Fine-grained image classification of microscopic insect pest species: Western Flower Thrips and Plague Thrips

Don Chathurika Amarathunga[a,*], Malika Nisal Ratnayake[a], John Grundy[b], Alan Dorin[a]

[a] *Computational and Collective Intelligence, Department Data Science and AI, Faculty of Information Technology, Monash University, Wellington Rd, Clayton VIC 3800, Australia.*
[b] *Humanise Lab, Department of Software Systems and Cybersecurity, Faculty of Information Technology, Monash University, Wellington Rd, Clayton VIC 3800, Australia*

Abstract

Accurate identification of insect pests is essential in crop management as they are one of the primary causes of yield losses. However, differences between insect species demand different pest control strategies. Hence, research on new technology for fine-grained classification of insect pests is potentially important. Morphologically similar microscopic pest species classification has received little attention in the literature, and is targeted by this study as a means to address the need for agricultural pest management. We propose a novel computational method for deep learning-based, fine-grained classification of microscopic insects using the Vision Transform (ViT) architecture. This architecture employs an attention mechanism motivated by domain knowledge. The proposed approach consists of two main modules, a Data Preprocessing Module to segment relevant insect features and split the insect into body segments to inform identification, and a Domain Knowledge-Driven Stacked Model based on ViT to generate the prediction from each body segment and to fuse predictions for each segment into an accurate species-level classification. We tested the approach using an image dataset of two economically devastating thrip species - Western Flower thrips (*Frankliniella occidentalis*) and Plague thrips (*Thrips imaginis*). These insects are small ($\sim 1mm$), exhibit minute inter-species differences, and require different pest control strategies. We compared our model with the original ViT model, RestNet101, and RestNet50. Experimental results achieve an F1-score of 0.978, a 3.27% improvement over the baselines. This is important in the horticultural context given the yield losses that these pest insects are known to cause if their populations remain incorrectly quantified.

Keywords: Pest Monitoring, Integrated Pest Management (IPM), Image Classification, Insect Classification, Machine Learning

*Corresponding author
Email addresses: don.amarathunga@monash.edu (Don Chathurika Amarathunga), malika.ratnayake@monash.edu (Malika Nisal Ratnayake), John.Grundy@monash.edu (John Grundy), Alan.Dorin@monash.edu (Alan Dorin)

1. Introduction

Global food demand is expected to increase 56% by 2050, posing serious challenges for food security[1]. To help meet this demand, insect pests, a primary cause of crop yield loss (García-Lara and Saldivar, 2016), should be detected early, before they cause major damage to food (Cui et al., 2018). The diversity of insects generally is organized into a hierarchy of 28 orders (Gullan and Cranston, 2014), then into more fine-grained family, genus and species categories. In an agricultural context, the identification of insect pests at coarse-grained taxonomic levels may be insufficient, because pest control strategies often differ between the finer insect classification levels. However, species-level insect classification can be particularly difficult as inter-specific differences gradually decrease as we move from taxonomic rankings at order level down to species level (Ratnayake et al., 2022). Fine level identification of tiny pest insects is also challenging in the field without good magnification, since the closeness of related species' morphology can be challenging to discern outside controlled lab conditions. Consequently, to improve insect pest management, in this work we focus on classifying tiny insect pests to species level.

Thrips have been a major worldwide crop and plant pest since around 1793, negatively affecting vegetables, fruits and ornamental plants (Lewis et al., 1973; Marsham, 1797). Thrips are small in size ($\sim 1mm$) and differences in appearance between some key species can be difficult to discern. However, their accurate identification to species level is essential to agriculture (Lewis et al., 1973; Loomans et al., 1995): (1) multiple thrip species are often found on a single host plant; (2) not all thrip species are harmful, some are beneficial; (3) damage inflicted on crops by different thrip species varies; and (4) pest management and control strategies vary between thrip species. For example, *Haplothrips spp.*, Plague thrips (*Thrips imaginis*) and Western Flower thrips (*Frankliniella occidentalis*) (WFTs) can be found on Australian strawberry farms (Steiner and Goodwin, 2005). While Plague thrips and WFTs are harmful to strawberries, *Haplothrips spp.* are beneficial predators of other tiny pest species such as spider mites (*Tetranychidae*)(Luong, 2008; Steiner and Goodwin, 2005). Plague thrips can be controlled using chemicals, but WFTs are controlled using biological control agents (predatory bugs) as they are resistant to many chemicals (Sampson et al., 2014; Mouden et al., 2017). Such issues justify the importance of the study of thrip species identification, as explored in this research.

Various techniques are used for accurate and effective identification of tiny pest species including thrips. In laboratories, insects are manually classified by observing their morphological traits - *keys* under a microscope. This is time-consuming and labour-intensive. The process is also potentially error-prone if observers lack domain knowledge on insect taxonomy, and clear identification keys that account for specific distinguishing features and subtle differences in insects' morphological structure. Even skilled entomologists find classification to be difficult due to the vast number of insect species (Gaston and O'Neill, 2004). In response to these challenges, recent research has focused on automating insect classification using computer vision, image processing, and other computer science technologies (Amarathunga et al., 2021).

Following recent advancements in computer vision and computer processing power, and the popularity of high-quality image capture devices, there has been an increasing interest in automating the insect classification task using their images. Recently, several review and survey papers have been published that analyse the literature on existing image-based insect or/and pest identification systems and report the gaps in this line of research (Martineau et al., 2017; Júnior and Rieder, 2020; Amarathunga et al., 2021). Based on their analysis, species classification methods can be divided into shallow and deep learning based approaches (Amarathunga et al., 2021). Existing insect classification techniques that extract predefined, hand-engineered features from insect images, and then adopt machine learning models with few hidden layers for image classification, are labelled shallow learning methods. These typically adapt shallow machine learning models such as Support Vector Machine (SVM) (Noble, 2006), K-Nearest Neighbours (KNN) (Cunningham and Delany, 2021), and Naïve Bayes (Murphy et al., 2006) for classification. However, it is challenging to derive discriminating features from an insect image for classification since there are many species and variants of similar size and shape. In contrast, deep learning approaches (e.g., AlexNet (Krizhevsky et al., 2012), ResNet (He et al., 2016), Vision Transformer (ViT) (Dosovitskiy et al., 2020)) adopt neural network architectures with many hidden layers. Recently, deep learning techniques, such as Convolution Neural Networks (CNN), have attracted greater research interest than shallow machine learning methods in the domain of species-level pest classification (Amarathunga et al., 2021). This is partly due to their ability to automatically extract relevant species identification features from images, and partly due to their state-of-the-art performance for image classification in the domain (Wu et al., 2019; Amarathunga et al., 2021). ViT is a recent advancement in computer vision used as an alternative to CNN-based architectures. The approach is capable of assigning varying importance to different image regions (i.e., patches) when performing classification tasks (Dosovitskiy et al., 2020). Recent studies demonstrate ViT's

[1] the World Resources Institute

ability to outperform CNN-based architectures for general image classification tasks, but the approach is not currently well studied for insect classification specifically. Nevertheless, in general, these insect classification models can perform poorly in differentiating morphologically close tiny insect species, such as the thrips that are the subject of our study. Hence, although there are a few studies (Xia et al., 2015; Li et al., 2021; Espinoza et al., 2016) particularly aimed at generalizing such models for tiny insects, they have to date only been conducted at a coarse-grained level.

Thrips' small size puts them into the class of microscopic insects. Insect identification systems proposed in (Solis-Sánchez et al., 2011; Xia et al., 2015; Espinoza et al., 2016; Ebrahimi et al., 2017; Lu et al., 2019; Rustia et al., 2020; Li and Yang, 2020; Li et al., 2021) considered these tiny insects as a target category, but have the ability to distinguish them only from other insect orders. Some studies (Fedor et al., 2008, 2014) analyse thrip morphometric details (the qualitative and quantitative traits of an insect's body) from microscope images for species level classification. These studies though, required manual measurement or calculation of morphometric traits from the microscope images which were then fed as features to an Feed Forward Neural Network (FFNN). The measurement process takes considerable time and effort if it is to be precise, and remains potentially subject to measurement errors caused by the thrips' small size. These are issues that automated, computational, image-based measurement algorithms could potentially address.

A properly labelled image dataset is also required to test and evaluate the performance of any image-based insect classification system. While several datasets have been published for insect identification tasks (Amarathunga et al., 2021), many either consist of low-resolution images or they do not cover morphologically similar, tiny insect species such as thrips (Amarathunga et al., 2021). Thus, the existing datasets do not meet the requirements of this study to train classification models of morphologically similar thrips. This is our motivation for publishing an image dataset for future reference along with the present article.

To address the requirements outlined above, **we propose a novel Vision Transformer (ViT)-based architecture, with an attention mechanism motivated by domain knowledge, to conduct species level classification of thrips. The approach has the potential to be extended to other tiny pest species such as fire ants, an important invasive species that could severely damage the environment, and the agriculture (Tschinkel, 2013), and flea species that are pests of livestock (Iannino et al., 2017).** The key contributions of this research can be listed as follows:

- We provide a new public dataset containing microscopic images of two harmful and morphologically close thrip species – WFTs and Plague thrips;
- We propose a novel image-processing pipeline to segment relevant insect features from an image background, and to split the components of the insect into constituent body segments that inform identification of inter-species morphological differences;
- We propose a Domain Knowledge-Driven Stacked Model based on ViT to conduct an accurate species-level classification of morphologically close insect species; and
- We evaluate our model using the thrip dataset to demonstrate its performance improvements against the previous state-of-the-art.

2. Methodology

In this section, we present our methodology for the trip classification. We first construct an image dataset of two thrip species (Section 2.1). Our classification approach consists of a Data Processing module and a Domain Knowledge-Driven Stacked Model. The Data Preprocessing Module segments relevant insect features and splits the insect into body segments to inform identification (Sections 2.2, 2.3, and 2.4). The Domain Knowledge-Driven Stacked Model generates the prediction from each body segment and fuses predictions for each segment into an accurate species-level classification (Section 2.5.2).

2.1. Data Collection

We collected thrip samples from the Sunny Ridge strawberry farm located at Boneo, Victoria, Australia during December 2020 to May 2021. Strawberry flowers were collected in plastic specimen tubs in 70% of ethanol and transported to the lab. Thrips were then extracted from flowers and mounted on microscope slides. A handheld DinoLite digital microscope (USB) was used to capture images of individual thrip specimens. Thrip species have a very close morphological structure and domain knowledge is a crucial factor for the visual identification of thrip species. Entomologists generally use the diagnostic aids called 'keys' to distinguish thrip species. Those keys represent different morphological characters, which can be used to discriminate two or more

species from each other (Mehle and Trdan, 2012). For example, a WFT has banded antenna segments, while a plague thrip does not (Figure 1)[2]. We followed the guidelines provided by the website of the NSW Department of Primary Industries to identify images of WFTs and Plague thrips using their keys to construct our labelled dataset [3]. We labelled the images and 10% of labeled images were verified by the domain experts for the correctness. Due to the high magnification, some body parts of the specimens may not be focused while taking images causing not enough visible identification keys to recognise WFTs and plague thrips. Hence, we took multiple images of each specimen focusing different body parts and then analysed each image for identification keys to decide the thrip species (i.e., WFT or plague) that the specimen belonging to. Then one image from the corresponding image set of a specimen is randomly selected to create the dataset. Table 1 provides a summary of the final dataset (the dataset is available in https://drive.google.com/drive/thrip-images-dataset).

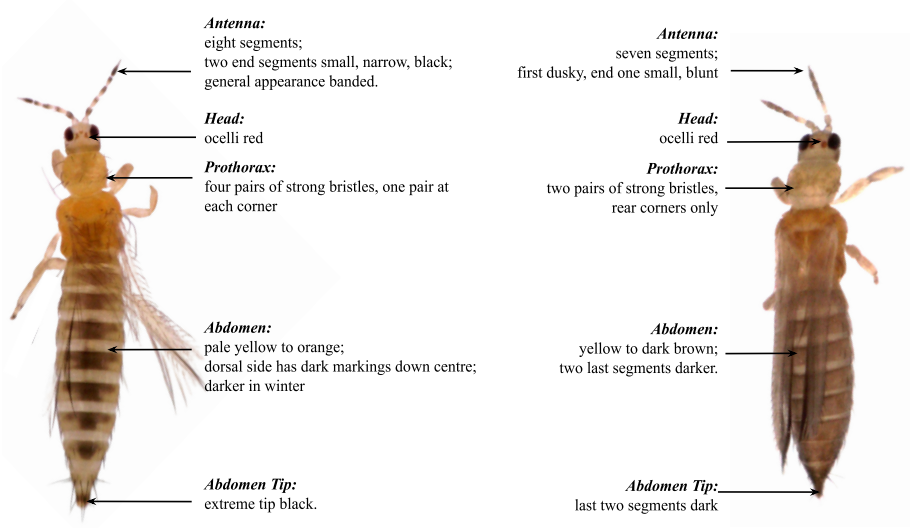

Figure 1: Some Differentiating characteristics (keys) for WFT and plague thrips under low power magnification (40x)

Table 1: Dataset Properties

Method	WFTs	Plague Thrips
Flower Sampling	210	215

Figure 2: Randomly selected images from the dataset; first row and second row contain images that are labelled as WFTs and plague thrips respectively

2.2. Early Data Preprocessing (EDP)

There are several reasons why image preprocessing is beneficial in our application. In some images, the thrip body only covers a small portion of the image. Thus, classification model performance can potentially be

[2]http://www.dpi.nsw.gov.au/data/strawberry-thrips.pdf
[3]https://www.dpi.nsw.gov.au/search?collection=&query=WFT

improved by feeding it only the region of interest (ROI) containing the thrip. Also, image background colour can vary depending on the source of the image and model predictions may be incorrectly biased by this. Lastly, sticky trap images are especially prone to include small non-insect particles such as pollen, dust and leaf parts. Thus, we propose an image preprocessing pipeline to extract the ROI from the image and remove unwanted background and debris (Figure 3).

First, the image is converted to greyscale, and binary thresholding is performed on this to separate the foreground from the background. The morphological dilation operation is applied to expand the boundary of the foreground pixels which will be used to join broken thrip body parts in the resultant image[4]. Then, the contours of the binary image are detected. The parent contour (see Fig. 3(b)) of the image gives the coordinates of the boundary pixels of the foreground (i.e., the thrip's body). Then, a mask image is generated that contains white pixels inside the parent contour and black pixels in the rest of the image as shown in Fig. 3(c). By multiplying the original image with its mask, we separated the ROI from the original image which results in the image of Fig. 3(d). Subsequently, the bounding rectangle of the parent contour and its rotated angle are calculated (Figure 3(e))[5]. The rotated angle provides the angle of rotation of the thrip body with respect to the X-axis to be used to horizontally align the thrip as shown in Fig. 3(f).

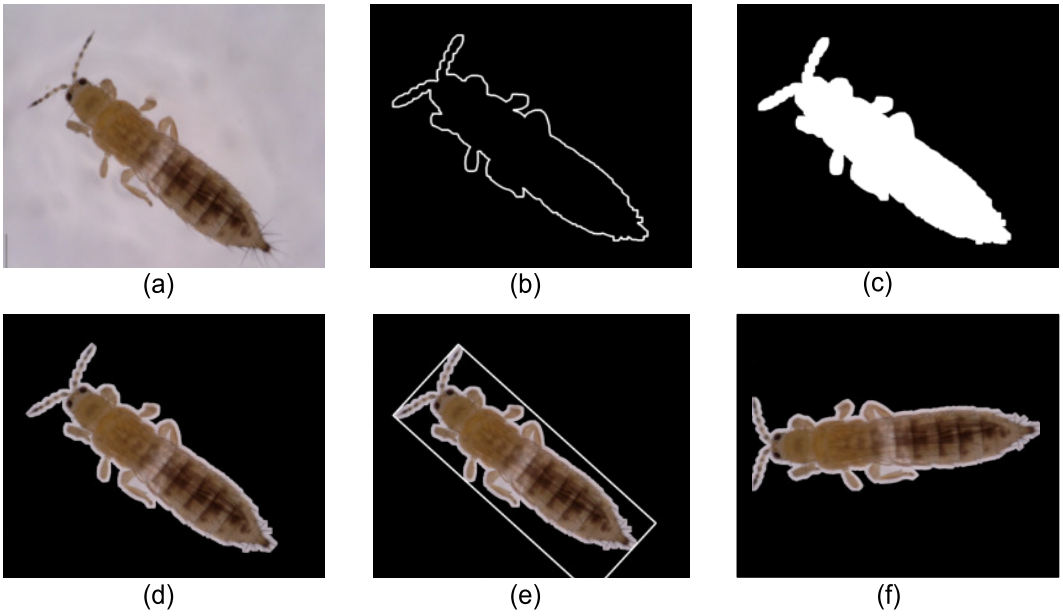

Figure 3: Early Data Preprocessing (EDP). (a) original image; (b) parent contour; (c) mask generated from the contour; (d) image after masking; (e) bounding rectangle; and (f) rotated image.

Thrip heads in the horizontally aligned-images can be directed towards the left or right. To support the subsequent module in our pipeline, we must rotate all images in our dataset consistently. Here, we propose a novel *directional classifier* to identify thrip heads and abdomens for consistent alignment.

2.3. Directional Classifier (DC)

For directional classification we propose two types of geometric features based on the planar insect surface area and its silhouette, as represented by foreground pixel distributions in the head/abdomen regions that will assist discrimination of the head from the abdomen of thrips.

As shown in Figure 4(a), the foreground pixel count presented as a changing measure along the direction of the positive horizontal axis, is smoother for the abdomen than the head region due to the presence of insect antennae, thorax and forelegs. To capture this difference in level of fluctuation, we fit second degree polynomial curves for the pixel plots of the left and right image quarters and generate one feature (fea_1) as follows:

$$fea_1 = \begin{cases} 1, & \text{if } MSE_{left} > MSE_{right} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

[4]https://docs.opencv/morphologica_operation.html
[5]https://docs.opencv.org/contour_properties

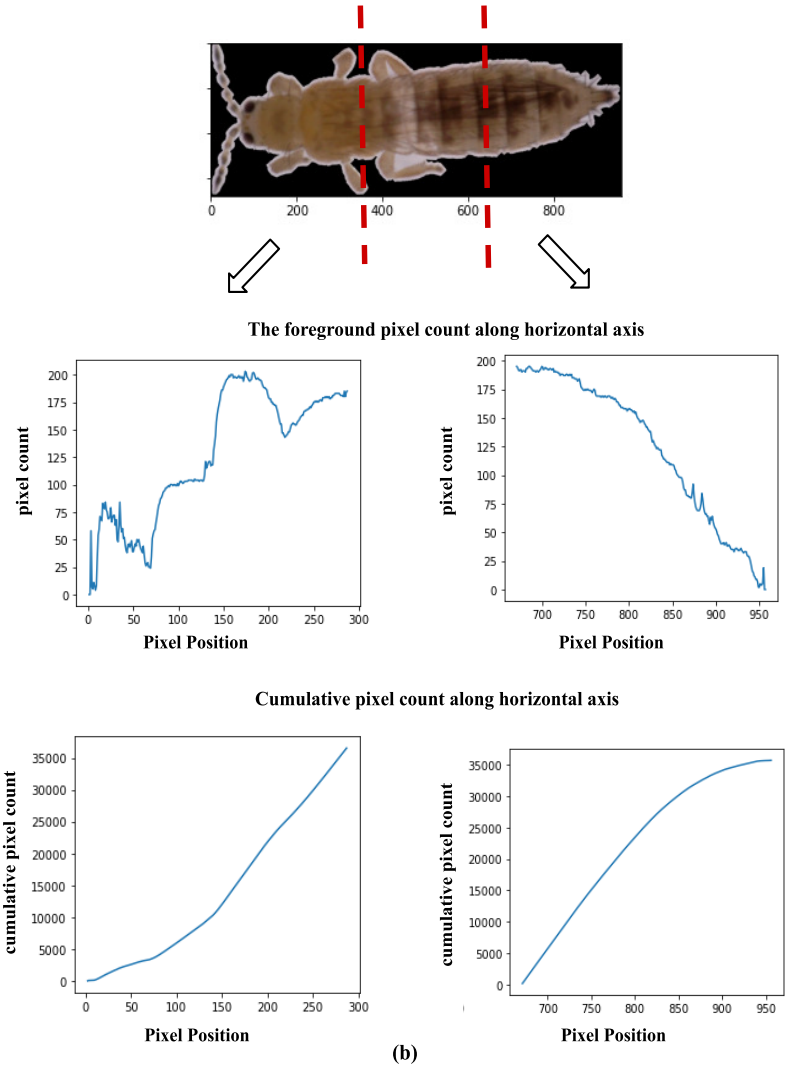

Figure 4: The foreground pixel count along the direction of the positive horizontal axis (a), and the cumulative pixel count along the direction of the positive horizontal axis (b) for head and abdominal areas.

where MSE is the *mean square error* of the curves fitted for the left (R^2_{left}) and right (R^2_{right}) sides. This measures how well the foreground pixel fluctuations can be approximated using a second-degree polynomial. From analytical evidence, MSE_{head} is usually greater than $MSE_{abdomen}$ due to the sudden fluctuations in the silhouette of the head region. Thus, we set fea_1 to 1 if the region with sudden fluctuations (i.e., the head region according to our evidence) is located in the left part of the image.

A second feature (fea_2) we implement can be interpreted as number of foreground pixels in head and abdomen regions. Due to the two antennas, head region of a thrip generally has less number of pixels compared to abdomen area. Hence, we set fea_2 to 1 if lesser foreground pixel count in left part of the image ($Total\_Pixels_{left}$) compared to right ($Total\_Pixels_{right}$), and 0 otherwise.

$$fea_2 = \begin{cases} 1, & \text{if } Total\_Pixels_{left} < Total\_Pixels_{right} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

To formally assess the effectiveness of selected two features individually and as an ensemble, we manually labelled resultant image dataset after applying EDP module as follows:

$$direction\_label = \begin{cases} 1, & \text{if the head is in the left side} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

In order to calculate our features, we first divide the left-most and right-most $p\%$ of the horizontally-aligned images for analysis to compute the features of these regions that will assist discrimination of the head from

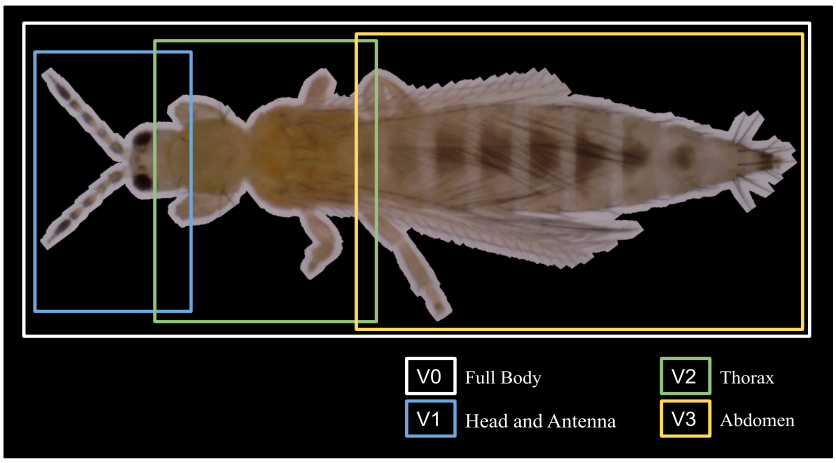

Figure 5: The segments of the thrip body used to generate augmented subsections from the original image

the abdomen. To find the optimum $p\%$, we alter the $p\%$ within the range of $10\% - 50\%$ and calculate the f1-score for the direction labels predicted by each feature for the training dataset. Based on the results, we select $p\% = 33\%$, which gives the highest f1-score in this experiment.

Next, we trained a decision tree classifier (called a *directional classifier* from hereon), using the training dataset in our experiments to predict a strong label for the head region using our two features $\{fea_1, fea_2\}$. See the ablation study related to the directional classifier in Section 3 for detail. In our pipeline, we adopt this directional classifier to identify the head and abdomen and rotate all images such that the thrip's head is directed leftwards.

2.4. Domain Knowledge-driven Data Augmentation (DDA)

Thrip species identification from images is difficult for many reasons. Thrips typically show large intra-species variations e.g., colour and size differences between males and females (sexual dimorphism) and colour variation due to seasonal changes[6]. They also potentially exhibit pose variation when they adhere "naturally" to microscope slides or sticky traps. The fact that different thrip species may exhibit small inter-species variation, can also make it difficult to tell species apart, especially by reference to only a single feature. Since thrips are small, high magnification is required to capture them in clear images. Under high magnification, some specimen body parts may be unfocused due to the short lens depths-of-field typical in macro-photography, unless complex focus stacking techniques are employed (Cremona, 2014). Also, the specimen may have deteriorated or been damaged. Therefore, some identification features may not be visible in some images. These obstacles to identifying thrip species are evident in our dataset. Hence, we augment the dataset by splitting each image into sections containing a segment of the thrip body so that our classification model can attend to them individually.

We split each image into the three main insect body sections: i) antenna and head; ii) thorax, the midsection of the body; and iii) abdomen (Figure 5). Conveniently, each segment contains at least one feature from the key for thrip identification (Figure 1). For the images in the training dataset, we measured the length of each segment along the x-axis and calculated the ratio of segment length to total image length. We calculated the mean (m) and standard deviation (sd) of the distribution of each ratio. Finally, we cropped the images to obtain each body segment using the calculated mean ratios with the margin of sd. After DDA, each sample thrip instance in the dataset contains four images: the original image of the whole thrip body, and three augmented subsections extracted from the original showing the antenna and head, thorax, and abdomen.

2.5. Fine-grained Thrip Classification Model

2.5.1. Problem Statement

Consider an image dataset of thrips $D = \{(x_1, x_1^1, x_1^2, ..., x_1^M, y_1), ..., (x_N, x_N^1, x_N^2, ..., x_N^M, y_N)\}$, where each instance consists of the original image and M different augmented subsection inferred using domain knowledge from the original image (see Section 2.4). Here, $x_i \in \mathbb{R}^{H \times W \times C}$ is the original image of the i^{th} instance, and $x_i^m \in \mathbb{R}^{H \times W \times C}$ is the m^{th} augmented subsection of x_i. H, W, and C denotes the height, width, and number of channels of the images respectively. The label y_i of each instance (i) in D defines whether the instance belongs to the WFT category ($y_i = 1$) or not ($y_i = 0$).

[6]https://www.dpi.nsw.gov.au/search?collection=&query=WFT

The aim is to learn the mapping function $f : (x_i, x_i^1, x_i^2, ..., x_i^M) \rightarrow y_i$ that reveals the label y_i of a given data point i using its original image x_i and their augmented subsections $\{x_i^1, x_i^2, ..., x_i^M\}$.

2.5.2. Overview

Our model adopts the popular Vision Transformer (ViT) (Dosovitskiy et al., 2020), a pre-trained image classification module, that can emphasise interesting/informative regions of images to classify them correctly. We adopt transfer learning and fine-tuning using our dataset to make ViT suitable for thrip classification (see Section 2.5.4 for detail). In our model, $M + 1$ different ViT image classification modules are used to predict the label of instances using their original images and M different augmented subsections. Subsequently, there is another feed-forward neural network in our model for creating an ensemble of the predictions from the $M + 1$ ViT image classification modules to generate a strong label.

2.5.3. Vision Transformer Background

In this section, we briefly review the concept behind the Vision Transformer (ViT) (Dosovitskiy et al., 2020), and its emphasis of image regions important for identifying its class label. As shown in Fig. 6, the ViT initially divides an image $x \in \mathbb{R}^{H \times W \times C}$ into a sequence of N flattened 2D patches $x_p \in \mathbb{R}^{N \times (P^2 C)}$, where (P, P) is the resolution of a patch. Here, $N = \frac{HW}{P^2}$ as the sequence of non-overlapping patches covers the whole image. The first layer of ViT performs the following operation on the inputs:

$$z_0 = [x_{CLASS}; x_p^1 E; ...; x_p^N E] + E_{POS} \quad E \in \mathbb{R}^{P^2 C \times D} \quad E_{POS} \in \mathbb{R}^{(N+1) \times D} \tag{4}$$

where E and E_{POS} are trainable projection matrices that embed visual information (patch embeddings) and positional information (positional embeddings) in the patches to low-dimensional vectors (D-dimensional) respectively. Here, x_{CLASS} is a special token added to the start of the patch sequence, which learns the representation of the complete sequence (i.e., the whole image). Following this layer, ViT has multiple blocks with the same operation stacked on top of one another. Each block takes the hidden representation of the previous block z_{l-1} as input and performs the following operations to generate the block's hidden representation z_l:

$$z_l' = MSA(LN(z_{l-1})) + z_{l-1} \tag{5}$$

$$z_l = MLP(LN(z_l')) + z_l' \tag{6}$$

where $LN()$ and $MLP()$ indicate the LayerNorm operation and multi-layer perceptron with a single hidden layer respectively. $MSA()$ denotes the multi-headed self-attention operation proposed in (Furfari, 2002). This operation effectively learns different weights for different patches such that it can signify the importance of the informative regions (i.e., patches) in the image when generating the hidden representation of a particular block. Following the number of blocks L with the operations in Eq. 5 and Eq. 6, ViT generates the final hidden representation of the whole image z as the hidden representation of the last layer corresponding to the x_{CLASS} token (Z_L^0). In order to predict the label of an image, z can be fed through another MLP:

$$y = MLP(LN(z)) \tag{7}$$

2.5.4. Our Model - Domain Knowledge-Driven Stacked Model for Thrip Classification

In our dataset, each sample includes four different subimages – i.e., the original image and three augmented images. Thus, we adopt four different ViT models, one to encode each subimage. We employ the ViT architecture- ViT-B/16 of 12 layers with 16x16 input patch size (Dosovitskiy et al., 2020). Since we have a relatively small dataset, training the whole ViT model from scratch is infeasible. Instead, we adopt transfer learning from a pre-trained ViT model[7] pre-trained using ImageNet[8], a very large image dataset with more than 1M images. We initialize the whole pipeline in our ViT models using the aforementioned pre-trained model and freeze all the weights except those in the last layer (Eq. 7) during training. In this way, the number of parameters that should be learned during training can be drastically reduced. We denote all operations (Eq. 4, Eq. 5 and Eq. 6) related to the frozen weights in a ViT as $VIT_F() : x \rightarrow z$, which maps an image to a D-dimensional vector.

As mentioned above, our model has four different ViT encoders $\{VIT_F^0, VIT_F^1, VIT_F^2, VIT_F^3\}$ to embed each of the different image subsections of each sample. These encoders generate a hidden representation for each

[7]https://github.com/timm/models/vision_transformer.py

[8]https://image-net.org/download

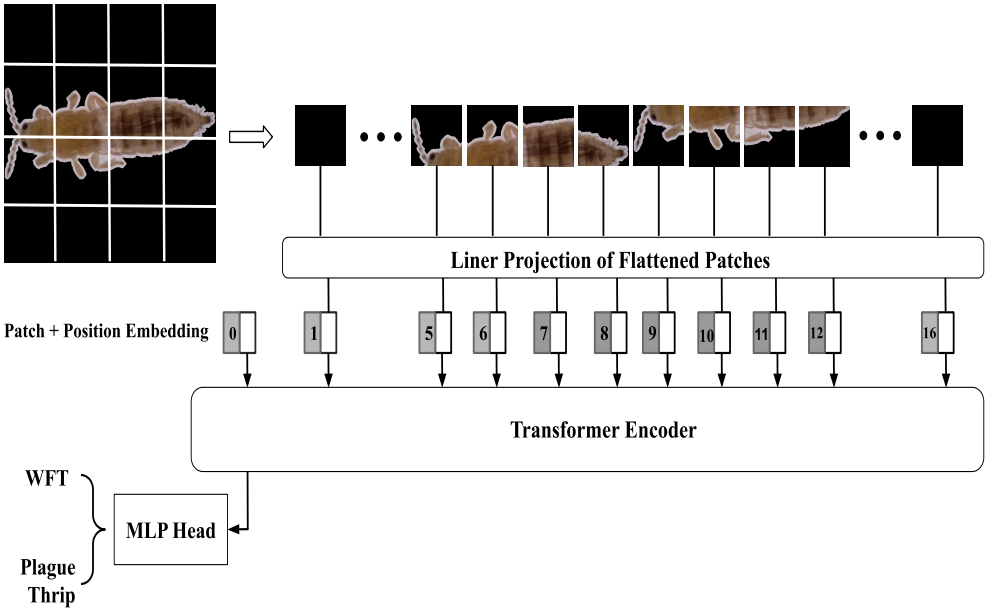

Figure 6: ViT model architecture. Please refer to the original paper on ViT (Dosovitskiy et al., 2020) for detail.

subsection as follows:

$$z_i^v = VIT_F^v(x_i^v) \quad forall\ v \in \{0,1,2,3\} \tag{8}$$

Our model (Figure 7) adopts an ensemble approach using these hidden representations to generate a final label for the images which can be elaborated as follows:

$$\hat{y}_i^v = MLP_v^V(z_i^v) \quad \forall v \in \{0,1,2,3\} \tag{9}$$

$$\hat{y}_i^E = MLP^E(\{\hat{y_i}^0, \hat{y_i}^1, \hat{y_i}^2, \hat{y_i}^3\}) \quad \forall v \in \{0,1,2,3\} \tag{10}$$

where $MLP_v^V()$ is a multi-layer perceptron that generates the label of an image using the v^{th} subsection and $\hat{y_i}^v$ is the predicted label of the v^{th} subsection. Likewise, each augmented subsection has its own classifier that predicts the label of the image just using that particular subsection. Then, predictions from each subsection are fed through another multi-layer perceptron $MLP^E()$ to aggregate them and make a strong prediction. Since the goal of each multi-layer perceptron in Eq. 9 and Eq. 10 is to predict the correct label of an image, our model performs end-to-end learning using the following loss function to train all trainable parameters $\{MLP_0^V, MLP_1^V, MLP_2^V, MLP_3^V, MLP^E\}$:

$$Loss = \frac{1}{N}\sum_{i=1}^{N}[\sum_{v=0}^{3} CE(y_i, \hat{y}_i^v)] + CE(y_i, \hat{y}_i^E) \tag{11}$$

where $CE()$ is the conventional cross-entropy loss and N is the total number of instances (thrip samples) in the training dataset.

2.6. Baseline Comparisons

We compare our model to four baselines:

- **SVM**: this baseline trains a support vector machine (SVM), a conventional machine learning model, to predict sample labels from their raw images. Here, the raw images are flattened before feeding to the machine learning model as SVMs cannot handle 2-dimensional inputs. Thus, this baseline is unable to capture the spatial dependencies in the images effectively;
- **ResNet50**: this baseline adopts the pre-trained image classification model ResNet (He et al., 2016), a 50 layer deep convolutional neural network (CNN). The pre-trained model contains parameters trained using the ImageNet database. We can fine-tune the last hidden layer in the model (similar to the fine-tuning process with ViT) to make it applicable for any downstream image classification tasks. With the

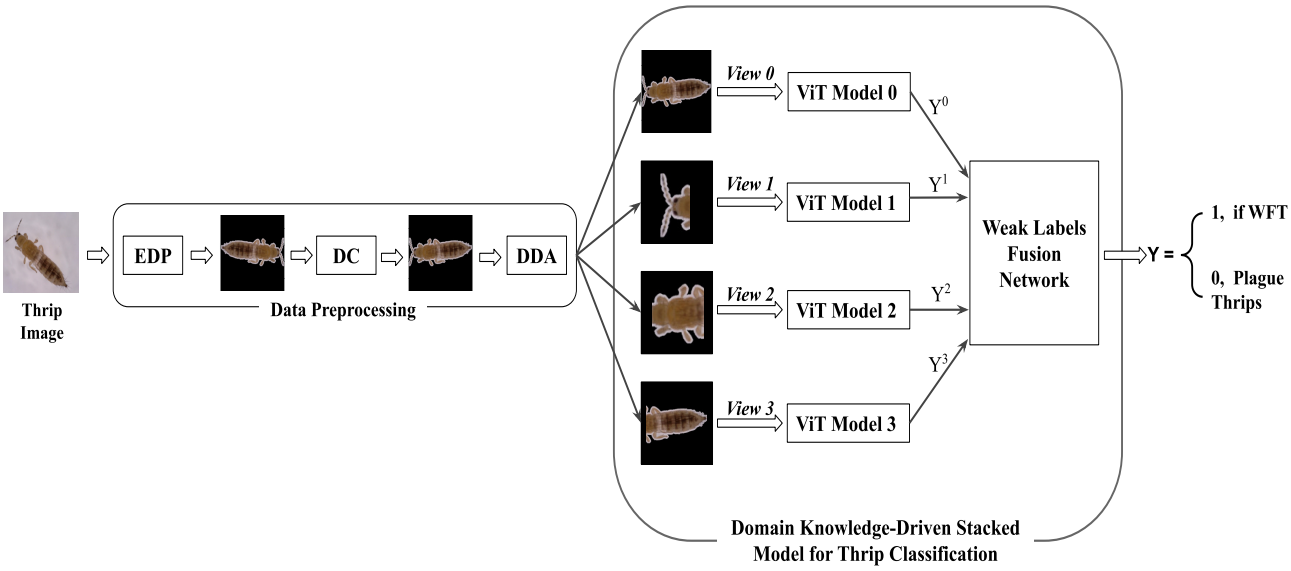

Figure 7: Proposed model for Thrip Classification

ability of CNNs to learn spatial dependencies in images, ResNet is considered a strong baseline for image classification tasks;

- **ResNet101**: this model extends ResNet50 by adding layers to its architecture making 101 in total. This is also pre-trained using the ImageNet database;
- **ViT-B/16**: here, we test Vision Transform model explained under Section 2.5.3 using the full body images of thrips.

For a fair comparison, the baselines are tested using the thrip image dataset that was preprocessed using the proposed EDP module in this work.

In addition, we conduct an ablation analysis where we compare our full model's performance to variants of our model to demonstrate the positive contribution of each model components:

- **Only {Full Body, Abdomen, Thorax, Head} subsection**: this variant only uses a single subsection of the image and adopts the proposed model to predict the label;
- **Without {Full Body, Abdomen, Thorax, Head} subsection**: this variant removes one subsection at a time from the complete model and adopts the same learning process.

By comparing our final model with these two sets of baselines, we can identify the importance of each subsection of the images and the importance of our weak label fusion network.

2.7. Hyper-parameter Settings

We conducted an hyperparameter analysis to find the optimal values for the number of epochs, batch size, and learning rate for both ViT and ResNet models. Please see Appendix B for more details about this parameter sensitivity analysis. The selected values for the ViT and Resnet models from this analysis can be found in Table 2. For other hyperparameters, we used the parameters proposed in the original papers.

Table 2: Hyperparameter Settings

Hyperparameter	ViT	Resnet
epochs	30	40
batch size	40	40
learning rate	0.01	0.05
optimiser	Adam	Adam

We adopted two random augmentation techniques during training: (1) vertical flipping; and (2) horizontal flipping, which effectively increases the size of the training dataset by four times. For evaluation, we adopt five fold cross validation to alleviate the overestimation of the results due to the selection biases of the testing dataset. (The code is avialble in https://github.com/ChathurikaA/fine-grained-classification-of-thrips).

2.8. Evaluation Metrics

We use four well-known image recognition performance evaluation metrics namely accuracy, precision, recall and f1-score to measure the performance of the models (Hossin and Sulaiman, 2015; Akosa, 2017). For a binary classification problem, these metrics are:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall(Sensitivity) = \frac{TP}{TP + FN} \tag{14}$$

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{15}$$

where TP is the number of positive samples predicted as positive classes, TN is the number of negative samples predicted as negative classes, FN is the number of positive samples predicted to be positive classes, and FP is the number of negative classes predicted to be negative classes.

3. Results

3.1. Model Performance Comparison

In Table 3, we compare our model with the baselines. As can be seen, **our proposed model outperforms the baselines by as much as 3.27% in accuracy and F1-score**. The original ViT model shows the best performance from the three baselines. Of the baselines, as expected, SVM has the weakest performance due to their inability to classify small differences in 2D images effectively. ResNet101 is the second best baseline, possibly due to its use of a deep CNN-based architecture compared to ResNet50 that allows it to learn complex patterns. However, CNN-based models alone are unable to emphasise informative image regions. By contrast, ViT has attention layers that allow our model to emphasise informative image regions. Since the images in our dataset are quite similar to one another, the ability to emphasise pixels corresponding to small insect features is useful here. This may be the reason our approach of integrating attention-based ViT models yields superior performance for thrip classification. We further analysed the accuracy of the models using one-way ANOVA. The results verified that there is a statistically significant difference in mean accuracy values between our model and the baselines ($p < 0.0001$).

Table 3: Experimental Results for Thrip Classification

Method	Accuracy	Precision	Recall	F1-score
ResNet101	0.921 ± 0.021	0.923 ± 0.038	0.921 ± 0.035	0.921 ± 0.020
ResNet50	0.873 ± 0.039	0.870 ± 0.046	0.877 ± 0.045	0.872 ± 0.0310
ViT-B/16	0.947 ± 0.024	0.939 ± 0.039	0.955 ± 0.023	0.947 ± 0.024
SVM	0.729 ± 0.040	0.788 ± 0.021	0.691 ± 0.011	0.693 ± 0.038
Our Model	$\mathbf{0.978 \pm 0.013}$	$\mathbf{0.969 \pm 0.021}$	$\mathbf{0.987 \pm 0.015}$	$\mathbf{0.9780 \pm 0.012}$

Table 4: Ablation Study for Thrip Classification ($mean \pm SD$). The mean accuracy values followed by the same letter do not differ significantly in One-way ANOVA ($P < 0.05$)

Method	Accuracy	Precision	Recall	F1-score
Only Full Body subsection	$0.947 \pm 0.024a$	0.939 ± 0.039	0.955 ± 0.023	0.947 ± 0.024
Only Abdomen subsection	$0.934a \pm 0.022a$	0.932 ± 0.029	0.934 ± 0.039	0.933 ± 0.023
Only Thorax subsection	$0.946a \pm 0.014a$	0.946 ± 0.027	0.948 ± 0.028	0.946 ± 0.014
Only Head subsection	$0.924 \pm 0.027b$	0.914 ± 0.035	0.935 ± 0.038	0.924 ± 0.027
Without Head subsection	$0.968 \pm 0.016c$	0.970 ± 0.021	0.965 ± 0.025	0.967 ± 0.016
Without Thorax subsection	$0.969 \pm 0.014c$	0.960 ± 0.027	0.979 ± 0.020	0.969 ± 0.014
Without Abdomen subsection	$0.968 \pm 0.015c$	0.961 ± 0.022	0.974 ± 0.017	0.968 ± 0.016
Without Full Body subsection	$0.965 \pm 0.016d$	0.956 ± 0.028	0.976 ± 0.020	0.965 ± 0.014

Table 5: Importance Analysis of Preprocessing Pipeline

Method	Accuracy	Precision	Recall	F1-score
Our Model	0.978 ± 0.013	0.969 ± 0.021	0.987 ± 0.015	0.9780 ± 0.012
(-) EDP	0.917 ± 0.0135	0.926 ± 0.021	0.904 ± 0.031	0.915 ± 0.014
(-) DDA	$0.947 \pm 0.024a$	0.939 ± 0.039	0.955 ± 0.023	0.947 ± 0.024

Table 6: Performance of the directional classifier for the test set

Method	Accuracy	Precision	Recall	F1-score
Full Model with $\{fea_1, fea_2\}$	0.976	0.980	0.974	0.976

3.2. Ablation Study

Contribution of the different image subsections. Table 4 shows the comparison between our complete model and its variants without elements of the pipeline. The results shows that Our complete model outperforms all these weaker variants. We conducted a one-way ANOVA to determine whether the differences between mean accuracy values of our model and its variants are statistically significant. The results revealed no significant difference between mean accuracy values of our model and its variant models which rely on the prediction from the full body and two other subsections (full_body + two other subsections) ($p = 0.2088$), but the improvements of our complete model are significant relative to the variants that only consider the prediction from one subsection or the prediction from the combined models without full body image (head + thorax + abdomen). This observation verifies the importance of each subsection in our model. Of the different subsections we selected, the full body image is the most informative, yielding 0.947 accuracy alone. However, our complete model outperforms this model by 3.27% in accuracy. Hence, we can conclude that the predictions from other subsections introduce additional knowledge. To further verify the positive contribution from our model, we perform a deep error analysis, which also verifies the importance of the ensembling approach in our model. We present the results of this error analysis in detail in Figures 8 and 9. We have performed a correlation analysis between the prediction from the complete images and the predictions from the other subsections. None of the correlations is higher than 0.75, confirming the importance of the different subsections and the label fusion network to merge the subsection predictions.

Importance of the different processing steps. In Table 5, we remove the preprocessing steps from our model and analyse the effect on performance. Early preprocessing steps (EDP) and domain knowledge-driven augmentation steps (DDA) **account for 6.6% and 3.27% improvements in F1-score**, respectively. This verifies the positive contribution of prepossessing in our complete model. Also, the proposed directional classifier has the ability to identify direction of thrip head with 97.6% of accuracy which directly impact the performance of the final thrip species classification results (Table 6).

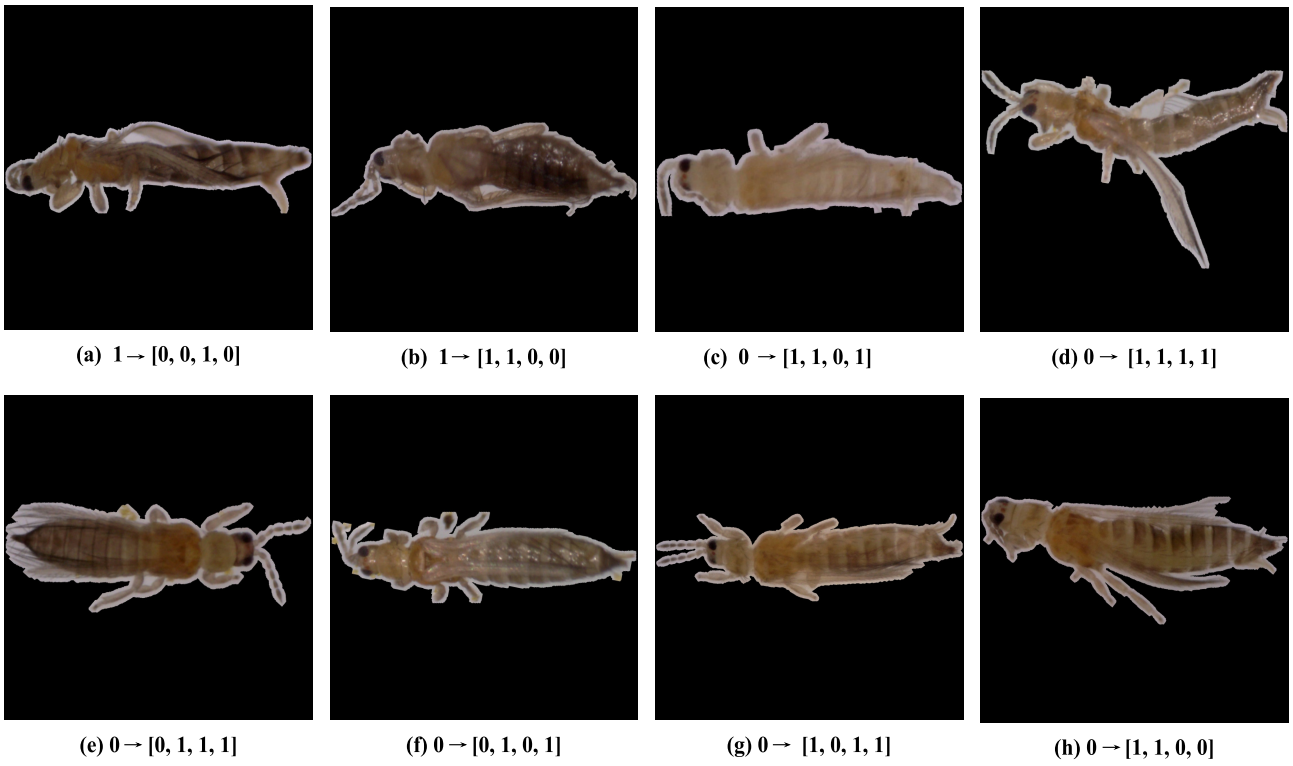

Figure 8: A set of images wrongly classified by the proposed model. The sub-caption of each image represents the ground truth label and the prediction generated by our model for each subsection, formatted as $\hat{y}_i \rightarrow [\hat{y_i}^{body}, \hat{y_i}^{head}, \hat{y_i}^{thorax}, \hat{y_i}^{abdomen}]$ for image id i. Most of the errors are due to highly disordered multiple parts in the images - e.g., (a) is highly disordered around its head and has bent wings and (d) is highly disordered throughout the body. Errors such as (e) are due to propagation from the preprocessing step of the directional classifier in our framework.

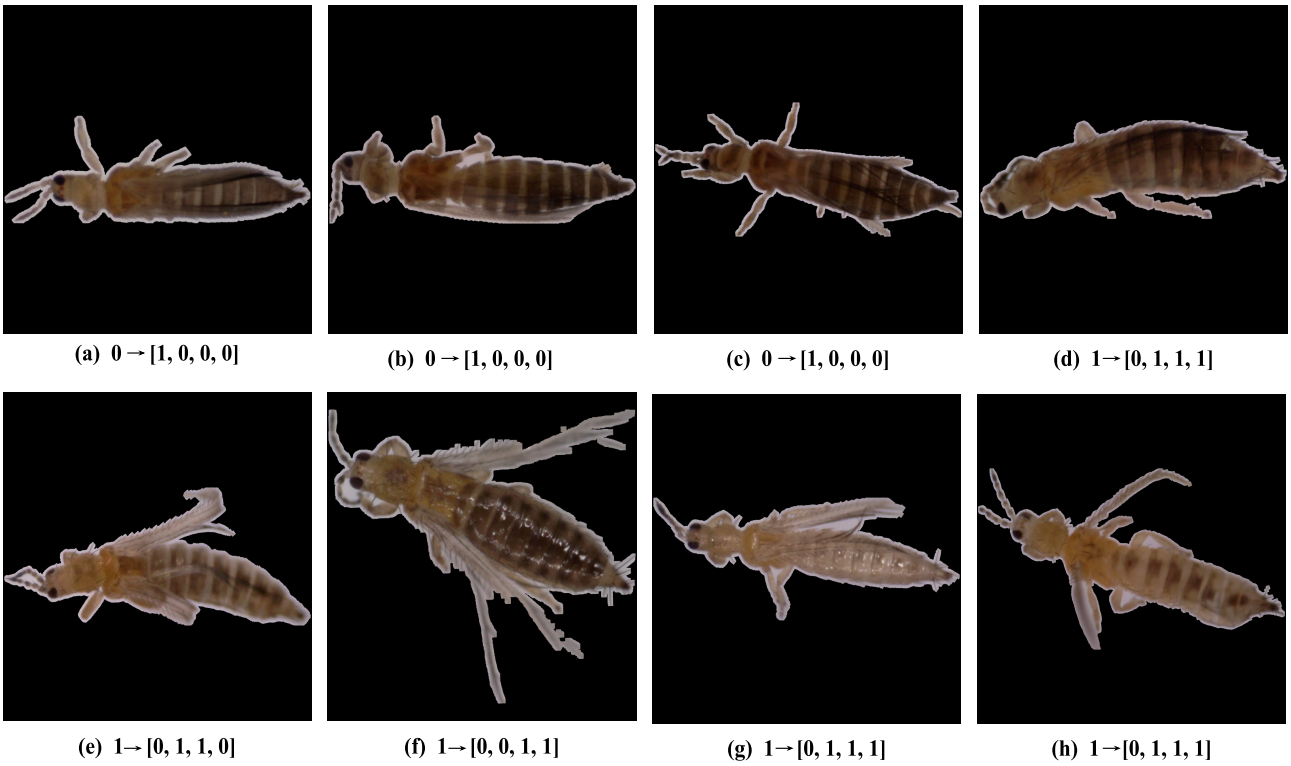

Figure 9: A set of images wrongly classified by the ViT-B/16 baseline which only relies on the full body images. The sub-caption of each image represents the ground truth label and the prediction from our model for each subsection, formatted as $\hat{y}_i \rightarrow [\hat{y_i}^{body}, \hat{y_i}^{head}, \hat{y_i}^{thorax}, \hat{y_i}^{abdomen}]$ for image id i. Here, we can see that most errors can be corrected by focusing on the individual subsections of the body. For example, (a-d), (g-h) can be correctly predicted even by focusing on a single subsection.

4. Discussion

Early detection and insect pest classification is important for preventing losses and improving crop yield. However, capabilities in these tasks are currently limited by the high labour costs and difficulty of species level insect pest classification, especially for microscopic insects. In this research, we presented a novel image-processing pipeline to segment insects from image backgrounds, and to split the image based on insect body segments. We introduced a part-dependent classification model based on ViT to conduct accurate species-level classification of morphologically close insect species, and a dataset of microscope images of WFTs and Plague thrips. Our method enables automated insect pest classification with the potential to inform crop management and have improved on previously available methods by up to 3.27% in accuracy achieving 97.8% accuracy.

Model Complexity: Since our approach has 4 parallel ViT models to focus on different thrip body parts, its complexity is around four times that of a baseline with a single model. We leave optimizing the complexity of our algorithm as future work. The proposed image processing approach for the Data Preprocessing (DP) module might not be robust for images with complex backgrounds. Hence, we could consider replacing the DP module using deep learning aided by existing object detection and segmentation algorithms such as YOLO (Redmon et al., 2016) and Mask R-CNN (He et al., 2017). However, this would probably require a large, rigorously labelled, training dataset for trials.

Thrip Species Variety: The current image dataset covers two thrip species commonly found in strawberry fields. As future work, we intend to increase the number of samples in the set and extend it to other thrip species. Having a large, diverse dataset will potentially help train our model to specialize its thrip classification and improve generalisation. However, an increased number of thrip species may inadvertently lower system performance due to high inter-species similarity. One possible solution is to introduce more informative body segments to the classification model, emphasising key thrip body features. Very high magnification images of different thrip body segments might also be captured separately, and then separate predictions from each image could be joined in an ensemble to make a strong prediction. Apart from thrip species, in future research our algorithm can potentially be applied to distinguish between other morphologically similar microscopic insect species such as ants, fleas or parasitoid wasps.

Economic Impact: The "economic threshold" is the pest density at which management action should be taken to prevent an increasing pest population from reaching a level causing economic injury (Steiner and Goodwin, 2005). In the thrip pest classification task, it is vital to have high precision and high recall as thrip numbers generally have a low economic threshold. For example, the economic threshold for WFTs in strawberries is 5 per flower, and for Plague thrips it is 10 (Steiner and Goodwin, 2005). The parameter "finite rate of increase (λ)" of an insect defines the rate of increase per individual per unit time. This value is high for WFTs, around 1.16 individuals per day at $25°C$ for strawberries (Gerin et al., 1994). A commercial strawberry grower may inspect and collect flower samples fortnightly to

estimate thrip numbers. Hence, during an inspection period of two weeks, a missed individual thrip –i.e., a false negative from an automated system – can generate approximately 16 new thrips. Identifying WFTs in the field with high recall can therefore be extremely important to reduce the economic impact of these pests.

Pest Thrip Treatment: WFTs are resistant to many pesticides (Reitz et al., 2020). Thus, growers typical use expensive biological control agents, introduced predator insects, to control them (Reitz et al., 2020). In contrast, pesticides can be used to control Plague thrips. This is cheaper than biological control. Thus, overestimating a thrip population (i.e., high false positive numbers) is not cost-efficient. This emphasises the importance of identifying thrips with high precision. In our model, we mainly focus on optimizing both recall and precision values using F1-score as an evaluation metric. Our model yields 3.27% improvements in F1-sore over the best of the selected baselines.

Potential for In-field Application: Sticky traps are an effective way to monitor and control insect pests in the field. Some growers set them to catch thrips, and the number of different species caught can be employed as an estimate of thrip population size (Sampson et al., 2014). Hence, it will be worthwhile in the future to test our proposed approach on image data collected directly from sticky traps, rather than from flower samples. To show scalability of our approach for this setting, we conducted a preliminary analysis using sticky trap images. See Appendix A for the results of this study, which shows promise for future research. Exploring an alternative cost effective, and simple image capture method, such as commercially available magnification lenses for mobile phones, would be worthwhile also to improve the usability of the proposed model in field.

Threats to Validity: There are a few factors that might affect the performance of our proposed approach. First, we captured thrip images in a laboratory under controlled lighting and against consistent backgrounds. Hence, the system requires further testing for images taken under variable lighting and backgrounds. Also, we tested our system for two thrip species commonly found on strawberries, but other crops can be threatened by different thrip species (Lewis et al., 1973). System testing with more thrip species would therefore be valuable. Finally, thrips are small and fiddly to handle and photograph, so human annotation errors may have occurred during image labelling. To minimise annotation errors we took multiple images of each thrip specimen focused on different body parts, and then analysed each image to correctly identify its species. However unlikely, annotation errors may still have crept in reducing slightly the accuracy of the method.

5. Conclusion

In this paper, we proposed a novel machine learning method to conduct fine-grained classification of morphologically similar microscopic insect species using their images. This method consists of a simple data processing pipeline to segment insect images into informative body segments (full body, head, thorax and abdomen), and a classification framework based on the Vision Transform architecture to generate a strong prediction for the insect species. The prediction is generated by combining information provided by weak labels produced from each of the insect's body segments. We constructed and published a new dataset containing microscopic images of two morphologically close, economically devastating thrip species. Using this dataset, we demonstrated that the proposed approach yields a 3.27% improvement in Accuracy against previous state-of-the-art image classification models, while achieving 97.8% accuracy for thrip classification. The ablation study showed that the potential of the proposed model to explicitly give attention to various body segments as it helps to identify minute morphological differences between the insects.

Our technique demonstrates the potential for improved farm pest management based on automated classification of even tiny insects with strong visual resemblances. Hence, extending the proposed approach to other morphologically similar insect species could be a promising future direction to explore. Also, in the future, we envisage that the deployment of automated insect classification systems on agricultural sites will facilitate a step-change in integrated pest management (IPM). The data they generate is readily interpreted to inform decisions about the relative benefits of biological control or insecticides when insect species are the determinant.

Acknowledgments

The authors would like to thank Dr Hazel Parry, CSIRO, Australia, for her intellectual input on the research and constructive feedback on the manuscript. The authors thank the management staff of Sunny Ridge Strawberry Farm, VIC, Australia, and P. Borse and K. Rusevski from Biological Services, Australia, for their support while collecting thrip samples and conducting species identification.

Funding: Amarathunga's contribution to this research was funded by the Australian Government through the Australian Research Council's ITRH for Driving Farming Productivity and Disease Prevention (IH180100002). Grundy is supported by ARC Laureate Fellowship FL190100035.

References

Akosa, J., 2017. Predictive accuracy: A misleading performance measure for highly imbalanced data, in: Proceedings of the SAS Global Forum.

Allan, S.A., Gillett-Kaufman, J.L., 2018. Attraction of thrips (thysanoptera) to colored sticky traps in a florida olive grove. Florida Entomologist 101, 61–68.

Amarathunga, D.C.K., Grundy, J., Parry, H., Dorin, A., 2021. Methods of insect image capture and classification: A systematic literature review. Smart Agricultural Technology , 100023.
Cremona, J., 2014. Extreme close-up photography and focus stacking. Crowood.
Cui, S., Ling, P., Zhu, H., Keener, H.M., 2018. Plant pest detection using an artificial nose system: a review. Sensors 18, 378.
Cunningham, P., Delany, S.J., 2021. k-nearest neighbour classifiers-a tutorial. ACM Computing Surveys (CSUR) 54, 1–25.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .
Ebrahimi, M., Khoshtaghaza, M., Minaei, S., Jamshidi, B., 2017. Vision-based pest detection based on svm classification method. Computers and Electronics in Agriculture 137, 52–58.
Espinoza, K., Valera, D.L., Torres, J.A., López, A., Molina-Aiz, F.D., 2016. Combination of image processing and artificial neural networks as a novel approach for the identification of bemisia tabaci and frankliniella occidentalis on sticky traps in greenhouse agriculture. Computers and Electronics in Agriculture 127, 495–505.
Fedor, P., Malenovskỳ, I., Vaňhara, J., Sierka, W., Havel, J., 2008. Thrips (thysanoptera) identification using artificial neural networks. Bulletin of entomological research 98, 437–447.
Fedor, P., Peña-Méndez, E.M., Kucharczyk, H., Vanhara, J., Havel, J., Doricova, M., Prokop, P., 2014. Artificial neural networks in online semiautomated pest discriminability: an applied case with 2 thrips species. Turkish Journal of Agriculture and Forestry 38, 111–124.
Furfari, F.A., 2002. Attention Is All You Need. IEEE Industry Applications Magazine 8, 8–15. doi:10.1109/2943.974352.
García-Lara, S., Saldivar, S.S., 2016. Insect pests, in: Caballero, B., Finglas, P.M., Toldrá, F. (Eds.), Encyclopedia of Food and Health. Academic Press, Oxford, pp. 432–436. URL: https://www.sciencedirect.com/science/article/pii/B9780123849472003962, doi:https://doi.org/10.1016/B978-0-12-384947-2.00396-2.
Gaston, K.J., O'Neill, M.A., 2004. Automated species identification: why not? Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 359, 655–667.
Gerin, C., Hance, T., Impe, G.V., 1994. Demographical parameters of frankliniella occidentalis (pergande)(thysanoptera, thripidae). Journal of Applied Entomology 118, 370–377.
Gullan, P.J., Cranston, P.S., 2014. The insects: an outline of entomology. John Wiley & Sons.
He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.
Hossin, M., Sulaiman, M.N., 2015. A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process 5, 1.
Iannino, F., Sulli, N., Maitino, A., Pascucci, I., Pampiglione, G., Salucci, S., 2017. species, biology and flea-borne diseases. Veterinaria italiana 53, 277–288.
Júnior, T.D.C., Rieder, R., 2020. Automatic identification of insects from digital images: A survey. Computers and Electronics in Agriculture 178, 105784.
Kandel, I., Castelli, M., 2020. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. ICT express 6, 312–315.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25, 1097–1105.
Lewis, T., et al., 1973. Thrips, their biology, ecology and economic importance. Thrips, their biology, ecology and economic importance. .
Li, W., Wang, D., Li, M., Gao, Y., Wu, J., Yang, X., 2021. Field detection of tiny pests from sticky trap images using deep learning in agricultural greenhouse. Computers and Electronics in Agriculture 183, 106048.
Li, Y., Yang, J., 2020. Few-shot cotton pest recognition and terminal realization. Computers and Electronics in Agriculture 169, 105240.
Loomans, A., Van Lenteren, J., Tommasini, M., Maini, S., Riudavets, J., 1995. Biological control of thrips pests. Wageningen, Netherlands: Wageningen Agricultural University.
Lu, C.Y., Rustia, D.J.A., Lin, T.T., 2019. Generative adversarial network based image augmentation for insect pest classification enhancement. IFAC-PapersOnLine 52, 1–5.
Luong, L., 2008. Investigations into aspects of biology of tubular black thrips, haplothrips victoriensis bagnall(thysanoptera; phlaeothripidae) in south australia [thesis submitted for the degree of master of science, discipline of ecology and evolutionary biology]. Adelaide (Australia): The University of Adelaide .
Marsham, T., 1797. Observations on the insects that infested the corn in the year 1795. in a letter to the rev. samuel goodenough, ll. dfrs tr. ls. Transactions of the Linnean Society of London 3, 242–251.
Martineau, M., Conte, D., Raveaux, R., Arnault, I., Munier, D., Venturini, G., 2017. A survey on image-based insect classification. Pattern Recognition 65, 273–284.
Mehle, N., Trdan, S., 2012. Traditional and modern methods for the identification of thrips (thysanoptera) species. Journal of Pest Science 85, 179–190.
Mouden, S., Sarmiento, K.F., Klinkhamer, P.G., Leiss, K.A., 2017. Integrated pest management in western flower thrips: past, present and future. Pest management science 73, 813–822.
Murphy, K.P., et al., 2006. Naive bayes classifiers. University of British Columbia 18, 1–8.
Noble, W.S., 2006. What is a support vector machine? Nature biotechnology 24, 1565–1567.
Pobozniak, M., Tokarz, K., Musynov, K., 2020. Evaluation of sticky trap colour for thrips (thysanoptera) monitoring in pea crops (pisum sativum l.). Journal of Plant Diseases and Protection 127, 307–321.
Ratnayake, M.N., Amarathunga, D.C., Zaman, A., Dyer, A.G., Dorin, A., 2022. Spatial monitoring and insect behavioural analysis using computer vision for precision pollination. arXiv preprint arXiv:2205.04675 .
Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.
Reitz, S.R., Gao, Y., Kirk, W.D., Hoddle, M.S., Leiss, K.A., Funderburk, J.E., 2020. Invasion biology, ecology, and management of western flower thrips. Annual review of entomology 65, 17–37.
Rustia, D.J.A., Lin, C.E., Chung, J.Y., Zhuang, Y.J., Hsu, J.C., Lin, T.T., 2020. Application of an image and environmental sensor

network for automated greenhouse insect pest monitoring. Journal of Asia-Pacific Entomology 23, 17–28.
Sampson, C., et al., 2014. Management of the western flower thrips on strawberry. Ph.D. thesis. Keele University.
Solis-Sánchez, L.O., Castañeda-Miranda, R., García-Escalante, J.J., Torres-Pacheco, I., Guevara-González, R.G., Castañeda-Miranda, C.L., Alaniz-Lumbreras, P.D., 2011. Scale invariant feature approach for insect monitoring. Computers and electronics in agriculture 75, 92–99.
Steiner, M.Y., Goodwin, S., 2005. Management of thrips (thysanoptera: Thripidae) in australian strawberry crops: within-plant distribution characteristics and action thresholds. Australian Journal of Entomology 44, 175–185.
Taylor, R., Ojha, V., Martino, I., Nicosia, G., 2021. Sensitivity analysis for deep learning: ranking hyper-parameter influence, in: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), IEEE. pp. 512–516.
Tschinkel, W.R., 2013. The fire ants. Belknap Press.
Wu, X., Zhan, C., Lai, Y.K., Cheng, M.M., Yang, J., 2019. Ip102: A large-scale benchmark dataset for insect pest recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8787–8796.
Xia, C., Chon, T.S., Ren, Z., Lee, J.M., 2015. Automatic identification and counting of small size pests in greenhouse conditions with low computational cost. Ecological informatics 29, 139–146.

Appendix A: Experiment Using Sticky Trap Images

Sticky traps are commonly used in greenhouses and crop fields to monitor flying insects, including thrips. The two most attractive trap base colours for thrips are blue and yellow (Allan and Gillett-Kaufman, 2018; Pobozniak et al., 2020). Typically, the traps are positioned in the field and transferred to the lab after a set period (e.g., a week or two) to identify captured insect species. Entomologists visually inspect each thrip on a sticky trap using a lab microscope (at least 40X magnification) to determine its species. Hence, we have explored the possibility of applying our proposed model to sticky trap images for thrip species classification.

A.1. Data Collection

We set sticky traps under strawberry polytunnels on the Sunny Ridge strawberry farm from February - May 2021 (Figure 10). The traps remained in situ for two weeks and were then transported to the lab. Images of the thrips on the traps were captured using a handheld digital microscope (Figure 11). We followed the guidelines provided by the website of the NSW Department of Primary Industries to identify images of WFTs and Plague thrips using their keys as explained in Section 2.1 and extended our thrip dataset. Table 1 provides a summary of the extended dataset (available online at https://drive.google.com/drive/thrip-images-dataset).

Figure 10: Sticky traps set up under strawberry polytunnels

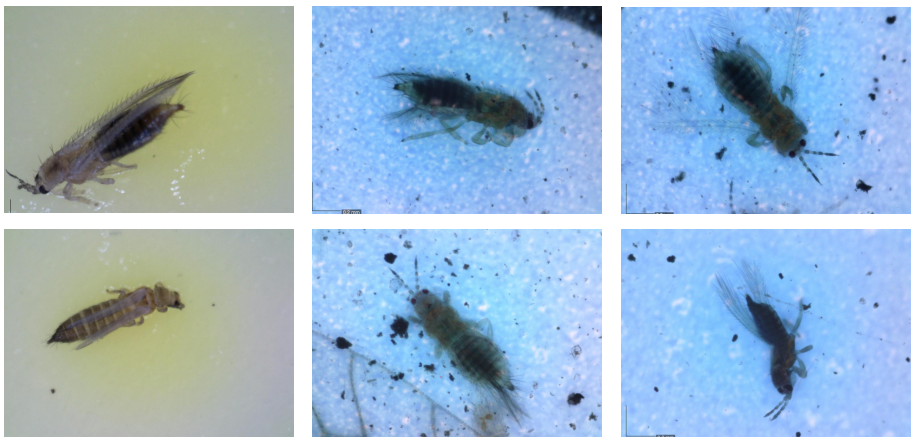

Figure 11: Randomly selected images from the sticky trap image dataset

A.2. Model Implementation, Results and Discussion

For the species-level classification, we followed the same approach as explained in Section 2.5. Since the number of sticky trap images per thrip species is highly imbalanced, the colour of the sticky traps can affect the performance

Table 7: Properties of the extended dataset using sticky trap images

Method	**WFTs**	**Plague Thrips**
Flower Sampling	210	215
Sticky Traps	116	20
Total	316	220

of the models. To minimise the effects of the sticky trap colour on the final results, we converted all the images into grey-scale at the end of the pre-processing stage and then fed them to classification models. We tested our model for the extended dataset and the classification results are presented in Table 8. As can be seen, **our proposed model shows improvements of 2.4% in F1-score compared to the Vit model which trained only using full body images**. We further analysed the F1-score of the models using a one-way ANOVA. The results verified that there is a statistically significant difference in mean F1-score values between our model and the baseline ($p < 0.0001$).

For this preliminary exploratory study, our model was tested on limited sticky trap data. It would be worth collecting more such images in the future to extend testing of model performance. However, the Data Prepossessing (DP) module of our model may work poorly if insects are greatly overlapped or occluded by dust particles on the sticky traps. In this case, the DP model could be replaced using deep learning aided by existing object detection and segmentation algorithms such as YOLO and Mask R-CNN. The integration of these algorithms with the rest of the model should be straightforward, but this remains as something to explore in the future.

Table 8: Experimental Results for Thrip Classification

Method	Accuracy	Precision	Recall	F1-score
ViT-B/16	0.943 ± 0.014	0.951 ± 0.020	0.951 ± 0.025	0.952 ± 0.013
Our Model	0.966 ± 0.001	0.960 ± 0.019	0.984 ± 0.017	0.972 ± 0.007

Appendix B: Hyperparameter Tuning

In this study, the three important hyperparameters, learning rate, epochs, and batch size, were varied individually while keeping others constant to select values giving the highest accuracy for the test dataset. This process was conducted for both ViT and ResNet101 for pre-processed full-body thrip images using the EDP module. We selected Adam as the optimiser and StepLR as the learning rate scheduler.

Effects of learning rate. Learning rate controls the size of the weights updating steps of the model with respect to the loss gradient. A learning rate that is too small may unnecessarily lengthen the training process and lead to a sub-optimal solution. A value that is too large can cause instability in the learning process. We trained ViT and ResNet101 with learning rates of 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, and 0.1. We set epochs and batch size to 30 and 40 respectively. As shown in Figure 12, the mean accuracy value is comparatively low for learning rate below 0.001 and the best accuracy is achieved when at 0.01 for ViT and 0.05 for ResNet101.

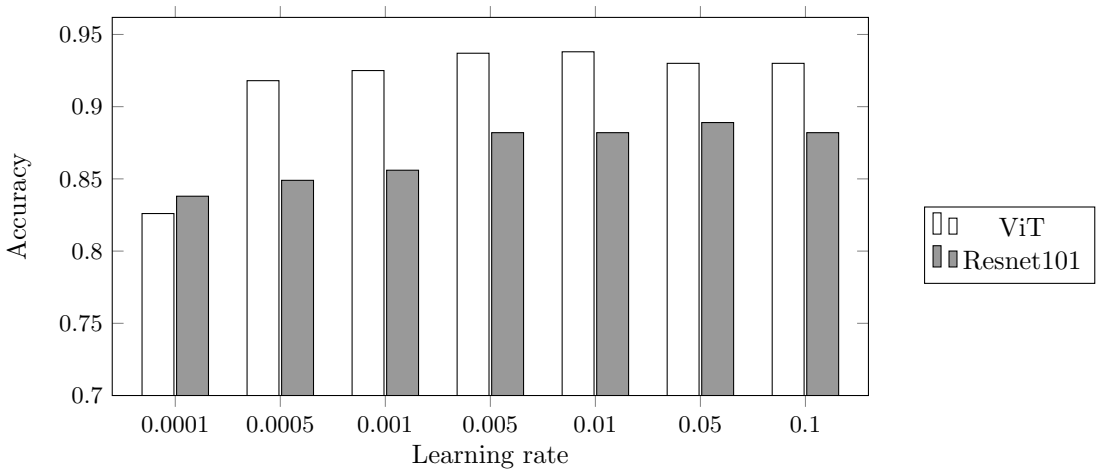

Figure 12: Variation of Accuracy with Learning Rate

Effects of the number of epochs. The number of epochs defines the number of complete passes of the training dataset during the training phase. If the number of epochs used to train a model is too high the model will lose generalization capacity by overfitting to the training data. The solution could be sub-optimal if the number of epochs is too small. We plotted the accuracy values for the test dataset over 40 epochs (Figure 13). The learning rates were 0.01 and 0.05 for ViT and ResNet101 respectively. The results show that after 21 epochs the classification accuracy does not change significantly for ViT, and this value is 34 for ResNet101.

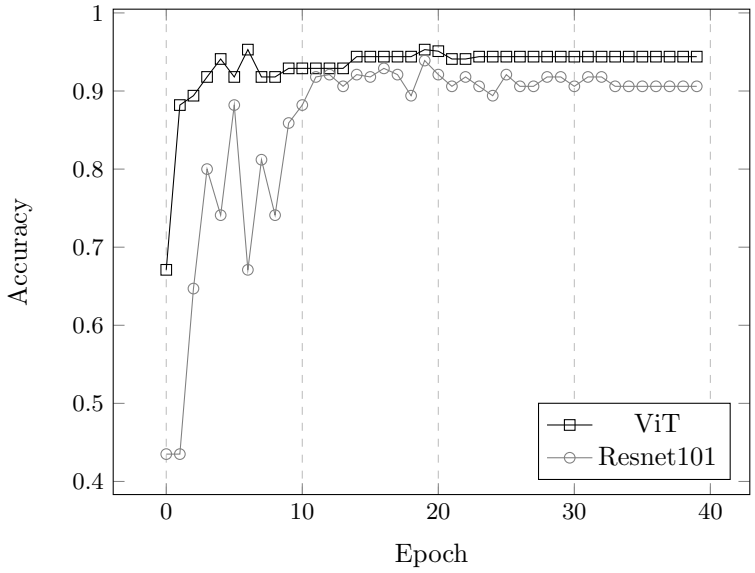

Figure 13: Variation of Accuracy with number of epochs

Effects of batch size. The batch size defines the number of training samples considered for each calculation of weight update of the model's internal parameters. Setting this hyperparameter too high can result in high memory requirements. A value that is too low can cause the model to bounce back and forth without converging (Kandel and Castelli, 2020). The hyperparameter sensitivity analysis conducted in the study (Taylor et al., 2021) also verifies that the batch size can largely influence the performance of a deep learning model. We evaluated the performance of ViT and ResNet with batch sizes of 10, 20, 40, 80, 160 and 360 (Figure 14). Here, we set the learning rate to 0.01 for ViT and 0.05 for ResNet101. Based on the results, the batch size of 40 gave the best accuracy for both models, and a further increase does not improve accuracy.

We adopted the same hyperparameter values chosen for the ViT-full_body model for the thrip body subsections (i.e., head, thorax, and abdomen) and the Resnet101 model's parameters for the Resnet50 model. By tuning the hyperparameters of each subsection separately within the combined model, the performance of the newly proposed model may potentially be further improved, an approach to be considered in future work.

For the SVM baseline, we tested four different kernel functions, radial basis function (RBF), poly, sigmoid and linear. The highest accuracy was given by the RBF kernel. Then, we set the kernel to RBF and changed the regularization

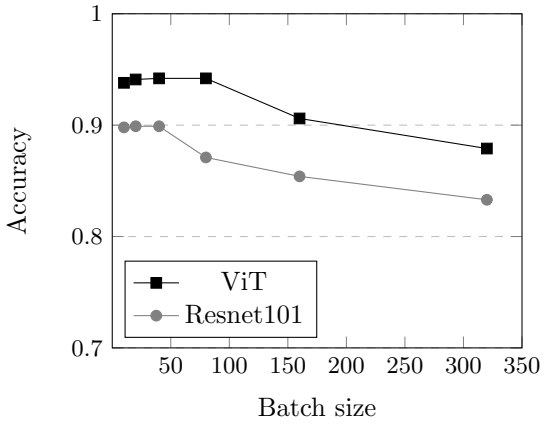

Figure 14: Variation of accuracy with batch size

parameter to 0.1, 1, 10, and 100. The value 10 gave the highest accuracy for the SVM classifier and is used for this study.