Hyperspectral Seed Imaging SFFF

Nick Davies, Marlon M. Reis, Federico Tomasetto

25th October 2024

Report for Seed Industry Research Centre Inc. "SIRC"

SFFF-22019

AgResearch report number: 12404

agresearch āta mātai, mātai whetū
SIRC SEED INDUSTRY RESEARCH CENTRE
Ministry for Primary Industries Manatū Ahu Matua

Contents

1. Executive Summary 1

Graphical abstract 2

2. Background 3

2.1 Introduction 3

2.2 Scope 4

2.3 Aim and objectives 4

3. Methods 5

3.1 Samples 5

3.1.1 Batch 1 5

3.1.1 Batch 2 6

3.2 Sampling 6

3.3 Hyperspectral imaging 7

3.4 Experimental design, imaging processing, modelling and computational resources 8

4. Results and Discussion 9

4.1 Milestone and outputs 1 (6 Months): 9

4.1.1 Summary of key observation of Milestone 1 10

4.2 Milestone 2 and outputs (12 Months): 11

4.2.1 Transfer of the initial methods and algorithms 11

4.2.2 Enhancement of methods for prediction of contamination 14

4.2.3 Remarks 16

5. Recommendations 18

References 20

1. Executive Summary

New Zealand's seed industry both exports and imports certified seed. Approximately 36,000 tonnes of pasture and vegetable seed are exported annually, with a direct export value of $194 million. The country provides out-of-season production for Northern hemisphere growers, and its seed quality assessment methods rely on International Seed Testing Association standards, including physical, physiological, biochemical, and molecular evaluation. While services provided by ISTA accredited labs in New Zealand are effective, they are time-consuming, labour intensive, and require highly trained seed experts. New tools are in demand to ensure the long-term suitability of this service.

This project aimed to leverage emerging technologies, including hyperspectral imaging and artificial intelligence to improve seed detection, increase throughput, accuracy, and reduce costs. The project investigated the potential for hyperspectral imaging and AI to be developed into a system which could be used as a pre-screening tool for the human seed technicians, the intention being that seed technician's time efficiency could be greatly increased by providing them with samples containing largely contaminants to identify rather than the status quo where their samples are almost entirely crop seed. This will enhance quality assurance for domestically grown seed for both domestic and export markets and provide an extra layer of biosecurity protection when seed is imported.

The validity of the inspection system requires that crop seeds (meant to be clear of contaminants) are detected as such and everything else is identified as contaminants. The system must cope with the presence of contaminant seed species which were not in the training or testing data during development while still requiring a low level of false negative results (where contaminants are not detected as contaminants). This project addressed these requirements in a stepwise approach: first characterizing the ability of hyperspectral imaging to differentiate seeds; then investigating more complex scenarios where contaminant seeds were very similar to the crop seed. Hyperspectral cameras were tested under a series of differing conditions for the collection of data: varying image resolution; type of illumination; speed of acquisition, and types of acquisition. Three types of image processing were utilized, and a series of machine learning algorithms were applied, including 3D-convolutional neural networks (3D-CNN). Data for individual seeds were collected for more than 25,000 seeds from 15 different species.

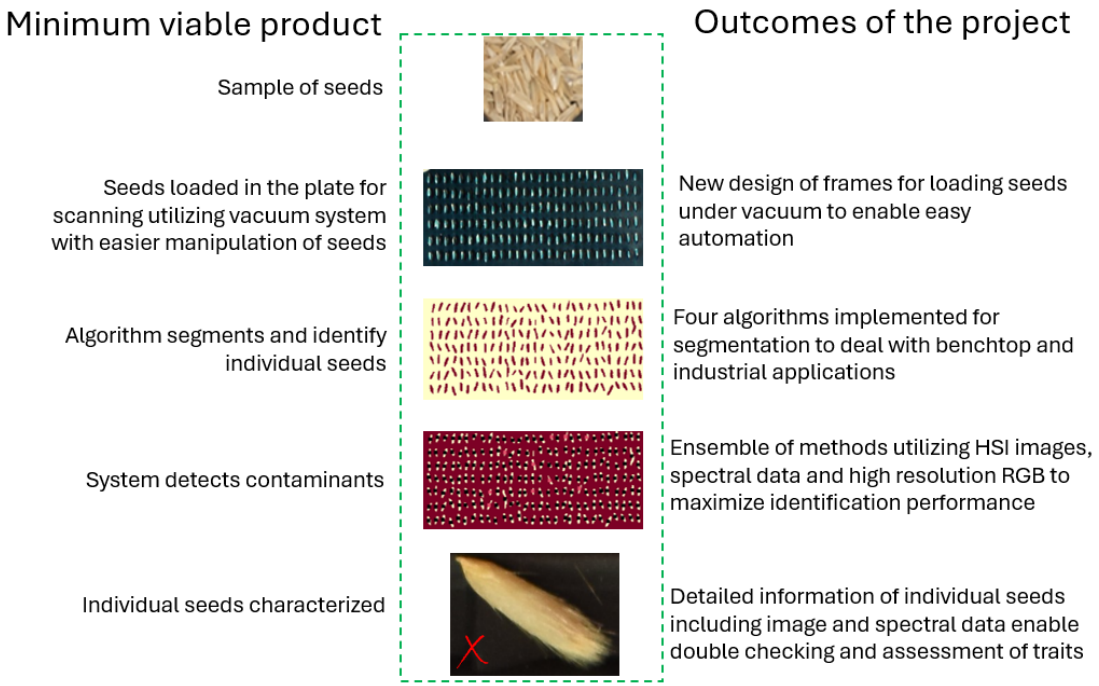
Overall, the project achieved its goal to assess the detection of wild oat seeds (contaminant) in the presence of ryegrass, wheat, and barley (crop seeds). The models based on hyperspectral data achieved up to 100% of correct classification depending on spectral range utilized. The use of morphological information enables the use of a three-tier approach for reducing the chance of false negatives. Our learning suggested that the best approach is to utilize an ensemble of machine learning techniques to provide extra layer of protection during the assessment of the seeds, lowering the risk of missing contaminant species.

A system for detecting contaminant seeds within crop seeds was developed and validated from both technical and commercial perspectives. Independent assessments were conducted to evaluate the technical capabilities of the system, gaining insight into potential risks and opportunities. Multiple scenarios for commercial implementation were explored in the study, which enabled the development of a risk-based plan for future development. A suite of capabilities was gained, including algorithms for data processing and modelling, techniques for scanning seeds, and a database of hyperspectral images of seeds. The

system is designed to collect data on individual seeds, enabling further characterization of seeds in regard to other traits. Additionally, the data from individual seeds could be used to assess seed quality, perform phenotypic characterization, or obtain other commercial traits. The outcome of this project has the potential to benefit the seed industry in New Zealand, particularly at borders, seed labs, and other commercial ventures. It will provide wider benefits by enhancing quality control and biosecurity.

For border security it is recommended that a minimum viable product with optimal conditions for imaging, spectral resolution, illumination and imaging processing to scan 120,000 seeds per hour is implemented. Data should be collected over at least one year aligned with the current testing to build a database covering seasonal batch variation and describing several scenarios of contamination risks. For seed lab usage, an extension to look at physical and genetic purity could be worked on along with other physiological traits. Of particular interest would be if genetic purity could be accurately identified for hybrid cultivars to detect cross contamination.

Graphical abstract



2. Background

2.1 Introduction

New Zealand has an important seed industry that both exports and imports certified seeds. Approximately 36,000 tonnes of pasture and vegetable seed is exported annually, with a direct export value of $194 million [1]. New Zealand is a disproportionately important piece of the global seed supply chain as it provides out of season production for Northern hemisphere growers. Traditional methods for seed quality assessment are based on International Seed Testing Association (ISTA) standards and Seed Field Production Standards, including physical [2], physiological, biochemical, and molecular evaluation. Testing is delivered by ISTA accredited labs in New Zealand, although seed testing is effective, it is also time-consuming, labour intensive, requires highly trained seed experts and some misidentifications occur [3]. One of these labour-intensive tasks in seed labs and at border inspection is identifying non-crop seed contaminants. Given seed lots generally have a low proportion of non-crop seed, labourer's time and expertise in identifying plant species from seed morphology is not optimally used. Seed assessment based on spectral imaging is currently being widely investigated [5] and this project aims at positioning New Zealand to take advantage of these emerging technologies. The adaption of current hyperspectral technology combined with modern artificial intelligence techniques has the potential to greatly increase the speed and volume of seed tested for contamination at lower costs by creating samples with high proportions of non-crop contaminants for seed technicians to identify, rather than the status quo requiring large proportions of time sifting through the expected crop seed within samples. Thus, this project has important ramifications for:

- Quality Assurance of domestically grown seed for domestic and export markets – Improve foreign (contaminant) seed detection: increase throughput, increase accuracy (improve detection rates of regulated species) and reduced cost.

- Border biosecurity – New Zealand imports and exports large quantities of seed, should the feasibility of this tool be proven, this work has the potential to reduce the chances of biosecurity incursions via seed contamination and reduce the risk of exporting contaminated seed shipments to other countries which has market access, reputational and financial consequences.

- Labour & demographics – seed testing is a laborious and specialist role currently covered in New Zealand by an aging workforce – variability between labs and researchers is significant (ISTA [3], Buddenhagen et al. [4]). Recruitment of new staff in this field is challenging. This project shows potential to change/automate aspects of seed certification to help alleviate this labour issue. Particularly by increasing the proportion of seed the technicians observe being non-crop seed, and by achieving more efficient use of their time and expertise.

- Reduced costs and domestic uptake – a seed certification technology/process which is cheaper could lead to greater use of seed certification for domestic operators being focused on the domestic market improving internal biosecurity, making weed management of crops easier, and improving productivity.

2.2 Scope

This project delivers a feasibility study to evaluate the use of hyperspectral imaging and modern artificial intelligence technologies to identify weed seed contamination of seed lots. The study investigated existing technologies to detect wild oat (*Avena* spp.) seeds within ryegrass, wheat and barley seed samples using the same sample sizes typically used by seed certification and boarder security technicians. Wild oats are regarded as a restricted contaminant during seed testing and field inspections in some seed crops [2]. This project:

- Developed an innovative workflow using an engineered platform to organize seeds for successful Hyperspectral Imaging (HSI) and used Artificial Intelligence (AI) to distinguish seed morphological differences. This could reduce processing time, increase contaminant detection rates, reduce error rates and time costs. Eventually such technology could help to alleviate the predicted seed testing bottleneck as testing capacity is reduced.

- Assessed the feasibility of AI and HSI to detect contaminants in samples of ryegrass, wheat, and barley seeds.

- Built a roadmap outlining what work would be required to bring this technology into use in seed labs and other purposes.

2.3 Aim and objectives

The project aimed to:

- Investigate the detection ability of different types of seeds by hyperspectral imaging and AI/machine learning techniques.

- Create known levels of contamination in seed samples and develop more sophisticated algorithms to identify contaminant seeds.

- Assess possibility of transferring initial methods and algorithms to an existing platform (i.e. Clarospec^{TM}) and provide early indications of the speed-predictive performance trade-off for the detection of wild oat seeds in ryegrass, wheat and barley seed lots.

- Assess if this new technique will meet the requirements of the modern industrial control and sorting systems of seeds

- Suggest further research and development directions

3. Methods

The approach utilized in this study employs crop and weed seeds to mimic a seed testing procedure. Trays of crop seeds with and without the presence of contaminant seeds were scanned by hyperspectral and colour (RGB) cameras under varying conditions (e.g. number of seeds, mixture of seeds). The resulting data was assessed with machine and deep learning methods to associate the data from hyperspectral and/or RGB cameras with the presence/absence of weed seeds (contaminants).

3.1 Samples

The main goal of the project was to assess three crop seeds against presence of one contaminant (wild oat). But the development of models only utilizing two types of seeds could result in optimistic methods (i.e. not capable of dealing with out-of-training-set contaminants either within the species but differing in physiological presentation, or from outside the species). Thus, the conditions of scanning were assessed utilizing additional seed types. Two batches of samples were utilized in the methodological development below.

3.1.1 Batch 1

Hyperspectral imaging involves a set of customizable parameters depending on (1) sample types and (2) image resolution (defined by type of lens and distance of the camera to the sample) which affects the amount of light necessary. Thus, preliminary investigation was required to assess the best scanning conditions.

As preliminary work to identify our initial conditions for scanning the seeds in our hyperspectral system, 13 seed types were used. The seeds varied in species, size, morphology and characteristics (e.g. coated; see Table 1).

Table 1 – Seeds utilized in the preliminary work to adjust initial conditions of the hyperspectral system.

Species	Seed image
***Avena fatua* – wild oat**	
***Avena sativa* – oat**	
***Avena strigosa* – black oat**	
***Bromus catharticus* – prairie grass**	
***Dactylis glomerata* – cocksfoot**	
***Hordeum vulgare* – barley**	
***Lolium arundinaceum* – tall fescue**	
***Lolium perenne* – perennial ryegrass**	
***Lolium perenne* – perennial ryegrass (Nui)**	
***Lolium* sp. - ryegrass**	
***Secale cereale* – ryecorn**	
***Triticum aestivum* - wheat**	
***Triticum aestivum* – wheat (with seed coating)**	

3.1.1 Batch 2

To develop preliminary models for hyperspectral imaging, 14 seed types formed the second batch. The seeds significantly varied in morphology (Table 2).

Table 2 – Seeds from batch 2 (Variation: 1 – mixed, 2 – big and small, 3 – small, big, mixed).

Row Labels	Variation	
Avena fatua – wild oat	1	
Avena sativa – oat (milling)	3	
Dactylis glomerata - cocksfoot	3	
Fumaria muralis – fumitory	1	
Hordeum vulgare – barley (Chertsey)	3	
Linum usitatissimum - linseed	3	
Lolium arundinaceum – tall fescue	3	
Lolium arundinaceum – tall fescue (Fortuna)	2	
Lolium multiflorum - annual ryegrass	3	
Lolium perenne - perennial ryegrass	2	
Lolium perenne – perennial ryegrass (Nui)	3	
Trifolium repens – white clover	3	
Triticum aestivum – wheat (Kowhai)	2	
Vulpia sp. – hairgrass	1	

3.2 Sampling

Samples were organized in a grid system with rows and columns properly labelled (Figure 1). This enabled tracking the position of each seed. This was done repeatedly in a series of experiments with one or with multiple seed types. Seeds were added manually and individually to the grid to ensure accuracy.

Figure 1 – Two grid systems utilized to scan the samples.

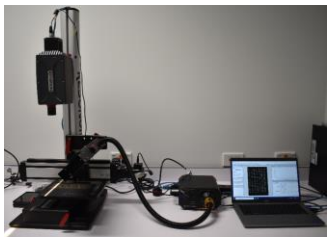
3.3 Hyperspectral imaging

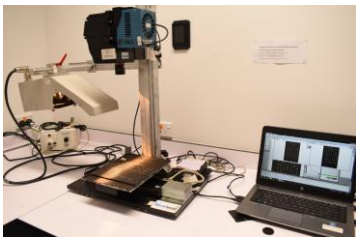
There were five hyperspectral camera systems used in this study (Table 3, Figure 2). Several scanning conditions were evaluated mainly varying the resolution of the image adjusting the distance of the camera to the sample and utilizing appropriate lenses along with optimizing the lighting system.

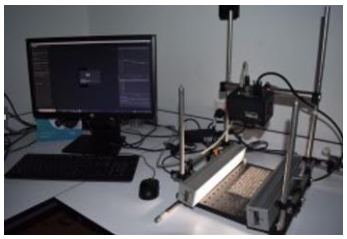
Table 3 – Hyperspectral cameras utilized in the study (and wavebands covered in nm).

Camera	Specification
Snapshot	25 channels (670-960)
Snapscan	150 channels (470-900 nm)
Resonon	168 channels (896-1713 nm)
Headwall	235 channels (550-1700 nm)
Specim IQ	204 channels (400-1000 nm)

Figure 2 – Three hyperspectral systems most used in the study, from left to right: Resonon, Headwall and Snapscan.







3.4 Experimental design, imaging processing, modelling and computational resources

The scanning of seeds was carried out either utilizing a mix of seed types on the same grid or with grids having only one type of seeds. This allowed the generation of several datasets for analysis.

Each hyperspectral image collected within a grid contained data for several seeds. Different algorithms were deployed to identify single seed (e.g. segmentation) and data for each seed were extracted. Three approaches were used for processing the images with focus on seed segmentation. Approach 1 was based on traditional image analysis which utilized a threshold of intensity distribution, i.e. if the detected intensity in each wavelength for the pixel is below a set threshold it is considered background otherwise it is considered a seed. Data pixels per seed in each grid cell were labelled accordingly. Approach 2 was based on hyperspectral images (with principal component analysis and thresholding based on scores of principal components) and approach 3 used a deep learning methods that identified all seeds in any image using a single model [6, 7]. After segmentation, seeds were identified individually and morphological information was extracted [6].

The machine learning methods utilized were support vector machine (SVM), XGBoost, random forest and convolutional neural networks (CNN) [8-9]. Models were based on two or more classes, where one class was the crop seed and the other contaminants. All calculation was carried in R or Python. R was utilized for data processing, SVM, XGBoost, random forest and visualization. Python was utilized for deep learning. R was utilized in Central Processing Unit CPU based computing and deep learning utilized Graphical Processing Units GPU as resources to reduce the time spent on optimization of models with large numbers of parameters.

4. Results and Discussion

4.1 Milestone and outputs 1 (6 Months):

Aim: Investigate the detection ability of different seed types by hyperspectral imaging and AI techniques and develop calibration methods.
Deliverable: A report detailing calibration methods / preliminary analysis from AgResearch spectral instruments for early indication of predictive performance of the detection of wild oat seed in ryegrass, wheat, and barley seed lots.

Preliminary work was carried out to assess the ability of hyperspectral imaging techniques to discriminate seeds from batch 1 and to identify initial conditions to operate Headwall hyperspectral system to scan seeds.

A database of hyperspectral scans of approximately 23,000 seeds was created, including several seed species (classified as small, big, mixed sizes within species). This ensured capturing variation in both crop and contaminants (Table 1, Table 2). The Headwall system would allow scanning of 120,000 seeds per hour with our speed of scanning if these were scanned continuously (i.e. if the constraining factor on speed was exposure time).

The segmentation method (Approach 1 – Section 3.4) based on intensity of selected bands from hyperspectral data was implemented and a method for quickly extracting morphological seed traits (e.g. size and shape, texture) from a single near infrared wavelength was developed, utilizing a deep learning model (Approach 3 – Section 3.4) [6].

The potential to identify contaminants based on hyperspectral data was assessed utilizing a series of machine learning models (i.e. SCM, XGBoost, RF) with two approaches: (I) based on average spectrum of each seed and (II) based on spectra of pixels per seed. In this case, SVM showed the best performance when compared to XGBoost and RF. For the Approach II, Hierarchical models with various dimensionality reduction techniques, including Principal Component Analysis and Linear Discriminant Analysis were investigated for pixel level classification. The prediction performance of SVM model increased when spectral data was combined with morphological data (aspect ratio, convex hulls etc). Overall, this preliminary model showed the potential of spectral data to discriminate targeted seeds.

Further investigation was carried out with four scenarios where the crop seed was contaminated with other three ‘weed’ types with higher morphological similarity, such as:

Crop: Wheat, contaminants: Barley, Wild oat, Milling oat

Crop: Barley, contaminants: Wheat, Wild oat, Milling oat

Crop: Perennial ryegrass, contaminants: Annual ryegrass, tall fescue, cocksfoot

Crop: Annual ryegrass, contaminants: Perennial ryegrass, tall fescue, cocksfoot

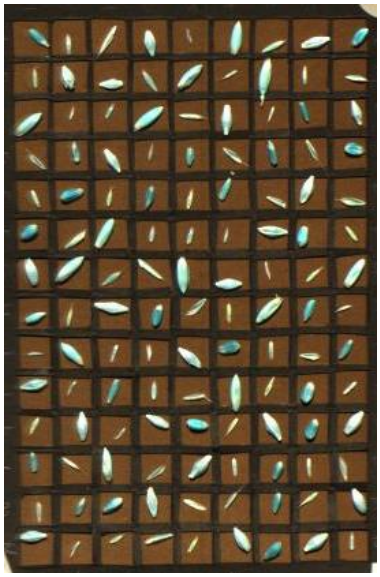
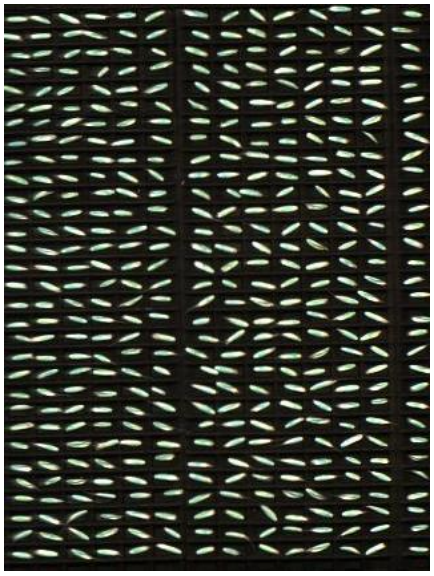
A summary of the SVM modelling for scenario 1 (Table 4 showed that morphological information contributed to performance improvement, where three of four cases the rate of miss classification were lower.

Table 4 – Accuracy estimates of mixed similar species estimated using SVM and data collected from Headwall system (refer to Table 3) – Wheat is the crop seed and each row corresponds to one type of contaminant.

Contaminant	Correct	Miss	Miss (%)	Correct	Miss	Miss (%)
	Spectra only			Spectra + morphology		
Barley mixed	381	24	5.9	386	19	4.7
Barley small	339	14	5. 0	337	16	4.5
Milling oat mixed	485	19	3.8	476	28	5.7
Milling oat small	383	59	13.3	390	52	11.8
Wild oat	438	47	9.7	449	36	7.4

It was also possible to optimize the system (Figure 3) where it would be possible to scan 151,200 seeds in two hours, and if pixel resolution was increased by 20% this would still scan 125,000 seeds per two hours. Assuming scan time is the bottleneck.

Figure 3 - Example of two seeds distribution in a grid design. It illustrates the importance of optimizing the scanning design.

[120 seeds/ frame] x [3 frames/min.] x [120 minutes] = 43,200 seeds in two hours	[420 seeds/ frame] x [3 frames/min.] x [120 minutes] = 151,200 seeds in two hours
	

4.1.1 Summary of key observation of Milestone 1

- The rate of false negatives needed to be decreased.
- The methods to maximize number of seeds scanned would be further investigation.
- The methods to estimate morphology was successful and improved classification performance.

- Benchmark data on performance was obtained.
- A database with HSI data over 23,000 seeds was produced.
- Milestone achieved successfully.

4.2 Milestone 2 and outputs (12 Months):

Aim: Create known levels of contamination in seed lots and develop more sophisticated algorithms to identify contaminant seeds
Deliverable: Write a report detailing the transfer of the initial methods and algorithms to an existing platform (i.e. ClarospecTM) and early indications of the speed-predictive performance trade-off for the detection of wild oat seed in ryegrass, wheat and barley seed lots.

4.2.1 Transfer of the initial methods and algorithms

Transfer of the initial methods and algorithms to Clarospec™

The ability to transfer the outcomes of the project in a commercial implementation depends on its ability to: (I) scan 5 × ISTA (150,000), or 1 × ISTA (30,000) seeds within a suitable time frame; (II) acquire images with enough resolution to maximize the detection of contaminant seeds; (III) provide a computational framework where models identified from this research project can be incorporated. In brief, ClarospecTM was not fit for purpose for rapid imaging of seeds for our machine learning based image recognition of contaminants. It was physically too large, provided redundant features, and image resolution was insufficient. It was agreed, rather than investing in trying to modify a version of ClarospecTM to fit this project, to use the learnings of the ClarospecTM builds, and consider a lower cost, lower complexity benchtop version that could be instead developed.

Pathway towards commercial system

As ClarospecTM was not suitable in its current format, we performed an assessment of five hyperspectral cameras available at AgResearch, to identify alternative pathways for commercialization. Each camera has distinctive specifications that impact commercial implementation (Table 5). Models were fitted mimicking the spectral resolution (and range) of the cameras (based on our exiting hyperspectral library) to identify how much the spectral range of each camera would contribute to improve performance of the detection of contaminant seeds (i.e. wild oat). Finally, a camera investigation was carried out regarding the value of image information. This was performed for the most challenging case (i.e. the detection of wild oat in ryegrass). We identified the Resonon and Headwall cameras showing the best performance across models (based on spectral information only), especially in detecting wild oat in ryegrass (Table 5). Headwall has wider spectral range (550-1700 nm) as compared to Resonon (896-1713 nm). The two cameras present an overlapping spectral range over 1000 nm, which suggests that spectral range above 1000 nm is important for models based on spectral information. The Resonon camera is less expensive than Headwall and provides the python support necessary for a commercial implementation, thus it was further investigated for deep learning.

Table 5 – Assessment of cameras regarding commercial implementation.

Camera	Specification	Python support*	Speed**	Optics***
Snapshot	25 channels (670-960)	Yes	Ultra-fast	Flexible
Snapscan	150 channels (470-900 nm)	Yes	Suitable	Flexible
Resonon	168 channels (896-1713 nm)	Yes	Suitable	Flexible
Headwall	235 channels (550-1700 nm)	No	Suitable	Flexible
Specim IQ	204 channels (400-1000 nm)	No	Suitable	Fixed

* Python support: This allows to run the camera using python, which enables customized commercial implementation where models can be incorporated leading to simple end-user interface.
**Speed: 'Ultra-fast' the faster camera allow collection of up 180 hyperspectral images in one second; 'Suitable' enable scanning of 200 seeds in 10 seconds.
***Optics: 'Flexible' allows different lens to increase image resolution
'Channel' corresponds to one spectral band.

Table 6 Assessment of importance of the spectral range and resolution of each camera on the detection of contaminant ('wild oat') as compared to crop seed. In this case, only spectral information of each seed was used in SVM (support vector machine) models to detect wild oat.

Crop	Corresponding spectral range for the instrument	Wild oat (number of seeds)	Crop (number of seeds)	PPV (Positive Pred Value)	NPV (Negative Pred Value)
Crop=Wheat	HeadWall	258	257	**1.00**	**0.99**
	Snapshot	258	257	0.95	0.96
	Resonon	258	257	**1.00**	**0.99**
	Snapscan	258	257	0.97	1
Crop=Barley	HeadWall	257	257	**0.99**	**0.98**
	Snapshot	257	257	0.98	0.96
	Resonon	257	257	**0.98**	**0.98**
	Snapscan	257	257	0.96	0.99
Crop=Ryegrass	HeadWall	257	257	**0.98**	**0.99**
	Snapshot	257	257	0.96	0.98
	Resonon	257	257	**0.99**	**0.99**
	Snapscan	257	257	0.95	0.98
Note:	$PPV = \frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}}$		$NPV = \frac{\text{True Negatives}}{\text{True Negatives}+\text{False Negatives}}$		

Industrial seed cleaning

Two categories of seed volumes were considered: low volume bench top devices (which sort seed at rates of grams to kilograms per hour) and high-volume cleaners which sorts seeds at tones per hour (Table 5).

For the low volume category, we concluded that with further work this would be feasible. The challenges to be solved in the future would include working at about 100-200g of ryegrass seed per hour. This could be extended to one kilogram or so an hour but would require lighting sources that enable fast image acquisition without overheating the system along with having high enough throughput in computational hardware. This is achievable through more efficient programming and high-end workstation hardware. For example, we have been

investigating the use of lower resolution on HSI, along with RGB sensors to generate images with very good spatial resolution. The merge of the two datasets could help maximize the benefits of faster HSI without losing spatial resolution for morphological characteristics.

Even in the low volume category, the automation of loading seeds into the hyperspectral cameras view in a fast and reliable way was needed. A possible solution could sit with a conveyer or similar systems which maintains a single layer of seed, and ideally little or no touching amongst seeds. The simple approach we developed involved the use of vacuum on specifically designed (Figure 4) and 3D printed trays (Figure 5) which could be placed over a seed sample. In this case, the seeds were separated and stayed in the tray locators or in the needles on the array. Further work is needed to implement an automated system for individual seed removal.

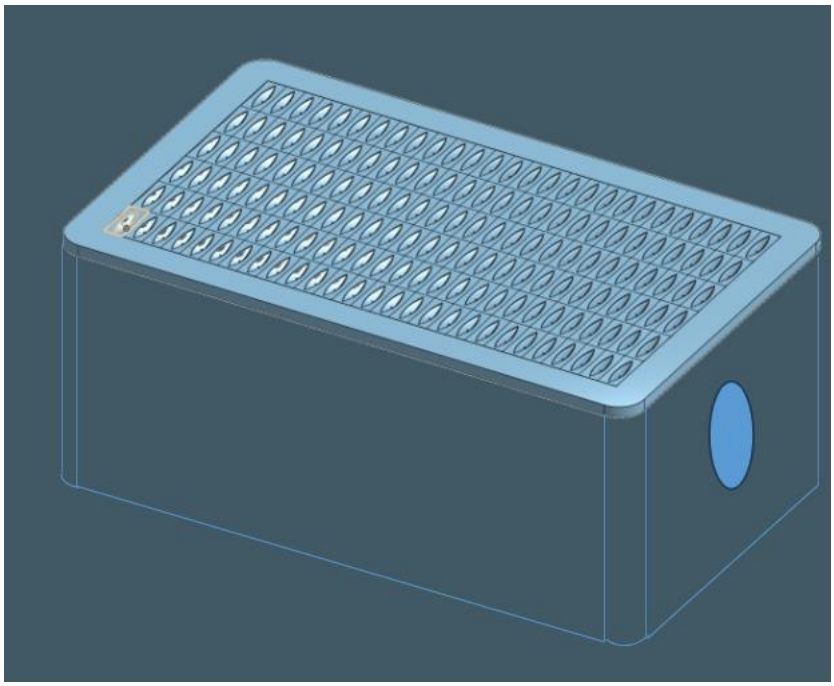


Figure 4 - Seed tray on vacuum frame. A vacuum pump applies negative pressure resulting in seeds being sucked into the wells. If seeds are too small and fall though the wells, they are captured in the box by filter paper on the outlet and identified by seed technicians. This separates seeds mostly into individual wells. The seed trays are software generated for any seed dimensions and shapes.

Figure 5 – Frames utilized for scanning seeds. The 'black dots' in the middle of each well corresponds to the hole to capture the seeds during the process of preparation of the seeds

For the high-volume category, the challenge would be substantially more difficult. We were not able to identify a low number of wavelengths which taken individually can provide high levels of detection of similar seeds. While for selected individual crop-weed pairs this may be possible, it does not appear that it could be generalised well to other situations. The reason for identifying a small number of wavelengths in this case (where the hardware is generalisable to multiple crops with only software changes, and at speeds of tones of seed per hour) is twofold. At the hardware level, a sensor must be exposed to each wavelength of interest. To achieve this, light is filtered to a specific waveband before it reaches the sensor. With a small number of wavelengths, either multiple sensors can detect each waveband simultaneously, or a single sensor can rotate through different filters and capture them sequentially (this is how hyperspectral imaging works). Because each waveband requires an exposure time (which is related to the amount of light illuminating the target), more broad-spectrum light needs to be applied (more light = faster exposure) to reduce the exposure time to be low enough so that seeds can be captured at a significant resolution to enable identification. However, these broad-spectrum lights produce significant levels of heat, while LED's do not produce much heat, they do not produce broad spectrum light, causing other limiting factors. Ultimately, with current technology, we concluded that sufficiently low exposure times to get the throughput volumes required without damaging the seed cannot be achieved. Moreover, because of the sheer volume of data which must be processed through non-trivial analytical pipelines, the computational requirements for processing this data would not be easily obtainable without specialist infrastructure. With current technology, a small number of bands would need to be identified for this to be feasible, which our preliminary analysis showed no evidence for.

Table 7 - Suggested approaches for different orders of magnitude of seed volumes.

Seeds per hour	Approach
Less than 50,000 seeds per hour	HSI with manual loading
50,000- 500,000 seeds per hour (0.1-1 kg ryegrass, 0.4-4kg wheat)	HSI + RGB with manual or simple loading mechanisms
1-100kg per hour	Ultra-fast snapshot HSI (with spectral range above 1000 nm that is available in the market) + high speed RGB Seeds are scanned in high-speed conveyer (4 cm/second) with belt design for seeds. Seeds are loaded with automated systems.
More than 100kg per hour	Probably not doable with current technology in a way which outperforms existing commercial products.

4.2.2 Enhancement of methods for prediction of contamination

As described in Section 4.2.1, the Resonon camera was selected for further investigation. Additional testing was carried out to enhance image resolution (assuming scanning of 200 seeds in 10 second) and deploying deep learning (3D-CNN) techniques in detecting wild oat amongst ryegrass seeds. The outcome showed that for the internal validation set utilized during training, a 2% misclassification of wild oat seed as ryegrass and a 0.7% of misclassification of ryegrass as wild oat seed were found. This is in line with models based

on spectral data only. While results were limited (as all the seeds of wild oat or ryegrass seeds were in individual trays and not mixed) it informed us that it could be possible to use an additional approach in the final system to improve the detection of wild oat. A second test was carried out where models were applied on independent sets of seeds including both ryegrass and wild oat as shown in Figure 6. In this case, the prediction performance was less effective, where there were eight contaminants' seeds within 201 seeds, and two were misclassified (Table 8 - HSI/3D-CNN).

A hyperspectral system for the assessment of seeds included: sample presentation, a camera, a lighting system and a processing unit. Image processing and assessment of performance for each individual seed was carried out. These components are interconnected where speed of data collection depends on the camera's ability to acquire a good enough signal in period of time that is closely connected with the lighting system. The success of assessment depends on the image resolution of the camera, that is dependent on speed of data collection. This project focused on a practical application, and we utilized scenarios to enable the most suitable design within a set of constraints. The commercial feasibility of the system can be constrained by the cost of the camera, which might require compromising the image quality, which is also affected by the speed required for the application. Three approaches were evaluated (with same dataset corresponding to Figure 6): 1) 3D-CNN, an algorithm that maximizes both information of morphology and spectral data on the same models; 2) spectral data; 3) high resolution RGB image, that maximizes the morphological information. Overall, the approach based on spectral data performed the best, followed by 3D-CNN and RGB (Table 8). These three approaches can be combined as 1 and 2 both utilize hyperspectral data and 3 (RGB) is not very expensive to add to the hyperspectral system. The main advantage of combining the three techniques is to mitigate the presence of false negatives.

The study carried out by Buddenhagen et al. showed that a seed lab inspections at an accredited ISTA lab detected 5–95% of spiked contaminants for perennial ryegrass [4]. They observed that Dactylis glomerata species were among the most difficult to detect, with <10% being detected by the ISTA seed analysts [3].

Figure 6 – Seeds of ryegrass with presence of wild oat seeds indicated by the presence of white arrows and within a magnified region (white rectangle).

Table 8 – Assessment of two approaches to detect presence of wild oat in ryegrass. There were 201 seeds in the scanned frame, where eight seeds were from wild oat.

Method	P	N	FP	FN
Spectra/SVM*	8/8(100%)	193/193(100)	0	0
HSI/3D-CNN	6/8 (75%)	189/193 (98%)	2/8 (25%)	4/193 (2%)
RGB/3D-CNN	3/8(37.5%)	187/193 (97%)	5/8 (62.5%)	6/193 (3%)
Note: P – positive – weed detected; N – Negative – seed detected; FP – False Positive – weed missed; FN – False Negative – seed missed				

4.2.3 Remarks

One of the original goals of the project were to transfer the findings of the research to a system designed to be operated in an industrial environment, and the initial proposal was to utilize ClarospecTM (an industrial hyperspectral system developed by AgResearch). ClarospecTM was determined to be an unsuitable due to its inability to obtain the required resolution and the physical size being substantially larger than what is required. To fulfil this gap, we investigated a series of trade-offs between speed of scanning, image resolution and conditions (e.g. illumination) needed to achieve the best trade-off. A knowledge basis was built to inform future commercial implementation.

As part of our work to optimize speed of scanning we observed the need for a very intense illumination system, which produces more heat than expected with potential to damage the seeds during the scanning process, potentially leading to incorrect estimates, or health and safety concerns. The amount of light delivered to each seed limits the speed of sample

imaging, particularly in where very high-speed imaging is required without damaging the seeds. In this case, the potential solution is the use of very focused light that does not increase the temperature of the system substantially. LED light sources are a posable solution, however the physical nature of the light produced from LED sources is that each diode produces a specific wavelength, meaning either specific individual wavelengths need to be identified as having high levels of information encoding potential and the LEDs built for those wavelengths, or a very large number of different LEDs are needed to produce a light source covering the hyperspectral system (Some examples use a small number of wavelengths such as the LED light source used for the automation of seed purity tests at Bayer Crop Science;

https://www.seedtest.org/api/rm/GB7538U558WW742/pur-application-of-msi-to-digitalize-seed-purity-a.pdf). We tested a fibreoptic tube combined with cylindrical lenses to move the light/heat sources further away from the imaging plane reducing heat near the seeds.

For border security or seed laboratory testing where significantly lower speeds are required, less light intensity is required and hence less heat, suitable speeds could still be met with automation of tray filling or other engineering solutions to increase the speed in other parts of the system, as the camera exposure time is of less importance. Thus, engineering solutions were investigated to mitigate some of these problems including: the filtering of seed which is substantially smaller than the target crop seed for human investigation; and the automation of tray filling for image capture as discussed above.

We investigated more complex analytical methods to increase accuracy of detection of contaminants which are morphologically close to the target crop seed in order to better inform future research directions. For example, very small seeds within wheat or barley samples cause challenges with image resolution as the contaminant seed is observable within the HSI as only a small number, or less than a single pixel. The trade-off is gaining a high enough resolution image to capture the required number of pixels to identify a very small seed drastically reduces the number of crop seeds which can be recorded in a frame, resulting in greatly reduced through-put for larger crop seed targets. Our preferred solution to this issue is to create a hybrid system of high resolution RGB imaging combined with HSI.

5. Recommendations

The techniques and methods implemented in this project are applicable to benchtop and industrial usage, both cases would follow different pathways due to technical requirements. Most of the investigation carried out in this project is closely related to seed laboratories and border security usage and it is recommended that further work is focused on benchtop solutions for seed laboratories and border security.

For border security it is recommended that a minimum viable product with optimal conditions for imaging, spectral resolution, illumination and imaging processing to scan 150,000 seeds per hour is implemented. Data should be collected over at least a year aligned with the current border and laboratory testing facilities to build a database covering batch variation across seasons and describing several scenarios of contamination risks.

For seed lab usage, an extension to look at physical and genetic purity could be worked on along with other physiological traits. Of particular interest would be if genetic purity could be accurately identified for hybrid cultivars to detect cross contamination.

References

1. Robertson, N. and K. Hurren, *Economic impact 2021* A.F.I. Council, Editor. 2022.
2. Industry, M.o.P., *Seed Field Production Standards*. 2014.
3. ISTA Method Validation for Seed Testing 2006. TCOM-P-01- ISTA Method Validation for Seed Testing Version 2.0
4. Buddenhagen, C.E., et al., *Factors impacting the detection of weed seed contaminants in seed lots.* Pest Management Science, 2023. **79**(2): p. 881-890.
5. Michelon, T.B., E. Serra Negra Vieira, and M. Panobianco, *Spectral imaging and chemometrics applied at phenotyping in seed science studies: A systematic review.* Seed Science Research, 2023. **33**(1): p. 9-22.
6. Dayrell, R.L.C., et al., *Automated extraction of seed morphological traits from images.* Methods in Ecology and Evolution, 2023. **14**(7): p. 1708-1718.
7. Kirillov, A., et al. "Segment anything." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015-4026. 2023.
8. Soni, A., et al., *Hyperspectral imaging and deep learning for quantification of Clostridium sporogenes spores in food products using 1D- convolutional neural networks and random forest model.* Food Research International, 2021. **147**.
9. Al-Sarayreh, M., et al., *Potential of deep learning and snapshot hyperspectral imaging for classification of species in meat.* Food Control, 2020. **117**.