Hyperspectral Imagery Combined with Machine Learning to Differentiate Genetically Modified (GM) and non-GM Canola

This thesis is presented for the degree of

Accelerated Research Masters with Training (Murdoch University)

by

Brianna Rose West (BSc)

April 2022

Murdoch University

Declaration

I declare that this thesis is my own account of my research and contains as its main content, work that has not previously been submitted for a degree at any tertiary education institution.

Brianna Rose West

aRMT candidate

April 2022

Acknowledgement

I would like to thank everyone who assisted and supported me through this challenging and exciting journey. Thank you to my supervisors Dr Manjree Agarwal and Professor Yonglin Ren. Your encouragement, optimism, and approachability are true assets to your character Manjree. You have both helped me develop an appreciation and passion for grain safety, contamination, and biosecurity, I hope to work with you both in the future. I have also very much enjoyed meeting and working with our whole team, I have always felt welcomed and valued by everyone. Thank you particularly to Pushpendra and Rongrong for teaching me how to use the hyperspectral camera.

The biggest thank you to Nick Berryman and my dad for helping me with learning coding and understanding data science. For an area I was completely new to, you both had so much patience in teaching me and explaining everything. I have learnt so much and have found a new passion for coding because of you both – a whole new world has opened up to me!

I would also like to acknowledge the State Agricultural Biotechnology Centre (SABC) for allowing me to use their research facilities, especially Dr David Berryman; Google Colab for the use of their online coding computers; Co-operative Bulk Handling (CBH) for providing me with the GM and non-GM canola grain I used in this project, particularly James Newman and Alannah Martin, and the Australian Government for the Research Training Program scholarship that allowed me to complete this degree. Thank you also to my work team at CBH, working over harvest this and last year is the reason I was even aware of this issue in the first place. Industry experience, awareness, and connections is incredibly invaluable. I look forward to seeing this technology change the way we sample grain in Australia and the world.

Lastly, a special thank you to my friends, family, peers, and colleagues; thank you for listening and your reassurance when I needed it. To the Murdoch Meat Science team, thank you for providing me with the opportunity to develop my skills as a researcher, especially in the last year, and all the support you all gave me this year. Bec, thank you for always making me laugh and Ben for being there for me when I needed you. Mum, as usual, thank you for being my rock in life. I couldn't have done it without you all.

Abstract

Canola, also known as rapeseed (*Brassica napus L.)*, is an oilseed that produces a healthy food-grade oil, canola meal by-product, and biofuel. It is the fourth most grown grain in Australia. Genetically modified (GM) canola currently represents approximately twenty percent of national canola production; hence, with clashing public and industry perceptions of genetically modified organisms (GMOs), transparency and traceability must be enabled throughout the supply chain to protect markets and relationships with consumers. GM canola must not cross-contaminate non-GM canola as our largest export market, Europe, has extremely strict protocols on GMOs. GM and non-GM canola cannot be differentiated by the human eye, with polymerase chain reaction (PCR) methods currently the main alternative, which is expensive and time-consuming. This thesis evaluates the potential to differentiate GM from non-GM canola using the novel, rapid, and non-destructive technique of hyperspectral imaging combined with machine learning.

Hyperspectral imagery captures and processes wavelengths beyond simply red, green, and blue. It has a pre-existing multitude of uses including the characterisation and variety identification of other grains. In this study 500 images each of non-GM and GM canola seeds were captured. Seeds were placed on a black background with two lights sources. Images were captured from the 400nm to 1000nm wavelengths, a total of 80 bands, at a 25-millisecond exposure time. These images were run through a convolutional neural network in Keras for analysis. The high dynamic range and raw files were combined into a NumPy file for the hyperspectral image generator. Contrary to expectations, however, the models using the bitmap image files performed similarly to the models receiving the hyperspectral images. Regardless, both produced high validation accuracies around 90%, indicating a detectable phenotypical difference between the two, and further studies could lead to the development of a new approach to GM canola detection.

List of Figures

List of Tables

Abbreviations

AI	Artificial intelligence
ANN	Artificial neural networks
AOF	Australian Oilseeds Federation
AUC	Area under the curve
BMP	Bitmaps
CBH	Co-operative Bulk Handling
CNN	Convolutional neural networks
DL	Deep learning
ELISA	**E**nzyme-linked immunosorbent assays
EU	European Union
FN	False negatives
FP	False positives
FPR	False positive rate
GDP	Gross domestic product
GM	Genetically modified
GMO	Genetically modified organisms
GPS	Geographical positioning systems
HDR	High dynamic range
HSI	Hyperspectral imagery
MCC	Mathew Correlation Coefficient
ML	Machine learning
NDVI	Normalised difference vegetation index
NIR	Near-infrared
PCA	Principal components analysis
PCR	Polymerase chain reaction
RGB	Red, green, blue
ROC	Receiver operating characteristics
RS	Remote sensing
SABC	State Agricultural Biotechnology Centre

SAP	Simultaneous amplification profiling
SD	Standard deviation
TN	True negatives
TNR	True negative rate
TP	True positives
TPR	True positive rate

Table of Contents

Chapter 1. Literature Review

1.1 Introduction: Canola in the Agricultural Industry

The agricultural sector in Australia is worth $67 billion (2019-2020 financial year), accounts for 11% of the total national exports, and significantly contributes to Australia's gross domestic product (GDP) and workforce (Weragoda & Duver, 2021). As a result of Australia's relatively small population and large production potential, around 70% of agricultural commodities are exported annually – a large proportion of which is made up of grains such as wheat, barley, oats, and canola.

Characterised by its bright yellow flower, canola (*Brassica napus L.)* was developed from a rapeseed cultivar in Canada in the 1970s. Traditional rapeseed contains high amounts of erucic acid (<55%) and glucosinolate compounds (>100μmol/g) which negatively affects colour, palatability, and nutritional value - as opposed to canola which contains less than 2% and 30μmol/g of the two respectively (ACIL-Tasman, 2007). As an oilseed, modern canola produces a healthy and popular food-grade oil and high protein feed source for livestock as a by-product. Canola oil is also particularly popular for use as an environmentally friendly and renewable biofuel due to its comparatively stable properties, high yield potential, and high oil content (Ge *et al.*, 2017; McKeon, 2016).

Canola is also commonly used as a winter rotation break crop for weed and disease management (ACIL-Tasman, 2007). In Australia, roughly three million tonnes are produced annually, 40-50% of which comes from Western Australia. Increased prices and optimum conditions experienced and forecasted recently, however, has seen the area planted and yield per hectare both increase, with the 2021/22 season production expected to exceed five million

tonnes (ABARES, 2021a). Known for its high quality, food safety, and for being in reliable supply, Australian canola contributes 15-20% of the global canola export trade (AEGIC, 2021) and has achieved record pricing, averaging over $A600/tonne in the first half of 2021 (ABARES, 2021b) before reaching highs around $A1000/tonne by the beginning of the harvesting season in 2021 (Brann & Prendergast, 2021).

Although there is a competitive global market for canola, demand for both canola oil and meal is predicted to continue to overtake supply and drive prices up, especially as Canadian and European Union (EU) supplies dwindle after an environmentally challenging year; Australian farmers are in a good position to take advantage of this (ABARES, 2021a). Concurrent to the climatic/seasonal, and macroeconomic conditions driving supply and demand, the uncertainly and unpredictability apparent in the current COVID pandemic will also mean biofuel demand will be largely determined by the journey to recovery (ABARES, 2021a).

The EU is the main market for Australian canola – particularly for use as a biofuel as it aligns with the sustainability criteria for the European Union Renewable Energy Directive (AEGIC, 2021). The EU market, however, has strict protocols on genetically modified organisms (GMOs); the complexity of GMOs and related biotechnology is suggested to be why many consumers reject it – they simply do not understand it (Phillips *et al.*, 2001). Regardless, it is therefore important to protect this market and all of Australia's trading relationships by maintaining and improving biosecurity standards, transparency, and traceability along the agricultural supply chain – and for canola, this includes keeping GM and non-GM canola separate (AEGIC, 2021).

After gaining approval by the Gene Technology Regulator in 2003, multiple varieties of GM canola became available and were adopted in most Australian states after most moratoria lifted in 2008 (ACIL-Tasman, 2007; Smyth *et al.*, 2014). Genetically modified canola now represents around 20% of the total national canola production (AEGIC, 2021). This figure is expected to increase as weed prevalence becomes an increasing issue in cropping systems – as most varieties of GM canola have been modified to be herbicide-tolerant (i.e. from glyphosate), allowing growers to spray their canola crop to kill the weeds but without significantly affecting the plant (*Genetically Modified (GM) Canola in Australia*, 2018). GM canola in Western Australia has historically exhibited slightly lower oil content and slightly higher moisture content (Paull, 2019). GM canola adoption has also been reported to provide greater management flexibility and crop performance in stressed conditions. Conversely, barriers to adoption include cost and market considerations, logistic issues, and product value (Hudson & Richards).

GMOs are characterised by having part of their genetic material, e.g., DNA strands are altered, removed, or added to (from an external source), in some way (ACIL-Tasman, 2007). There continues to be strong clashing public and industry perceptions of GMOs and hence, transparency and traceability must exist throughout the supply chain to protect markets and relationships with consumers. As our largest export market - the EU, has strict protocols on GMOs, GM canola must not cross-contaminate non-GM canola (AEGIC, 2021), and although the industry is currently fairly successful at separating the two, it will become more difficult as GM canola production increases and market regulations tighten.

In order to uphold Australia's status of high standards for food safety and consistency, biosecurity protocols, standards, and technologies must continue to develop and improve. Most of the grain sample checks at collection sites are completed by trained personnel which can be

very laborious, subjective, and often inconsistent (Ravikanth *et al.*, 2015). Fraught with limitations, the grain inspection processes illustrate a clear gap that could be filled with innovative technologies and artificial intelligence. For example, using hyperspectral imagery (HSI) and machine learning has the potential to overcome these issues by making grain sampling processes more automated. This would also make the process recordable – helpful in the event of a trade dispute and illustrates the extreme pre-existing and potential value in traceability and trade market relationship protection in the increasing globalisation of products (Igne, 2009).

Identifiable barriers to this and other relevant innovations include (bio)technological rules and government roles in the canola industry, as well as low market rates of return (of research and development investment); conversely, factors such as increasing available canola varieties encourage grower confidence and can support holistic technological industry growth (Smyth *et al.*, 2014).

1.2 Grain Exportation in Australia

Australia exports canola all over the world, particularly the EU (i.e. Germany, France, and Belgium), China, Japan, the U.A.E, Nepal, and Malaysia (AEGIC, 2021; Kingwell, 2020). It is evident after the recent tariff changes to Barley exports into China that disruptive market changes to agricultural systems can be detrimental. China has specifically banned the importation of barley from Co-operative Bulk Handling (CBH) – the largest exporter of grain in Western Australia, supposedly because of the excessive amount of weed seeds and contaminates (Central, 2020) – this highlights the need for improved biosecurity and contamination identification processes. Like barley, non-GM canola is also exposed by a single market dominance – namely the EU, so, the canola industry would be at risk if similar contamination claims surface (Kingwell, 2020). Conversely, canola prices and demand are also

extremely dependant on other trade country relationships. For example, the Canada-China trade dispute in 2019 saw the suspension of Canadian canola to China, with trading levels still below average (Wells & Slade, 2021). These international trading events open up opportunities for Australian grain markets to fill in.

1.2.1 Grain Sampling and Contaminants

International quality assurance procedures involve multiple steps throughout the supply chain, including collections sites such as the grains exporting company CBH. Australian canola quality standards for growers are published by the Australian Oilseeds Federation (AOF) and Grain Trade Australia (AEGIC, 2021). Some automated technology such as the Infratec Analysis is already performed, which measures protein, moisture, and temperature, however, other checks, such as for contaminates, screenings, and weight is manually captured. Contaminants and general grain assessment include storage pests, insects, soil/earth/rocks/sticks, weed seeds, cross-contamination, fungi, disease, and damaged seeds; all of which have upper count/amount limits and can be classified into levels of contamination based on their severity (CBH, 2020). Having control methods to remove these contaminants is vital to food safety and consumer satisfaction, as well as general seed quality and storage longevity i.e. to prevent infestations, mould growth, or quality degradation (Neethirajan *et al.*, 2007).

1.2.2 Current Methods of GM Canola Detection

Currently, GM canola grains are identified using molecular testing, such as polymerase chain reaction (PCR) detection methods. This process involves collecting and combining multiple samples to form a composite sample, before a small subsample is extracted and tested. Of course, this method lends itself to sampling error; hence, statistical approaches have been utilised to determine the minimum number of samples needed, often using the concept of

binomial distribution to provide confidence levels in the data outputted (Emslie *et al.*, 2007). Previous studies have explored statistical approaches to develop sampling protocols tailored to the different GM level thresholds of different export markets, including the use of enzyme-linked immunosorbent assays (ELISA) which have been validated to provide up to 99% accuracy in some cases (Emslie *et al.*, 2007). Efforts such as this are commendable but still do not seem to take enough consideration into overall time-efficiency and availability/accessibility for grain receival sites outside of the main ports, suppliers, and growers.

Other studies have tested the sensitivity of multiplex PCR procedures such as simultaneous amplification profiling (SAP), which have illustrated to provide reliable GM canola identification using a small amount of DNA (James *et al.*, 2003). This technology has been described as commercially applicable with detection limits around 0.013-0.025% (Kim *et al.*, 2015). However, methods available and possible at all receival sites that can be combined with other detection technologies have not been identified or explored to their full potential as of yet.

Regardless of the testing method used, it is likely that in the short term, segregation of GM and non-GM canola will become more significant as markets become less tolerant and as price penalties are expected to increase – historically around seven percent (Paull, 2019). Non-GM canola receival standards currently allow 0.9% GM canola cross-contamination after international markets recognised the difficulty in separating and differentiating the two (Emslie *et al.*, 2007). This threshold may decrease in response to market and consumer pressure, however (Paull, 2019).

1.2.3 Current Methods of Insect detection

Similar to cross-contamination, insect detection, specifically stored grain insects, is one of the most concerning biosecurity risks for the grains industry. Pests such as the khapra beetle, *Trogoderma granarium*, could compromise the whole grains industry if introduced and are unfortunately difficult to detect and identify (Agarwal *et al.*, 2020). Current detection techniques include grain probes and insect traps, pheromone level readers, visual lures, acoustical amplification, electrical conductance, funnel method, near-infrared (NIR) spectroscopy, machine vision, and x-ray. All of which have advantages and disadvantages - often relating to speed, effectivity, and detection limits relating to life cycle stages (Neethirajan *et al*., 2007). Previous research has indicated that hyperspectral imagery, coupled with deep or machine learning and similar technologies, has the potential to be a viable method for insect analysis (Agarwal *et al*., 2020), offering a promising alternative to taxonomical diagnosis. Although insect, specifically stored grain insect, detection and identification is extremely important, insects, whether field or stored grain, are also routinely fumigated at grain collection sites, reducing the risk of insect contamination and export concerns specifically. However, this same technology could be applied and tested against contaminations that cannot be neutralised via fumigation or simple methods such as sieving, such as grain cross-contamination – i.e., wheat in barley or GM in non-GM canola. Although this has not been tested specifically before, similar research has outlined its potential for identifying other factors in other grains, supporting the idea that hyperspectral imagery could be used to differentiate GM canola from non-GM canola. This would be even more valuable if the eventual development of the diagnostic tool were able to detect multiple contaminations such as grain and field insects, foreign material, and cross-contamination.

1.3 Hyperspectral Imagery

HSI and related technology involves capturing and processing images to predict or reveal information (Feng & Sun, 2012). Similar to a standard camera, a hyperspectral camera takes images that are composed of an array of pixels created by scattered or absorbed light at different wavelengths (Dale *et al.*, 2013). Standard images are often referred to as arrays of I rows and J columns with the intersections or coordinates having a grey value or red, green, blue (RGB) triplet colour pixel value. Colour images hence have three bands which provide different information; multivariate images split the information up into more wavelength bands which then becomes a 3D array (IxJxK) and often will include different modes of information; whereas hyperspectral images are characterised by having many bands – in the tens or hundreds and can express pixels as a spectrum (Geladi *et al.*, 2007). This three-dimensional data is often referred to as a 'hypercube' (Caporaso *et al.*, 2018). These numerous and narrow bands of information expressed as a spectrum can then be illustrated as a spectral signature which builds the foundations of object identification or categorisation and eventually data classification (Paoletti *et al.*, 2019). Hyperspectral images are analysed in a variety of ways including the use of a principal components analysis (PCA) framework. PCA decomposes these matrices, also known as dimension reduction, to increase interpretability and visualisation (Gallagher & Lawrence, 2020). The application of HSI includes the subcategories of visible, fluorescence, thermal, and spectral imaging (Moghimi, 2019). Most uses for HSI in agriculture uses wavelengths above 1000nm, but depending on the use, this can range, with some prediction models requiring very specific wavelengths (Erkinbaev *et al.*, 2017).

1.3.1 Near-infrared Hyperspectral Imagery

Near-infrared hyperspectral imagery covers the visible lights spectrum, but also wavelengths beyond that of what the human eye can see on the electromagnetic scale, including short wave NIR, near, middle, and far-infrared. NIR HSI can be used to determine chemical

and physical compositions (and their spatial distribution) – suggesting significant potential in grain inspection and contamination detection, from cross-contamination and insect or weed seed detection to variety and quality identification (Caporaso *et al.*, 2018). As opposed to molecular testing, near-infrared hyperspectral imagery is also favoured for its ability to be easily replicated for research and its object identification/detection capabilities in industry practice without needing to destroy or alter the subject (Caporaso *et al.*, 2018). Non-destructive analytic assessments are preferable in this, and many other industries, but especially in agriculture where the nature of the supply chain dictates multiple testing stages are necessary (Feng *et al.*, 2019). Comparative to other methods, it is also a lower cost and more rapid analytical method, making it ideal in industry and research. Additionally, HSI systems currently used in industries are already renowned for their low power use, heat emission, and size, as well as their speed of processing and reliability (Manley, 2014).

1.3.2 Data Collection

Hyperspectral imagery also addresses one of the greatest bottlenecks restricting the efficiency of grain production and processing - data collection and processing. This applies to every aspect of the supply chain, from plant breeding and on-farm collection of phenotypic data such as crop health, disease susceptibility, or biomass production, to sampling the grain at collection sites and ports (Walter *et al.*, 2017). The current laborious, inefficient, and inconsistent methods of sampling and analysis highlight the value of technologies, like HSI, in significantly increasing the speed and accuracy of processes along the supply chain - driving research behind, and the development of, these technologies (Dantes, 2020). This area of technology is important as it can protect trade markets and international relationships by reducing the risk of contaminations, and therefore disputes and possible market losses. \

1.3.3 Applications in the Agricultural Industry

Hyperspectral imagery allows for the identification of a material's spectral signature – made up of its physical and chemical properties (Ravikanth *et al*., 2015). It has been a significant focus in many studies as its versatility means it can be used across many areas, industries, and disciplines including a large range of applications in the whole agricultural supply chain and for food safety and security (Feng & Sun, 2012). As technologies rapidly change and develop, HSI technology continues to evolve and move very quickly. Current applications cover everything from plant breeding to the pre-farm gate, and then the post-farmgate.

1.3.3.1 Remote Sensing for Plants

Remote sensing (RS) was one of the first uses of hyperspectral imagery in and outside of agriculture. As the name implies, RS involves collecting spatial and temporal information, generally relating to crop productivity/health status, from a distance, without physically touching the object/s in question or having to conduct physical or chemical tests, whether this is from close range (like a hand-held device) or from far (satellite, drone, or plane) (Weiss *et al*., 2020). Being able to use multi- or hyperspectral imagery on a grand scale and then using those images to measure energy reflectance means RS is an ideal tool for researchers and farmers, because of its speed and accuracy in conveying information (Aggarwal, 2004).

RS is commonly used to monitor plant health in both urban and rural environments, often combined with geographical positioning systems (GPS). In particular, RS is commonly used to map normalised difference vegetation index (NDVI), the measure of greenness in cities to support council planning and in cropping enterprises (Khanal *et al*., 2020), which can help to identify higher and lower areas of productivity and estimate yield (Weiss *et al*., 2020). Doing so also supports the growing movement of precision agriculture - which entails using spatial

and temporal information to make more acute decisions about management practices such as fertiliser or irrigation input levels, time of sowing, spraying, or harvesting (Weiss *et al*., 2020).

The biggest barriers to the adoption of RS technologies generally concerns cost, complexity, and the lack of tools available to analyse the data collected (Weiss *et al*., 2020). Although acquiring and processing the large amount of data associated with HSI is still a challenge in this area, multiple products and developments have been introduced in the industry, mostly only picked up by early adopters and larger corporations however (Weiss *et al*., 2020). There has also been a plethora of studies using unmanned aerial vehicles to capture HSI, such as drones, or using close range HSI technologies for estimating biomass and soil properties, identifying invasive plant species, detecting disease, and phenotyping (Lu *et al*., 2020). Whilst this has been particularly exciting for researchers and farmers alike, on a more holistic level, this technology has the capability to completely change how the agricultural industry tackles the increasing threat of food insecurity and climate change by supporting decision making with the wealth of information it makes accessible (Weiss *et al*., 2020).

1.3.3.2 Food Content and Compositions

Another particularly valuable area of HSI use, presently and in future, is food content and composition examination. This is inclusive of pre-harvest (crops) to post-harvest (grain) to post-processing and production (food items). For example, there has also been research conducted to determine if hyperspectral imagery can support the prediction of nitrogen, water, and salt levels in wheat – three of the most significant growth determinants or potential growth constraints. A study by Bruning, *et al.* (2020) saw significant potential in using this technology, especially as it is non-destructive and efficient. Their data set and range of treatments limited the strength of their prediction model, however, and so further research should aim to have a larger range of treatments.

Similarly, post-harvest focused studies have explored the use of HSI in quality assessment. A review by Caporaso *et al.* (2018) highlights the scientific progress made in recent years using HSI, especially NIR, for the quality and character assessment of cereal grains. For example, studies have revealed the potential for use in identifying wheat contaminates (Ravikanth *et al*., 2015), grain quality (Caporaso *et al*., 2018), insect damage (Singh *et al*., 2010), and the ability to predict protein levels or sprouting in cereal grains – this, in particular, would be of utmost value along multiple stages of the supply chain, particularly as the current quality check at grain receival sites for sprouting includes a falling numbers test which requires physically making dough and is considered laboursome and inefficient (Shafqat, 2013). Multiple studies in this area have been conducted with mixed results, often contributed to misclassified kernels, but illustrates significant potential especially when models included PCA and short-wave infrared HSI.

Similar quality assessment studies have also been conducted in the horticultural sector, including colour, ripeness, bruising identification, and component distribution in fruit and vegetables such as tomatoes, apples, kiwifruits and peaches; as well as in safety and contamination including identifying surface defects, bacteria, faecal matter, and soil matter on apples (Baeten *et al*., 2007).

Other studies have focused on the application of HSI to analyse and identify food contaminants, in both grains and meat, and predict food compositions – specifically for any substances or compounds that risk the food safety of the product. Similarly, many of these studies saw potential in this technology but further work was needed to increase the accuracy. The reliability of the results can also be questioned in some of the research where HSI has only analysed small areas of the food product when contaminations/compositions are not always uniform (Dantes, 2020).

1.3.3.3 Inspection and Safety

Beyond food composition identification, near-infrared hyperspectral imagery is also used in inspection and safety - the sheer versatility of HSI highlights its value across the food supply chain. For example, NIR HSI technology can be used to detect food-borne pathogens and physical, chemical, or biological hazards. Studies have previously been conducted to identify a range of these factors in food using HSI as it is less limited than other technologies and has illustrated immense promise in its potential application (Feng & Sun, 2012).

Although small data sets continue to be problematic for research in this area, further research would be very valuable if applied to a larger variety of areas. For instance, there are numerous studies on utilising HSI to classify different varieties of grains such as wheat and barley (Bao *et al*., 2019). GM and non-GM maize have also been successfully and accurately differentiated through the use of HSI combined with chemometric data analysis (Feng *et al*., 2017). If the same technology could be used to identify cross-contamination - different grains in a sample (such as wheat in barley, or mixed varieties, or GM canola in non-GM canola) – quickly and accurately, it could replace the current time-consuming processes for personnel and completely evolve the biosecurity aspects of grain supply chains.

Cross-contamination is a very significant issue, for many reasons. In terms of trading, it can dictate if loads are acceptable to export – such as GMO detection for European bound shipments, or if the quality is within the acceptable ranges to be sent to a certain market; but cross-contamination is also very important for the end-consumers themselves. For example, cross-contamination can be life-threatening when considering gluten and non-gluten containing grains. A recent study was conducted to test if NIR HSI could be used to differentiate oats from other grains for quality control and assurance for non-gluten products (Erkinbaev *et al*., 2017). This study outlines the potential of reducing analysation time and

increasing accuracy; however, further research would need to be conducted to improve the model created as predictions were only based on limited wavelengths.

Access to technology that can quickly and accurately identify cross-contamination, particularly non-destructive methods like HSI, is particularly significant as the risk factors leading to cross-contamination events are likely to become more common in future. For example, cross-contamination is at a higher risk in circumstances when: growers include various cropping rotations in their farm management practices and/or the farm is mixed practice; when farmers store grain in response to the threat of natural disasters, such as droughts; when grain prices do not reflect a market they want to buy or sell in; and as a result of herbicide and pesticide resistance at an on-farm level (Malcolm *et al.*, 2009); all of which are becoming increasingly common situations.

1.3.3.4 Red Meat Quality Prediction

NIR HSI has also been used to develop a non-invasive quality analysis for beef and lamb. As eating quality is affected by a range of factors such as visual evaluations, i.e. meat and fat colour, and taste factors, such as tenderness, juiciness, and flavour, being able to predict these factors – without having to consume the meat – is vital to quality assurance and pricing schemes (Qiao *et al.*, 2015). Similar studies have utilised this technology to predict moisture content in beef, lamb, and pork and illustrate the significant potential for industry use – as similar to grains, a quick but accurate and non-destructive method will always be preferred (Kamruzzaman *et al.*, 2016). Dixit, *et al* (2021) successfully used deep convolution neural networks to predict intramuscular fat and pH (two of the main eating quality indicators of meat) in beef, lamb, and venison using HSI. Similarly, Hoonsoo, *et al.*(2018) were able to use HSI to determine total volatile basic nitrogen content in pork to a high degree of accuracy, indicating its potential to replace the traditional and chemical methods of measurement.

Imaging systems have even been tested for use in meat safety. There have been several studies demonstrating how HSI technologies could detect contaminations, tumours, defects, and diseases in meat, particularly in relation to poultry (Baeten *et al*., 2007); as well as detecting bone fragments and other foreign materials (Lim *et al*., 2020). A study by Zheng, *et al.* (2019) also reported the determination of coefficients of 0.98 when using HSI technology to detect duck meat in minced lamb. The multitude of studies illustrating high accuracies and rapid analysis of inputs emphasises the large scope and diversity that HSI technologies could have in the agricultural industry.

1.4 Machine Learning

Machine learning (ML) represents the step between data collection and task performance. Where HSI can be used to gather large amounts of important data, ML is the necessary tool to utilise this data and create technologies that can complete a task. ML and HSI, together, have already been successfully utilised in a number of areas, including agriculture for the development of technologies associated with remote sensing, precision agriculture, and mapping plant stress and soil erosion (Paoletti *et al*., 2019). ML is valuable as it is less instructional coding orientated, and more teaching inspired, it is the backbone of areas in automation, machinery, and artificial intelligence (AI), combining computer science and statistics (Patel, 2021). In ML, computer programming is used to sort, process, and learn data. ML is the first step of machines/computers imitating human decisions or skills and therefore has incredible potential in areas like agriculture that are so heavily reliant on labour, which can be especially physically demanding and time-consuming (Liakos *et al*., 2018). ML can be classified under either supervised or unsupervised learning which ultimately concerns if input data is labelled or not (Patel, 2021). A supervised ML environment will include a ‘supervisor’ inputting training data which is used to train the model to create an algorithm that can then be

used to make predictions which direct how the model will retrain itself on that data set and new data. Supervised learning is task-driven, grouping data sets using various statistical means and patterns by classification or regression to make predictions. Unsupervised learning is equally self-explanatory and is used for more complex tasks and when unlabelled data is involved. This involves clustering, looking for similarities in subsamples of data, or association, identifying unknown patterns (Patel, 2021).

1.4.1 Deep Learning

Deep learning (DL) is a subfield of machine learning, only it aims to take a step further in accuracy and precision, as well as the possible capabilities. Deep learning requires intense machine learning and training with large datasets, but once developed, is flexible and more adaptable to other similar technologies (Campesato, 2020). Where machine learning is fed input data and provided an algorithm or feature classification by the user to then determine outputs, deep learning will does not rely on user input for feature classification and instead will determine this itself (LeCun, 2015). Algorithms become increasingly complex as the task becomes more complex, but stem from the core concept of linear regression and finding correlations between data sets; regressions are further used to create a boundary between data points and enable decision making (Campesato, 2020).

DL can compose of multiple processing layers which allow for generalisations, data representations, and ultimately the artificial intelligence's ability to perform abstract tasks. It begins by entering training data images, into the system in conjunction with an algorithm, together forming a complex matrix or neural type network made up of many layers. Each layer is transformed or filtered and passed on to another layer for feature detection before classifications can be made. From here new data or images are presented to predict related outcomes. This process is encompassed under the umbrella term of artificial neural networks

(ANN) – inspired by the learning process of the human brain and its neurons (Campesato, 2020) having an input layer, an output layer, and hidden layers in between making connections like biological neurons or a bipartite graph (Campesato, 2020). Machine and deep learning support a vast array of technologies in our modern society (LeCun *et al.*, 2015).

1.4.2 Convolution Neural Networks

Convolutional neural networks (CNN) are one type of ANN that integrates spectral and spatial features particularly efficiently. Like other ANN, initial stages detect recognisable features and later stages combine to detect abstract features with the kernels supporting this extraction. CNNs filter down the image using kernels to reduce the number of parameters, therefore considering only the more significant parts of the image (determined by pixel placement and weight) and has only partially connected layers. Hence, CNNs require less computation training time than ANNs and also recognise spatial significance, using 'back propagation' to look back on the weights placed upon the neurons and recognise patterns. CNNs can be 1, 2, or 3D which defines how many directions the kernels move with the resulting output one dimension above that (i.e., a 2D CNN will move in 2 directions but has 3D in and outputs). It is important to note, however, that the lower dimensional CNN can degrade object shapes, avoidable by using a 3D CNN. Granted, CNN algorithms are one of the slower ML classifiers due to the computational complexity of the convolutional layers and needing more parameters, which is only exaggerated by using a 3D CNN (Paoletti *et al.*, 2019). CNN are the data processing components of artificial intelligence and learning whereby multilayered neural networks are used as pattern recognition – like 'connecting the dots' (LeCun *et al.*, 2015).

Many recent papers have combined HSI with deep learning to marry the large amounts of data HSI creates with the powerful data analytical power of DL (Paoletti *et al.*, 2019).

Specifically, deep neural networks (two or more hidden layers) are considered particularly beneficial for the flexibility in their model structure and the type of data entered into the system, as well as their ability to extract information from raw data and adapt to different learning strategies from fully supervised to unsupervised. Conversely, the limitations identified include their coding complexity and the associated training difficulties, as well as being computationally expensive and often requiring extremely high processing power and memory. Neural networks also tend to overfit when only having a few training parameters combined with frequent degradation of data between the convolution layers. Overfitting is the term given where training accuracy may be high, increasing over epochs, but validation or test accuracy is low or decreasing over the number of epochs. This happens when the model is memorising the training data, by fixating on incorrect or insignificant details and therefore not training itself, leading to poor performance with new data (Campesato, 2020).

Lastly, although they have illustrated to be a powerful AI technique, the hidden layers in DNNs create the problem of having a 'black box' nature. This is because this technique takes the inputted data and itself creates an algorithm to classify the data, the user or designer does not know how the model has analysed or combined the variables. Hence, there have been some ambitious efforts made to visualise the parameters of DNNs, as not knowing what variables or parameters the model is focusing on means making interpretations is difficult (Paoletti *et al*., 2019) as users are not necessarily aware of what features were used to differentiate image categories. Regardless, deep learning is an essential component of recent agricultural movements like 'smart farming' (Kamilaris & Prenafeta-Boldú, 2018) and will continue to provide substantial scientific and technological advances in future.

1.5 Conclusion

Agriculture is an economically, socially, and civilly significant industry that is multilayered, complex, and ever evolving. With a large exporting market of grains, to countries all over the world, Australia needs to maintain high standards and regulations regarding quality assurance (Weragoda & Duver, 2021). Hyperspectral imagery and the use of correlated technologies are becoming increasingly popular because of their superior accuracy, process speed, and reduced labour inputs (Dantes, 2020). Currently, there is a lot of research and development in this area, some of which is centred around grain exports as markets become increasingly more competitive and stricter with quality regulations (Central, 2020). Contamination of grain is a particularly concerning area. Currently, research has been conducted in multiple aspects of grain contamination, such as insect identification, foreign materials, and food safety and composition. While much of this research illustrates potential, there is a pattern of lower than viable accuracies, as a result of limiting factors including sample sizes, and a lack of focus on the potential time and resources saved from the use of this technology and the accuracies in relation to current best practices. This highlights the possibility for new or continued research in areas such as cross-contamination of similar and common grains, like GM and non-GM canola. Research regarding cross-contamination of grains and identifying grain varieties, in particular, suggests research in other areas, like the detection of genetically modified canola, has promising potential. Combined with the power of artificial intelligence tools such as deep learning and neural networks which can discriminate images based on features that humans may or may not be able to see, hyperspectral imagery has the potential to radicalise the food supply chain and support food safety and food security all over the world.

Research Gap

Currently, a large proportion of the grain biosecurity and quality assurance protocols at grain receival sites are manually conducted by trained professionals. The sampling process is often laborious and time-consuming and can be subject to human error and bias, as it involves physically collecting a representative sample of grain from the truck, measuring hectolitre and admixture material weights, running a chemical analysis, and visually checking for contaminants. For example, to the naked eye, GM and non-GM canola cannot be differentiated, and as a result, the risk of cross-contamination occurring between the two is high. This highlights the need for change and technological innovation to fill this gap.

Furthermore, the relative success of previous research experiments offers confidence in a project using HSI to differentiate between GM and non-GM canola – which to the researchers' knowledge has yet to be done, as most of the research and tools used to detect GM canola concerns molecular testing such as real-time polymerase chain reaction methods (Akiyama *et al*., 2010).

Research Aim

This project aims to assist in the eventual development of a diagnostic tool that can be used in the industry to identify different grain contaminants. This study will specifically investigate if near-infrared hyperspectral imagery, coupled with machine learning, can differentiate between genetically modified and non-GM canola. The main objective will be to develop an algorithm using binary coding from the images taken using the hyperspectral camera, before testing our hypothesis.

Research Hypothesis

GM canola exhibits different traits and phenotypic changes as a result of a change in the DNA structure and chemical composition. Unlike a normal camera or human eye, near-infrared hyperspectral imagery can detect these chemical constituents' vibrational changes. Hence, we hypothesise near-infrared hyperspectral imagery combined with machine learning will be able to differentiate GM and non-GM canola to a high degree of accuracy and precision.

Chapter 2. Hyperspectral Imagery Combined with Machine Learning to Differentiate Genetically Modified (GM) and non-GM Canola

2.1 Materials and Research Methodology

2.1.1 Canola Procurement and Storage

As a novel study of hyperspectral imagery being utilised as a diagnostic tool to differentiate GM and non-GM canola, a composite sample (mixed and blended from all places and varieties) of canola was used – only separated by its genetic modification status. The GM and non-GM canola were donated by Co-operative Bulk Handling (CBH) from composite stacks of canola acquired from across the state in the 2020/2021 harvest season. This ensures that the grains utilised in this experiment are from a variety of farms in differing climatic zones with a variety of soil types. There are many different varieties within these samples, with the main ones including the Hyola varieties, Bonito and Bonanza for the non-GM canola sample; and the GM Hyola varieties, Cobra, and Viper varieties for the GM canola sample. The non-GM canola sample came from CBH's CAN1 stack and the GM canola sample came from their CAG1 stack. Both GM and non-GM canola is separated into CAG1/CAG2 and CAN1/CAN2 respectively. The numerals 1 and 2 denote their quality grade based on chemical composition, seed quality, and foreign materials checks such as insects, admixture, or foreign seeds performed when sampled at a grading site (CBH, 2020). The Infratec results for protein and moisture were 20.8% and 6.2% respectively for the non-GM canola sample and 21.3% and 6% respectively for the GM canola sample. This illustrates no significant difference between the two samples in regard to protein and moisture that could be affecting the results. The canola samples were kept separately in labelled jars in a cool, dark area throughout the experiment.

2.1.2 Hyperspectral Imaging System and Image Acquisition

The visible near-infrared hyperspectral imaging system used to conduct this experiment was located at the State Agricultural Biotechnology Centre (SABC) within the Murdoch University, Perth, Western Australia, campus. Attached to a microscope to provide an additional ten times magnification, the spectral imager (Applied Science Imaging Model: CCD-1300DS, Germany) was connected to the SpectraCube Spectral imaging acquisition software. To view images, the spectraview software was used. Two light sources were used to illuminate the stage and subject matter as well as support extra wavelengths – a 100W intense basking spot/heat lamp and a halogen desk lamp. The desk lamp (globe component) was situated 18cm vertically from the bench and approximately 20cm from the stage horizontally. The basking spot lamp (globe) was 12cm vertically from the bench and approximately 10cm from the stage horizontally (Figure 2.1). 80 bands between the wavelength of 400 – 1000nm were captured using this system.

Camera and image optimisation was determined after experimenting with the software settings and environmental conditions. A black slide background colour was chosen as it mostly eliminated the problem of background reflectivity with each individual canola grain being placed on the stage using plastic forceps one at a time. Photographed grain was placed in a separate jar in order to ensure there were no repeat images of the same grain. Grain yet to be photographed was kept out of direct light, to reduce heat absorption, particularly when the basking light was on.

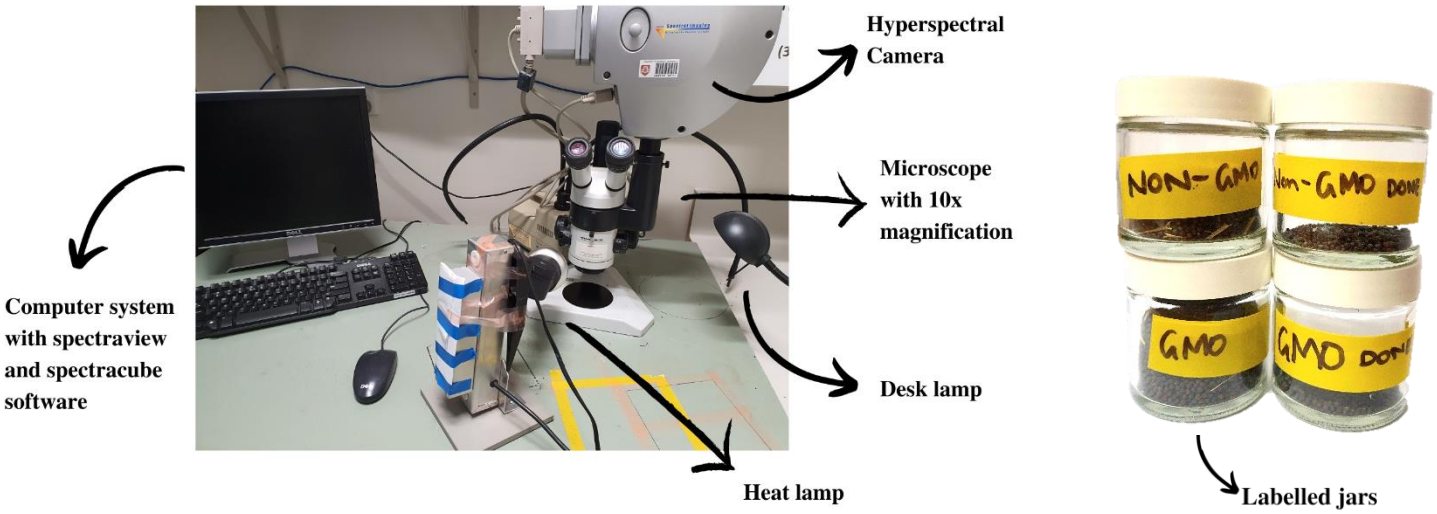

Figure 2.8: Hyperspectral Imaging System Setup for Image Acquisition

The image capture parameters are as follows: 256 frames, 45 steps, and with an exposure time of 25ms^{-1}. These conditions were used consistently for all captured images. Using the SpectraCube imaging acquisition system, images were saved under three file types: bitmaps (BMP), high dynamic range (HDR), and raw and stored on a hard drive to be analysed on the image processing software Scyven. The BMP images were 128 x 128 pixels and the hyperspectral images captured were 496 x 800 x 80 pixels (height by width by 80 wavelengths, representing a 3D image or 'hypercube' (Caporaso *et al.*, 2018)). A total of 1015 images were entered into the CNN (505 non-GM and 510 GM) after eliminating any incorrectly photographed/saved images from those captured. Images were generally captured in sets of 100 (50 of each grain type) on any given day.

2.1.3 Machine Learning

2.1.3.1 Dataset description

The full data set of roughly 1000 images, 500 of each grain type with files separated based on date and GM status, were divided randomly into training (70%), validation (20%), and test (10%) data subsets using a python script – ensuring that the GMO and non-GMO files were kept separate. The training data is used to teach the network the two classes before the network will then attempt to correctly identify the validation data, making alterations to the network in order to reduce the losses and incorrect classifications. This cyclical process is known as an epoch. Once the best model is identified, with the greatest validation data accuracy and lowest losses, it is used to run the test data set to determine the model's final accuracy. This was done this way as a study by Hänsch *et al.* (2017) highlighted concern surrounding the randomisation method to separate data by pixels into the train, validate, and test data sets in the model as it may compromise the model's reliability if the network has previously processed all the images and data sets are overlapping or very similar. This can be avoided by pre-processing the data set into train, validate, and test, and directing them as independent sources for the model to use, as done in this study (Hänsch *et al.*, 2017). Furthermore, Bitmap files were kept separately and used as the RGB reference. The HDR and RAW file types were combined and resized into a NumPy, which is a python library package for matrices and arrays, 224 x 224 x 80 array to be used as the hyperspectral images for the model.

2.1.3.2 Image Analysis

Analysis of bitmap and hyperspectral images were completed using Python code as well as in the Scyven software program, which analyse HSI under a principal component analysis. Analysing the images in Scyven allowed the light/wavelength intensities or colour values to be mapped using the pixel irradiance or 'illuminate' function which then graphs the images', or a single pixel's, spectral signature. This information was used to remove the

background, by determining the wavelength with the greatest intensity and removing pixels that had a low intensity of that wavelength and isolate the canola grain in Python so further image traits could be determined. Maximum vertical and horizontal diameters of the grain were determined to roughly estimate the centre of the grain so a 20 x 20-pixel square could be used for a basic investigation into the spectral signatures of the grains. This 400-pixel square provided numerical intensities at each wavelength for every image. From this data, a mean spectral signature for the GM canola images and the non-GM canola images was determined. These diameters were also used to calculate the grain size (area and diameter), with results mapped as a distribution curve (Figures 2.5 to 2.7).

2.1.3.3 Convolution Neural Network

A convolution neural network-based approach was used in this study to determine if machine learning can accurately differentiate GM and non-GM canola. CNNs are currently a popular tool in machine learning for areas such as pattern recognition and image classification as their deep hidden layers, such as convolution, pooling, and connected, prove highly successful (Albawi *et al*., 2017). The biggest advance of this network is its ability to automatically extract features from raw, pre-processed images as it trains itself to classify said images (Niu & Suen, 2012). Furthermore, a spatial and spectral classifier provides less 'noise' so accounting for the spectral data in relation to its spatial context creates more accurate feature maps (Paoletti *et al*., 2019) and hence should be beneficial for this project. The coding software, Python3, was run on Google Colab, an online platform to write and execute python coding whilst accessing the powerful google computers, using the various opensource machine learning libraries NumPy, TensorFlow, and Keras, to develop this binary classifier model. Two different scripts were written to create the CNN models for the hyperspectral and bitmap images. The skeleton code is based on Francois Chollet's image classifier script that was altered on the GeeksforGeeks.org website. Chollet is a French software engineer who created the

machine learning library Keras. Both the HSI and bitmap model included 2D convolution layer functions (3 x 3 weighted matrices/kernels move across the image) and rectified linear (relu) activation functions with max-pooling (subregions are mapped and the largest existing value is outputted) to create a total of three layers with the final fully connected dense layer (all neurons from the previous layer are connected to the next) having a dropout rate (randomly chosen neurons) of 0.5 to avoid overfitting (Figure 2.2). In the bitmap model, images were rescaled to 1/255 with a zoom factor of 0.2 and randomly horizontally flipped in the image data generator as a way of augmenting the original data set. Augmentation is used to prevent overfitting, where the network memorises the training data set and can classify those images well but cannot accurately classify new data, by artificially increasing the data set by adding altered versions of the original images. These pre-processing tools, as well as other alterations between designs, improved the performance of the CNN model. Augmentation was not used in the final HSI model, although a hyperspectral image generator code based on an article by Nilesh (2018) was sourced and written, because of a lack of computational power. Chollet's code was designed for jpeg files and so whilst image generator scripts for standard images are common, HSI generators/augmentation is not as readily available. A total of 100 epochs (number of times the dataset is run through the neural network) were run on both models with batch size (how many images were run at once) up to 100 depending on the available ram.

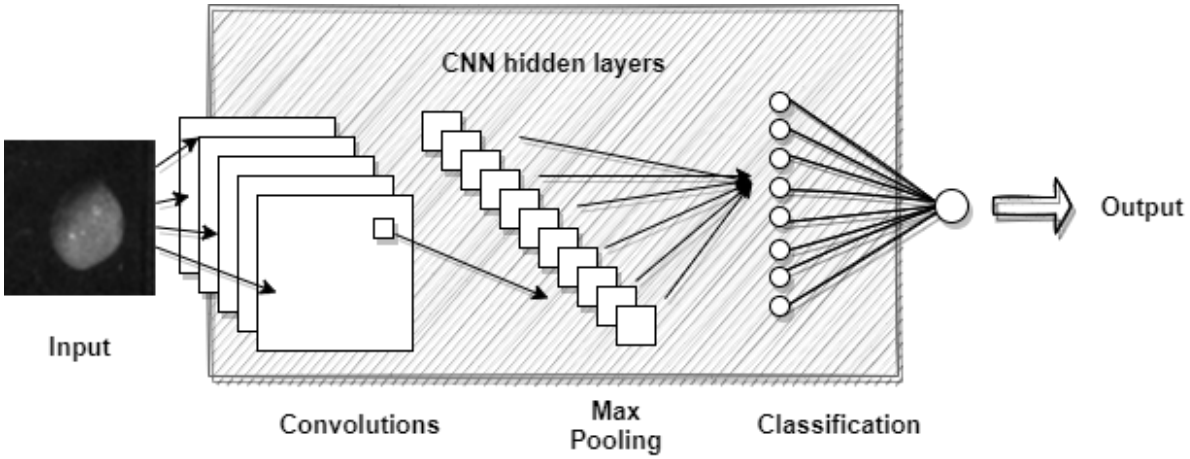

Figure 2.8. Simplified diagrammatic representation of a convolution neural network.

2.1.4 Performance Measurement

2.1.4.1 *Confusion Matrix*

Performance measurement is vital in defining the effectiveness of a program. Confusion matrixes are a common evaluation tool used in machine learning (An, 2020). Generally, they consist of a n x n table plotting actual class against predicted class (n denoting the number of classes, so a binary classifier would utilise a 2 x 2 table), to which the true and false (determined by the actual classes) positives and negatives (determined by the predicted classes) fit within (Visa *et al.*, 2011) (Figure 2.3). True positives (TP) and true negatives (TN) signify the model correctly identified the class, or in this case the genetic modification status of the canola grain in the image. False positives (FP) and false negatives (FN) signify that the model incorrectly identified the class (Zhu *et al.*, 2010). The data within can then be used to mathematically calculate various performance measurements such as precision, the proportion of identified positive cases (GMOs) that were correct; recall/sensitivity, the proportion of actual positive cases (GMOs) that were correctly identified; and specificity, the proportion of actual negative cases (non-GMOs) that were correctly identified (An, 2020) (Figure 2.3). High sensitivity equates to a high true positive rate which is conservative in predicting negatives (non-GM canola). High specificity is the opposite and will be conservative in predicting a positive (GM canola) class (Zhu *et al.*, 2010).

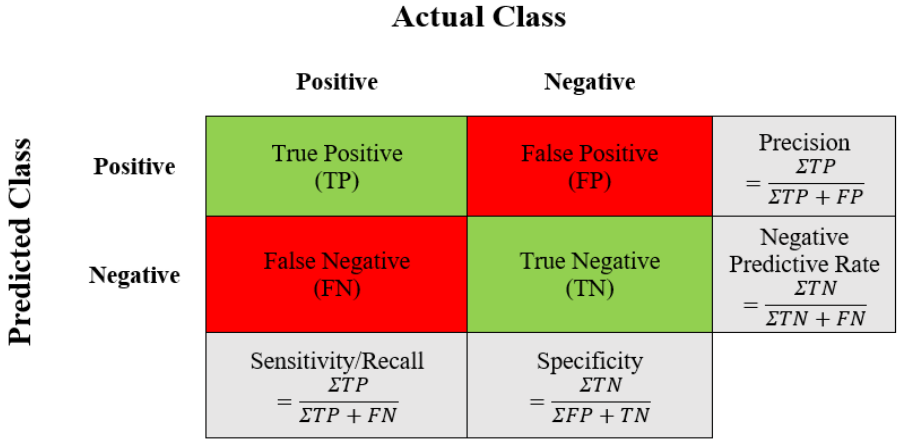

Figure 2.9. Breakdown of a confusion matrix including the formulas used to determine their respective performance measure.

2.1.4.2 *Measures of Accuracy*

Further evaluations can be made using the data from the confusion matrix and these calculations themselves. Accuracy and F1, or F-score, are two common performance indicators, especially in machine learning.

- Accuracy $= \frac{\Sigma TP+TN\ (correct\ identifications)}{\Sigma TP+TN+\ FP\ +FN\ (Total\ assessments)}$
- F1 score $= \frac{2 * TP}{2 * TP + FP + FN}$ = A balancing metric to find a harmonic mean between precision and recall between 0 and 1.

Accuracy (total correct divided by the total number of assessments), however, does not consider the significance of misidentified classes (Halimu *et al.*, 2019) and tends to be an overly optimistic performance indicator. Likewise, F1 scores do not consider the amount of correctly identified false, or in this case non-GMO, classes (Chicco & Jurman, 2020), only the precision and recall proportions which are plotted against each confidence threshold (cut-off for a decision to be made; between 0 and 1). However, both measures do not consider differences in the number of samples in the different classes and hence cannot account for sample size imbalances.

2.1.4.2.1 Mathew Correlation Coefficient

$$\frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

The Mathew correlation coefficient (MCC) is used to overcome sample size imbalances (Chicco & Jurman, 2020) in that it is a measure of strength between the actual and predicted classes with 1 denoting perfect strength/prediction, 0 indicating there is no class separation/it is random, and -1 meaning an inverse prediction is evident (Halimu *et al.*, 2019). The MCC accounts for the number of positively versus the number of negatively classed inputs in the event they are unbalanced, in this case, GM versus non-GM canola images, and will only

produce a high result (close to 1) if the model produced good scores in all categories of the confusion matrix (high TP and TN and low FP and FN). It has been suggested as the most reliable indicator of performance for binary classification (Chicco & Jurman, 2020).

2.1.4.2.2 *Area Under the Receiver Operating Characteristic*

Lastly, the area under the curve (AUC) for the receiver operating characteristics (ROC) curve illustrates the true and false positive rates for every confidence threshold between 0 and 1 of binary classifiers (An, 2020). It maps sensitivity, also known as the true positive rate (TPR), (y-axis), against 1-specificity, also known as the true negative rate (TNR), (x-axis). 1-specificity/TNR could also be defined as the false positive rate (FPR) calculated by the number of false positives divided by the sum of the false positives and true negatives (Figure 2.4).

If a model correctly classifies all the positives, then the AUC would be 1 and the ROC would run up the y-axis and go through the 'ideal coordinate' of (0,1) (Figure 2.4). If a model can only randomly guess between binary classes, then a linear diagonal line between the coordinates (0,0) and (1,1), would exist with an AUC of roughly 50%. AUCs are good indications of performance and are commonly used in machine learning (Halimu *et al.*, 2019). Studies have suggested that AUCs are much better than a simple accuracy formula and its consistent and discriminatory nature lends it to be a better measure than MCC when comparing model algorithms in a more holistic sense (Halimu *et al.*, 2019). Strictly speaking, however, AUC and MCC results cannot be compared directly as AUC is measured on a spectrum from 0 to 1, with the classification thresholds outlined in Table 2.1 and MCC results are measured on a spectrum of -1 to 1.

These various performance evaluations can yield significantly different estimations and are only a few examples of the many alternative performance metrics used in biostatistics and machine learning. Hence, it is important to consider more than one performance measurement to have a solid indication of a model's true ability to make predictions.

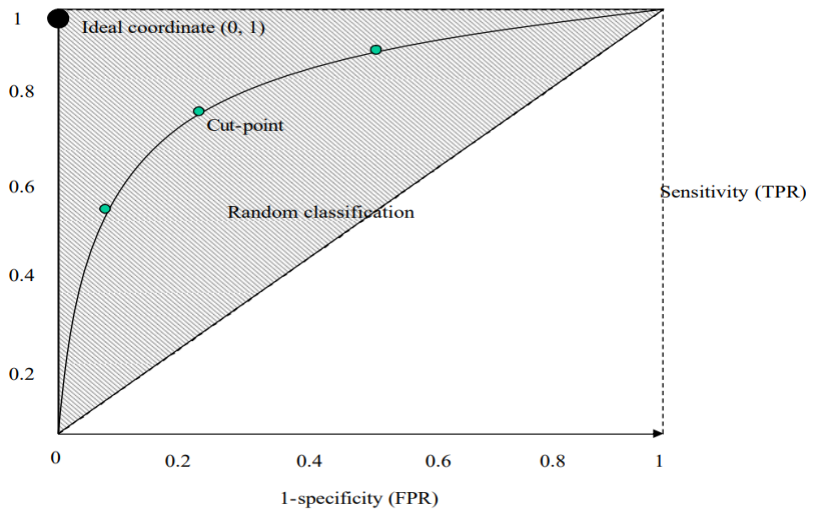

Figure 2.10. ROC diagram mapping sensitivity against 1 - specificity. The diagonal line represents a model performing as a random classifier, the grey space represents anything better than a random classifier, and the ideal coordinate represents where a perfectly performing model would go through. Sourced from Zhu et al. (2010).

AUC Range	Classification
0.9 < AUC < 1.0	Excellent
0.8 < AUC < 0.9	Good
0.7 < AUC < 0.8	Worthless
0.6 < AUC < 0.7	Not good

Table 2.3. Classifications describing the various area under the curve (AUC) values achieved by a model

2.2 Results

2.2.1 Image Analysis

Using the mean spectral signatures for the middle 400 pixels to analyse the images, through Python and Scyven, revealed the wavelength with the greatest intensity on average for all the images was 766nm – light intensity of 306 for non-GM and 313 for GM. The average light intensity for the GM canola grains was equal to or greater than the non-GM canola at every wavelength. Whilst there appeared to be a minimal visual difference between the two average spectral signatures (Figure 2.5), the wavelength with the greatest intensity difference between the two classes (713.536nm) was found by using the mean wavelength intensities and used to run an ANOVA and t-test with the values from all 1010 images for that wavelength to check for a statistically significant difference. Against a null hypothesis whereby there is no difference in the spectral signature, the associated p-value for the data set from this wavelength was 0.041 which is less than 0.05 and, hence, it can be concluded that there is a statistically significant difference between the two spectral signatures for GM and non-GM at that specific wavelength (Appendix Table B1). All other wavelengths did not provide a statistically significant difference, however.

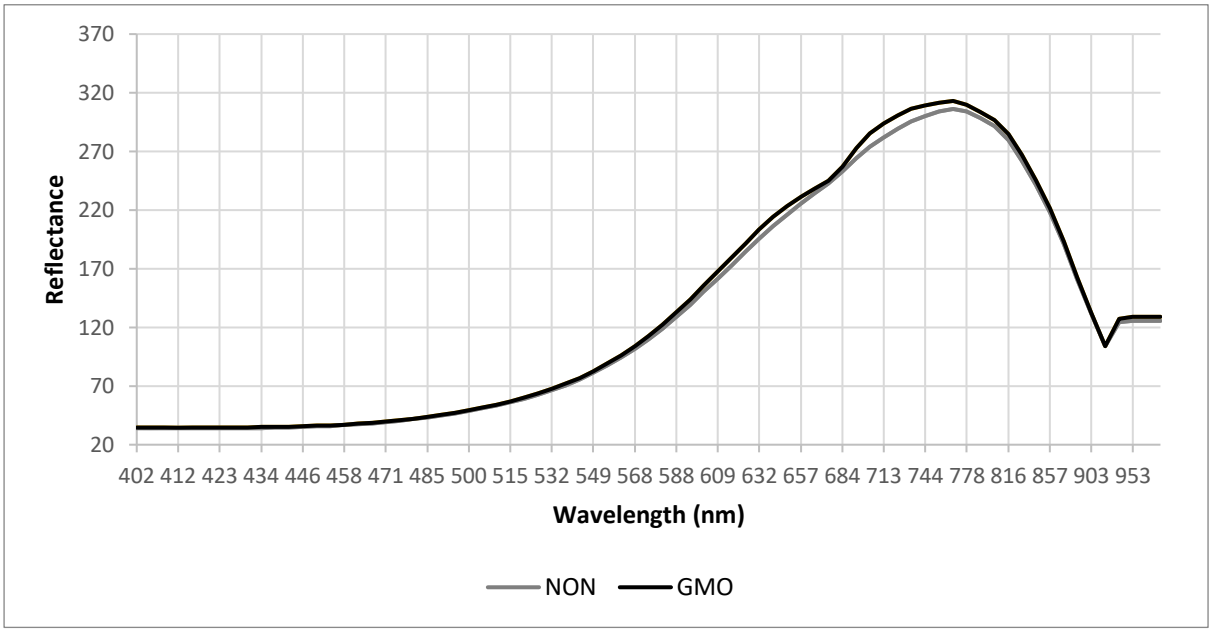

Figure 2.11. Average spectral signatures of non-GM canola in grey and GM canola in black.

In terms of grain size, there was a slightly greater distinction between the two classes, but still a significant cross-over. In terms of diameter, non-GM canola had an average diameter of 219.14 pixels (standard deviation (SD) of 19.00) versus GM canola, which had an average diameter of 228.55 pixels (SD of 20.92) – illustrated in the normal distribution curve in Figure 2.6. Regarding area, non-GM canola had an average area of 3.15×10^{-4} (SD of 5.51×10^{-3}) versus GM canola with an average area of 3.48×10^{-4} pixels (SD of 6.80×10^{-3}). Irrespective of the cross over, the p-values for both diameter and area of GM and non-GM were well below 0.01 (Appendix Table B2 and B3), indicating statistical significance.

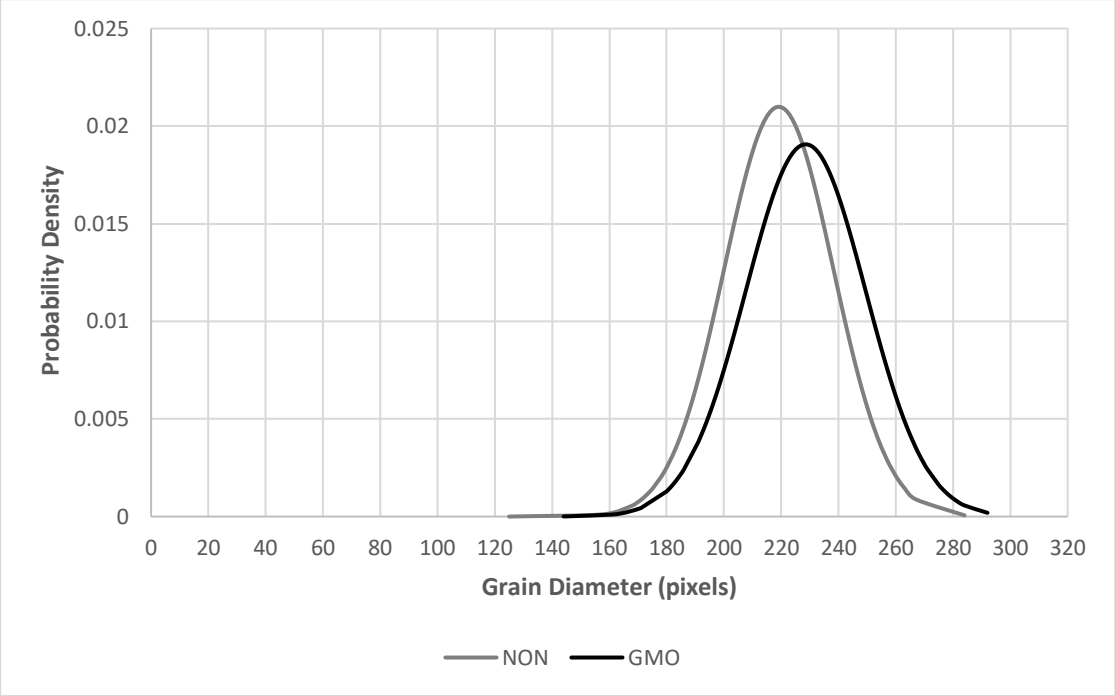

Figure 2.13. Normal distribution curve of non-GM canola (grey) and GM canola (black) for grain diameter in pixels.

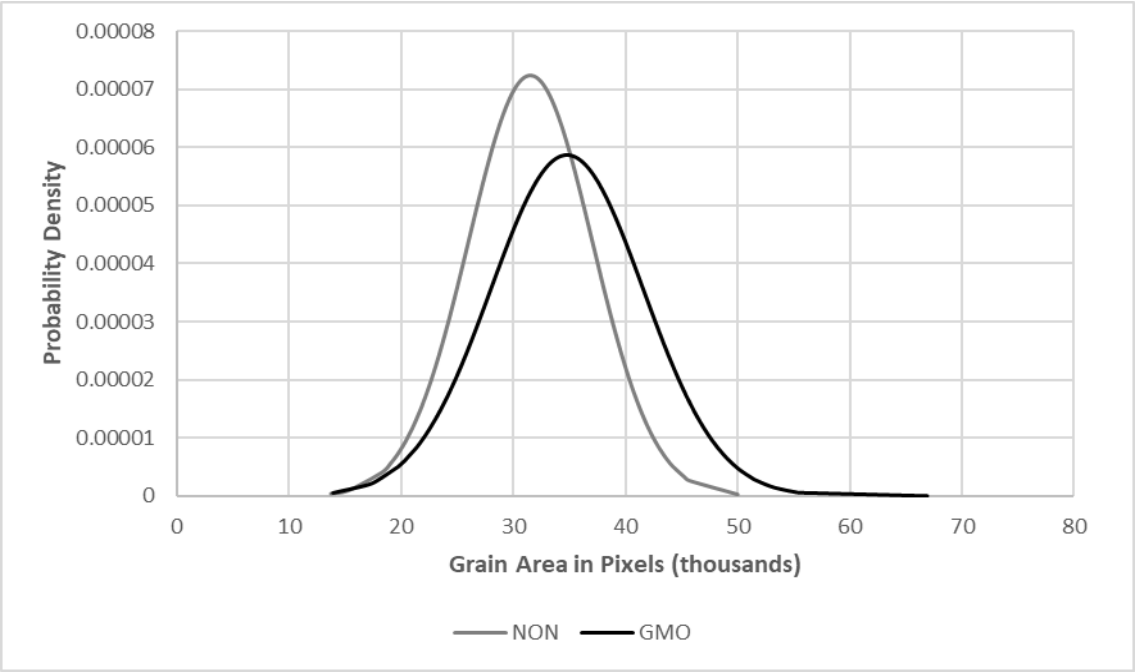

Figure 2.15. Normal distribution curve of GM canola (black) vs non-GM canola (grey) for grain area in pixels (thousands).

2.2.2 Convolution Neural Network

2.2.2.1 Hyperspectral Image Classifier

Figure 2.8 illustrates the loss (prediction error) graph over epochs. This suggests minimal loss can be quickly achieved using a CNN and also signifies that the network can adapt quickly and produce reliable results within a short amount of time. Concurrently, Figures 2.9 and 2.10 indicate that accuracy can be achieved quickly for both the training and validation data set, plateauing between the 50^{th} and 60^{th} epoch at a reasonable validation accuracy percentage. Figure 2.11 highlights the relationship between precision and recall over the confidence thresholds which can be used to determine the F1 score. Figures 2.12 and 2.13 illustrate the ROC on the validation model and the test model respectively. The test model has an AUC of 87% which can be considered a 'good' result (Table 2.1).

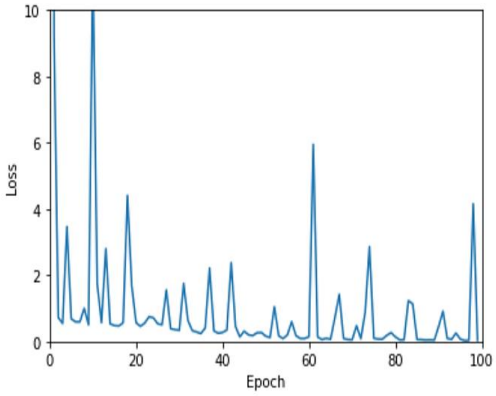

Figure 2.8. Hyperspectral image classifier loss over epochs.

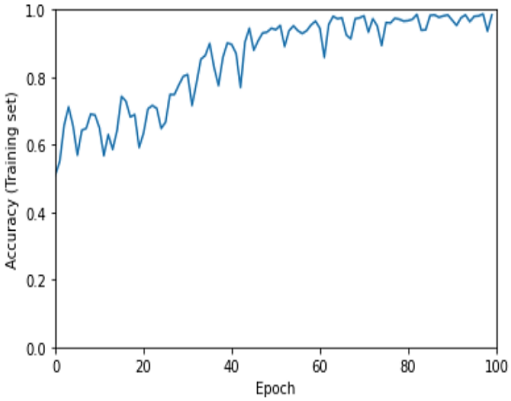

Figure 2.9. Hyperspectral image classifier accuracies over epochs for the training data set.

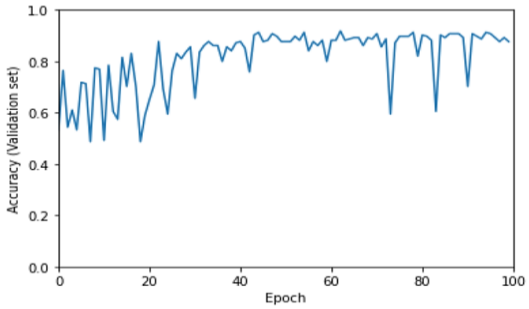

Figure 2.10. Hyperspectral image classifier accuracies over epochs for the validation

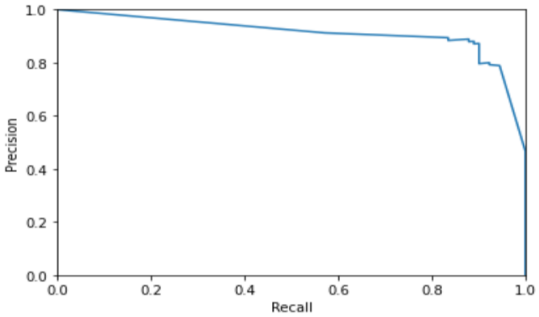

Figure 2.11. Hyperspectral image classifier Precision against recall for the validation data set with an AUC of 91.21%.

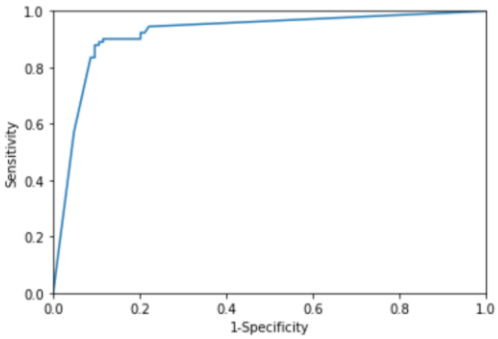

Figure 2.12. Hyperspectral image classifier sensitivity against 1 - Specificity for the validation data set with an AUC of 90.54%.

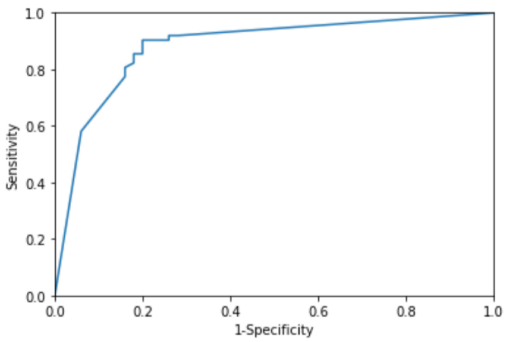

Figure 2.13. Hyperspectral image classifier sensitivity against 1 - Specificity for the test data set with an AUC of 86.47%.

The test set model outputs are illustrated in a confusion matrix (Figure 2.14) that includes 54 TP, 40 TN, 10 FP, and 8 FN. These values give an overall accuracy of 84% and precision, sensitivity/recall, and specificity all above 80%. These values were also used to calculate the F1 score (0.86) and MCC (0.67) (Table 2.2).

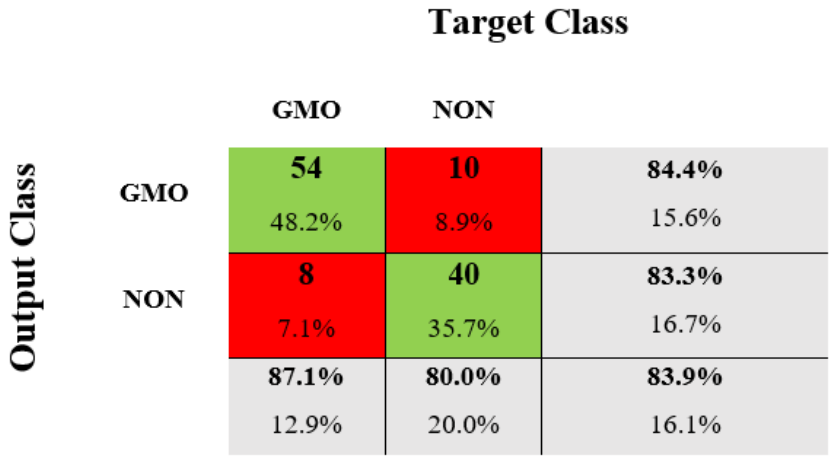

Figure 2.14. Confusion matrix for the hyperspectral image classifier against the test data set. Target class against the outputted class, outlining the true and false positives and negatives. Values are also used to calculate precision, sensitivity, and specificity and overall accuracy.

Table 2.4. F1, Mathews correlation coefficient, and validation area under the curves for the validation and test data set for the bitmap and hyperspectral image classifiers.

Model	BMP	HSI
F1 score	0.8974	0.8571
MCC	0.8256	0.6740
Validation AUC (Sensitivity x 1-specificity)	98.23	90.54
Test AUC (Sensitivity x 1-specificity)	97.06	86.47

2.2.2.2 Bitmap Image Classifier

With regards to the image classifier CNN designed for the bitmap images, Figure 2.15 illustrates the loss graph over epochs, which, from the first epoch, were less than 1. Figures 2.16 and 2.17 indicate accuracy levels beginning to plateau roughly around the 50th epoch. Figure 2.18 shows the relationship between precision and recall over confidence thresholds which can be used to determine the F1 score for the bitmap image classifier evident in Table 2.2. Figures 2.19 and 2.20 show the ROC (sensitivity versus 1-specificity) for the validation and test data sets respectively. Mapped against the HSI results Table 2.2 also outlines the validation and test data set AUCs, 98% and 97% respectively. These AUC values would be classified as excellent according to Table 2.1.

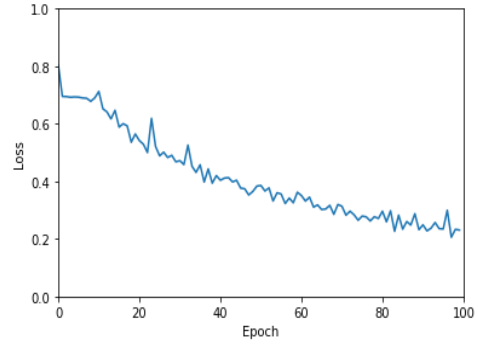

Figure 2.15. Bitmap image classifier loss over epochs.

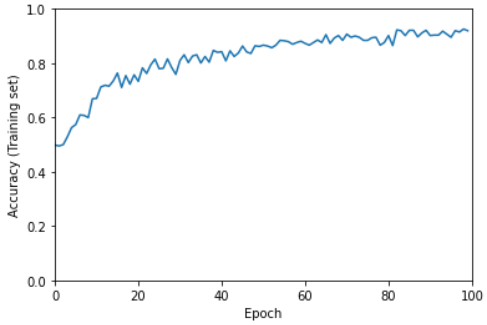

Figure 2.16. Bitmap image classifier accuracies over epochs for the training data set.

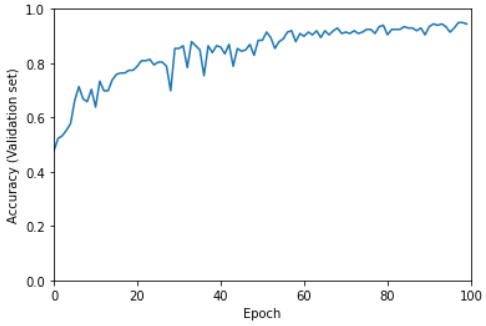

Figure 2.17. Bitmap image classifier accuracies over epochs for the validation

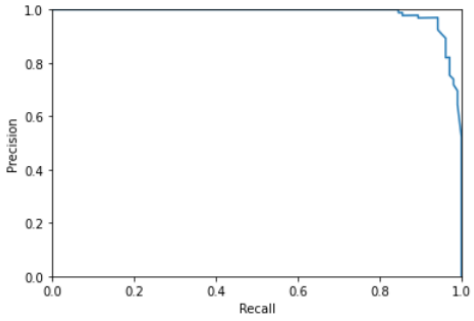

Figure 2.18. Bitmap image classifier. Precision against recall for the validation data set with an AUC of 98.49%

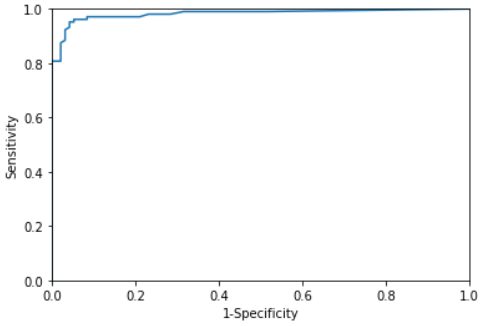

Figure 2.19. Bitmap image classifier sensitivity against 1 - Specificity for the validation data set with an AUC of 98.23%.

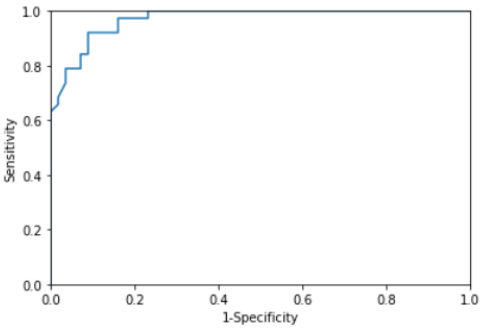

Figure 2.20. Bitmap image classifier sensitivity against 1 - Specificity for the test data set with an AUC of 97.06%.

		Actual Class: GMO	Actual Class: NON	
Predicted Class	GMO	**35** 37.2%	**5** 5.3%	**87.5%** 12.5%
	NON	**3** 3.2%	**51** 54.3%	**94.4%** 5.6%
		92.1% 7.9%	**91.1%** 8.9%	**91.5%** 8.7%

Figure 2.21. Confusion matrix for the bitmap image classifier against the test data set. Target class against the outputted class, outlining the true and false positives and negatives. Values are also used to calculate precision, sensitivity, and specificity and overall accuracy.

The test set model outputs for the bitmap image classifier convolution neural network are illustrated in a confusion matrix (Figure 2.21) that includes 35 TP, 51 TN, 5 FP, and 3 FN. These values give an overall accuracy of 92% and precision, sensitivity/recall, and specificity all above 87%. These values were also used to calculate the F1 score (0.9) and MCC (0.83) (Table 2.2).

2.3 Discussion

This study aimed to investigate if a machine learning model fed hyperspectral images would be able to accurately differentiate genetically modified and non-GM canola grain. Achieving accuracies close to 90% using a hyperspectral image classifier supports the research hypothesis that machine learning can be used to accurately differentiate hyperspectral images of GM and non-GM canola and indicates a significant performance standard that suggests this area of technology could be used as a novel, non-destructive and rapid tool in the agricultural industry. With further training and development, the machine learning aspect could effortlessly analyse enormous quantities of data quickly and efficiently, and at low cost. Contrary to expectations, however, the CNN model trained on the standard bitmap images performed similarly to the HSI CNN, suggesting the hyperspectral components of the images are not necessary for accurate classification and defining features are more likely to be spatial than spectral, since bitmap images primarily account for spatial differences. This is supported by Figure 2.5 where there is little significant difference shown between the spectral signature of GM and non-GM canola, the only wavelength with a p-value less than 0.05, and suggesting any significance, was at 713nm; but again, this is unlikely to be a major defining feature as there are substantial crossovers in reflectance values at that wavelength in GM and non-GM grains which would not support the high accuracies obtained by the HSI CNN or explain the high performance produced by the bitmap image classifier. However, this spectral signature is also very similar to the heat lamp's wavelength output provided on the packaging of the globe; hence, further research should be conducted to determine a more accurate spectral signature of GM and non-GM canola grain by either using a different or combination of light sources and increasing the spectral wavelengths captured beyond the small range of 400 to 1000nm as this only captures visible light and a small proportion of the NIR wavelengths.

Similarly, although the data sets for grain diameter and area are statistically significantly different with p-values below 0.01 (Figures 2.5 and 2.7), the cross over evident in their distribution curves also does not completely explain the model's ability to differentiate the two classes to the extent illustrated in the confusion matrixes for the bitmap and HSI CNN (Figures 2.14 and 2.21). Whilst this suggests GM canola is larger than non-GM canola grains on average, it also means the model would be able to predict class based on two grains of the same size. Since grain size is partly correlated to the environmental conditions in its respective growing season, it is important the model does not heavily rely on grain size as this would vary between seasons and climatic areas.

This emphasises that whilst the automatic feature selection abilities of CNN models are often hailed as their greatest advantage, the downfall is the difficultly in deciphering what the model is analysing. Although the results from this study are exciting and promising, replicated experiments with different grain and different coding CNN models are necessary to verify that this technology is able to predict GM and non-GM grains of the same size and different varieties.

2.3.1 Model's Performance Against Current Standards

Since current guidelines only allow non-GM canola to have cross-contamination levels up to 0.9% of GM canola, detection methods need to have extremely high accuracies (Emslie *et al*., 2007). Current methods are thought to provide up to 99% accuracy, hence, for another technology to be implemented in the industry, it would either have to match the accuracy or be significantly cheaper to implement. It is surprising that both models produced accuracies around 90%, with limited grain variability, strongly highlighting the potential for this technology to achieve the commercial 99% benchmark with further research. For instance, a larger data set, different circumstances, or a more complex neural network may have more

success. In saying that, methods such as PCR rely on taking a small subsample from a large composite stack which contributes to a level of uncertainty surrounding the associated results (Emslie *et al.*, 2007). Since image classification is much more rapid and does not require sample preparation or destruction, it would be practical to have multiple assessments conducted which would increase overall reliability and confidence in the collected data i.e., a camera periodically taking photos of grain as it passes on a conveyor belt. Hence, statistically speaking, even an accuracy of 90% would be favourable if this technology could cover a much larger proportion of the grain than the very small subset that PCR tests, thus increasing the likelihood of detecting GMOs overall, especially as grain handling companies such as CBH largely rely on the honesty of growers to declare GM canola loads when delivering grain to their sites. Being able to identify truckloads quickly and easily using something as simple as a camera, as either GM or non-GM, also reduces the risk of loads being delivered under an incorrect classification – whether purposely or accidently and helps to prevent the contamination of stacks altogether.

2.3.2 Performance Evaluation of Models

As there is not a sole performance measurement used in machine learning (Chicco & Jurman, 2020), it is important to consider multiple measures, and their advantages and disadvantages, when evaluating and comparing models. Most related studies in grain contamination use accuracy as an overarching term, generally referring to the accuracy as the number of correct identifications over total assessments or the F1 score. In saying that, the MCC is considered to be more holistically truthful as it incorporates all aspects of the confusion matrix and accounts for any data imbalances (Chicco & Jurman, 2020). This is important as the focus for identifying contaminations is less so on how many a method can get right and more so on how many were incorrectly identified or completely missed in the quality check process. With MCC scores of 0.83 and 0.67 for the BMP and HSI CNN respectively, there is

clear promise within this technology, but more work needs to be done to reduce the false positive and negative classifications.

Similarly, on a more specific level, it is extremely important for grain handling companies like CBH to strongly consider both sensitivity (the percent of true positives identified) and specificity (the percent of positive identifications that were correct) as the priority of either would vary based on who the target group is. For example, false negatives – missing GM canola, would be of greater concern to exporters as opposed to growers who would be more concerned with a false positive and unnecessarily losing money on a truckload of canola. It is important to quantify this balance as using misclassifications rates as a measure of performance can be misleading when the wrong assumptions are made (Hand & Till, 2001).

2.3.3 Convolution Neural Networks

Concurrent to having a large dataset of training images, a considerable determiner of accuracy and performance relates to the type and design of the machine learning tool used. This is evident both in this study itself and also in literary sources evaluating the strength of different types of neural networks currently used for image classification (Paoletti *et al.*, 2019). Convolution neural networks are a relatively recent development in machine learning and have illustrated superior performance to previous networks. One aspect that facilitated this improvement was the replacement of the sigmoid activation function with a rectified linear (relu) activation function which simplifies learning and reduces computational time (Wu, 2017). Relu functions makes all values zero when they are less than zero and does not activate all neurons at the same time (Paoletti, 2019).

2.3.3.1 Convolution Neural Networks with Hyperspectral Images

It is fairly universally agreed in literature that the combination of HSI with powerful data analysis tools such as machine or deep learning has extremely significant potential to improve efficiency and reduce resource use (time, labour, and money) in a large scope of areas in agriculture (Lu *et al.*, 2020); although the costs (financial and computational) and complexity associated with HSI continues to be an obstacle to larger uptake in the industry (Weiss *et al.*, 2020). Out of the main features extracted from an image, e.g., colour, morphological, and textural, colour is often highlighted as having the greatest and most straightforward discrimination potential for ML classifiers (Singh *et al.*, 2010). For many studies this has certainly been the case, with colour features dominating the classificational parameters used, for example, in a study by Singh, *et al.*(2010) using HSI and neural networks to identify insect damage in wheat; however, detection using this method frequently resulted in a high number of false positives – which is not reflected in their high accuracies, often above 90%.

Another study by Agarwal, *et al.* (2020) also used HSI but relied on unsupervised CNN and experienced similar accuracies levels (above 90%), such as identifying grain insect bodies and fragments. This study also compared these results against a capsule network (a variant of CNN that preserves structures/capsules during training) – which produced higher accuracies; and could hence, be a promising option to try for other material identification studies in grain biosecurity. Other studies involving insect detection in cereal grains also produced results of around 90% (Neethirajan *et al.*, 2007). The identified limitations, however, include lower accuracies at low levels of infestation, the effect of moisture levels in samples, and the difficultly sometimes associated with calibration of the needed equipment; alternative imagery methods of insect detection include x-ray (Neethirajan *et al.*, 2007).

This is fairly consistent with the accuracy levels achieved in this study using a HSI CNN (84%). Looking at the graphs produced by the model for loss, accuracy, and sensitivity over 1-specificity (Figures 2.8 to 2.13), a HSI CNN shows promise as a diagnostic tool in the grains industry. The drop in AUC between the validation data set to the test data set for the ROC space (Figures 2.12 and 2.13) is to be expected but it would be interesting to see the effect of changing the proportion of training, validation, and test samples.

A study similar to this one by Feng, *et al.* (2017) also used NIR HSI to distinguish GM maize from non-GM maize. The combination of using principal component analysis with a least-squares discriminant analysis illustrates the use of machine, as opposed to deep, learning; regardless, results were also significantly promising with accuracies close to one hundred percent. This study also saw slightly different average spectral reflectance between the GM and non-GM grains, however, in the case of maize, the non-GM kernels reflected more, whereas GM canola in this study reflected more than non-GM canola.

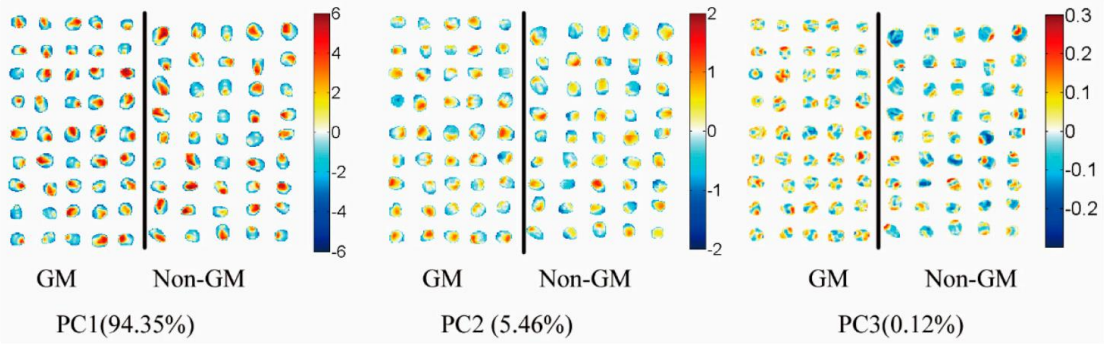

Figure 2.22. Visualisation of the PCA scores for the first three principal components of the maize kernels from a study by Feng, et al. (2017).

Using PCA and associated scores, they were able to create a visualisation of the difference between the GM and non-GM maize – evident in PC3 (Figure 2.22). Attempts to map similar differences in GM and non-GM canola grain for this study did not find a similar pattern, possibly because of the light sources used, hence, it would be interesting for further research to endeavour with confirming if there is a spectral difference between the two classes.

Lastly, it is important to note, however, the additional processing challenges involved with HSI CNN as opposed to standard images. Whilst the wealth of information within a hyperspectral image is unparalleled and has contributed to outstanding advances, especially in areas such as remote sensing, the larger dimensions not only take up more storage space and time to process but also add complexity to the model by increasing the number of parameters to consider and hence could explain why the HSI CNN performed slightly poorer than the bitmap CNN (Paoletti *et al.*, 2019). The lack of image augmentation (flipping or zooming in on sections of the image to artificially increase the data set) in the HSI CNN that was present in the bitmap CNN is likely to be another explanatory factor in this difference as it artificially increases the number of training samples and improves the architecture of the model (Paoletti *et al.*, 2019).

2.3.3.2 Convolution Neural Networks with Standard Images

Although the HSI CNN performed well, the similar and better performance exhibited through the use of the black and white bitmap images only in a similarly constructed CNN suggests the main differentiating parameters identified by the model are spatial, not spectral. The bitmap model showed more consistent and lower losses over the 100 epochs and greater scores in all the performance measures used in this study. This is particularly exciting as it suggests that differentiation could be done using a normal camera and hence, in the future could lead to the possibility of something as simple as using a phone application to identify if the grain is genetically modified or not. The possibility of an extremely accessible and user friendly, portable, easy to install, and cheap detection option is extremely exciting. The bitmap image model was also significantly quicker to run as well. The greatest challenges within this study included acquiring the images on the hyperspectral camera and running them in the model as processing the large and complex files is very time-consuming. Hence, if normal

images can be used instead, it would be beneficial on many fronts, particularly as it decreases the need for image processing, making the data easy to handle.

2.4 Conclusion

Canola is an important crop nationally and internationally. Keeping genetically modified canola from non-GM canola separate, and identifying how successfully this has been done, is an important step along various aspects of the supply chain. This research project aimed to determine if machine learning and hyperspectral imagery could be used as a rapid and non-destructive tool in discriminating GM from non-GM. Over 1000 images were used to train and test a convolution neural network model for standard bitmap images and hyperspectral images. Surprisingly, both models produced extremely promising results for a novel study, with the bitmap CNN outperforming the HSI CNN. Further research explaining why there is a clear difference between the two classes is needed, as is further projects testing the differentiation under different conditions such as different lighting, deep learning models, piles of grain instead of individual grains, and from different seasons and geographical areas. Although the data suggests that using HSI is not theoretically necessary in this instance, since many advances have been made in other areas of grain contamination, it would be beneficial – in the long term, to see a product that could detect and identify a multitude of contaminants in a sample, most likely using a HSI camera anyway. Regardless, just considering GM and non-GM canola, however, these results highlight the potential for identification to be possible using something as simple as an application on a mobile phone. For instance, there would be a significant benefit in future studies focusing on developing models that could identify the genetic modification status of a handful of grain for use in sampling huts, at ports for export, and by farmers, and assess the neural network's ability to accurately predict GM grains whilst they are moving on a conveyor belt. Overall, this illustrates that artificial intelligence, specifically machine/deep learning is the way forward in biosecurity for market export and safety reasons.

References

ABARES. (2021a). *Agricultural commodities: March quarter 2021*. Canberra Retrieved from https://doi.org/10.25814/r3te-d792

ABARES. (2021b). *Agricultural Commodities: September quarter 2021*. Canberra Retrieved from https://daff.ent.sirsidynix.net.au/client/en_AU/search/asset/1032541/0

ACIL-Tasman. (2007). *GM Canola: An Information Package*. Department of Agriculture, Fisheries and Forestry Retrieved from https://www.agriculture.gov.au/sites/default/files/sitecollectiondocuments/ag-food/biotech/gm-canola-info-package.pdf

AEGIC. (2021). *Australian Canola: Quality, safety, and reliability*. https://aegic.org.au/wp-content/uploads/2021/03/AEGIC-Grain-Note-canola_LR.pdf

Agarwal, M., Al-Shuwaili, T., Nugaliyadde, A., Wang, P., Wong, K. W., & Ren, Y. (2020). Identification and diagnosis of whole body and fragments of Trogoderma granarium and Trogoderma variabile using visible near infrared hyperspectral imaging technique coupled with deep learning. *Computers and Electronics in Agriculture*, *173*, 105438. https://doi.org/https://doi.org/10.1016/j.compag.2020.105438

Aggarwal, S. (2004). Principles of Remote Sensing. In M. V. K. Sivakumar, P. S. Roy, K. Harmsen, & S. K. Saha (Eds.), *Satellite Remote Sensing and GIS Applications in Agricultural Meteorology* (pp. 23-38). World Meteorological Organisation. http://www.wamis.org/agm/pubs/agm8/Paper-2.pdf

Akiyama, H., Makiyama, D., Nakamura, K., Sasaki, N., Minegishi, Y., Mano, J., Kitta, K., Ozeki, Y., & Teshima, R. (2010). A Novel Detection System for the Genetically Modified Canola (Brassica rapa) Line RT73. *Analytical Chemistry*, *82*(23), 9909-9916. https://doi.org/10.1021/ac102434q

Albawi, S., Abed Mohammed, T., & Alzawi, S. (2017). *Understanding of a Convolutional Neural Network*. https://doi.org/10.1109/ICEngTechnol.2017.8308186

An, J. (2020). How to Remember all these Classification Concepts forever. https://medium.com/swlh/how-to-remember-all-these-classification-concepts-forever-761c065be33

Baeten, V., Pierna, J. A. F., & Dardenne, P. (2007). Hyperspectral Imaging Techniques: an Attractive Solution for the Analysis of Biological and Agricultural Materials.

Bao, Y., Mi, C., Wu, N., Liu, F., & He, Y. (2019). Rapid Classification of Wheat Grain Varieties Using Hyperspectral Imaging and Chemometrics. *Applied sciences*, *9*(19), 4119. https://doi.org/10.3390/app9194119

Brann, M., & Prendergast, J. (2021, 18/09/2021). Canada's tough canola harvest driving prices up for Aussie farmers. *Landline*. https://www.abc.net.au/news/2021-09-18/canola-crop-canada-prices-high-australia-farmers/100469038

Bruning, B., Berger, B., Lewis, M., Liu, H., & Garnet, T. (2020). *Hyperspectral sensing for the prediction of nitrogen, water and salt content in wheat*. G. R. a. D. Corportation. https://grdc.com.au/resources-and-publications/grdc-update-papers/tab-content/grdc-update-papers/2020/02/hyperspectral-sensing-for-the-prediction-of-nitrogen,-water-and-salt-content-in-wheat

Campesato, O. a. (2020). *Artificial intelligence, machine learning, and deep learning*. Mercury Learning and Information. https://go.exlibris.link/JjSXkJxp

Caporaso, N., Whitworth, M. B., & Fisk, I. D. (2018). Near-Infrared spectroscopy and hyperspectral imaging for non-destructive quality assessment of cereal grains. *Applied Spectroscopy Reviews*, *53*(8), 667-687. https://doi.org/10.1080/05704928.2018.1425214

CBH. (2020). Canola Receival Standards Limits. In. Western Australia: CBH.

Central, G. (2020). Barley, wheat reset in wake of China trade restrictions: Rural Bank. https://www.graincentral.com/markets/barley-wheat-reset-in-wake-of-china-trade-restrictions-rural-bank/

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Chollet, F. (2016). *Building powerful image classification models using very little data*. In https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html

Dantes, P. T. G. (2020). *Nir Hyperspectral Imaging for Animal Feed Ingredient Applications* (Publication Number 28028579) [Ph.D., Iowa State University]. ProQuest One Academic. Ann Arbor. http://libproxy.murdoch.edu.au/login?url=https://www.proquest.com/dissertations-theses/nir-hyperspectral-imaging-animal-feed-ingredient/docview/2443581623/se-2?accountid=12629

Dixit, Y., Al-Sarayreh, M., Craigie, C. R., & Reis, M. M. (2021). A global calibration model for prediction of intramuscular fat and pH in red meat using hyperspectral imaging. *Meat Science*, *181*, 108405. https://doi.org/https://doi.org/10.1016/j.meatsci.2020.108405

Emslie, K. R., Whaites, L., Griffiths, K. R., & Murby, E. J. (2007). Sampling Plan and Test Protocol for the Semiquantitative Detection of Genetically Modified Canola (Brassica napus) Seed in Bulk Canola Seed. *Journal of agricultural and food chemistry*, *55*(11), 4414-4421. https://doi.org/10.1021/jf070267i

Erkinbaev, C., Henderson, K., & Paliwal, J. (2017). Discrimination of gluten-free oats from contaminants using near infrared hyperspectral imaging technique. *Food Control*, *80*, 197-203. https://doi.org/https://doi.org/10.1016/j.foodcont.2017.04.036

Feng, L., Zhu, S., Liu, F., He, Y., Bao, Y., & Zhang, C. (2019). Hyperspectral imaging for seed quality and safety inspection: A review. *Plant methods*, *15*(1), 91-91. https://doi.org/10.1186/s13007-019-0476-y

Feng, X., Zhao, Y., Zhang, C., Cheng, P., & He, Y. (2017). Discrimination of Transgenic Maize Kernel Using NIR Hyperspectral Imaging and Multivariate Data Analysis. *Sensors (Basel, Switzerland)*, *17*(8), 1894. https://doi.org/10.3390/s17081894

Feng, Y., & Sun, D.-W. (2012). Application of Hyperspectral Imaging in Food Safety Inspection and Control: A Review. *Critical reviews in food science and nutrition*, *52*, 1039-1058. https://doi.org/10.1080/10408398.2011.651542

Gallagher, N., & Lawrence, L. (2020). *Introduction to Hyperspectral and Multivariate Image Analysis and Principal Components Analysis for Multivariate Images*.

Ge, J. C., Yoon, S. K., & Choi, N. J. (2017). Using canola oil biodiesel as an alternative fuel in diesel engines: A review. *Applied sciences*, *7*(9), 881. https://doi.org/10.3390/app7090881

Geladi, P. L. M., F.Grahn, H., & Burger, J. E. (2007). Multivariate Images, Hyperspectral Imaging: Background and Equipment. In *Techniques and Applications of Hyperspectral Image Analysis*. John Wiley & Sons, Incorporated.

Genetically Modified (GM) Canola in Australia. (2018). Department of Health, Office of the Gene Technology Regulator Retrieved from https://www.ogtr.gov.au/sites/default/files/files/2021-06/12_-_genetically_modified_gm_canola_in_australia.pdf

Halimu, C., Kasem, A., & Newaz, S. H. S. (2019). *Empirical Comparison of Area under ROC curve (AUC) and Mathew Correlation Coefficient (MCC) for Evaluating Machine Learning Algorithms on Imbalanced Datasets for Binary Classification* Proceedings of the 3rd International Conference on Machine Learning and Soft Computing, Da Lat, Viet Nam. https://doi-org.libproxy.murdoch.edu.au/10.1145/3310986.3311023

Hand, D., & Till, R. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Hand, The*, *45*, 171-186. https://doi.org/10.1023/A:1010920819831

Hänsch, R., Ley, A., & Hellwich, O. (2017, 23-28 July 2017). Correct and still wrong: The relationship between sampling strategies and the estimation of the generalization error. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS),

Hudson, D., & Richards, R. *GM Canola Impact Survey: Information for Growers, Advisors and Industry* https://grdc.com.au/__data/assets/pdf_file/0021/74721/gm-canola-impact-survey-pdf.pdf.pdf

Hyperspectral Image Generator. In. (2020). https://github.com/mabdelhack/hyperspectral_image_generator/blob/master/hyperspectral_image_generator.py

Igne, B. (2009). *Intra and inter-brand calibration transfer for near infrared spectrometers* (Publication Number 3355508) [Ph.D., Iowa State University]. ProQuest One Academic. Ann Arbor. http://libproxy.murdoch.edu.au/login?url=https://www.proquest.com/dissertations-theses/intra-inter-brand-calibration-transfer-near/docview/304902000/se-2?accountid=12629

http://SR4LP2WR3C.search.serialssolutions.com?ctx_ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF-8&rfr_id=info:sid/ProQuest+Dissertations+%26+Theses+Global&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&rft.genre=dissertations+%26+theses&rft.jtitle=&rft.atitle=&rft.au=Igne%2C+Benoit&rft.aulast=Igne&rft.aufirst=Benoit&rft.date=2009-01-01&rft.volume=&rft.issue=&rft.spage=&rft.isbn=978-1-109-14336-2&rft.btitle=&rft.title=Intra+and+inter-brand+calibration+transfer+for+near+infrared+spectrometers&rft.issn=&rft_id=info:doi/

James, D., Schmidt, A. M., Wall, E., Green, M., & Masri, S. (2003). Reliable detection and identification of genetically modified maize, soybean, and canola by multiplex PCR analysis. *J Agric Food Chem*, *51*(20), 5829-5834. https://doi.org/10.1021/jf0341159

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, *147*, 70-90. https://doi.org/https://doi.org/10.1016/j.compag.2018.02.016

Kamruzzaman, M., Makino, Y., & Oshita, S. (2016). Parsimonious model development for real-time monitoring of moisture in red meat using hyperspectral imaging. *Food chemistry*, *196*, 1084-1091. https://doi.org/10.1016/j.foodchem.2015.10.051

Khanal, S., KC, K., P.Fulton, J., Shearer, S., & Ozkan, E. (2020). Remote Sensing in Agriculture—Accomplishments, Limitations, and Opportunities. *Remote Sensing*, *12*(22), 3783. https://doi.org/http://dx.doi.org/10.3390/rs12223783

Kim, J.-H., Park, S.-B., Hong, Y., & Kim, H.-Y. (2015). Detection of eight genetically modified canola events using two event-specific pentaplex PCR systems. *Food Control*, *51*, 183-189. https://doi.org/https://doi.org/10.1016/j.foodcont.2014.11.027

Kingwell, R. (2020, 7 October 2020). *Part 2: crop diversification - which markets and classes?"* GIWA Forum 2020, Crown Towers, Perth.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

Lee, H., Kim, M. S., Lee, W.-H., & Cho, B.-K. (2018). Determination of the total volatile basic nitrogen (TVB-N) content in pork meat using hyperspectral fluorescence imaging. *Sensors and Actuators B: Chemical*, *259*, 532-539. https://doi.org/https://doi.org/10.1016/j.snb.2017.12.102

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors (Basel, Switzerland)*, *18*(8), 2674. https://doi.org/10.3390/s18082674

Lim, J., Lee, A., Kang, J., Seo, Y., Kim, B., Kim, G., & Kim, S. (2020). Non-Destructive Detection of Bone Fragments Embedded in Meat Using Hyperspectral Reflectance Imaging Technique. *Sensors*, *20*(14), 4038. https://doi.org/http://dx.doi.org/10.3390/s20144038

Lu, B., Dao, P., Liu, J., He, Y., & Shang, J. (2020). Recent Advances of Hyperspectral Imaging Technology and Applications in Agriculture. *Remote Sensing*, *12*, 2659. https://doi.org/10.3390/rs12162659

Malcolm, B., Sale, P., Leury, B., & Barlow, S. (2009). *Agriculture in Australia: an introduction* (2 ed.). Oxford University Press.

McKeon, T. A. e. (2016). *Industrial oil crops*. Academic Press and AOCS Press. https://doi.org/10.1016/C2015-0-00068-5

Moghimi, A. (2019). *Integrating Hyperspectral Imaging and Artificial Intelligence to Develop Automated Frameworks for High-throughput Phenotyping in Wheat* (Publication Number 13806257) [Ph.D., University of Minnesota]. ProQuest One Academic. Ann Arbor. http://libproxy.murdoch.edu.au/login?url=https://www.proquest.com/dissertations-theses/integrating-hyperspectral-imaging-artificial/docview/2202997635/se-2?accountid=12629

Neethirajan, S., Karunakaran, C., Jayas, D. S., & White, N. D. G. (2007). Detection techniques for stored-product insects in grain. *Food Control*, *18*(2), 157-162. https://doi.org/https://doi.org/10.1016/j.foodcont.2005.09.008

Nilesh. (2018). *Custom Keras Generators: Customize Your Data Genrators for Faster Training*. In https://towardsdatascience.com/writing-custom-keras-generators-fe815d992c5a

Niu, X.-X., & Suen, C. Y. (2012). A novel hybrid CNN–SVM classifier for recognizing handwritten digits. *Pattern Recognition*, *45*(4), 1318-1325. https://doi.org/https://doi.org/10.1016/j.patcog.2011.09.021

Paoletti, M. E., Haut, J. M., Plaza, J., & Plaza, A. (2019). Deep learning classifiers for hyperspectral imaging: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *158*, 279-317. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2019.09.006

Patel, G. S. e. (2021). *Smart agriculture: emerging pedagogies of deep learning, machine learning and Internet of Things* (First ed.). CRC Press/Balkema/Taylor & Francis Group. https://go.exlibris.link/Q7B7q3YN

Paull, J. (2019). Genetically Modified (GM) Canola: Price Penalties and Contaminations. *Biomedical Journal of Scientific & Technical Research*, *17*(2). https://doi.org/10.26717/BJSTR.2019.17.002965

Phillips, P. W. B., Khachatourians, G. G., & Ebrary, I. (2001). *The biotechnology revolution in global agriculture: innovation, invention, and investment in the canola industry* (Vol. 24.). CABI Pub. https://doi.org/10.1079/9780851995137.0000

Qiao, T., Ren, J., Yang, Z., Qing, C., Zabalza, J., & Marshall, S. (2015). Visible hyperspectral imaging for lamb quality prediction. *Technisches Messen*, *82*(12), 643-652. https://doi.org/10.1515/teme-2015-0043

Ravikanth, L., Singh, C. B., Jayas, D. S., & White, N. D. G. (2015). Classification of contaminants from wheat using near-infrared hyperspectral imaging. *Biosystems engineering*, *135*, 73-86. https://doi.org/https://doi.org/10.1016/j.biosystemseng.2015.04.007

Shafqat, S. (2013). EFFECT OF DIFFERENT SPROUTING CONDITIONS ON ALPHA AMYLASE ACTIVITY, FUNCTIONAL PROPERTIES OF WHEAT FLOUR AND ON SHELF-LIFE OF BREAD SUPPLEMENTED WITH SPROUTED WHEAT.

Singh, C., Jayas, D., Paliwal, J., & White, N. (2010). Identification of insect-damaged wheat kernels using short-wave near-infrared hyperspectral and digital colour imaging. *Computers and Electronics in Agriculture - COMPUT ELECTRON AGRIC*, *73*, 118-125. https://doi.org/10.1016/j.compag.2010.06.001

Smyth, S. e., Castle, D. e., & Phillips, P. W. B. e. (2014). *Handbook on agriculture, biotechnology and development*. Edward Elgar. http://murdoch.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwfV1LTwIxEJ4IXjAeUDACYvoDxLB979EQCQeO3snstjVeloTw_-N07bKLUY6TNm06aefxZeYrgOCvy8Uvm5AHx3kwIVZdFKWQFj2SJbRlGQPWumGuLVNt80YyIn_yZTcABvnvSD3Yg57MY-ZltrwDsGSU2Oua81HRLbRCmo6gVKLfaQbFuRwb8btfCiWPsx5CP3Yh3MGVr-7h5u3zkDgyPEkdDsER6A1WLobKbF8xbOe9sOJrfzwB54xmMddWCI2Brd8_VptF2nyXcJxdc1rNH-AWY_17daz75NwjMAy5Us4anxVWohE5ar-UpHGOWakLPYHJv-tNL4zNYEAhgvwBHZ7gOtCj8POTZp5rnX8Da0OG_w

Visa, S., Ramsay, B., Ralescu, A., & Knaap, E. (2011). *Confusion Matrix-based Feature Selection* (Vol. 710).

Walter, J., Edwards, J., McDonald, G., & Kuchel, H. (2017). *The application of precision phenotyping technologies to a wheat breeding program*. G. R. a. D. Corporation. https://grdc.com.au/resources-and-publications/grdc-update-papers/tab-content/grdc-update-papers/2017/02/the-application-of-precision-phenotyping-technologies-to-a-wheat-breeding-program

Weiss, M., Jacob, F., & Duveiller, G. (2020). Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment*, *236*, 111402. https://doi.org/https://doi.org/10.1016/j.rse.2019.111402

Wells, J., & Slade, P. (2021). The effect of the Canada–China canola trade dispute on canola prices. *Canadian journal of agricultural economics*, *69*(1), 141-149. https://doi.org/10.1111/cjag.12258

Weragoda, A., & Duver, A. (2021). *Snapshot of Australian Agriculture 2021*. Australia: ABARES:Australian Bureau of Agricultural and Resource Economics and Sciences: Department of Agriculture, Water and the Environment Retrieved from https://daff.ent.sirsidynix.net.au/client/en_AU/search/asset/1031521/0

Wu, J. (2017). Introduction to Convolutional Neural Networks.

Zheng, X., Li, Y., Wei, W., & Peng, Y. (2019). Detection of adulteration with duck meat in minced lamb meat by using visible near-infrared hyperspectral imaging. *Meat Science*, *149*, 55-62. https://doi.org/https://doi.org/10.1016/j.meatsci.2018.11.005

Zhu, W., Zeng, N., & Wang, N. (2010). Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS ® Implementations. *NorthEast SAS users group, health care and life sciences*.

Appendix A

HSI CNN SCRIPT

Original file is located at
https://colab.research.google.com/drive/1r759Yj1TNgeuRcas3j0PNrSRs6-BTNlw

Based off Francois Chollet's Image Classifier script (Chollet, 2016)

Set up model and parameters

```
from google.colab import drive drive.mount('/content/drive')
```

Import all necessary libraries

```
from keras.preprocessing.image import ImageDataGenerator
from keras.models import Sequential
from keras.layers import Conv2D, MaxPooling2D
from keras.layers import Activation, Dropout, Flatten, Dense
from keras import backend as K
import os

img_width, img_height = 224, 224

train_data_dir = '/content/drive/MyDrive/combinednumpy/train'
validation_data_dir = '/content/drive/MyDrive/combinednumpy/valid'
test_data_dir = '/content/drive/MyDrive/combinednumpy/test'
nb_train_samples = len([i for i in os.listdir(train_data_dir+'/GMO') if 'npy' in i]) +
len([i for i in os.listdir(train_data_dir+'/nonGMO') if 'npy' in i])
nb_validation_samples = len([i for i in os.listdir(validation_data_dir+'/GMO') if 'npy'
in i]) + len([i for i in os.listdir(validation_data_dir+'/nonGMO') if 'npy' in i])
nb_test_samples = len([i for i in os.listdir(test_data_dir+'/GMO') if 'npy' in i]) + len([i
for i in os.listdir(test_data_dir+'/nonGMO') if 'npy' in i])

epochs = 100
batch_size = 60

valid_x = []
print('Train Samples:',nb_train_samples)
print('Validation Samples:',nb_validation_samples)
print('Test Samples:',nb_test_samples)

input_shape = (img_width, img_height, 80)
```

Convolution Neural Network Model

```
model = Sequential()
model.add(Conv2D(32, (3, 3), input_shape=input_shape, dtype='float16'))
```

```
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Conv2D(32, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Conv2D(64, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Flatten())
model.add(Dense(64))
model.add(Activation('relu'))
model.add(Dropout(0.5))
model.add(Dense(1))
model.add(Activation('sigmoid'))

import os
import tensorflow
try:
  shutil.rmtree('checkpoints')
except: pass
try:
  os.mkdir('checkpoints')
except: pass
model_checkpoint_callback = tensorflow.keras.callbacks.ModelCheckpoint(
    filepath='./checkpoints/model(Nilesh).h5',
    save_weights_only=False,
    monitor='val_accuracy',
    verbose=1,
    save_best_only=False)
csv_logger = tensorflow.keras.callbacks.CSVLogger('training.log')

!pip install pympler
```

Custom hyperspectral image generator

Based off script written by Nilesh (2018).

```
import time
import gc
from skimage.transform import AffineTransform, SimilarityTransform, warp
from numpy import deg2rad, flipud, fliplr
from numpy.random import uniform, random_integers
from random import choice
import numpy as np
def hyper_generator(files, batchSize, augment = False, rotation_range = 0,
scale_range = 0, transform_range=0, hori_flip = False, verti_flip = False, shear_range
= 0):
  epochIndices = []
```

```
  while True:
    if len(epochIndices) == 0:
      epochIndices = list(range(len(files)))
      #print('Epoch reset!!!!')
    batchIndices = np.random.choice(a = epochIndices, size =
min(batchSize,len(epochIndices)), replace=False)
    for i in batchIndices: epochIndices.remove(i)
    batchX = []
    for inputI in batchIndices:
      img = (np.load(files[inputI], mmap_mode='r')/255).astype(np.float16)
```

Augmentation (not run for final results)

Based off script published on Git Hub (*Hyperspectral Image Generator*, 2020)

```
      if augment:
        if hori_flip and choice([True, False]): img = flipud(img)
        if verti_flip and choice([True, False]): img = fliplr(img)
        rotation_angle = uniform(low=-abs(rotation_range), high=abs(rotation_range))
        shear_angle = uniform(low=-abs(shear_range), high=abs(shear_range))
        scale_value = uniform(low=abs(1 / scale_range), high=abs(scale_range))
        translation_values = (random_integers(-abs(transform_range),
abs(transform_range)),
                              random_integers(-abs(transform_range), abs(transform_range)))
        transform_toorigin = SimilarityTransform(scale=(1, 1), rotation=0, translation=(-
img.shape[0], -img.shape[1]))
        transform_revert = SimilarityTransform(scale=(1, 1), rotation=0,
translation=(img.shape[0], img.shape[1]))
        transform = AffineTransform(scale=(scale_value, scale_value),
rotation=deg2rad(rotation_angle),
                                    shear=deg2rad(shear_angle), translation=translation_values)
        img = warp(img, ((transform_toorigin) + transform) + transform_revert,
mode='edge', preserve_range=True)

      batchX.append(img)
    batchY = [int(not ('non' in files[inputI].lower())) for inputI in batchIndices]

    batchX = np.array(batchX)
    batchY = np.array(batchY)

    yield(batchX, batchY)

import os
import numpy as np
import random
import gc
from pympler.asizeof import asizeof

def loadGMOData(directory, printOutput=True, maxSize = None):
        outData = []
        gmPath = directory+'/GMO/'
```

```
        noGmPath = directory+'/nonGMO/'
        imageCount = len([a for a in os.listdir(gmPath)+os.listdir(noGmPath) if 'npy'
in a])
        if printOutput: print("Getting",imageCount,"images from",directory)
        for gm in os.listdir(gmPath):
                if 'npy' in gm:
                        if printOutput: print("\tLoading:",gm)
                        loaded = (np.load(gmPath+gm)/255).astype(np.float16)
                        outData.append((loaded,1))
                        if printOutput: print("\t",asizeof(outData)/1024/1024,"MB")
                        if maxSize != None:
                                if asizeof(outData)/1024/1024 > maxSize / 2: break
        for nogm in os.listdir(noGmPath):
                if 'npy' in nogm:
                        if printOutput: print("\tLoading:",nogm)
                        loaded = (np.load(noGmPath+nogm)/255).astype(np.float16)
                        outData.append((loaded,0))
                if printOutput: print('\t',asizeof(outData)/1024/1024,"MB")
                if maxSize != None:
                                if asizeof(outData)/1024/1024 > maxSize: break
        random.shuffle(outData)
        x = []
        y = []
        for i in outData:
                x.append(i[0])
                y.append(i[1])
        return np.array(x), np.array(y)
valid_x = []
```

Training the Model

```
model.compile(loss='binary_crossentropy',
                        optimizer='rmsprop',
                        metrics=['accuracy'])

if len(valid_x) == 0: valid_x,valid_y = loadGMOData(validation_data_dir)
print("Validation GMO number:",len([a for a in valid_y if a == 1]))
print("Validation Non-GMO number:",len([a for a in valid_y if a == 0]))


trainFiles = [train_data_dir+'/GMO/'+f for f in os.listdir(train_data_dir+'/GMO/') if
'npy' in f ] + [train_data_dir+'/nonGMO/'+f for f in
os.listdir(train_data_dir+'/nonGMO/') if 'npy' in f ]
validFiles = [validation_data_dir+'/GMO/'+f for f in
os.listdir(validation_data_dir+'/GMO/') if 'npy' in f ] +
[validation_data_dir+'/nonGMO/'+f for f in
os.listdir(validation_data_dir+'/nonGMO/') if 'npy' in f ]
trainGen = hyper_generator(trainFiles, batch_size, False, 0, 0.2, 0, True, False)
#trainGen = hyper_generator(trainFiles, batch_size)
validGen = hyper_generator(validFiles, batch_size)
```

```
model.fit(x=trainGen,
        batch_size = batch_size,
        steps_per_epoch=nb_train_samples / batch_size,
                                        validation_steps=nb_validation_samples /
batch_size,
                                        validation_data=validGen,
                                        epochs=epochs,
                                        callbacks=[model_checkpoint_callback,
csv_logger])

!cp training.log "/content/drive/MyDrive/combinednumpy/training.log"
```

Training Graphs

Loss graph

```
from matplotlib.pyplot import *

x = []
y = []
with open("/content/drive/MyDrive/combinednumpy/training.log") as trainFile:
  trainFile.readline()
  for line in trainFile:
    lineParts = line.strip().split(',')
    x.append(float(lineParts[0]))
    y.append(float(lineParts[2]))

plot(x,y, label = 'Loss')
xlabel('Epoch')
ylabel('Loss')
xlim(0,epochs)
ylim(0,10)
show()
```

Accuracy (Training set) graph

```
x = []
y = []
with open("/content/drive/MyDrive/combinednumpy/training.log") as trainFile:
  trainFile.readline()
  for line in trainFile:
    lineParts = line.strip().split(',')
    x.append(float(lineParts[0]))
    y.append(float(lineParts[1]))

plot(x,y, label = 'Loss')
xlabel('Epoch')
ylabel('Accuracy (Training set)')
xlim(0,100)
ylim(0,1)
show()
```

Accuracy (Validation set) graph

```
x = []
y = []
with open("/content/drive/MyDrive/combinednumpy/training.log") as trainFile:
  trainFile.readline()
  for line in trainFile:
    lineParts = line.strip().split(',')
    x.append(float(lineParts[0]))
    y.append(float(lineParts[3]))

plot(x,y, label = 'Loss')
xlabel('Epoch')
ylabel('Accuracy (Validation set)')
xlim(0,100)
ylim(0,1)
show()
```

Save Model

```
from shutil import copyfile
BEST_MODEL = "95" (based on the best model produced in the 100 epochs)
copyfile("checkpoints/model"+BEST_MODEL+".h5",
"/content/drive/MyDrive/combinednumpy/Best_Model.h5")
```

Metrics Function

```
from keras.models import load_model
import numpy as np

def calcMetrics(eval_x, eval_y, conf_thresh, printOutput):
  fp = 0
  tp = 0
  fn = 0
  tn = 0

  i = 0
  for x in eval_x:
    gmoLabels = model.predict(np.array([x]))
    if eval_y[i] == 1:
      if gmoLabels[0][0] >= conf_thresh: tp += 1
      else: fn += 1
    elif eval_y[i] == 0:
      if gmoLabels[0][0] < conf_thresh: tn += 1
      else: fp += 1
    i+=1

  sensitivity = tp/(tp+fn)
  specificity = tn/(tn+fp)
  precision = tp/(tp+fp)
  accuracy = (tp+tn)/(tp+tn+fp+fn)
```

```
  if printOutput:
    print("True Positives:", tp)
    print("True Negatives:", tn)
    print("False Positives:", fp)
    print("False Negatives:", fn)
    print("Sensitivity/Recall:",sensitivity)
    print("Specificity:",specificity)
    print("Precision:", precision)
    print("Accuracy:", accuracy)

  return sensitivity, specificity, precision, accuracy
```

Check the model on the validation set

```
valid_dir = validation_data_dir
model = load_model("/content/drive/MyDrive/combinednumpy/Best_Model.h5")
steps = 100

sen_rec = []
spec = []
prec = []
eval_x, eval_y = loadGMOData(valid_dir, False)
for i in range(0,steps+1):
  sen, sp, pr, acc = calcMetrics(eval_x, eval_y, i/steps, False)
  sen_rec.append(sen)
  spec.append(1-sp) #Is actually 1-specificity
  prec.append(pr)
  print("Confidence:",i/steps,"| Metrics:",sen,sp,pr,acc)

from matplotlib.pyplot import *
from sklearn.metrics import *
```

Plot ROC and calculate validation AUC

```
plot([1]+spec+[0], [1]+sen_rec+[0], label = 'Canola')
xlabel('1-Specificity')
ylabel('Sensitivity')
xlim(0,1)
ylim(0,1)
show()

print("Validation AUC",auc(spec, sen_rec))
```

Precision vs Recall graph

```
from matplotlib import *

plot([1]+sen_rec+[0], [0]+prec+[1], label = 'Canola')
```

```
xlabel('Recall')
ylabel('Precision')
xlim(0,1)
ylim(0,1)
show()

print("Validation AUC:",auc([1]+sen_rec+[0], [0]+prec+[1]))
```

Check the model on the test set

```
conf_thresh = 0.45 (based on validation ROC)

test_dir = test_data_dir
model = load_model("/content/drive/MyDrive/combinednumpy/Best_Model.h5")

test_x, test_y = loadGMOData(test_dir, True)
sen, spec, prec, acc = calcMetrics(test_x, test_y, conf_thresh, True)

from matplotlib.pyplot import *
from sklearn.metrics import *

sen_rec = []
spec = []
prec = []
eval_x, eval_y = loadGMOData(test_dir, True)
for i in range(0,steps+1):
  sen, sp, pr, acc = calcMetrics(eval_x, eval_y, i/steps, False)
  sen_rec.append(sen)
  spec.append(1-sp) #Is actually 1-specificity
  prec.append(pr)
  print("Confidence:",i/steps,"| Metrics:",sen,sp,pr,acc)
```

Test ROC

```
plot([1]+spec+[0], [1]+sen_rec+[0], label = 'Canola')
xlabel('1-Specificity')
ylabel('Sensitivity')
xlim(0,1)
ylim(0,1)
show()
```

Test AUC

```
print("Validation AUC",auc(spec, sen_rec))
```

BMP Image Classifier Script

Original file is located at
https://colab.research.google.com/drive/1ernRNCveup_VrB2S36KBRbCOb0H6LXXn

Based off Francois Chollet's Image Classifier script (Chollet, 2016)

Set up model and parameters

```
from google.colab import drive drive.mount('/content/drive')
```

Import all necessary libraries

```
from keras.preprocessing.image import ImageDataGenerator
from keras.models import Sequential
from keras.layers import Conv2D, MaxPooling2D
from keras.layers import Activation, Dropout, Flatten, Dense
from keras import backend as K

img_width, img_height = 128, 128 #changed

train_data_dir = 'drive/MyDrive/bmpall/trainbmp'
validation_data_dir = 'drive/MyDrive/bmpall/validbmp'
test_data_dir =  'drive/MyDrive/bmpall/testbmp'
nb_train_samples = 722
nb_validation_samples = 199
nb_test_samples = 94
epochs = 100
batch_size = 100

if K.image_data_format() == 'channels_first':
        input_shape = (3, img_width, img_height) #Should be 80 bands in
hyperpectral
else:
        input_shape = (img_width, img_height, 3)

import os
import tensorflow
try:
  os.mkdir('checkpoints')
except: pass
model_checkpoint_callback = tensorflow.keras.callbacks.ModelCheckpoint(
   filepath='./checkpoints/model{epoch:02d}.h5',
```

```
    save_weights_only=False,
    monitor='val_accuracy',
    verbose=1,
    save_best_only=False)
csv_logger = tensorflow.keras.callbacks.CSVLogger('training.log')
```

Convolution Neural Network Model

```
model = Sequential()
model.add(Conv2D(32, (3, 3), input_shape=input_shape))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Conv2D(32, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Conv2D(64, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Flatten())
model.add(Dense(64))
model.add(Activation('relu'))
model.add(Dropout(0.5))
model.add(Dense(1))
model.add(Activation('sigmoid'))

model.compile(loss='binary_crossentropy',
                    optimizer='rmsprop',
                    metrics=['accuracy'])
```

Augmentation

```
train_datagen = ImageDataGenerator(
                shear_range=0.2,
                zoom_range=0.2,
                horizontal_flip=True,
        rescale=1. / 255)

test_datagen = ImageDataGenerator(rescale=1. / 255)

train_generator = train_datagen.flow_from_directory(
        train_data_dir,
        target_size=(img_width, img_height),
        batch_size=batch_size,
        class_mode='binary')
```

```
validation_generator = test_datagen.flow_from_directory(
        validation_data_dir,
        target_size=(img_width, img_height),
        batch_size=batch_size,
        class_mode='binary')

model.fit_generator(
        train_generator,
        steps_per_epoch=nb_train_samples // batch_size,
        epochs=epochs,
        validation_data=validation_generator,
        callbacks=[model_checkpoint_callback, csv_logger])
!cp training.log "/content/drive/MyDrive/bmpall/training.log"

from shutil import copyfile
BEST_MODEL = "98" (best iteration number)
copyfile("checkpoints/model"+BEST_MODEL+".h5",
'/content/drive/MyDrive/bmpall/model.h5')
```

Training Graphs

Loss graph

```
from matplotlib.pyplot import *

x = []
y = []
with open("/content/drive/MyDrive/bmpall/training.log") as trainFile:
  trainFile.readline()
  for line in trainFile:
    lineParts = line.strip().split(',')
    x.append(float(lineParts[0]))
    y.append(float(lineParts[2]))

plot(x,y, label = 'Loss')
xlabel('Epoch')
ylabel('Loss')
xlim(0,epochs)
ylim(0,1)
show()
```

Accuracy (Training set) graph

```
x = []
y = []
with open("/content/drive/MyDrive/bmpall/training.log") as trainFile:
  trainFile.readline()
  for line in trainFile:
    lineParts = line.strip().split(',')
```

```
    x.append(float(lineParts[0]))
    y.append(float(lineParts[1]))

plot(x,y, label = 'Loss')
xlabel('Epoch')
ylabel('Accuracy (Training set)')
xlim(0,100)
ylim(0,1)
show()
```

Accuracy (Validation set) graph

```
x = []
y = []
with open("/content/drive/MyDrive/bmpall/training.log") as trainFile:
  trainFile.readline()
  for line in trainFile:
    lineParts = line.strip().split(',')
    x.append(float(lineParts[0]))
    y.append(float(lineParts[3]))

plot(x,y, label = 'Loss')
xlabel('Epoch')
ylabel('Accuracy (Validation set)')
xlim(0,100)
ylim(0,1)
show()
```

Metrics Function

```
from keras.models import load_model
import numpy as np

def calcMetrics(eval_x, eval_y, conf_thresh, printOutput):
  fp = 0
  tp = 0
  fn = 0
  tn = 0

  if calcMetrics.lastModel != model or conf_thresh == 0:
    print("Performing predictions...")
    calcMetrics.labelConfs = []
    for x in eval_x:
      gmoLabels = model.predict(np.array([x]))
      calcMetrics.labelConfs.append(gmoLabels)
    calcMetrics.lastModel = model

  i = 0
```

```
  for gmoLabel in calcMetrics.labelConfs:
    if eval_y[i] == 1:
      if calcMetrics.labelConfs[i][0][0] >= conf_thresh: tp += 1
      else: fn += 1
    elif eval_y[i] == 0:
      if calcMetrics.labelConfs[i][0][0] < conf_thresh: tn += 1
      else: fp += 1
    i+=1

  sensitivity = tp/(tp+fn)
  specificity = tn/(tn+fp)
  precision = tp/(tp+fp)
  accuracy = (tp+tn)/(tp+tn+fp+fn)

  if printOutput:
    print("True Positives:", tp)
    print("True Negatives:", tn)
    print("False Positives:", fp)
    print("False Negatives:", fn)
    print("Sensitivity/Recall:",sensitivity)
    print("Specificity:",specificity)
    print("Precision:", precision)
    print("Accuracy:", accuracy)

  return sensitivity, specificity, precision, accuracy

calcMetrics.lastModel = None
```

Check the model on the validation set

```
valid_dir = validation_data_dir
model = load_model("/content/drive/MyDrive/bmpall/model.h5")
steps = 100


import os
import numpy as np
import random
import gc
import cv2

def loadGMOData(directory, printOutput=True, maxSize = None):
        outData = []
        gmPath = directory+'/GMO/'
        noGmPath = directory+'/nonGMO/'
        imageCount = len([a for a in os.listdir(gmPath)+os.listdir(noGmPath) if 'bmp'
in a])
```

```
            if printOutput: print("Getting",imageCount,"images from",directory)
            for gm in os.listdir(gmPath):
                if 'bmp' in gm:
                    if printOutput: print("\tLoading:",gm)
                    loaded = cv2.imread(gmPath+gm)/255
                    outData.append((loaded,0))
            for nogm in os.listdir(noGmPath):
                if 'bmp' in nogm:
                    if printOutput: print("\tLoading:",nogm)
                    loaded = cv2.imread(noGmPath+nogm)/255
                    outData.append((loaded,1))
            random.shuffle(outData) #Make sure to randomise, otherwise training is
        lopsided
            x = []
            y = []
            for i in outData:
                x.append(i[0])
                y.append(i[1])
            return np.array(x), np.array(y)

        sen_rec = []
        spec = []
        prec = []
        eval_x, eval_y = loadGMOData(valid_dir, True)
        for i in range(0,steps+1):
          sen, sp, pr, acc = calcMetrics(eval_x, eval_y, i/steps, False)
          sen_rec.append(sen)
          spec.append(1-sp) #Is actually 1-specificity
          prec.append(pr)
          print("Confidence:",i/steps,"| Metrics:",sen,sp,pr,acc)

        from matplotlib.pyplot import *
        from sklearn.metrics import *

Plot ROC
        plot([1]+spec+[0], [1]+sen_rec+[0], label = 'Canola')
        xlabel('1-Specificity')
        ylabel('Sensitivity')
        xlim(0,1)
        ylim(0,1)
        show()

Validation AUC
        print("Validation AUC",auc(spec, sen_rec))

        from matplotlib import *
        plot([1]+sen_rec+[0], [0]+prec+[1], label = 'Canola')
        xlabel('Recall')
```

```
ylabel('Precision')
xlim(0,1)
ylim(0,1)
show()
```

Test AUC

```
print("Validation AUC:",auc([1]+sen_rec+[0], [0]+prec+[1]))

test_dir = test_data_dir
model = load_model("/content/drive/MyDrive/bmpall/model.h5")
conf_thresh = 0.5 #CHANGE THIS BASED ON VALIDATION ROC

test_x, test_y = loadGMOData(test_dir, True)
sen, spec, prec, acc = calcMetrics(test_x, test_y, conf_thresh, True)

test_dir = test_data_dir
model = load_model("/content/drive/MyDrive/bmpall/model.h5")
steps = 100

import os
import numpy as np
import random
import gc
import cv2

def loadGMOData(directory, printOutput=True, maxSize = None):
    outData = []
    gmPath = directory+'/GMO/'
    noGmPath = directory+'/nonGMO/'
    imageCount = len([a for a in os.listdir(gmPath)+os.listdir(noGmPath) if 'bmp'
in a])
    if printOutput: print("Getting",imageCount,"images from",directory)
    for gm in os.listdir(gmPath):
        if 'bmp' in gm:
            if printOutput: print("\tLoading:",gm)
            loaded = cv2.imread(gmPath+gm)/255
            outData.append((loaded,0))
    for nogm in os.listdir(noGmPath):
        if 'bmp' in nogm:
            if printOutput: print("\tLoading:",nogm)
            loaded = cv2.imread(noGmPath+nogm)/255
            outData.append((loaded,1))
    random.shuffle(outData) #Make sure to randomise, otherwise training is
lopsided
    x = []
    y = []
    for i in outData:
```

```
                    x.append(i[0])
                    y.append(i[1])
            return np.array(x), np.array(y)

    from matplotlib.pyplot import *
    from sklearn.metrics import *

    sen_rec = []
    spec = []
    prec = []
    eval_x, eval_y = loadGMOData(test_dir, True)
    for i in range(0,steps+1):
      sen, sp, pr, acc = calcMetrics(eval_x, eval_y, i/steps, False)
      sen_rec.append(sen)
      spec.append(1-sp) #Is actually 1-specificity
      prec.append(pr)
      print("Confidence:",i/steps,"| Metrics:",sen,sp,pr,acc)
```

Plot ROC

```
    plot([1]+spec+[0], [1]+sen_rec+[0], label = 'Canola')
    xlabel('1-Specificity')
    ylabel('Sensitivity')
    xlim(0,1)
    ylim(0,1)
    show()
```

Calculate test AUC

```
    print("Test AUC",auc(spec, sen_rec))
```

Appendix B

Table B1: Single factor ANOVA test for GM and non-GM wavelength reflectance at 713nm illustrating a p-value below 0.05.

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Non GMO	505	142503	282.1841584	6610.99578
	507	149047	293.9783037	10192.03313

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	35192.63417	1	35192.63417	4.187077073	0.04099178	3.850682268
Within Groups	8489110.635	1010	8405.060034			
Total	8524303.269	1011				

Table B2: Single factor ANOVA test for GM and non-GM diameter illustrating a p-value below 0.01.

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
GMO D	507	115877	228.5542406	438.4767952
NON D	505	110667	219.1425743	361.8169338

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	22410.51685	1	22410.51685	55.99510768	1.57748E-13	3.850682268
Within Groups	404224.993	1010	400.2227654			
Total	426635.5099	1011				

Table B3: Single factor ANOVA test for GM and non-GM diameter illustrating a p-value below 0.01.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
GMO A	507	17629700	34772.58383	46276697.46
NON A	505	15911100	31507.12871	30422330.03

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2697778329	1	2697778329	70.31834959	1.68E-16	3.850682268
Within Groups	38748863252	1010	38365211.14			
Total	41446641581	1011				