Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2100024

Title: A Hornet Prediction and Assessment Model based on Data Analysis

Summary

To analyze and investigate the likelihood of the Vespa mandarinia, this paper proposes a data-motivated model based on an original definition of entropy weight method (EWM) appropriate for the department of agriculture.

First, due to the relations between the distribution of the Vespa mandarinia and the life cycle of it, we use cellular automaton (CA) to predict the spread of this pest over time. In the model, the space is divided into cells by a grid of some form, and each cell or grid point is assigned a certain discrete value to represent the state of that point, and due to the CA model predictions can be made on the validation set.

Then, we create a model that predicts the likelihood of a mistaken classification by using entropy weight method (EWM). After filtering out some error data, we train the model with detection date, latitude & longitude, number of reports monthly, image as input and lab status (Positive/Negative) as output. For the spatial distribution, the possible action range is divided according to the nesting range of hornets as 30km to limit the discovery location; for the temporal distribution, the discovery time is filtered according to the life cycle of hornets; for the image processing, it is performed by Transfer Learning (Inception-v3 model) to discriminate whether it is a hornet or not. The expected values of the three factors are input into the model for classification prediction to obtain the estimated probability of prediction, to decide which the priority order can be set. If there are other new reports over time, they should be added to the training set to retrain our constructed model.

The probability that the hornet has been eradicated in Washington State can be explored later by predicting the historical data until the point in time when the predicted number of reports is zero and the pest is verified as the exact pest. The pest is considered to have been eradicated if it has not been reported at the point in time when the predicted number of reports is 0 and the pest is verified.

Keywords: hornet prediction, evaluation model, cellular automaton, entropy weight method, transfer learning

Memorandum

From: Team #2100024
To: The Washington State Department of Agriculture
Date: 8th February 2021

Subject: The result of a model for prediction and assessment of hornets

Dear Sir/Madam,

we would like to remind you the result of our model. As required by the MCM Contest, we have done a hornet prediction and assessment model based on data analysis. The result showed that we can predict the spread of the Vespa mandarinia and can deal with how to prioritize the investigation of the reports.

First, as to how to predict the spread of this pest over time, we use cellular automaton (CA) to build a model of hornet prediction. In the model, the space is divided into cells by a grid of some form, and each cell or grid point is assigned a certain discrete value to represent the state of that point. In a $L \times L$ space in Washington state, we assumed an initial point of hornet nest. When $c = 0.2, a = 0.3, d = 0, L = 300$, with the high precision of the prediction, the spread of nesting is very rapid, if we do not try to control and eliminate the hornet.

Second, we create a model that predicts the likelihood of a mistaken classification by using entropy weight method. After filtering out some error data, we train the model with detection date, latitude & longitude, number of reports monthly, image as input and lab status (Positive/Negative) as output. For the image processing, it is performed by Transfer Learning (Inception-v3 model) to discriminate whether it is a hornet or not. The expected values of the four factors are input into the model for classification prediction to obtain the estimated probability of prediction, to decide which the priority order can be set. The higher score the report can get, the higher priority has the report. And the precise of the evaluation model is 96.3921% with 3-monthly-upgraded data. The lab and government do not need to update the data frequently. As a result, we choose three months as a update periodicity.

Third, the probability that the hornet has been eradicated in Washington State can be explored later by predicting the historical data until the point in time when the predicted number of reports is zero and the pest is verified as the exact pest. When the death rate of hornet nesting d= 0.5, the pest is considered to have been eradicated because it has not been reported after about 20 months when the predicted number of reports is 0. So if the local government takes every action to stop the spread, the hornet can be eradicated in 20 months.

We hope the spread of hornet can be stopped as soon as possible. And we wish our model of hornet prediction and report analysis model can help you!

Yours truly,

Team #2100024

Content

1 Introductions

1.1 Problem Background

In September 2019, a colony of Vespa mandarinia (also known as the Asian giant hornet) was discovered on Vancouver Island in British Columbia, Canada. Since that time, several confirmed sightings of the pest have occurred in neighboring Washington State, as well as a multitude of mistaken sightings.

Due to the potential severe impact on local honeybee populations, the presence of Vespa mandarinia can cause a good deal of anxiety. The State of Washington has created helplines and a website for people to report sightings of these hornets. Based on these reports from the public, the state must decide how to prioritize its limited resources to follow-up with additional investigation. While some reports have been determined to be Vespa mandarinia, many other sightings have turned out to be other types of insects.

Fig. 1. 1. The processed picture of the Vespa mandarinia[1]

The primary questions for this problem are how to interpret the data provided by

[1] Photo'd by Filippo Turetta - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=80323907

the public reports and what strategies we can use to prioritize these public reports for additional investigation given the limited resources of government agencies.

1.2 Our Work

After analyzing the descriptions and demands of the Problem C, we find a few key points in this question:

- The data given is large and not organized. How to do the normalization of the data.
- How to predict the spread of the hornet according to the spatial and time distribution given by the data lists.
- How to analyze the reports sent by citizens and predict the likelihood of a mistaken classification with updating data.
- How to prioritize the investigation of the reports.
- How to know whether the hornet has been eradicated in Washington State.

Based on above discussion, the tasks can be boiled down to the following steps to prevent the spread of the hornet:

- First, due to the relations between the distribution of the Vespa mandarinia and the life cycle of it, we choose the confirmed reports (with positive ID) and use cellular automaton (CA) to predict the spread of this pest over time and then the predictions can be made on the validation set.
- Second, after filtering out some error data, we create a model that predicts the likelihood of a mistaken classification by using the entropy weight method (EWM). We train the model with detection date, latitude & longitude, image as input and lab status(positive/negative) as output. For the image processing, Transfer Learning (Inception-v3 model) is performed by deep learning to discriminate whether it is a hornet or not. The expected values of the three factors are input into the model for classification prediction to obtain the estimated probability of prediction, to decide which the priority order can be set. If there are other new reports over time, they should be added to the training set to retrain our constructed model.
- Third, the probability that the hornet has been eradicated in Washington State can be explored later by predicting the historical data until the point in time when the predicted number of reports is zero and the pest is verified as the exact pest. The pest is considered to have been eradicated if it has not been reported at the point in time when the predicted number of reports is 0 and the pest is verified.
- Further analysis and discussion of the model.

2 Assumptions

- Ignore the influence of the climate and weather. The climate in Washington State will be suitable for the survive of the Vespa mandarinia.
- Ignore the terrain of the detection places. The hornets will live on the plain ground with zero elevation.
- Assume that the found hornet nests have not been destroyed, which means there are possibility of new nesting from 14 “Positive ID” position.
- The hornets’ action and spread will not be affected by their natural enemies and human beings.
- Assume that all the lab status and comments were correct.
- Assume that the detection time was not mistaken. The appear of the hornet just happened in the detection date.
- The region of the detection will be flat. That means distances in longitude are not affected by changes in latitude.

3 Hornet Prediction Model

In this part, we focus our attention on how to predict the spread of the hornet according to the spatial and time distribution given by the data lists. In the agriculture sector, many scholars have studied the spread of plant pests and diseases through reaction diffusion models and cellular automaton. According to the concept, we use the available data to construct a hornet prediction model appropriate for the agencies and labs.

3.1 Concept of Cellular Automaton (CA)

The basic components of a cellular automaton include cells, cell space, neighbors, and rules, and the state of cells. To study the spatial distribution of pest and disease spread, the three-dimensional space of actual terrain is simplified into a two-dimensional $L \times L$ rectangular grid of hornet spread model according to the spatial distribution of nesting, and the distance between two nests is considered.

Also, we drew the elevation map of part of Washington state. The hornet prefers to live on the plain ground with low elevation. Therefore, the nesting rate is related to the elevation and we need to consider this factor.

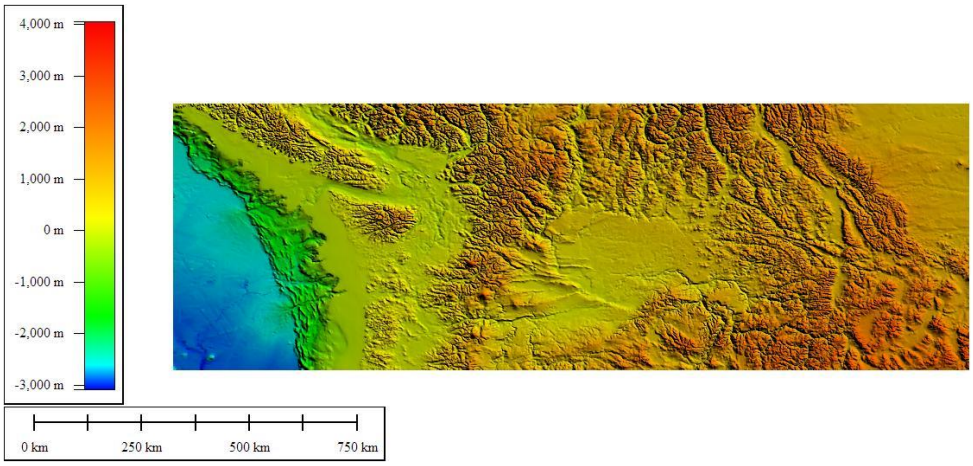


Fig. 3. 1. Elevation map of Washington state

Based on the individual cellular automaton model[1], the local dynamics and interactions of hornet transmission are depicted in the form of Von Neumann neighbors, assuming that the change in the state of a cell is only related to the state of its neighboring cells, according to the law of hornet nesting transmission [2]. To eliminate the boundary effect, we adopt the period boundary, which is closest to infinite space. This means the spread of hornet between neighboring nests is considered. The status of a potential nest is determined by the condition of itself and its four surrounding neighbors at that moment and is updated simultaneously.

For each cell in the $L \times L$ cell space, the covariates also contain the location variables of the detection (i, j). There are three possible states $S_l(i,j)$ for each cell: $S_l(i,j) = 0$ means open space without hornets. $S_l(i,j) = 1$ the status of detection is "Negative ID", and the insect is not a hornet. $S_l(i,j) = 2$ means the status of detection is "Positive ID".

Under the assumption that nesting are not inherently resistant and that no external conditions affect the rate of hornet transmission, the cell-specific evolutionary rules are shown in the table4.1, where: c is the probability of the open space being planted; a is the probability of being infected by hornets; d is the rate of the elimination or the death of the nesting location, and $A(i, j)$ is the number of state 2 of the four adjacent cells with state 1.

Table 3.1. The law of state evolution of each cell in CA

$\boldsymbol{S_l(i,j) \rightarrow S_{l+1}(i,j)}$	Probability
0→0	*1-c*
0→ **1**	*c*
1 → **1**	$(1-a)^{A(i,j)}$
1 → **2**	$1-(1-a)^{A(i,j)}$
2 →0	*d*
2 → **2**	*1-d*

On this basis, the main steps of the modeling process of hornet nesting transmission are as follows.

Step 1: The initial state is given by the calculator based on coordinate values with latitude and longitude from the data.

Step 2: Each cell is transformed according to the Von Neumann neighborhood form and the evolutionary law that has been set, and a new cell space is created after the first evolution. We record the number of cells in different states after transformation and outputting the corresponding dendrograms.

Step 3: The cells in $L \times L$ are updated step by step, and then the simulation is performed in step 2. The simulation continues until the number of cells in each state reaches a stable state.

3.2 Simulation Results and Analysis

The transmission process of the nesting is simulated by computer. In this case, the simulated initial conditions were based on a reasonable planting and space distribution, with the initial space ratio controlled at a certain percentage, and the computer generated no more than 1% of infected locations on an ad hoc basis. At the end of the simulation, the simulation results were compiled and analyzed, and the time-varying distribution of plants in different states during the nesting transmission process and the time-varying distribution of plants in three states were discussed.

To facilitate the study and observation of the tree array of hornet nesting transmission processes, we decide $c = 0.2, a = 0.3, d = 0, L = 300$.

Considering that the nesting process comes from the hornet itself, the iterative initial matrix is chosen as S, and we choose one of the “Positive ID” nest as the initial point, $\boldsymbol{S}_1(\boldsymbol{Lat}, \boldsymbol{Lon}) = 2$, $\boldsymbol{S}_1(\boldsymbol{i}, \boldsymbol{j}) = 0, 1$ ($i \neq$ Latitude, $j \neq$ Longitude).

Table 3.2. The initial point

Latitude	Longitude
49.14939	-123.9431

After 300 times of iteration, we can get the number of the total nests, the spread rate in this $L \times L$ area and the distribution of the nesting.

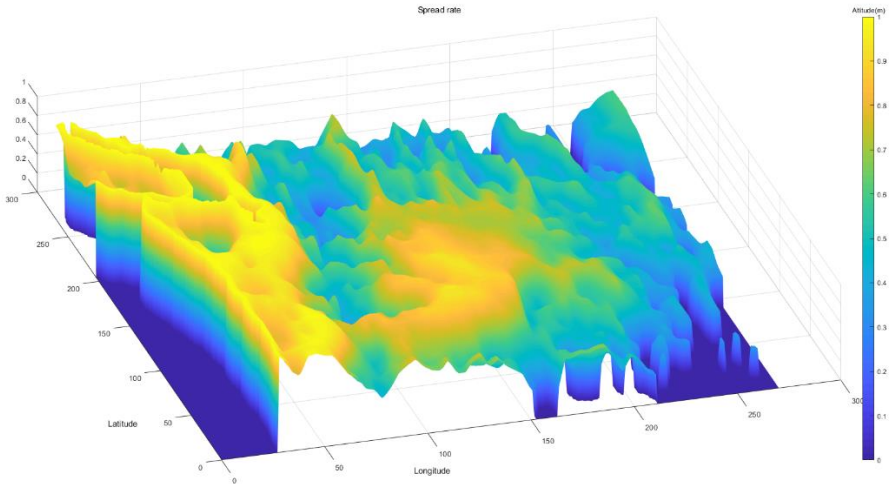
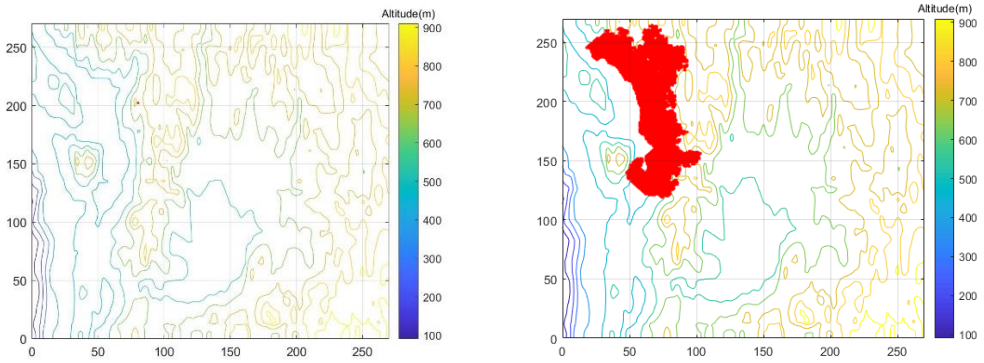


Fig. 3. 2. The spread rate in $L \times L$



(a) The initial point (b) The result

Fig. 3. 3. Distribution of the nesting

As it is shown above, with the high precision of the prediction, the spread of nesting is very rapid, if we do not try to control and eliminate the hornet. With the prediction of the new nesting, the lab and local department can act immediately to pay attention to the unverified nest nearby.

4 Hornet Report Evaluation Model

4.1 Overall Idea of the Model

In this part, we focus our attention on how to analyze the reports sent by citizens and predict the likelihood of a mistaken classification with updating data. Then with the output we can decide how to prioritize the investigation of the reports.

First, we divide the data given in the question into text-based information and picture-based information and preprocess them separately. For the detection date, location coordinates and the number of reports in the current month in the submitted

reports as inputs for modeling, we use the entropy weight method (EWM) to assign weights separately to get the comprehensive evaluation score of text-based information[4]; for the picture information in the reports, we use Transfer Learning (Inception-v3 model) to recognize the images with machine learning. It will output the probability of whether each image is a hornet or not in [0,1], where 1 is exactly a hornet and 0 is not a hornet at all.

After obtaining the textual information score and picture information score corresponding to each report, we analyze when these two should be prioritized. When the picture scores are close to 1 and 0, Lab can very clearly determine whether the hornets were truly found, and the priority should be high at that time. So, we decided to use a normal distribution to assign weights to these two ratings, and finally form a complete and comprehensive evaluation system.

4.2 Textual Information Evaluation

4.2.1 Score for each sector

We choose three different sectors from data to decide the scores of every report separately. It is worth noting that the submission date in the data is generally later than the detection date and at varying intervals. It cannot be used to determine the time of occurrence of the hornets.

What is more, we found that the notes submitted by citizens generally did not contain enough useful information. For example, among the 14 notes with “Positive ID”, the only key word with a common point was "hornet", and the positive sample is too small to be processed by natural language processing (NLP) to be used as a rating factor. So, we did not use the note sector.

A. Spatial Location Score

It is said in the question that a new queen has a range estimated at 30km for establishing her nest. We draw a circle with the point marked as “Positive ID” as the center and with a radius of 30km. We also draw a circle with a radius of 30km around the point marked as “Unverified”, because we cannot tell whether the point marked as “Unverified” is a real hornet or not.

For the range of activity made by the proven point, we make the spatial location score at the center of the circle (i.e., detection point) 1, and the fraction at the edge of the circle 30km away 0, decreasing from the center of the circle to the surrounding area.

In the same way, for the range of activity made by the unverified point, we make the spatial location score at the center of the circle 0.5, and the fraction at the edge of the circle 30km away 0, decreasing from the center of the circle to the surrounding area. By this way we can get a range and probability map of hornets.

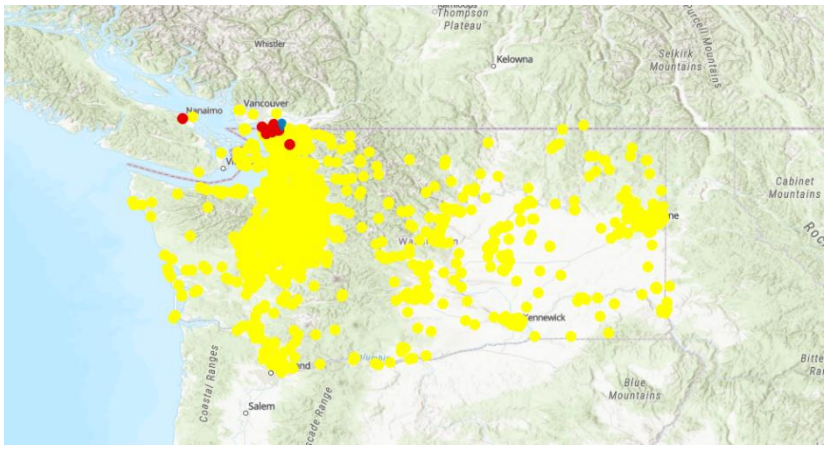


Fig. 4. 1. Range and probability map of hornets

For every new detection of the insect, each new input coordinate can get the corresponding [0, 1] score from the distribution map.

B. Life Cycle Score

A search of the Vespa mandarinia shows that[3], They first appear in early to mid-April and begin feeding on the sap of Quercus (oak) trees. Males and new queens take on their responsibilities in mid-September and mid-October. After this period, they will enter hibernation. We decide that the life cycle of the hornet is from April 1st to October 31st, in which people can see this specie. When it is not in the period of its life cycle, especially in the winter, it is more difficult to find this creature. So, we made a table to give a life cycle score to every detection date.

Table. 4. 1. Monthly life cycle score

Detection Month/Date	*Life Cycle Score*
April 1st – October 31st	1
November	0.7
December	0.4
January	0.1
February	0.4
March	0.7

For every new detection date, each new input can get the corresponding [0, 1] score from the table.

C. Report Number Score

We have analyzed the data and found that most reports were concentrated in 2020, and the number of reports received per month was different, with summer and fall being the main times of insect outbreaks, and the number of reports directly reflects the degree of concentration, which can affect the priority of the processing performed by the lab. So we can normalize the number of reports as one of the factors of the scoring index.

Table. 4. 2. Report number and score in 2020

Month	*Report Number*
Jan	1
Feb	3
Mar	22
Apr	188
May	566
Jun	316
Jul	980
Aug	1346
Sep	637
Oct	176
Nov	2
Dec	1

4.2.2 Data Normalization

After filling all the missing data, we should do the data normalization to facilitate the further analysis of the data. It can provide an approach for comparison of different kinds of data and reflect the combined results of different factors.

Here we use Min-max normalization, doing linear transformation of the raw data, mapping data set

$$x = \{x_1, x_2, \dots, x_n\}$$

into [0,1] interval. Let *x'* be the normalized value, the formula is:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

While x'_{min}= 0 and x'_{max}= 1. And all the data hereinafter is normalized data.

4.2.3 Information Entropy

According to the definition of information entropy in Information Theory[5], the information entropy of a set of data is

$$E_j = -\frac{1}{ln\, n}\sum_{i=1}^{n} p_{ij} ln\, p_{ij} \tag{2}$$

And

$$p_{ij} = \frac{x'_{ij}}{\sum_i^n x_i} \tag{3}$$

If $p_{ij} = 0$, then

$$\lim_{p_{ij}\to 0} p_{ij} ln\, p_{ij} = 0 \tag{4}$$

4.2.4 Weight of Each indicator

According to the formula of information entropy, the information entropy of each index is calculated as

$$E = \{E_1, E_2, \ldots, E_n\}$$

We can calculate the weight of each index by information entropy:

$$W_i = \frac{1 - E_i}{k - \sum E_i}, i = 1,2,\ldots,k \tag{5}$$

For every indicator, the entropy and weight will be

Table. 4. 3. Entropy and Weight of sector

Entropy			Weight		
Spatial Location Score	Life Cycle Score	Report Number Score	Spatial Location Score	Life Cycle Score	Report Number Score
0.660806	0.999422	0.980084	0.943022	0.001608	0.055370

And we can get the combined score for every textual data information.

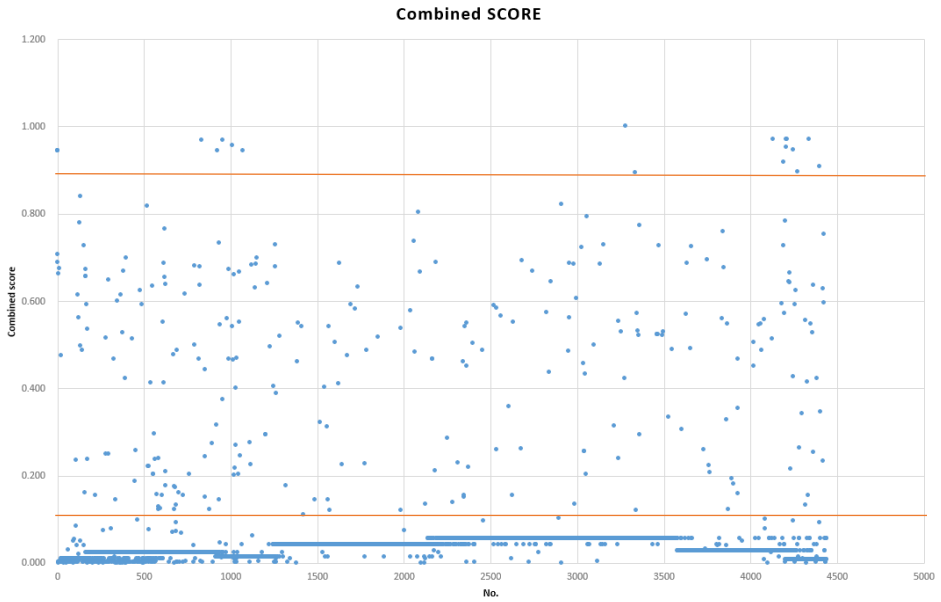


Fig. 4.2. Combined score for each data

It shows vividly that only about a dozen reports get scores over 0.9, which have a "Positive ID" lab status; and plenty of reports get scores under 0.1, which have a "Negative ID" lab status. The rest are distributed in the interval [0.1, 0.9]. The textual information evaluation meets the requirements.

4.3 Picture Information Evaluation

4.3.1 Selection basis of Model

We have obtained the standardized and unified positive and negative sample images of hornet. The next step is to classify the data samples, referring to previous research[6], mostly using “Pyradiomics” library to extract features, and then use the classifier to do the classify training. We can select classifier such as SVM classification, “Random Forest” classification, etc. Although the results given by the classifier have differences, basically they are similar, and the correct rate of classification is less than 90%. So the bottleneck of improving the correct rate of classification is not in the selection of classifier, but in the feature extraction.

“Pyradiomics” is a feature extraction package for Python with the following five categories of extractable features.

- First Order Statistics features
- Shape-based features
- Gray Level Cooccurrence Matrix (GLCM)
- Gray Level Run Length Matrix (GLRLM)
- Gray Level Size Zone Matrix (GLSZM)

These features are calculated based on the sample data and have obvious physical meaning, such as first-order features, shape features, etc. The unsatisfactory final results of classification training using these features also indicate the problem that features with obvious physical meaning have limited ability to characterize the wasp category, so a new idea is naturally to seek a different way to extract new features, our attention is focused on convolutional neural networks (CNN).

Currently, the mainstream approach proposed for training convolutional neural networks on small data sets is transfer learning[7]. The idea of this method is to use a large data set for model training, and then use the trained parameters to fine-tune the parameters of small samples, so that the final trained model has the dual advantages of small data set and high accuracy.

There are many models available on the web for migration learning. Taking the number of model parameters, model performance, model availability, model training computational resource requirements, model ease of use, etc. into account, this paper selects Google's open-source “Inception-v3” model as the migration learning framework, which is trained on the ImageNet dataset, which has 1000 The model

achieves excellent performance on the ImageNet dataset, which has more than 1 million images in 1000 categories.

4.3.2 Model Construction

After choosing the “Inception-v3” model, the approach in this sector is to use the output of the bottleneck layer of this model as the extracted features, and then put these features through a fully connected layer for classification, and the schematic diagram of the improved Inception-v3 model is as follows[8].

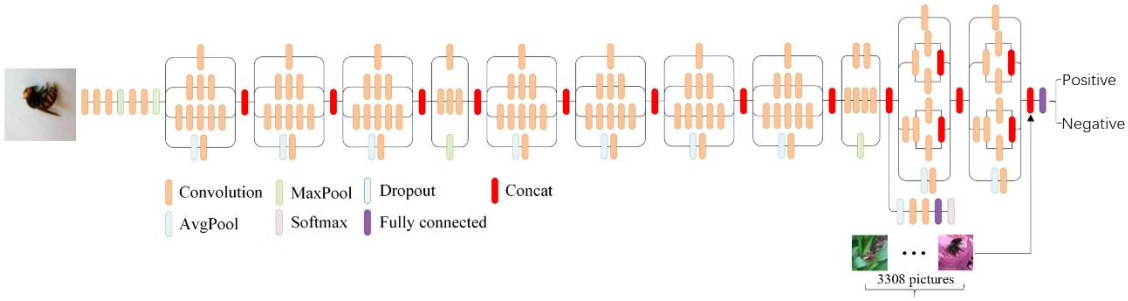


Fig. 4.3. Network of the Inception-v3 model

In the above figure, the rightmost arrow position is the location of the feature output, and then the result of the bottleneck layer is used as the feature extraction result of the new model in this paper, which is a 2048-length feature vector.

After obtaining this feature vector, it needs to be fed into a fully connected layer for classification, and the model schematic of the fully connected layer is as follows.

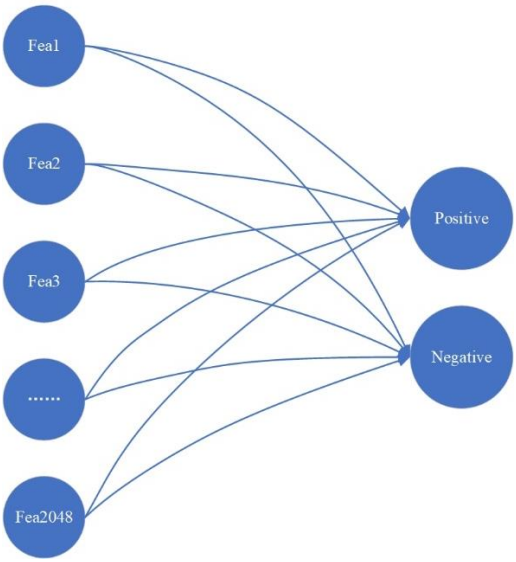


Fig. 4.4. Classification Model

The above two networks are connected together to form the complete model, where the input of the model is the image of hornet and the output is the judgment of wasp

species.

4.3.3 Data Enhancement

Image data preparation has an important impact on the training of neural network and convolutional neural network models. When the sample space is insufficient or the number of samples is not enough, the training is seriously affected, or the trained models are not generalized enough and the recognition rate and accuracy are not high. Data enhancement methods can greatly increase the sample size of the training data set and increase the generalization ability of the network model. In essence, data augmentation is the process of artificially increasing the sample size of a dataset through data transformation. Based on the previous pre-processing of image data and the requirements of the Inception-V3 model, the image enhancement step is a necessary step to identify and classify the hornet images.

Here we use the “Imgaug” library in python. “Imgaug” is a library for image augmentation in machine learning experiments. It supports a wide range of augmentation techniques, allows to easily combine these and to execute them in random order or on multiple CPU cores, can not only augment images, but also key points/landmarks, bounding boxes, heatmaps and segmentation maps. It shows as follows after the data enhancement of hornet images with “Positive ID”.

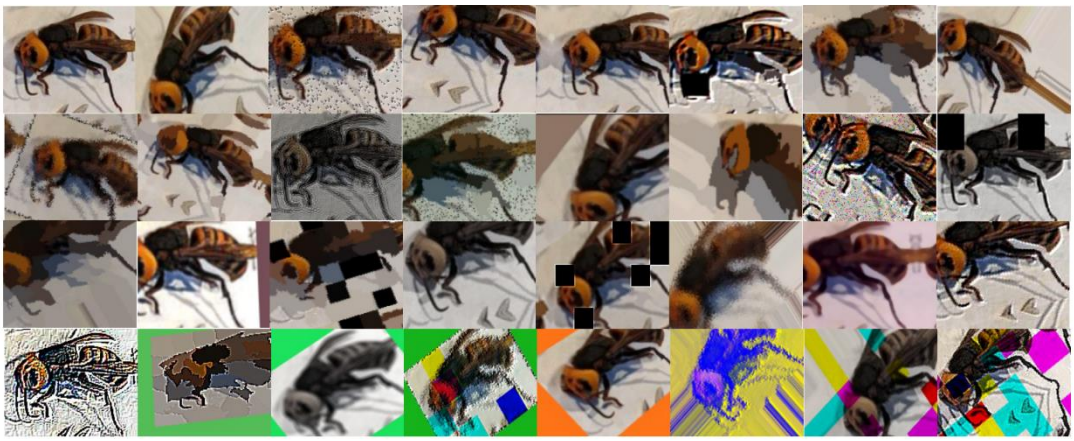


Fig. 4.5. Result of image enhancement

This increases the original number of positive samples from 14 to 576.

4.3.4 Model Training and Analysis

In order to utilize the computational resources, save time and improve the model training effect, it is necessary to design the model structure reasonably and set reasonable training parameters to improve the model training results and reduce the training times.

The model parameters mainly include the following: validation set and test set percentage, learning rate, number of input images at a time, and number of training sessions.

- The validation set and test set percentage: the validation set is used to test the training effect of the model in real time during the training process, while the test set is used to test the correctness of the model at the end. Considering that the validation set and test set need to be able to test the effect of the model, so the amount of data cannot be too small, and at the same time, in order to effectively use the data for training, it is not possible to divide too much data out, so after comprehensive evaluation, we choose to use 10% of the data as the validation set and test set.
- The number of training sessions: the number of training sessions is the number of times the model is trained. Too few training sessions are not effective in classification, but too many sessions are a waste of time. In this paper, we use a small amount of data for trial training, and we find that 10,000 times of training can make the model stable, so we choose 10,000 times as the training number.
- The learning rate: the learning rate is a scaling factor that adjusts the weights during the training process. A too large learning rate will cause the model to fluctuate and not converge, while a learning rate that is too small will make the model converge too slowly and waste training time and computational resources. A large learning rate of 0.1 was first used for testing and then continuously reduced, and finally an exponentially decreasing learning rate was chosen with a decay formula of

$$LR = BLR \times e^{-\frac{i \times ln(0.1/MLR)}{STEPS}} \tag{6}$$

LR is Learning Rate.

BLR stands for Base Learning Rate, and BLR=0.1.

MLR is Minimum Learning Rate, and MLR=0.001;

STEPS means the number of training sessions, and we choose STEPS=10000.

i is the current training times from 0 to 9999.

The learning rate shows as below.

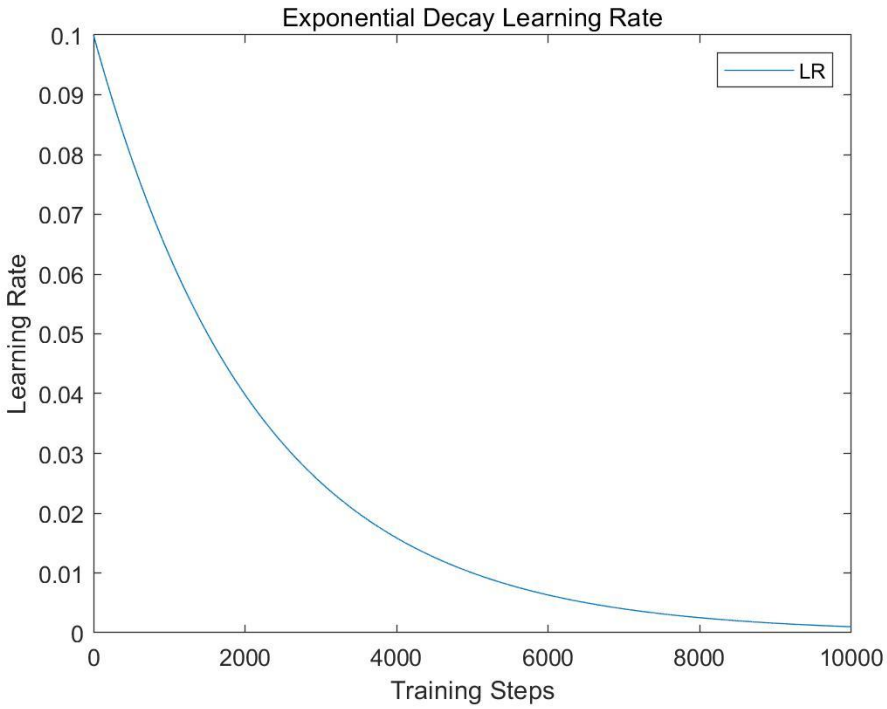
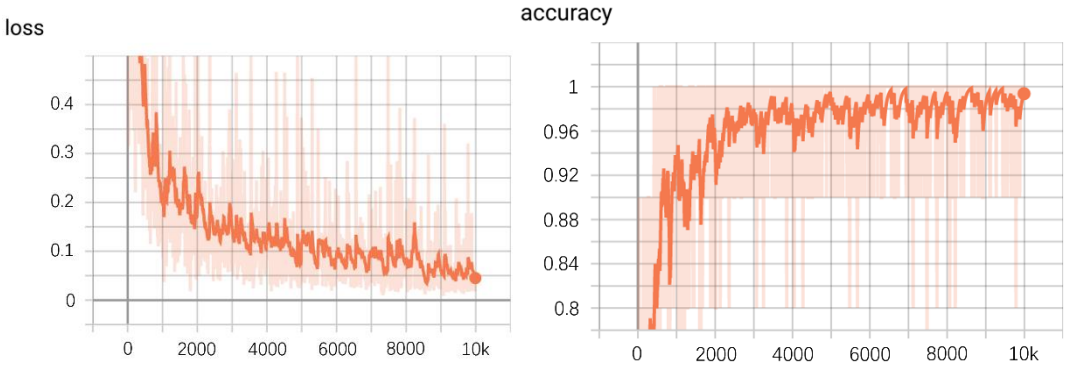


Fig. 4.6. Change of learning rate

The network accuracy and loss function during the training process are as follows.



(a) The loss (b) The accuracy

Fig. 4. 7. The network training process

As you can see, the accuracy of the training to the final test set can reach about 96%.

4.4 Comprehensive Evaluation System

As it is said in the overall idea, after obtaining the textual information score and picture information score, we analyze when these two should be prioritized. When the picture scores are very close to 1 and 0, Lab can very clearly determine whether the hornets were truly found, and the priority should be high at that time. Meanwhile, the report with a medium picture information score cannot tell vividly whether the insect hornet is. So the lab ought to deal with them later.

We decided to use normal distribution $f(x|\mu,\sigma)$ to assign weights to these two ratings, and finally form a complete and comprehensive evaluation system.

For a normal distribution,

$$y = f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{7}$$

We decide $\sigma = 0.199474$, $\mu = 0.5$, and normalize y.

The weight of picture information score ($0.5\times a$) is the normalized y, and the weight of textual information score is (1-$0.5\times a$).

$$Score_{total} = 0.5 \times a \times Score_{pic} + (1 - 0.5 \times a) \times Score_{text} \tag{8}$$

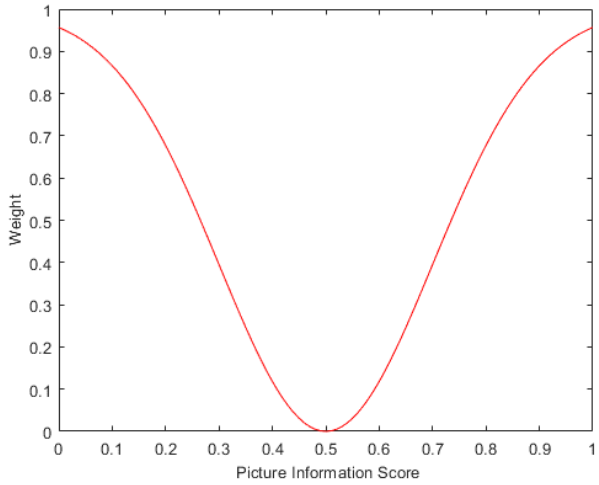


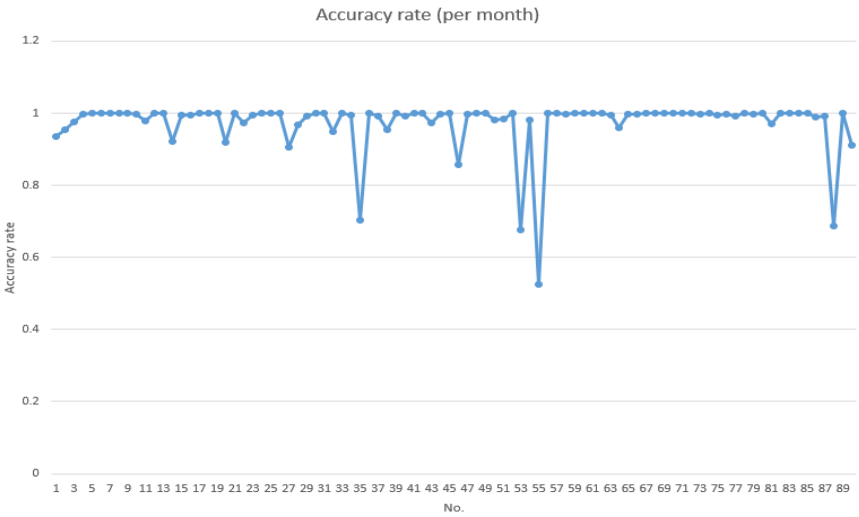
Fig. 4.8. Weight of picture information score

4.5 Model Discuss and Accuracy Analysis

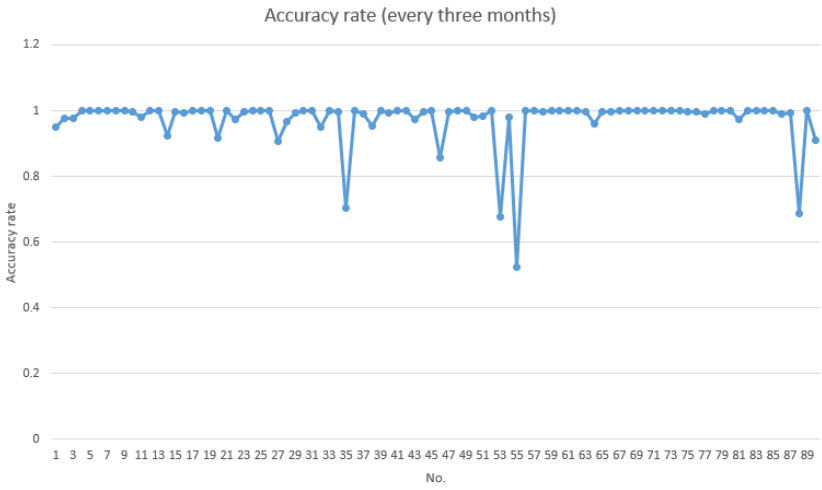
With the comprehensive evaluation system, we can deal with how to prioritize the investigation of the reports. When the report is sent by citizens, the analysis model can evaluate it immediately and give a comprehensive score. The lab researchers will prioritize the report with the highest score, which means it is most likely or least likely the hornet was found.

As to how and how often the updates should occur, we built up a testing group with 91 pieces of data in October 2020 (1 Positive ID and 90 Negative ID) to test the accuracy of the evaluation model. We decide to update the data monthly and every three months. The accuracy rates are separately calculated.

The average accuracy rate of the model is shown as below.



(a) Per month



(b) Every three months

Fig. 4.9. Accuracy rate of test

The average accuracy rate with a per month update is 96.3522%.

The average accuracy rate with a per three months update is 96.3921%.

It shows that the update frequency has little influence in the model itself. The lab and government do not need to update the data frequently. As a result, we choose three months as an update periodicity.

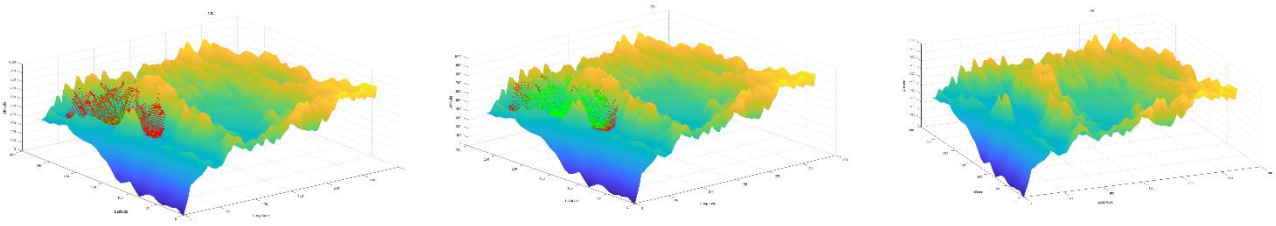
5 Evidence of Hornet Eradication

In this part, we will deal with how to know whether the hornet has been eradicated in Washington State. Just like the disease outbreak, the new nesting of hornet should be stopped and eradicated, otherwise the spread of it will be very rapid, like in Chapter 3.

The model Cellular Automaton (CA) will still be used to predict the evidence of hornet eradication. In cellular automaton, there is a parameter d. And d is the rate of the elimination or the death of the nesting location. In Chapter 3 we decided $d = 0$, which

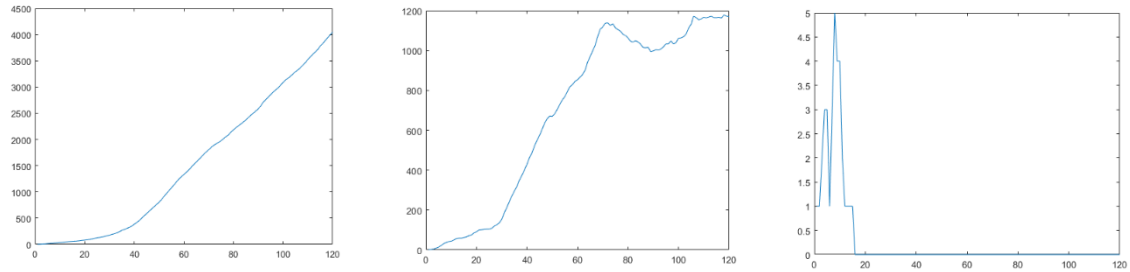
means no action will be taken to stop the spread of nesting. When the local government wants to eradicate the hornets, measures such as searching nests, clearing adult worms and destroying nests must be taken.

In the following section, we let d= 0.005, 0.05, 0.5 respectively and run the model to know how well the eradication is. The other parameters and the initial point are the same as in Chapter 3. The results are shown as below.



(a) d= 0.005 (b) d= 0.05 (c) d= 0.5

Fig. 5.1. Distribution of nesting by eradication



(a) d=0.005 (b) d= 0.05 (c) d= 0.5

Fig. 5.1. Number of nesting by eradication

The green points mean the destroyed nests after finding, and the red points mean the nests which are not eradicated. As it is shown above, when d= 0.5, the pest is considered to have been eradicated because it has not been reported after about 20 months when the predicted number of reports is 0. So, if the local government takes every action to stop the spread, the hornet can be eradicated in 20 months.

6 Conclusions

In our paper, we have done lots of data processing work, built up the prediction and the evaluation model system for the hornet report and constructed three kinds of model. We use many theories and methods such as CA, EWM, Inception-v3 and other geographic theories to complete our work.

6.1 Strengths

- Comprehensive: We comprehensively use almost all the data except less useful Notes, making evaluation criteria more credible.
- Technical Supporting: We use the advanced transfer learning theory for computer vision, which is highly precise with small amount of data.

- Flexible and Extendable: To include additional factors, the model could be extended to apply to many situations. And the model can be upgraded with instant data if they wish.

6.2 Weaknesses

- Lack of Data Support: The data given by the question is not big enough. There are only 14 positive samples and too many negative samples, which did a bit of difficult to us.
- Ignoring the note factor: The notes sent by citizens are mostly clutter and not useful. We did not use the natural language processing (NLP) to take notes into consideration, which can be not so integrity.
- Simplifying Assumptions: Simplifying assumptions had to be made to create a solvable model. Thus, some valuable data and information are unable to use.

7 Future Work

Due to limitation of time, lack of accessible data and experience in related field, our model is not robust based on the real world. And from the results we can see, the chosen university is not very ideal. We cannot figure out intuitively why we should invest them. So future work may process in this field.

- Use natural language processing to consider notes
  Documents contain elements such as words, phrases, sentences and paragraphs. In most text classification methods, these elements appearing in the text are used as text features, and as the level of the elements increases, the clearer the semantics they express and the richer the information attached to them, but the number of feature combinations will also be larger, therefore, sentences and paragraphs are rarely used as features. According to researchers' experiments, the common feature item representations are: words, phrases (Phrase) and N-gram items, etc.
  The bag-of-words model treats a piece of text as a separated word and classifies the text by the difference of words that may occur in different classes of text. It is important to note that for text classification, the context has a strong relevance for its true class. Personally, I understand that such methods may selectively discard the contextual information in the text for the sake of simplicity of the algorithm, or it may be possible to obtain the required text classification results by focusing only on word frequency information.
- More data is needed.
  First, more real data should be filled in the blank of the missing data, which can reduce the error resulting from the data imputation.
  Second, we should find more data for more information of hornet and the state. Having more data means we can rethink whether the attributes we have selected are suitable and replaced the improper ones. Having more data means we can do more detailed analysis from the dimension of time.

References

[1] Vasmer Michael,Browne Dan E. & Kubica Aleksander.(2021).Cellular automaton decoders for topological quantum codes with noisy measurements and beyond. Scientific Reports(1),. doi:.

[2] Andrew A, Ramon A S, Lawniczak A, et al. Automata-2008: Theory and applications of cellular automata[M]. London, UK; Luniver Press, 2008.

[3] Matsuura, Makoto; Sakagami, Shôichi F. (1973). "A Bionomic Sketch of the Giant Hornet, Vespa mandarinia, a Serious Pest for Japanese Apiculture (With 12 Text-figures and 5 Tables)". 北海道大學理學部紀要. 19 (1): 125–162. hdl:2115/27557. 11111111111

[4] Wang Wanyu,Li Haochen,Hou Xueliang,Zhang Qian & Tian Songfeng.(2021).Multi-Criteria Evaluation of Distributed Energy System Based on Order Relation-Anti-Entropy Weight Method. Energies(1),. doi:.

[5] Yongjun Li,,Qian Zhang,,Lizheng Wang & Liang Liang.(2020).Regional environmental efficiency in China: An empirical analysis based on entropy weight method and non-parametric models. Journal of Cleaner Production(),. doi:.

[6] Liu, Q. L. , Jiang, P. , Jiang, Y. H. , Ge, H. J. , & Li, Y. X. . (2019). Prediction of aneurysm stability using a machine learning model based on pyradiomics-derived morphological features. Stroke, 50(9).

[7] Feng Zeng,Xingcun Li,Xiaoheng Deng,Lan Yao & Guanghui Lian.(2021).An image classification model based on transfer learning for ulcerative proctitis. Multimedia Systems(prepublish),. doi:.

[8] Mehrotra Rajat,Ansari M.A.,Agrawal Rajeev & Anand R.S..(2020).A Transfer Learning approach for AI-based classification of brain tumors. Machine Learning with Applications(),. doi:.