

3.1.2 Model evaluation

model performance using receiver operating characteristic(ROC) curves. ROC analysis was developed in signal processing and is widely used in clinical medicine[6]. The main advantage of ROC analysis is that area under the ROC curve (AUC) provides a single measure of model performance, independent of any particular choice of threshold.[5]

By definition, species relative density distribution function $\mu_i(X)$ is the proportion of established hornets in all bees. It can be seen as the probability of a report at position X to be positive output by our model. We evaluate the model by mean ROC of a 5-fold cross-validation. For the positive report is sparse in both time and space. We evaluate the model on a long timespan from 2019-09 to 2020-11. Parameter setting is $\gamma = 7$ and for we only have 1 time window, w dosn't matter. **Figure 5** shows comparison between species density $\hat{f}_i(X,1)$, $\hat{f}_i(X,0)$ and hornets' relative density distribution $\mu_i(X)$

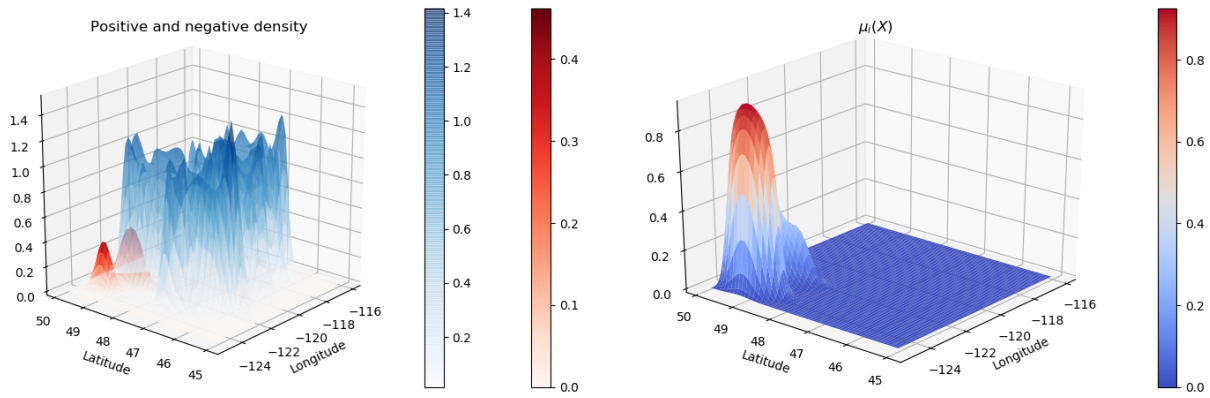

Figure 5: Positive and negative density and $\mu_i(X)$ with $\gamma = 7$

Figure 6 shows $\mu_i(X)$ and ROC curve of our model. The mean area under the ROC curve (AUC) is 0.75 which provides a level of precision of our species distribution model.

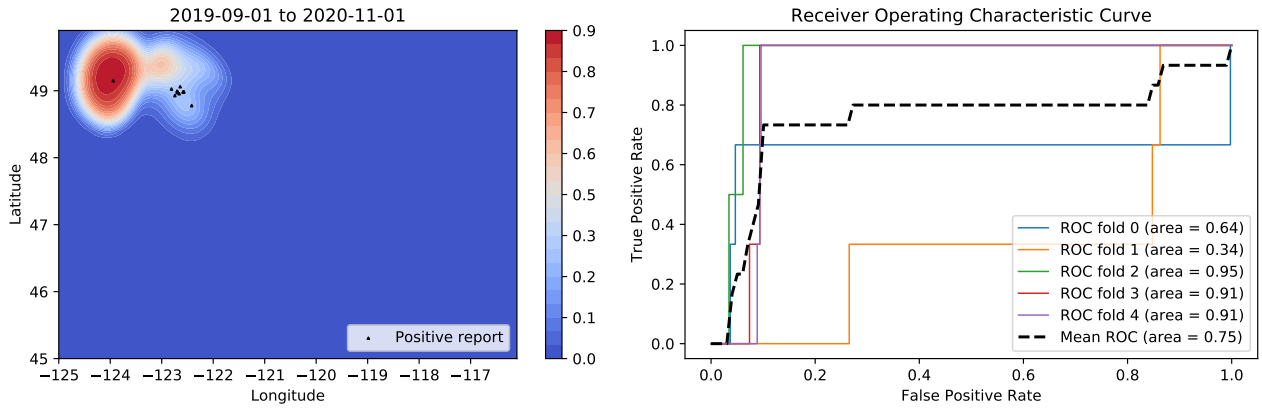

Figure 6: $\mu_i(X)$ and ROC curve with $\gamma = 7$

3.2 EfficientNet Convolutional Neural Network for Image Classification

When it comes to image classification problem, we usually use Convolutional Neural Network to train a model to help us.Efficientnet is a popular neural network that can take into account both the speed and accuracy of the model, making it faster and more efficient for image processing tasks.

3.2.1 Low Quality Image Filtering

Before using EfficientNet for training, we can see that in some of the images, the part of bee is so indistinct that it is impossible to distinguish any of the detailed features of the bee. It's hard for a neural network to learn anything about the difference between a hornets and a normal bee from these pictures. Therefore, before training, we need to get rid of these low quality pictures.

We use three image classification networks vgg16, resnet50 and mobilenet pretrained on Imagenet[7] as picture filter. ImageNet uses a list of 1000 non-overlapping classes, including bees. ImageNet can be applied in object recognition, image classification and automatic object clustering. For each input image, our picture filter give a confidence probability for all 1000 classes.

Intuitively, the more bee features in the image, the greater the confidence probability of the CNN network output bee class. Therefore we can evaluate the image quality by the minimum confidence probability that the input image is classified as bee. Then we can manually choose a threshold p_t and drop images whose quality is lower than the threshold. In our experiments we found that the model does perform better to classify the hornet species using filtered image dataset.

3.2.2 Classification Target and Evaluation

A convolutional neural network $\mathcal{N}$ can be represented by the whole convolution layers

$$\mathcal{N} = \mathcal{F}_k \odot \ldots \odot \mathcal{F}_2 \odot = \mathcal{F}_1 (X_1) \tag{9}$$

In fact, multiple convolutional layers with the same structure are usually referred to as a stage. The convolutional layer structure in each stage is the same (except the first layer is a downsampling layer). The convolutional layer can be convolved in the unit of stage. Network $\mathcal{N}$ is expressed as:

$$\mathcal{N} = \odot_{j=1\ldots s}\mathcal{F}_i^{L_i}\left(X_{\langle H_i,W_i,C_i\rangle}\right) \tag{10}$$

and $< H_i, W_i, C_i >$ represents the dimensions of the ith input tensor,i (from 1 to s) represents the sequence number of the stage $\mathcal{F}_i^{L_i}$ represents the ith stage It consists of a convolutional layer F_i repeated L_i times.

we change the output dimensions into 2 class numbers reprensents the prediction label of hornet species $Y = \{\hat{p}_0, \hat{p}_1\}$ then we choose the max possibility of P. After we have received the predicted hornet label, we can get four statistical metrics.

1 TP(True Positive):The true value is positive and the model considers the number positive

2 FN(False Negative):The true value is positive and the model considers that the number is negative

3 FP(False Positive):The true value is negative, the number considered positive by the model

4 TN(True Negative):The true value is negative, and the model considers the number negative.

we can use the Confusion Matrix and F_1 score to choose the best model for image classfication.

$$\begin{cases} \mathcal{N} & = \odot_{j=1\ldots s}\mathcal{F}_i^{L_i}\left(X_{\langle H_i,W_i,C_i\rangle}\right) \\ Y & = \mathcal{N}(X_1) \\ \hat{y} & = \arg\min Y \\ P & = \frac{TP}{TP+FP} \\ R & = \frac{TP}{TP+FP} \\ Accuracy & = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \\ F_1 & = \frac{2\times P\times R}{P+R} \end{cases} \tag{11}$$

3.2.3 EfficientNet Structure

EfficientNet[8] provides a new model scaling approach is used for an effective network. To extend the network from depth, width and resolution, it utilizes a simple and powerful composite coefficient. It does not arbitrarily scale the dimensions of the network like the conventional approach. The optimum group of parameters (composite coefficient) can be obtained based on the neural structure search technology. Not only is EfficientNet much faster than other networks, it also has greater accuracy.

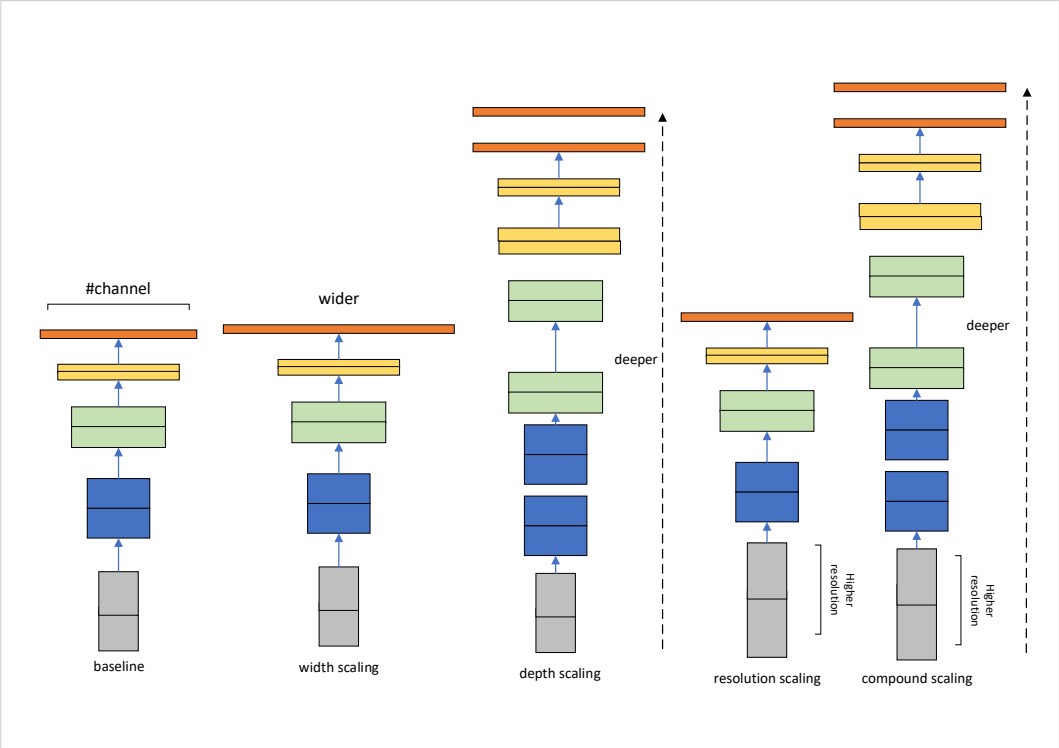

Figure 7: EfficientNet scales technology

Table 2: EfficientNet-B0 baseline network

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	Channels $\hat{C}_i$	Varied Layers $\hat{L}_i$
1	Conv3x3	224×224	32	1
2	MBConv1,k3x3	112×112	16	1
3	MBConv6,k3x3	112×112	24	2
4	MBConv6,k5x5	56×56	40	2
5	MBConv6,k3x3	28×28	80	3
6	MBConv6,k5x5	14×14	112	3
7	MBConv6,k5x5	14×14	192	4
8	MBConv6,k3x3	7×7	320	1
9	Conv1x1,Pooling,FC	7×7	1280	1

In EfficientNet, a compound expansion method is proposed. α, β, γ a set of parameters that we need to solve, and the optimal parameter with constraints is solved. They measure the proportions of depth, width, and resolution respectively. β, γ have a Square constraint, because if you increase the width or resolution twice, the amount of calculation will increase by four times, but if you increase the

depth twice, the amount of calculation Will be doubled.

$$\begin{aligned} &\text{depth: } d = \alpha^{\phi} \\ &\text{width: } w = \beta^{\phi} \\ &\text{resolution: } r = \gamma^{\phi} \\ &\text{s.t. } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \\ &\alpha \geq 1, \beta \geq 1, \gamma \geq 1 \end{aligned} \tag{12}$$

The calculation amount of convolution operation (FLOPS) is proportional to d, β^2, γ^2 We can adjust the hyper-parameters ϕUnder this constraint, after ϕ is specified, the calculation amount of the network will be approximately 2^{ϕ} times the previous .

Therefore, our network calculations can be controlled under the specified hyperparameters ϕ. The model will affect the with depth resolution of the network through the adjustment of other parameters, and find the best parameters in the grid search.

3.2.4 Focal Loss Function

There are only 14 positive samples, which is less than one-tenth of the negative samples. Our image data is extremely unbalanced. The traditional cross-entropy loss function cannot effectively solve this problem. Data filtering is only a partial solution to the problem of sample imbalance as is shown in **Figure 8**. Except for the case of $p_0 > 0.6$, the selected images are still at least 5 times larger than the original positive samples, and there are 3 samples in the positive examples that are useless, so we only have 10 valid positive samples.

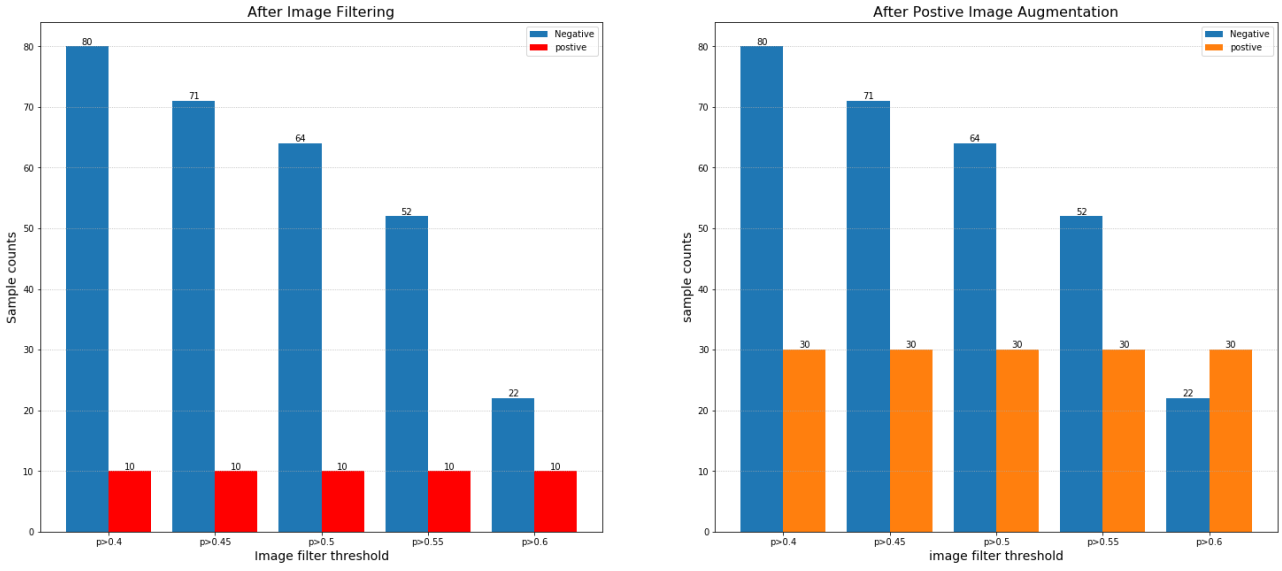

Figure 8: Data filtering can partialy solve class imbalance

We hope to further solve the problem of sample imbalance by modifying the loss function. **Focal loss** has been proved effective to calibrate deep neural network[9]. We introduce the focal loss starting from the cross entropy (CE) loss for binary classification.

$$\mathrm{CE}(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{otherwise} \end{cases} \tag{13}$$

In the above $y \in \{\pm 1\}$ specifies the ground-truth class and $p \in [0, 1]$is the models estimated probability for the class with label $y = 1$. For notational convenience, we define p_t:

$$p_{\mathrm{t}} = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise} \end{cases} \tag{14}$$

and rewrite $\mathrm{CE}(p, y) = \mathrm{CE}(p_{\mathrm{t}}) = -\log(p_{\mathrm{t}})$

in order to reshape the loss function to down-weight easy examples and thus focus training on hard negatives.we can add a modulating factor $(1 - p_t)^{\gamma}$ to the cross entropy loss, with tunable focusing parameter $\gamma \geq 0$. We define the focal loss as:

$$\mathrm{FL}(p_{\mathrm{t}}) = -(1 - p_{\mathrm{t}})^{\gamma} \log(p_{\mathrm{t}}) \tag{15}$$

3.2.5 K-fold Cross Validation

Due to the small number of samples and imbalance, we should use K-Fold cross-validation[10] to split the image dataset. Because there are too few positive samples in the data set, cross validation can make the evaluation to the performance of the model trained under certain hyperparameters more accurate. Furthermore, we can adjust k in different values to find the best k value for our datasets maximizing the performance of the network.

Ariginal data is devided into K groups (K-Fold), makes a validation set for each subset of data, and uses the remaining K-1 subset of data as the training datasets, so that K models will be obtained. The K models are evaluated in the validation dataset,then use an average loss and accuracy for model evaluating.

$$\mathrm{Loss}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Loss}_i \tag{16}$$

$$macro - P = \frac{1}{n} \sum_{i=1}^{n} P_i \tag{17}$$

$$macro - R = \frac{1}{n} \sum_{i=1}^{n} R_i \tag{18}$$

$$macro - F_1 = \frac{2 \times macro - P \times macro - R}{macro - P + macro - R} \tag{19}$$

We select different k values, perform multiple test experiments on the data, and extract the various metrics of each fold, and calculate their average value,then we can choose the best k for our datasets.

3.2.6 Result and analysis

After many tests, we found that when k=10 and the threshold of p0 is set to 0.5, the effect is the best, and when fold=2, we obtain the best model effect, and give his various metrics and Confusion matrix.

Table 3: Metrics of the best model

metrics	value	metrics	value
Recall	0.8850	Precision	0.8403
Accuracy	0.7434	F_1	16

Table 4: Positive Predicticon for Unverified image of the best model

GlobalID	possibility
882EC093-D907-454C-A9EE-8D1DC46A4291	0.535679
05F3D0A6-EAC3-472F-A591-D2D21D7A91B2	0.659352
9A5CB940-8951-4FE7-8619-DAF4A8FE1850	0.599529
D8F21BBA-7ED8-41EE-B068-1C2030580FBD	0.503856
13B67BCB-AFCE-4100-AD2B-76EF178BA228	0.543146
B95A5C78-19CE-4C75-8763-1A8BB952F141	0.573015
98D817B8-BB8B-4C31-9A42-DEED4491599C	0.58357
BBBA5BA0-CAFB-43D3-8F1D-FB2D9CF777E0	0.756872

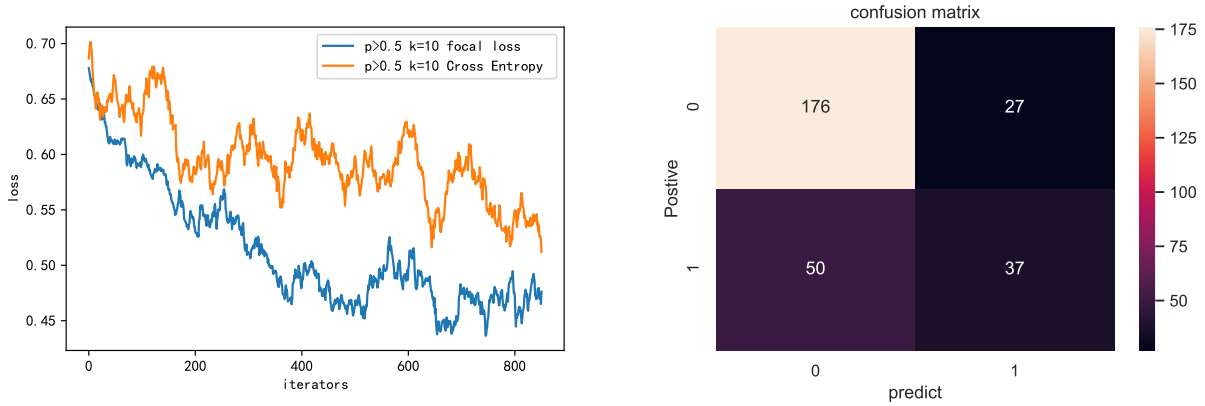

Figure 9: Training Loss and Confusion Matrix of the best model

3.3 Unverified Report Prediction

3.3.1 Feature Engineering

We first perform feature engineering on the dataset. The predicted results of the EfficientNet is used as a feature. Due to the life habits of the hornet are related to the month. We take the month of the

sighting date as a feature. Then we looked at how to extract information from eyewitness' description and laboratory comments.

The hornet's length and color is a distinguishing feature compared to other bees. But because of the different eyewitness' description way. It is difficult to extract information directly representing the characteristics of bees from the text.So let's simplify the problem.As long as the bee's length and color are mentioned in the sighting report, we assume that the eyewitness was deeply impressed by the length and color of the bees in the sighting which suggests that the sighting is more likely to be positive. Therefore, whether the words related to length and color appear or not is extracted as a binary feature. Also, we assume that the length of the report also reflects the credibility of the report to some extent.

Laboratory reviews are the responses that experts give to witnesses after reading information about the reports. The strength of its positive emotions may reflect the credibility of the report in one way. We use pretrained **FINBERT**[11], a language model based on BERT[12], as the sentiment analysis model.

For missing values, we simply fill them in with zero.

Table 5 includes few examples' engineered feature.

Table 5: Few examples of our dataset after feature engineering

lenth	note sentiment	lab sentiment	color	bee lenth	EfficientNet pred	Season	Lab Status
0.00	0.00	0.04	0	0	0.58	6	1
0.00	0.00	0.09	0	0	0.73	6	1
0.70	0.06	0.07	1	1	0.54	7	0
0.42	0.41	0.22	0	1	0.63	7	0
0.30	0.07	0.10	0	1	0.76	6	0

3.3.2 Xgboost Classification Model

The idea behind XGBoost's multi-classifier system is to build a decision tree in a gradient way, and then iterate gradually, adding a tree in each iteration and adjusting it so that it can reduce the current prediction error, gradually forming a strong evaluator integrating multiple tree models[13].

We use Python API **xgboost.XGBClassifier** to build the classifier with hyperparameterăspecified by **Table 6** .

3.3.3 Result and analysis

For this model, we still use recall, accuracy, precision and f1-score as the metrics. Confusion matrix is shown in **Figure 10**. There are 28 unverified report samples are predicted to be positive samples by the model. **Table 8** shows 5 examples in these 28 reports

Table 7: Metrics of the Xgboost classifier

metrics	value	metrics	value
Recall	0.5	Precision	0.2778
Accuracy	0.988	F_1	0.3571

Table 6: hyperparameter of Xgboost classifier

hyperparameteră	value
learning_rate	0.1
n_estimators	100
max_depth	10
min_child_weight	2
gamma	0.1
subsample	0.8
colsample_bytree	0.8
objective	multi:softprob
scale_pos_weight	1
num_class	2
n_jobs	-1

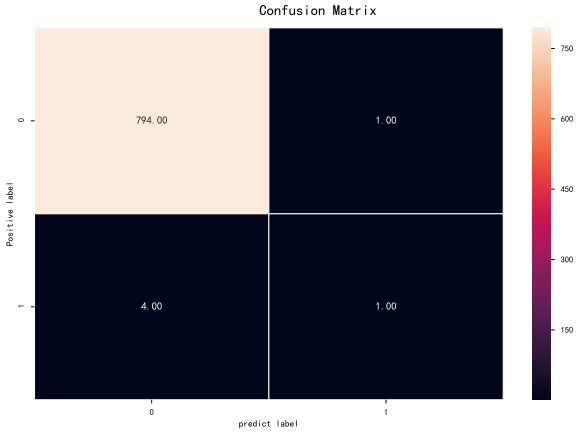

Figure 10: Confusion Matrix of Xgboost classifier

Table 8: Positive predictions of Xgboost classifier

GlobalID	Possbility
076DEDA1-899E-4FED-AB26-16FDC985985D	0.583
11EDE286-85BA-4433-8AF2-7F60CF8D7409	0.532
141066CC-E400-401B-A5A3-976BE391849C	0.526
19314731-2AAD-4F4E-B425-EF03601A51A7	0.732
1B4BD010-A445-4190-9B9D-6B81889B4C7A	0.535

For a binary classifier, the XGBoost model performs poorly. This is mainly because the sample data are too unbalanced. As the positive sample is too sparse, it is difficult for the model to learn effective information about the data distribution of positive example. In addition, the model has a large number of missing values. There are many samples that are missing one or more of the eyewitness

notes, lab comments and images.

Xgboost is already a very complex model for binary classification tasks. So until we have more effective data, it is very difficult to improve model performance by tuning the model itself.

Due to the poor performance of our report prediction model, we do not use it to modify the species distribution function and use the version independent of unverified report prediction given by **Equation 4**.

3.4 Priority Level of Unverified Reports

We measures the priority level of a unverified report i by following 2 principle.

- The more uncertain we are about $\mu(X_i, t_i)$, the higher the priority.
- According to information attached toreport i, the more likely that report i is a positive sample, the higher the priority.

The probability of report i being positive is $p_i = p(Image_i, Note_i, LabComment_i)$. Therefore, we measure the priority of unverified report i by

$$\varrho_i = \sigma_i(X) \cdot p_i \tag{20}$$

3.5 Eradicated Evidence

We define the eradication of hornets as the maximum value of $\mu_i(X)$ is smaller than a certain threshold T:

$$\max_X \mu_i(X) < T. \tag{21}$$

In the current time window, newly received negative samples can reduce the value of the distribution function.And positive samples can dramatically increase the value of the distribution function around them. If no new reports are received within a period of time.The half-life of the distribution function can be given by the number of time Windows according to

$$halfLife = \frac{\ln 2}{\ln \frac{w+1}{w}}. \tag{22}$$

4 Senitivity Analysis

Sensitivity analysis is a method to study and analyze the sensitivity degree of model state or output change to system parameters or surrounding conditions change In this section we choose some manually given important parameters and analyze the effects of changing these parameters on the model performance.

4.1 Sensitivity of $\mu_i(X)$'s AUC to parameter γ

According to **Equation 2**, γ determines how fast a sample's contribution to the surrounding density decays with distance. It will affect the density distribution prediction of species $\hat{f}_i(X)$, and then affect the hornets' relative density distribution $\mu_i(X)$ which stands for the probability of a report at position X to be positive output by our model.

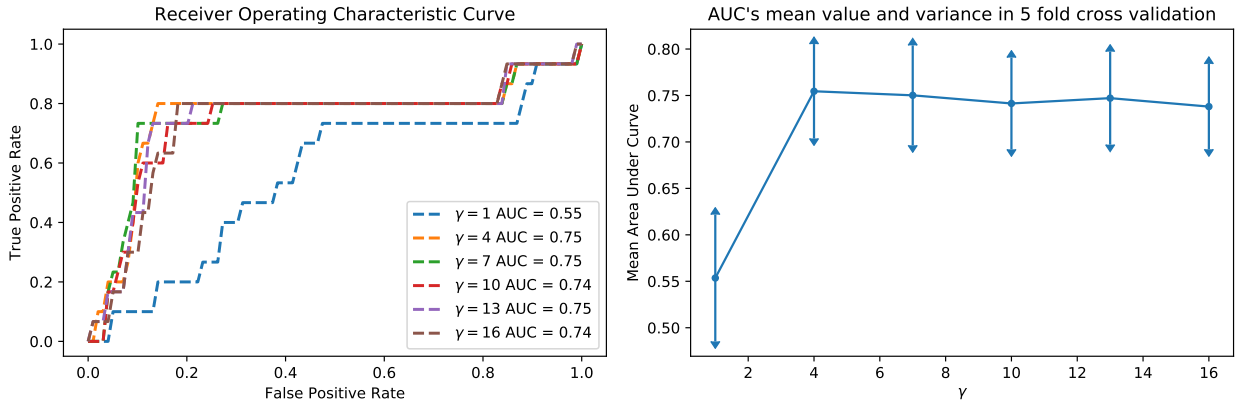

Figure 11: Sensitivity of AUC to parameter γ

Our analysis results is shown in **Figure 11**. According to the analysis, as long as γ is not too small, the performance of our model $\mu_i(X)$ is stable.

4.2 Sensitivity of EfficientNet Metrics

4.2.1 Image filter threshold p_t

Because we only use images of quality above a certain threshold for training.The threshold of image filtering will affect the performance of the model.If the threshold is too high, the training set will be too small, and CNN cannot obtain enough data for training.If the threshold is too low, the image data set will have too much noise, which is not conducive to the learning and convergence of the network.

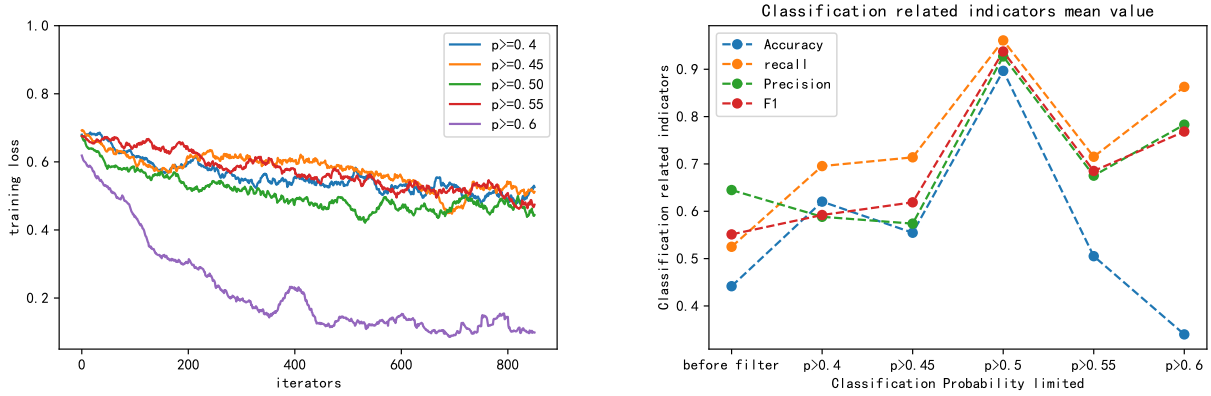

Figure 12: Training Loss and mean value of Metrics in best fold

Figure 12 shows the training process and final metrics of the model under different image filtering thresholds. We can see p_t have a great influence on the performance of Efficientnet.The performance of the model reached the highest at at $p_t = 0.5$.

4.2.2 Number of folds in the cross-validation k

Due to insufficient sample data of image data set, different cross-validation sampling may lead to deviation of data distribution, thus affecting model performance. **Figure 13** shows the final loss of the model in different cross-validation.

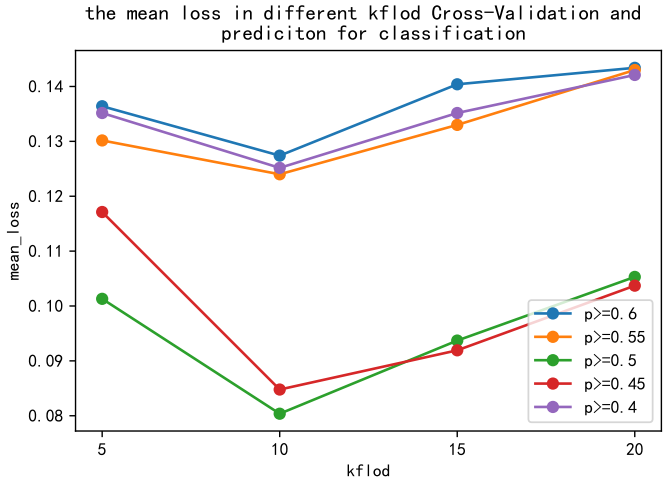

Figure 13: Training Loss and mean value of Metrics in best fold

We can find that k does not have a big influence on the performance of the model, and roughly the best performance is achieved when $k = 10$.

5 Strengths and Weaknesses

5.1 Strengths

- Our species relative density estimation model uses a continuous kernel density estimation algorithm. It doesn't depend on meshing the space. There is only one parameter γ needs to be specified manually. Moreover, it can be proved that the performance of the model is stable relative to γ within a wild range.
- We use an open source pretrained model to filter the image data set. It can be proved that image filtering greatly improves the performance of the classification model.
- We did our best to consider all the data available to us. Our binary classifier for unverified witness report takes into account the output of the image classifier, description from the witness, laboratory comments and the seasonal habits of hornets.

- Our model can be easily implemented if more detailed geographic data and report information is available which shows a strong robustness.

5.2 Weaknesses

- Our species density estimation model can not give an accurate estimate of species density for areas with sparse reports.
- For eyewitness's description, as well as laboratory comments, we only utilize simple features through regular matching and output of natural language sentiment classification model. We didn't dig deeper into information in the text and compare it with the real features of hornet.

6 conclusion

In our proposal, we describe the spread of hornet by **species relative density distribution** $\mu_i(X)$ which is the possibility of a report to be positive in the ith time window at location X. $\mu_i(X)$ is obtained by **species density function** estimated by **KDE-method**. In order to synthesize the historical distribution and report information in the current time window, $\mu_i(X)$ is updated by **Equation 7**.

A credibility estimation model for undetected reports p_i can not only be used to rank the priority of unverified reports, but also enable us to use unverified reports to modify hornets' relative density distribution function $\mu_i(X)$. **EfficientNet** is an image classifier. Given the image, it outputs a confidence probability measure of the likelihood that the image is a hornet. NLP is a **emotion classification model** based on natural language processing , whose input is the expert's evaluation of the report, and output a positive emotion intensity to measure the expert's attitude towards the report. The two features above, along with other features extracted from the report data, such as season information and whether the report provides bee length, are extracted as features. All these features are fed into the machine learning model which give the results of **Unverified Report Prediction**.

The denser the reported samples is near location X, the lower the uncertainty of the model's prediction. The uncertainty level of the prediction at location X is given by **Equation 8**. The **unverified report priority** given by **Equation 20** takes both its credibility and uncertainty level of our model at position X into account.

We define that the hornet species have disappeared in this area if the maximum value of the distribution function is less than threshold T which can be interpreted as the maximum proportion of hornets in all bees is negligible.

So far, our hornet species distribution model based on **KDE** density estimation has been completely established.

Memo

The Washington State Department of Agriculture:

Thank you very much for your time. We are sorry that your area has been plagued by an invasion of hornets species. We believe that you have fully awared of the serious impact that hornets can have on the safety of local species and resident. We share your urgency to address hornets invasion problem.

Our hornet species invasion surveillance system is based on species density estimation algorithm. It aims to use all available information from eyewitness' reports. Our model can predict and visualize the distribution and spread of hornet species. It is based on sound mathematical principles and uses the state of the art deep convolutional neural network structure and natural language emotion classification model. Also the unique behavior and living habits of hornets is taken into account.

we have some suggestions for you:

1. We suggest you to encourage local residents to make as much reliable observations and reports as possible. Includes photos (videos) that **show bees' detailed feature and descriptions of their characteristics**. There can even be financial rewards for verified, high-quality reports.

2. We suggest you to mobilize as many government resources as possible to verify sighting reports, because more verified reports will make the species distribution model's prediction more accurate. If in practice the resources available are not enough to validate all sightings, we strongly recommend that you **prioritize sightings according to the priority level given by the model**. Our prioritization takes into account both the reliability of the reports and the uncertainty of the distribution of current model predictions of sighting locations.

3. We recommend that **experts who comment to residents' reports give a score on the reliability of the report** as well as their message responses. We realize that human resources are the most expensive resources. But on one hand the current machine learning algorithms are still unable to completely replace the experience of human experts. On the other hand, since the expert has commented, he or she must has made an assessment of the credibility using available information. Therefore, giving a numerical score will not dramatically increase the workload of experts, but the accuracy of machine learning algorithm prediction will be greatly improved if the opinions of human experts are taken into account.

4. We would like to obtain **detailed geographic data for the state of Washington** (including but not limited to elevation, forest cover density, rainfall and monthly mean temperature) and statistics on the distribution of local species. All of these features may affect the habitability of bumblebees in the region, but at present, our model's prediction of species density only considers longitude and latitude. If we get more detailed geographical data, the model is bound to give more accurate predictions. In addition, the distribution of hornets can interact with local species. For example, hornets hunt down other bees in September and October. The Japanese honey bees (Apis cerana japonica) coevolved with the Asian bumblebee and is defensible against it.

5. We suggest that you **increase the publicity of hornet knowledge** such as its appearance and living habits. On one hand, this can improve the quality of the report and reduce the proportion of negative report. At present, most sightings are negative. Due to the imbalance of samples, the model's performance is also affected. On the other hand, reducing the rate of false positives could save

the government's investigative resources and allow more valuable eyewitness reports to be verified. In addition, the understanding of hornet can improve residents' awareness of self-protection and reduce possible human injuries and economic losses.

We sincerely hope that the problem of hornet species invasion can be solved as soon as possible so that you and local residents will not be bothered by it anymore.

References

[1] Hwang, Jenq-Neng, Shyh-Rong Lay, and Alan Lippman. "Nonparametric multivariate density estimation: a comparative study." IEEE Transactions on Signal Processing 42.10 (1994): 2795-2810.

[2] Guisan, Antoine, Thomas C. Edwards Jr, and Trevor Hastie. "Generalized linear and generalized additive models in studies of species distributions: setting the scene." Ecological modelling 157.2-3 (2002): 89-100.

[3] ASPINALL, RICHARD. "An inductive modelling procedure based on Bayes' theorem for analysis of pattern in spatial data." International Journal of Geographical Information Systems 6.2 (1992): 105-121.

[4] Hirzel, Alexandre H., et al. "Ecologicalniche factor analysis: how to compute habitatsuitability maps without absence data?." Ecology 83.7 (2002): 2027-2036.

[5] Phillips, Steven J., Robert P. Anderson, and Robert E. Schapire. "Maximum entropy modeling of species geographic distributions." Ecological modelling 190.3-4 (2006): 231-259.

[6] Hanley, James A., and Barbara J. McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." Radiology 143.1 (1982): 29-36.

[7] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[8] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." International Conference on Machine Learning. PMLR, 2019.

[9] Mukhoti, Jishnu, et al. "Calibrating deep neural networks using focal loss." arXiv preprint arXiv:2002.09437 (2020).

[10] Bengio, Yoshua, and Yves Grandvalet. "No unbiased estimator of the variance of k-fold cross-validation." Journal of machine learning research 5.Sep (2004): 1089-1105.

[11] Araci, Dogu. "Finbert: Financial sentiment analysis with pre-trained language models." arXiv preprint arXiv:1908.10063 (2019).

[12] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[13] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.