Evaluation of data processing platforms for camera trap data

E. Oyston and J. Tinnemans. August 2022. Department of Conservation

Introduction

The primary objective of this study was to determine the uptake of 1080 sausage baits on feral cats and the resulting kill efficacy; but in doing so an opportunity presented itself to run a comparison trial of processing platforms for camera trap footage. This section reviews the methods used for processing and management of camera trap footage and provides a comparison of relative efficiency and efficacy of artificial intelligence (AI)/machine learning (ML) platforms with manual classification.

It has previously been identified that processing images and data from landscape-scale camera networks is prohibitively labour intensive and not realistically scalable. The proposed scale required for detection and eradication of cats on the Auckland Islands is not feasible with current management tools. For the proposed methodology for the eradication of cats to be possible, the feasibility report (DOC, 2021) identified that

a) the automated processing of image data is required to exclude falsely triggered images (no animal present) as a minimum and preferably identify images where cats are present

and

b). Ultimately this would be coupled with a technology that has the ability to detect and identify cats, capture footage, and remotely report detections to a base computer.

This form of technology would save significant amounts of labour and reduce data management, and more importantly enable a rapid response to detected target animals – potentially reducing the time until eradication can be assumed, and the project concluded. Methods for the automatic detection, classification, and conversion of data into relevant information are urgently needed and have the potential to broaden and enhance ecology and wildlife conservation in scale and accuracy outside the Maukahuka project (Tuia et al., 2021).

In 2020, an internal report (DOC, 2020) was commissioned reviewing the availability and performance of the AI and ML market in relation to camera trap footage processing. The report reviewed mostly prototype solutions provided by software developers in regards to object detection and cat classification with the aim of reducing workload in terms of manual reviewing – but no solutions provided the close to 100% recall level whilst maintaining high enough levels of precision to be useful for an eradication project.

The 1080 sausage bait registration trial on Maukahuka (Feburary-March 2022) has provided another opportunity to review developments in AI and ML camera trap processing platforms since the 2020 review.

Methodology

The dataset used to test platforms consisted of camera trap footage of 10 and 12 second video clips, depending on which of the two models of trail cameras (Browning Dark Ops and Bushnell Aggressors) were used at the camera trap sites (n=111). The dataset was initially processed using a manual classification platform, the results of which acted as a baseline to assess the performance of the object detection model platform (MegaDetector) and machine learning classification platform (eVorta).

Various terms used in the assessment of data are outlined in **Table 1** below. Results on efficacy and resourcing could be compared relative to baseline manual classification. Efficacy for each platform was assessed using the metrics of precision and recall. Recall is of primary importance for eradication projects, where detection of every individual can be the difference between success and failure. Precision is also important, but a small amount can be sacrificed if it means ensuring detections are not missed. However, the lower the precision rate, the more manual input and reclassification will be required - potentially reducing the time resource and cost saving benefits of automation.

Recall and precision are useful metrics, but are focussed on individual frame detection and do not necessarily represent the ability of a platform to detect individuals/events for an eradication scenario. When processing results, we examined the false-negatives produced by both MegaDetector (at 0.5 sensitivity) and eVorta (for cat detections classed at 0.99 confidence), in order to determine any patterns or obvious reasons why the platforms missed these detections. We then reviewed immediate adjacent footage from false-negatives to determine if these false-negatives would have resulted in missing a cat 'event' at a site. If immediate corresponding footage of that cat had a true-positive detection by the relevant platform, we deemed the event was still detected.

Table 1: Definition of terms used in assessment

Term	Description
Detection	Platform identifying target species from footage
Event	A visitation of an individual target species to a given site. e.g *A cat visits a camera site for 5 minutes eating a sausage bait over 5 minutes moving in and out of a camera frame repeatedly resulting in 23 true positive frames, but is counted as a single event.*
True-positive	When a relevant item in footage is correctly marked as present
False-negative	When a relevant item in footage is incorrectly marked as absent
Recall	A percentage of relevant items retrieved
Precision	A percentage measure of accuracy measured by how many retrieved items are relevant
Machine learning classification	The use of machine learning algorithms that learn how to group observations into categories. This typically results in footage being tagged to species or group level.
Object detection model	**A** computer vision technique for locating instances of objects in images or videos. This typically results in footage being tagged either 'empty' or 'not empty'

An outline of the platforms and workflows used for each are detailed as follows.

Manual classification platform (customised Microsoft Access Database, Joris Tinnemans)

A customised Microsoft access database with a simple Graphical User Interface (GUI) was used for manual classification of the dataset. Camera information and site location for each camera were entered as metadata which was linked to footage that was classified from that particular camera when footage was processed. This metadata was also used to automate a folder structure when downloading footage from cameras. After importing footage, a human viewed and manually classified each video or image in real-time using a GUI format (**Figure 1)** where users selected species present and could add additional comments. The resulting data consisted of a row per video/image stored in a data-table in Microsoft access that can be filtered, queried, and exported for analyses with other software. Data was then summarised by number of videoclips containing specific species.

The dataset being used by this trial consisted of video clips and was processed as such for manual classification, but was extracted and processed as keyframes (still images) by the object detection and classification platforms. Acknowledging that there will be different time requirements between manually classifying videos and still images, we calculated an average manual classification time for images using the Microsoft Access platform of 3600 images per hour based on an experienced user processing at dataset of 26,000 images with the same environmental context and species diversity.

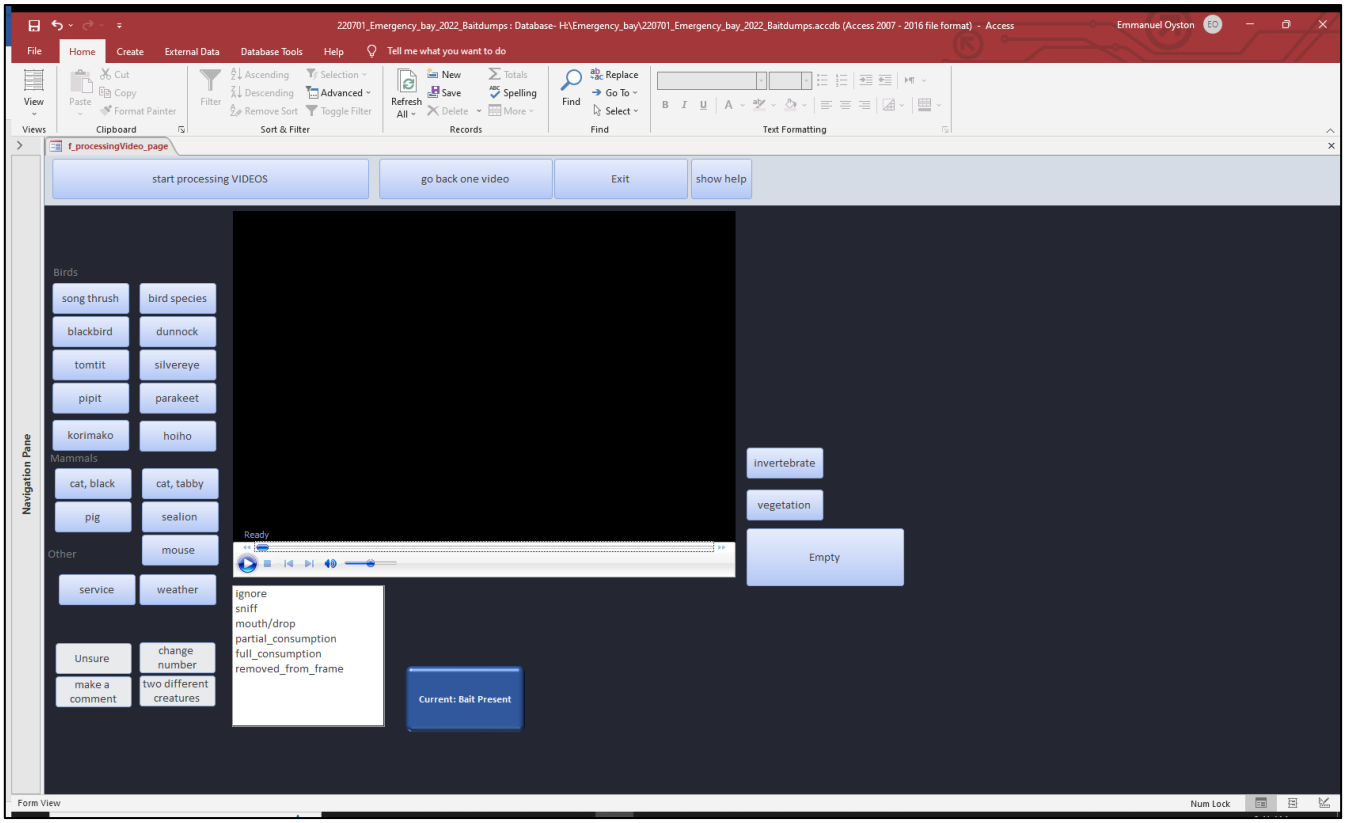

Figure 1: GUI of MS Access database used for manual classification of footage

Object detection model platform (Megadetector, Dan Morris/Microsoft AI for earth)

The open-source package 'MegaDetector' version 5a (released 06/22) was used to test the ability of the relatively well-known AI object detection platform. The MegaDetector uses a species agnostic object detection model to classify images into one of five broad object classes (human, animals, vehicle, multiple, and empty). Thus it can be used to exclude images that are blank and save time through manual classification only being required on footage where animals have been detected. The system has been in development for a number of years with its first release in 2018. The platform consists of a static model, so improvements in recall and precision can only be achieved through updated releases from the developer. MegaDetector requires users to have basic familiarity with programming as it requires users to initially setup packages from a command prompt environment using the programming language Python; and to execute the processing of footage from a Python command prompt with a series of appropriate parameters. The platform is run locally on the user's machine, and benefits significantly from using a dedicated Graphics Processing Unit (GPU). Specifications for the machine used to run this trial consisted of an AMD Ryzen 5900x 4.8Ghz CPU, XPG Gammix S70 SSD storage (7400MB/s read, 6400MB/s write), 128GB RAM, and a Geforce RTX 3060 with 12GB RAM.

MegaDetector only processes still images, so video footage from the camera traps was converted to still images using the open-source software package 'av'. This package splits video clips into a series of 'child frames' associated with the 'parent video clip', resulting in 10-13 images per clip. This resulted in a total of 489,135 child frames, from 45,427 parent video clips. These frames were then processed using a batch command of MegaDetector using a minimum setting of 0.5 sensitivity.

The output of processed images by the MegaDetector package is a JSON file containing metadata of images processed, including category of object detected, location of the detection (through a bounding box), and confidence level of the detection for a particular image. The JSON can then be loaded into software such as Timelapse2 in conjunction with the original images and detection bounding boxes can be reviewed, and tags of images validated or changed (see **Figure 2**).

Figure 2: Timelapse2 interface, used in conjunction with MegaDetector to review batch outputs of object detection

The statistical computing package 'R' was used to read and unnest the metadata that MegaDetector produced from processing the images. Using 'R', data from processing child frames was condensed to parent video clips, assigning the value of the highest confidence level detection from child frame to its particular parent video clip. 'animal' detections were prioritised over 'human' detections, and 'vehicle' detections were ignored and treated as 'empty'. Data from the manually classified footage in Microsoft Access was then reclassified as either 'animal', 'human', or 'empty' to use the same values as assigned by MegaDetector. The manually classified data tags were then compared with the MegaDetector tags for each video and compiled into a true/false positive/negative table, as well as a precision/recall outcome. The R code used for this process to validate data is provided in **Appendix 1** (not included in this excerpt).

Machine learning classification platform (eVorta/Hamesh Shah)

The cloud-based platform 'eVorta' was used to assess the performance of a machine learning classification platform.

eVorta requires users to upload their footage through a portal in a browser-based cloud interface. Footage is then automatically classified using a model that has been developed through users manually tagging and training data which is processed by machine learning algorithms to improve on the detection and classification of target species.

Uploaded data is processed on request by the developer. The results include a cloud database of images that have detections and target species classified with bounded boxes accompanied with confidence scores relating to their respective classification (see **Figure 3**). eVorta is a commercial platform, with costs being on a per image basis. As well as detection and classification, the platform also provides the user with analytics such as spatial/temporal mapping, a workflow for rapidly reviewing and training footage, summary statistics, a range of basic and advanced filters, and the ability to export results and metadata in a variety of formats.

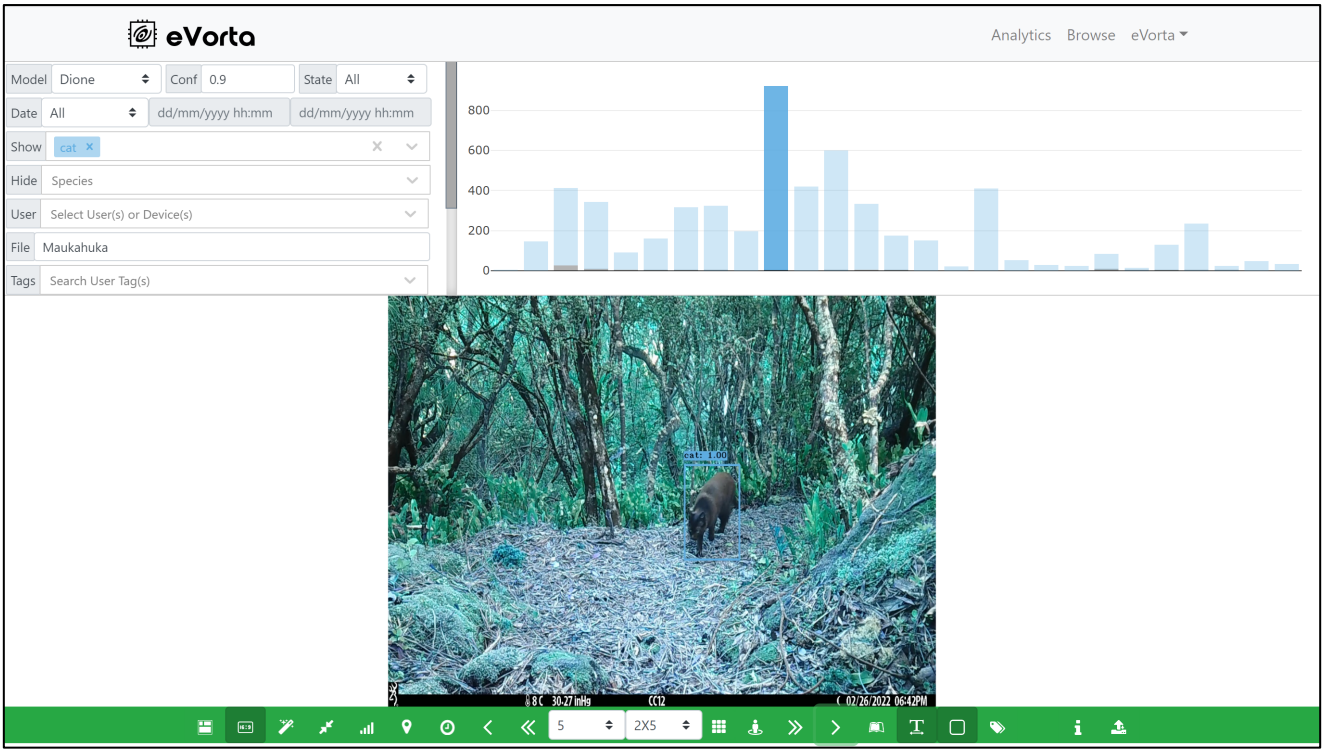

Figure 3: The online interface of the eVorta platform showing search/filter criteria (top left), histogram of search results (top right), and image with bounding box and confidence of classification.

eVorta processes still images but is able to process video footage through automated extraction of key frames for processing – reducing the need to do this manually. 45,427 parent video clips were uploaded to the platform, which eVorta then split into 509,268 child frames for processing.

Initial processing of the footage is run using a pre-existing model based off all previous data processed from other projects and locations. In theory, when datasets contain new environments and species, the initial results from this process should benefit significantly from a small amount of training where users use a tool provided by the platform to train a small subset of the data for unknown species, and localised environmental and fauna contexts. The developer of eVorta advised this step as the training data used to build the current model for eVorta at the time ('Aegir') was only built from 0.02% video sourced cat data. It was advised that a small amount of additional training from our dataset would result in significant improvements to recall and precision. Four hours of training was completed on the initial results using the eVorta platform to draw a bounding box over classified species throughout the dataset, and confirming or retagging classifications on a small subset of the data (<0.15%).

The output of this training generated a new model, which was then used to reprocess the whole dataset.

The eVorta interface provides the ability to export metadata of the results in a database/spreadsheet format. The metadata of the results from eVorta's image processing were exported in CSV format using a tool provided on the eVorta interface.

Using the software package 'R', the eVorta results were filtered to show frames where cats were detected. This data was then condensed from child frames to parent video clips through assigning the value of the highest confidence level cat detection from a given child frame to its parent video clip, and removing remaining child frames from that parent video clip from the dataset. The resulting list of video clips with cat detections was then validated with the results from manual classification.

This process was run in iterations using confidence scores for cats ranging from 0.01-1.00, using a programming loop in R.

Results were summarised in terms of precision and recall.

Results

Manual classification

We used manual classification results as the baseline for a comparison of results with MegaDetector and eVorta, and assumed perfect recall and precision for the detection and classification of cats by humans. It took 74 human hours to classify the video footage, with a further estimated six hours associated with processes of importing footage to the database, setting up site surveys, and exporting metadata.

MegaDetector

Converting 45,427 video clips to 489,135 frames to enable processing with MegaDetector took approximately 48 hours.

Batch processing the frames to produce a JSON file (containing bounding box locations and confidence scores but not replicating imagery with this overlaid) took approximately 19 hours (approximately 7.2 frames per second; or the equivalent of approximately 2350 video clips per hour). Results for precision and recall for object detection where any animal was present in a frame is shown in **Figure 4**. Review of the false-negative data from MegaDetector showed the platform detected animals in 98.3% of cat events from the dataset (173 of 176 events) at a 0.5 sensitivity setting.

As MegaDetector is an object detection platform, sorting images into one of five broad categories, a further six hours were required for human classification of the 'animal' and 'multiple' categories in order to determine frames where cats were present.

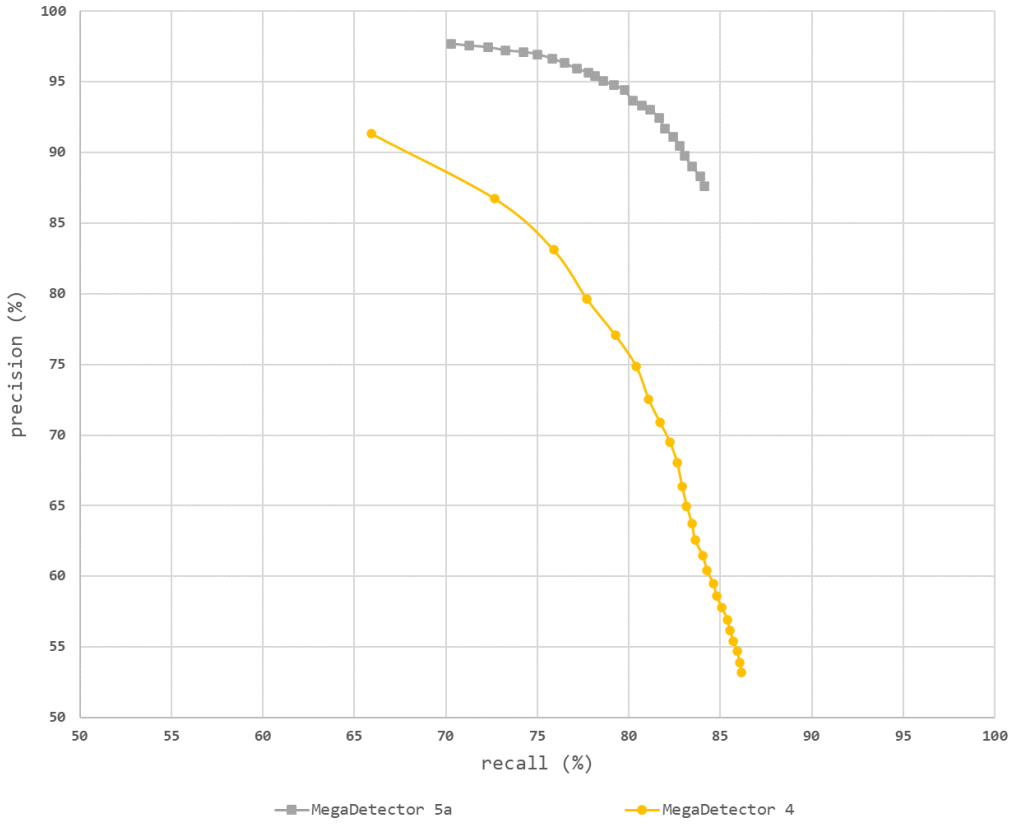

Figure 4: Recall and precision results of MegaDetector for object detection of frames with an animal (any species) present

eVorta

Due to local bandwidth issues, we were unable to test the time taken to upload the dataset to the eVorta cloud platform, and the dataset was physically sent via HDD to the developers to process. Predictions are that it would have taken approximately 55-110 hours to upload the 1.1TB dataset on a 50/10Mbps fibre connection.

Video clips were processed and split into key frames and processed at the approximate rate of 20 frames per second, equating to approximately 6,545 video clips per hour. Initial results were based on the model 'Aegir' - a pre-existing model based on the mostly Australian and north American datasets.

After training a subset of the initial results, a new model 'Dione' was created and the dataset was re-processed under the new model. Results for precision and recall for cat detection and classification with eVorta is shown in **Figure 5**. Review of the false-negative data from eVorta (Dione) showed the platform detected 97.7% of cat events from the dataset (172 of 176 events) when run at 0.99 sensitivity.

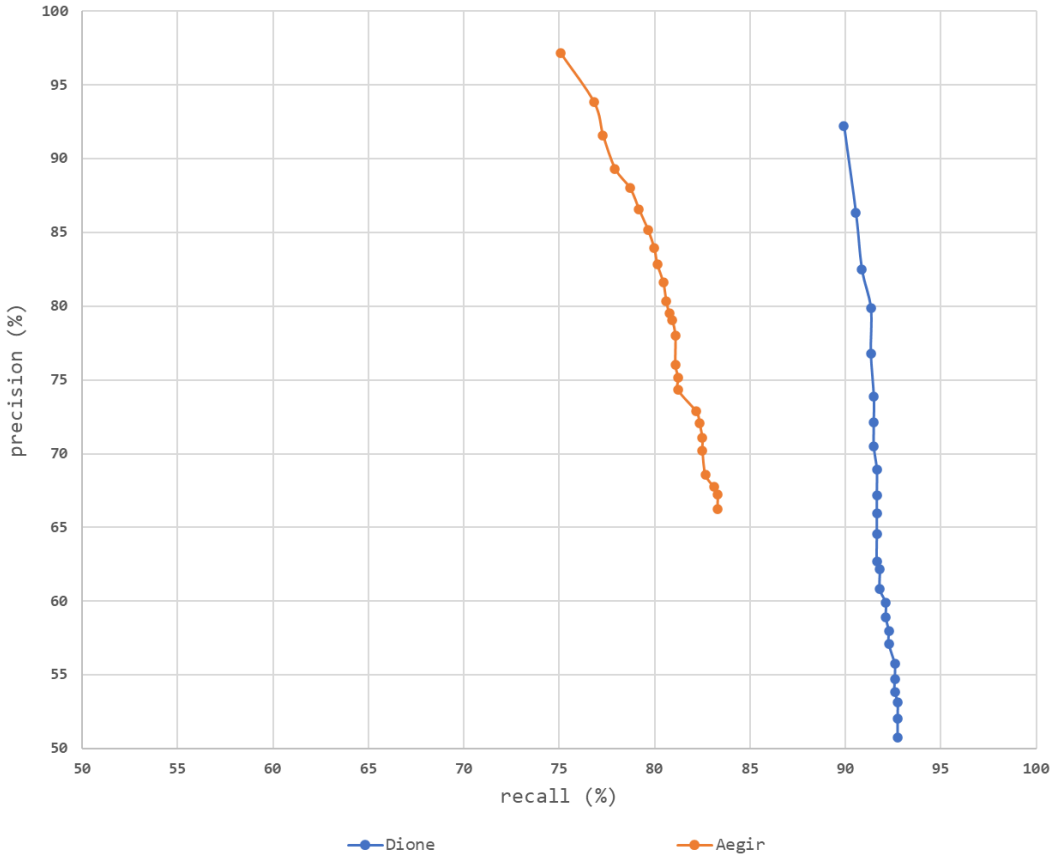

Figure 5: Recall and precision results of eVorta for the detection and classification of cats from footage (using the confidence score range of 0.75-0.99)

Costs and resourcing across platforms

Table 2 provides a summary of approximate cost and resourcing for each platform to reach an output of a classified dataset.

Table 2: Cost and hours associated with each platform

Platform	Hours	Cost	Notes
Manual classification (MS Access GUI)	**136** (labour)	$4760	• Based on persons classifying 3600 images per hour for 489,000 frames. • Hours include 132 hours of classifying time and 4 hours of time to allow for database interaction and exporting of results • Costs based solely on labour of $35 per person hour
MegaDetector 5a	19 (processing) + 8 (labour) **27**	$280	• Based on MD processing rate of 25,920 images per hour for 489,000 images • Hours include processing time of MD, setup and export of metadata and results, and six hours of classifying time based on manual classification rate of 3600 images per person hour • Costs based on $35 per person hour (MD processing time excluded)
eVorta	7 (processing) + 1 (labour) **8**	$4925	• Based on eVorta processing rate of ~72,000 images per hour for 489,000 images • Hours include eVorta processing time, and setup/export of results • Costs based on $0.01 NZD per image processed, and $35 per person hour

Discussion

In the two years since the 2020 review of what various software developers could offer in terms of AI and ML processing of camera trap data (DOC, 2020), there have been significant improvements in 'market ready' platforms.

In terms of processing speed, both of the assessed platforms have improved several-fold (thirty-fold in some instances) relative to some of the AI and ML solutions assessed in 2020, and are significantly faster than what manual classification alone allows. This ultimately provides the ability to upscale to large and intensive projects such as the Auckland Island cat eradication, where management and processing of camera trap data at scale has been considered unmanageable when done manually.

There are advantages and disadvantages between using the two platforms which are outlined in **Table 3**, but ultimately MegaDetector and eVorta serve different purposes. As an object-detection model, MegaDetector can be used to significantly reduce manual classification time by footage where any animal has been detected. For the dataset used in this trial, the use of MegaDetector equated to a reduction of 96.5% of manual classification time required (based on figures in **Table 2**) compared to manual classification alone as a processing method, whilst identifying 98.3% cat events from the dataset. This high rate of detection is likely to satisfy eradication requirements when applied over increased temporal scales and further detections of an individual are likely.

Table 3: Advantages and disadvantages of footage processing platforms

Platform	Advantages	Disadvantages
eVorta	• Detects and classifies to species specific level, rather than a broader object class, therefore saving significant time needed for manual classification. • Significantly faster than other platforms in terms of processing time but also when factoring no manual classification of broader classes is required (as with object detection models). • Scalability is almost unlimited, with cost and speed of uploads the main limitations rather than processing time. • The platform provides a 'one-stop-shop' for the entire camera trap data processing work flow. Data is easily uploaded, manipulated, queried, filtered, analysed, mapped, graphed, and exported all within the browser the platform. • Platform is not static – users can continue to improve precision and accuracy of models and apply these changes to their datasets. New iterations of the models incorporating training are released by developer based on training from all datasets. • Workflow for validating images and detections is very efficient compared to manual classification. • Automatically extracts keyframes if uploading video footage. • Low level of computer literacy required – easy GUI. • Acts as cloud storage repository for data.	• Cannot be run locally - requires internet connection and enough bandwidth to upload large amounts of data. • Footage not processed live or on demand by users – dependent developers activate processing of data. Could potentially lead to dead time between users uploading data and data being processed. • Cost associated with processing (one-off of $0.01 per image).
MegaDetector	• Can be installed and run on a local machine • Can use multiple local machines concurrently to process datasets faster • Is free to use	• Requires a higher degree of computer literacy to setup appropriate environment and run batch commands, as well as being familiar with the different software involved in the workflow. • Requires the use of multiple programmes and commands to produce full workflow. • Requires higher end system specifications, and a CUDA compatible GPU to achieve the higher processing speeds. This effectively means more expensive equipment. Processing speed on a standard DOC Lenovo are 0.4-0.5 images per second, and on a high-spec DOC Lenovo are 6-10 images per second. • Effectively a means of significantly speeding up manual classification by removing human or empty footage, but still requires manual classification of broad 'animal' class results. • Versions have static models so rely on new version releases (approximately every 2 years to date). Users cannot train and improve results unless pre-tagged datasets are sent to the developer and they incorporate them in their model development for future version releases. • For DOC users - Programs are required to be run outside of AWS, posing a risk to users that do not back-up their data on DOC approved data storage platforms such as OneDrive or S3 Browser.

Being a classification platform, eVorta serves the purpose of complete classification of footage without the need for an additional manual classification process. It has the additional benefit of not only classifying cats, but other species present in the dataset. The results of eVorta improved significantly in terms of speed, recall, and precision relative to solutions reviewed in 2020 (see **Figure 6**). Note that solutions assessed in 2020 were based on significantly smaller datasets, and were effectively binary classification models (cat present vs anything else) whereas the eVorta platform classifies into a broad range of species which would be of benefit to most conservation applications. For example, for no extra cost, training, or running time – eVorta classified mouse, pig, and bird, and penguin detections the dataset used for this assessment.

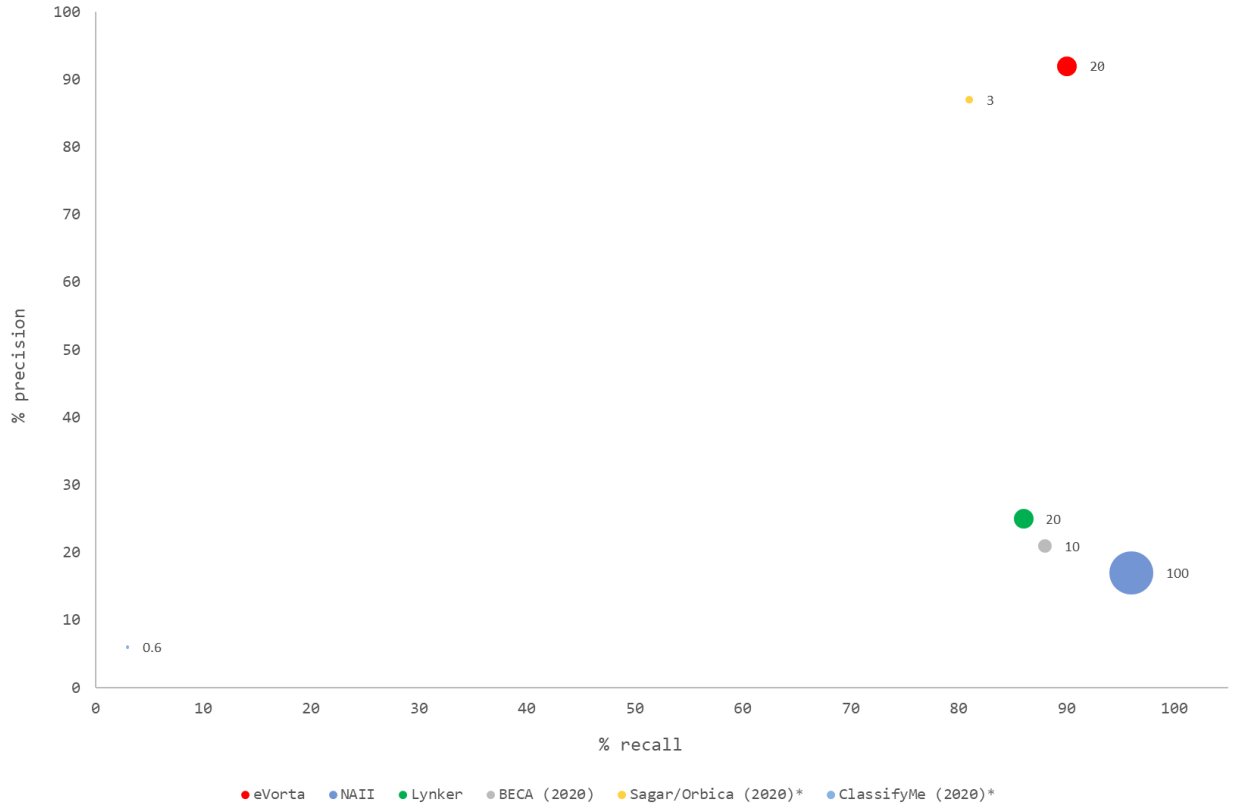

Figure 6: Comparative results of classification solutions reviewed in 2020 with eVorta (0.99 sensitivity settings). Symbology size correlates to processing speed in images per second (also annotated)

The most obvious improvement from eVorta against the other reviewed classification solutions is the significant improvement in recall (89.2%) while maintaining high precision (92.2%). More importantly, it captured 97.7% of cat events that occurred when filtering the dataset with the high confidence setting (0.99). A full scope of advantages and disadvantages we found with the platform are shown in **Table 3**. With 4 hours of manual training, recall and precision results for cat classifications were significantly improved (see **Figure 5**) between the initial model (Aegir) and the model produced from additional training (Dione). If used for a project such as the Auckland Island cat, pig, and mouse eradication – we were advised that a small amount of additional training would be of benefit to the ML algorithms and smaller increments of improvement would occur to precision and recall. The developers also advised that using footage from video clips was a disadvantage compared to still images as keyframes from videos are of a much lower definition than that of still images on the trail cameras used, and the models and ML algorithms have largely been trained off patterns from still images.

Recommendations

The review of MegaDetector and eVorta in this trial have shown the significant improvements in capability that Artificial Intelligence and Machine Learning platforms have made in the past few years. In the context of removing cats from the Auckland Islands, these platforms are performing at a level where they would be a valuable component of the tools used in an eradication context.

These learnings are not necessarily restricted to the Auckland Island project, and both platforms have the potential to be used to significantly reduce costs and staff time on camera trap applications across the New Zealand conservation context; whilst enabling the possibilities of higher intensity of monitoring at scale.

Based on the results from this trial, we make the following recommendations:

- **Encourage investment and familiarity with both tools within DOC for medium to large size camera trapping projects.**

  There are significant resource savings to be made for all medium to large scale camera trapping projects by using either MegaDetector or eVorta.

  If DOC adapts and leads on this, other groups (PF2050, local government, community groups) are likely to follow and benefit also.

- **Invest and collaborate with industry to produce autonomous detection units**
  AI and ML algorithms are at a stage that satisfy the first criteria from the feasibility report concerning the automated processing of image data and identifying cats where present. The challenge now lies in creating autonomous units that can detect and report detections remotely, using the classification algorithms developed. eVorta are currently working on a prototype product – the eV mini –which is connected to a trail camera through its SD card slot and communicates to users through a hub/mesh and/or satellite link. The unit holds a static version of a model from eVorta to detect and send images of a specified target species on detection satellite. In the context of the Auckland Island mouse, cat, and pig eradication (and island biosecurity in general) – there is significant benefits in investing in development of such a device, and improving the models it uses.

References

Department of Conservation 2020. File note: Analysis of the artificial intelligence and machine learning market in relation to camera trap footage review. Unpublished memo, Lindsay Chan. DOC-6306763.

Department of Conservation 2021: Technical feasibility study report for eradication of pigs, mice and cats from Auckland Island. Department of Conservation Te Papa Atawhai, Invercargill, New Zealand, 123 p

Tuia, D., Kellenberger, B., Beery, S. *et al.* Perspectives in machine learning for wildlife conservation. *Nat Commun* **13,** 792 (2022). https://doi.org/10.1038/s41467-022-27980-y