Development of Image Analysis Tools to Quantify Potato Tuber Quality Traits

A THESIS
SUBMITTED TO THE FACULTY OF
THE GRADUATE SCHOOL OF THE UNIVERSITY OF MINNESOTA
BY

Michael Donald Miller

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Dr. Laura Shannon, Advisor

August 2022

Acknowledgements

There are many people I'd like to thank for their contribution to the work covered in this thesis:
Dr. Laura Shannon for taking me on as a graduate student, providing countless hours of excellent guidance, as well as teaching me the joys of polyploid genetics and linear algebra.
Dr. Candice Hirsch and Dr. Aaron Lorenz for serving on my committee and offering valuable input and advice during the development of this thesis.
Dr. Darrin Haagenson for creating the grant for this project, providing needed equipment, and esteemed insight. I also thank the other members of the USDA-ARS Sugarbeet and Potato Research Unit for collecting images of pressure bruises.
Dr. Susie Thompson and Dr. Max Feldman for providing images and data from other breeding programs as well as crucial suggestions about analyses to include in this work.
Dr. Cari Schmitz Carley for creating the original scripts for TubAR functions and coming up with an excellent R package name.
Dr. Maria Caraza-Harter and the Endelman Lab at the University of Wisconsin-Madison for providing the imaging protocol and the image analysis work which led to the creation of TubAR.
Rachel Figueroa, Sophia Fitzcollins, Jessi Huege, Adynn Stedillie, Ian Spencer, and others for taking many pictures of potatoes.
Keith Mann, Ron Faber, and the past and present members of the Shannon lab at the University of Minnesota including Husain Agha, Katelyn Filbrandt, Sophia Fitzcollins, Jessi Huege, Colin Jones, John Larsen, Thomas McGehee, Dr. Xiaoxi Meng, Laura Schulz, Dr. Thomas Stefaniak, Heather Tuttle, and Muyideen Yusuf for their work in designing, planting, maintaining, harvesting, and grading the field experiments I have relied on for image data.
Tuberosum Technologies for funding and help in the design of a skin finish function and for providing me an international internship opportunity despite exceptional logistical difficulties.
The United States Department of Agriculture and Minnesota Department of Agriculture for funding this research.

Dedication

I dedicate this thesis to the many mentors and teachers throughout my life including:

Wayne Coburn, who taught me long division and who had an acronym for all of life's questions.

Gail Petroff, who told me I can do great things, but I need to finish my spelling homework.

Jordan Pollack, whose lecture about Norman Borlaug had major impact on at least one student.

Patt Ligman, who taught me the writing skills needed to write a thesis, among other things.

Adam Gilbertsen, Jenn McCurtain, Eric Watkins, Garett Heineck, and Brian Steffenson, who helped me find my interest in research, as well as my research interests.

Finally, I dedicate this thesis to God and my family, who have done innumerable things for me.

Abstract

Potato is the most popular non-cereal food crop and a major staple crop. Despite the importance of potato, it has seen little yield improvement through breeding over the past century when compared to other crops. One difficulty in potato breeding is the large number of quality traits that must be accounted for in order to create marketable potato varieties. These quality traits are often measured using imprecise, subjective scales. This thesis covers my work in improving the tools available for use in measuring and breeding for potato tuber quality traits. In Chapter 1, I review the literature relevant to a selection of quality traits and their measurement. I discuss machine learning and its use in identifying more intricate tuber quality traits, as well as efforts to perform genomic selection in autotetraploid potato as a possible application for highly quantitative quality trait data. Chapter 2 covers the mechanics and capabilities of the potato tuber image analysis program, TubAR. I compare the quantitative measurements provided by TubAR to human visual scores for analogous traits. In Chapter 3, I discuss efforts to expand the scope of traits able to be measured with image analysis by employing machine learning image classification, using the pressure bruise and skin finish traits.

Table of contents

List of tables

List of figures

Chapter 1
Review of Literature

Summary

Potato is the most popular vegetable crop, grown as a staple crop in many regions of the world. Despite its essential role in global agriculture, potato has not attained the yield increases seen in other crops. One difficulty in breeding is the large number of quality traits that must be accounted for in order to create a marketable potato variety. Despite the importance of quality traits, they are often measured using imprecise, subjective scales. Image analysis seeks to provide more precision and accuracy in many applications including the measurement of potato tuber traits. Various image analysis designs have been applied to tuber traits using different camera types and platforms. These may be able to be further enhanced, especially for use in difficult to define traits, by using machine learning models for trait detection. This high accuracy, high precision phenotype data could improve breeding decisions on its own but may be able to be best utilized in genomic selection. Genomic selection of tetraploid potato is a growing area of research for which many different types of population and genotype data have been used to create selection models for a multitude of traits. Precise and accurate phenotype data may further improve the reliability of these models and allow for accelerated improvements through potato breeding.

Introduction

The cultivated potato (Solanum tuberosum L.) is the fourth most widely grown food crop and the most popular non-cereal food crop (FAO 2021). It is a staple crop that is adaptable to a variety of growing conditions (Campos and Ortiz 2020). Despite this, Potato has seen little to no gain in genetic yield potential since the 19th century. One reason for this lack of progress is the many quality traits that must be considered when breeding for new varieties (Douches et al. 1996).

Quality traits are factors that influence a potato's appeal to consumers and utility in processing. These include tuber skin color, flesh color, shape, size, skinning, skin finish, and eye depth (Carputo et al. 2004). Meeting quality trait requirements is an important component of potato breeding and makes introduction of new varieties more difficult. Very old varieties, such as the 108-year-old Russet Burbank, which conform with many quality expectations while not excelling in areas such as yield, continue to be cultivated (Douches et al. 1996). New varieties must also meet market requirements for quality traits for any meaningful genetic improvement in potatoes to occur. However, quality traits are often evaluated subjectively without the precision and accuracy needed for meaningful selection. Ways to improve the accuracy of quality trait measurements at a low cost are therefore a vital component of progress in potato breeding.

Skin Color

Potato tuber skin color is an important factor in consumer appeal and marketing (Jemison et al. 2008). Red or purple skin color is caused by the presence of anthocyanins in the tuber periderm (Liu et al. 2019, Roe et al. 2014). High anthocyanin concentrations result in intense colors that tend to appeal to consumers and may increase nutritional value by acting as an antioxidant (Han et al. 2006, Jemison et al. 2008, Roe et al. 2014). Anthocyanin concentration in potato can be influenced by a variety of factors including soil type, temperature, fertilization, growth regulators, harvest time, storage, and genetics (Bonar et al. 2018, Caraza□Harter and Endelman 2020, Liu et al. 2019, Jones et al. 2021, Roe et al. 2014, Stefaniak et al. 2021, Thornton et al. 2013).

Categorical tuber skin color (i.e. red, purple, white) is largely controlled by a limited set of loci, known as the D, R, and P loci, all of which are involved in

anthocyanin synthesis. The dominant D allele is necessary for any anthocyanin production while the R and P alleles control whether the anthocyanins have a pink, red, or purple color (Jung et al. 2009). Quantitative color intensity is a more complex trait. Zhang et al. (2009) found three QTL related to color intensity and color homogeneity, though these only account for a portion of the variation. Pigmentation can also vary across the skin of each tuber. Bonar et al. (2018) found micro-RNA associated with anthocyanin levels both between clones and in differently pigmented regions of the same tuber. These mechanisms operating upon different aspects of skin color demonstrate a trait complexity that must be accounted for in breeding.

Shape in potatoes

Tuber shape must meet different requirements depending on application. For russet processing potatoes, a long narrow tuber is favored for compatibility with commercial fry cutting equipment and creating a long French fry. Chipping potatoes are expected to be either spherical or cubic in order to create many large, uniform chips. Qualitative shape is often visually selected for very early in a potato breeding program, however quantitative aspects of shape are more difficult to differentiate (Van Eck et al. 1994).

Qualitative tuber shape is controlled by the Ro locus on chromosome 10, which is linked to the D (called I in diploid potatoes) locus that impacts skin color (De Jong and Burns 1993, Van Eck et al. 1994). However, this locus does not explain quantitative differences in length between similarly shaped tubers or other differences in shape that are observed (De Jong and Burns 1993). Several studies have found minor QTL for shape; however, these vary depending on the population studied, with no agreed upon QTL, indicating the trait is controlled by many loci (Lindqvist-Kreuze et al. 2015, Prashar et al. 2014, Śliwka et al. 2008).

Skin set in potatoes

Tuber skinning, or excoriation, is an undesirable trait in potato tubers which leads to shorter storage life, an entry point for disease, and lower consumer appeal (Boydston et al. 2018, Jemison et al. 2008, Lulai and Corsini 1998). Skinning occurs when the phellem

layer of the tuber periderm separates from the phellogen layer, also called the cork cambium (Lulai and Orr 1993). While skinning is the term commonly used to describe this phenomenon in industry, Lulai and Freeman (2001) contest that this term is misleading as it implies the entire periderm separates from the tuber and propose excoriation as a more precise term. Here I will use both terms synonymously, each referring to the separation of the phellem cells. The cause of this separation is the rupture of the phellogen cell walls that border the phellem layer. This rupturing is more likely to occur in young tubers where the cell walls are thinner and the phellogen cells are more active. Frequency of ruptures decrease as the tuber matures and the cell walls thicken and cell activity decreases (Lulai and Freeman 2001).

Given this physiological connection between tuber skinning and maturity, the primary method of skinning control is to ensure tubers mature before harvest. Full tuber maturity is generally accomplished by interrupting above ground plant growth in the practice of vine kill. Either physical means, such as flailing, or chemical desiccants are used to vine kill a few weeks before harvest. Vine kill stops the bulking stage and causes tubers to mature at their current size, avoiding premature harvest. Both Boydston et al. (2018) and Krupek et al. (2021) found various herbicides and physical vine kill methods to be mostly equivalent to each other in their ability to reduce skinning and avoid tuber desiccation in storage, as compared to no vine kill. However, genetics also contribute to skinning susceptibility (Caraza-Harter and Endelman 2022, Lulai and Orr 1993, Neubauer et al. 2013, Vulavala et al. 2019).

At the molecular level, several genes have been identified that influence the maturation rate and thickness of phellogen cell walls, suggesting skinning is a complex trait controlled by many genes (Neubauer et al. 2013, Vulavala et al. 2019). More distant genetic factors, such as the onset of vine maturity, also influence skin maturity (Caraza-Harter and Endelman 2022). Lulai and Orr (1993) tested four popular varieties of different market classes for skin set and found that they varied in both skin strength at maturity and rate of skin maturation after harvest.

Skin Finish

Skin finish, also called skin quality or skin texture, is a tuber quality trait that lacks clear definition. Generally, skin finish is the presence and degree of skin roughness, russeting, cracking, and netting (Clulow et al. 1995, McCord et al. 2011, Siano et al. 2018). Other skin characteristics are sometimes included, and aspects are emphasized differently across studies. While the exact components of this trait are not universally agreed upon, skin quality, as interpreted by poll respondents, has been found to be one of the most important traits that influence potato consumers (Jemison et al. 2008). Therefore, its importance can't be ignored due to imprecise definition and perhaps demands better attempts at definition. Many of the studies below use differing definitions, while some focus on specific factors such as skin roughness and russeting.

Skin finish has many potential influences. Genetics play a large role as suggested in the widespread use of "russet" as a description for varieties (Johansen et al 1988, Novy et al. 2008, Pavek et al. 1992). However, potatoes often exhibit skin finishes considered uncharacteristic of the variety, indicating environmental influences as well. In varieties that are characteristically smooth skinned, rough skin and russeting can result under conditions including hot temperatures, drought, and low soil potassium or calcium (Ginzberg et al. 2009, Ginzberg et al. 2012, Keren-Keiserman et al. 2019, Koch et al. 2019, Raimo et al. 2018).

Limited studies on the underlying genes responsible for skin finish are available. De Jong (1981) proposed that russeting in diploid potato is controlled by three loci based on segregation ratios in crosses between russeted and smooth skinned parents. Ginzberg et al. (2009) found differentially expressed genes in tuber skin at high temperatures were mostly related to heat-shock and the formation of protective membranes. The protective membranes in this case likely form the russeting seen in heat stressed and drought situations. Paradoxically, this group also found that russeted skin surface can lead to increased water loss (Ginzberg et al. 2012). McCord et al. (2011) performed a QTL mapping study of many traits including skin finish, which was associated with more genetic regions than any other trait studied besides yield, though one QTL on chromosome 3 explained 20% of the variation. They also discovered a QTL for skin netting but the phenotype was only expressed in one year of the study. Another study by

da Silva Pereira et al. (2021) found QTL on chromosomes 4 and 9, with no detection of the chromosome 3 QTL.

Studies of skin finish in the breeding context have found it to be relatively easy trait to select for. Russeting has been estimated in one population to have a heritability of 0.8 and is recommended to be selected for earlier in a breeding program than most traits (Love et al. 1997). Russeting has also been found to be among the easier traits to breed for in a tetraploid true seed population (Clulow et al. 1995). However, studies on breeding for or against other skin finishes besides russeting seem to be unavailable.

Bruise in potatoes

Physical impact during tuber harvest, grading, and storage can lead to multiple tuber disorders including scuffing, splitting, internal fissures, and bruising. Bruising, also called blackspot, is a darkened area of the tuber flesh due to melanin formed because of damage to tuber cell membranes (Baritelle et al. 2000, Dean et al. 1993). It causes potatoes and derivative products to be less aesthetically appealing to consumers and is associated with a "lumpy" texture when cooked (Kaack et al. 2002). Mechanical damage and bruising has been estimated to cost the US potato industry the equivalent of over $460 million a year (Brooke 1996).

Some studies of the underlying genetics of bruise in the context of mechanical damage have been conducted. Some research has found a correlation between tyrosine, a precursor to melanin, and bruising (Dean et al. 1993). However others have found bruising only weakly correlated with tyrosine, and attribute the majority of bruise resistance to intracellular compartmentation (Lærke et al. 2002). Urbany et al. (2011) conducted an association study to locate significant genes affecting bruising and enzymatic discoloration. They found a correlation between specific gravity and bruising, which is supported elsewhere (Baritelle and Hyde 2003), as well as 33 candidate genes which may influence bruising. In a follow up study, they found several differentially expressed protein families in bruised tissue, many of which have also been observed in genotypes subject to darker frying colors (Urbany et al. 2012).

Pressure Bruise

The studies described above addressed bruising from impact while the tuber is being harvested or processed. Another factor that can result in tuber bruising is the pressure exerted by the weight of other tubers over long periods of storage, called pressure bruise. This often occurs near the bottom of stacks of potatoes over three meters high (Schaper and Yaeger 1982, Lulai et al. 1996). In this case, bruising is caused by the crushing of phelloderm cells, leading to moisture loss and subsequently leading to the oxidation of tyrosine. Unlike some situations where blackspot symptoms lessen over time, pressure bruise is likely to worsen once tubers are removed from storage. Increased bruising is due to higher oxygen levels causing tyrosine oxidation after tubers are removed from the low oxygen environment at the bottom of storage crates, where pressure bruising often occurs (Lulai et al. 1996). Because moisture loss is a major contributing factor to pressure bruise, efforts to preserve moisture content in storage have been proposed as a control method (Shetty et al. 1991, Kelderman 2017). There is also some evidence that slowing changes in storage temperature can lessen pressure bruise (Muthukumarappan et al. 1994).

Studies on the genetics of pressure bruise are even more limited than studies on the genetics of bruises due to mechanical damage, however not entirely absent. Effectiveness of moisture control in preventing pressure bruise varies between russet clones (Castleberry and Jayanty 2012). Pressure bruise also correlates with the force needed to deform the tuber surface, as measured with a penetrometer (Castleberry and Jayanty 2017). While these studies give evidence to a genetic component of pressure bruise, the causative genes or loci are still unknown.

Common Scab

Common scab of potato, caused primarily by the bacteria Streptomyces scabeii along with other species, is a disease which results in lesions and pits along the surface of potato tubers (Agrios 2005). The identity of the causal pathogen is complex in that it is a set of species and genotypes in the primarily saprophytic Streptomyces genus. Taxonomic classification fails to perfectly describe which bacteria produce the disease, as this is determined by a block of genes, called the pathogenicity island, which allows

infection of living potato tubers (Wanner 2006). The ability of the pathogen to live saprophytically in the soil allows it to remain in a field indefinitely once introduced (Agrios 2005).

The complexity of the genetics of the common scab pathogen is also seen in the resistance exhibited by potatoes. Genetic variance in resistance to common scab can be observed across potato genotypes (Braun et al. 2017, Dees and Wanner 2012, Driscoll et al. 2009). However, this appears to be highly quantitative resistance without the gene-for-gene action that is observed in many other plant diseases (Flinn et al. 2005). Hosaka et al. (2000) examined 1800 potato wild relative accessions for resistance and while they classified some as resistant; no genotype showed total resistance across all trials. Bradshaw et al. (2008) found two QTL for scab resistance on chromosomes 2 and 6, the latter of which was also related to fry color. These traits were in repulsion, with good fry color being associated with scab susceptibility and vice versa. They do not discuss whether this is due to two linked genes or pleiotropy. Braun (2013) found a QTL for scab resistance on chromosome 11. None of these QTL had effect sizes above 15%, and detected QTL vary across studies, suggesting much of the variation is in small effect loci.

Another factor that complicates breeding for common scab resistance, is large environmental and genotype-by-environment effects. Wilson (2001) observed strong environmental effects for scab in russet potatoes, though genetic effect was still larger. Haynes et al. (2010) observed large genotype-by-environment effects in each field year studied. The environmental and genotype-by-environment effects could be due to the pathogen's sensitivity to soil conditions, such as pH, and the diversity of strains seen in different geographical areas (Agrios 2005, Wanner 2006).

Chip and Fry color

Processed potatoes, such as potato chips and French fries, account for 10% of the global potato consumption and over a third of potatoes consumed in North America and Europe (Keijbets 2008). Marketing processed products brings in additional requirements for the color of the flesh after frying. At high temperatures, such as those experienced when deep frying, the glucose and amino acids in the potato undergo Maillard reactions which are responsible for the distinct flavor and color of fried potatoes. However, this can

result in excessive darkening of chips and fries. The Maillard reaction of glucose and asparagine also leads to the formation of acrylamide, which has been studied as a potential carcinogen and correlates with darker fry and chip colors (Bethke and Bussan 2013, Pedreschi et al. 2006a). While the exact color of chip or fry favored by consumers varies by culture and personal preference, in the United States light colored chips and fries are considered ideal, so breeders and processors tend to aim for very limited browning when potatoes are fried. The main factors in the potato that encourage fry browning are high glucose levels and low moisture content (Amrein et al. 2006). These factors can be influenced by potato storage conditions, as well as genetics (De Wilde et al. 2005, Kaur et al. 2008).

The genetics of fry and chip color are primarily studied in the context of low temperature storage, which typically leads to the phenomenon of cold sweetening. Cold sweetening is the increased conversion of starch into simple sugars seen in tubers stored at temperatures below approximately 8°C. This is not easily avoided as storage temperatures above 4°C require the use of chemical sprays to avoid sprouting and pathogen growth (Byrne et al 2020, Sowokinos 2001). Therefore, there is strong incentive to find genetic factors that affect the severity of cold sweetening. Lynch et al (2003) proposed a three-locus model for inheritance of cold sweetening susceptibility. A later GWAS study by Byrne et al (2020) does not seem to support this model, finding only one major locus on chromosome 10. In addition to storage temperature, storage time is a major factor in glucose formation. Another GWAS of chip color after six months of storage at 8°C and found major QTL at chromosomes 2 and 9 (Rak et al. 2017). The QTL on chromosome 9 could be related to the apoplastic invertase protein, which is involved in the conversion of starch to sugar (Zhu et al. 2014). Surprisingly, neither of these studies detected the invertase loci on chromosomes 3 and 10, or the UGPase locus on chromosome 11, for which the biochemical pathway that contributes to cold sweetening is well described (Chen et al. 2001, Sowokinos 2001). Innate® potatoes minimize cold sweetening by RNAi suppression of the vacuolar invertase gene on chromosome 3 (VInv), however this is achieved through genetic modification rather than utilizing existing genetic variation in potato (Richael 2021, Zhu et al. 2014). Given issues of

market receptivity to genetic modification, further work on creation of cold sweetening resistant potatoes through genetic recombination will be needed.

Human ratings and image analysis

Quality traits are often quantified using human rated visual scales. While these have many applications in breeding programs and even QTL mapping (Bradshaw et al. 2008, Brown et al. 1984, McCord et al. 2011, Poland and Nelson 2011, Topcu et al. 2021), their lack of precision and potential issues of rater bias raise questions about their reliability.

An increasingly common way of improving measurement of traits that are difficult to quantify or labor intensive to measure throughout the biological sciences is image analysis. The first step of image analysis is image acquisition. Image acquisition can be accomplished through a variety of remote sensing devices and techniques including hyperspectral, infrared, and standard RGB cameras. These can be deployed on a variety of platforms including aerial drones, satellites, carts in the field, and controlled lightbox environments (Heineck et al. 2019, Li et al. 2014, Maimaitijiang et al. 2020, Oberholzer et al. 1996, Segarra et al. 2020, Su et al. 2020, Weiss et al. 2020). In the context of evaluating the yield of a crop post-harvest, as is the case with potato tubers, controlled environment image acquisition is the most practical option. While many remote sensing devices are able to be used in this scenario, we will primarily focus on low cost, widely available RGB image capture.

Because visual scales are widely deployed in assessing disease severities, especially in contexts where measurement precision is highly important such as QTL mapping, studies in plant science evaluating the value of visual scales compared to image analysis tend to be in the context of disease scales. Parker et al. (1995) evaluated human rater reliability compared to image analysis in cereal diseases. Human ratings were inaccurate and imprecise compared to an image analysis baseline in a controlled classroom-like setting and also in a more realistic experimental setting. Experimental results of a seed treatment trial would lead to different interpretations between the human ratings, where no significant difference between treatments was found, and image analysis, where one treatment was found to be twice as effective as the other. Poland and

Nelson (2011) compared image analysis, human rating using a percentage estimate, and human rating with an ordinal scale in performing QTL mapping of northern leaf blight resistance in maize. While the same QTL were generally found across different methods, the effect sizes between methods and between raters varied by as much as three times. The use of the percentage estimates and scoring by experienced raters increased accuracy and precision, but experimental results were still heavily dependent on the individual rater.

These studies assumed image analysis to be a suitable baseline for rating accuracy without directly providing evidence that image analysis should be accepted as the most reliable source of data. Bock et al (2008) tested image analysis reliability as well as that of human raters in citrus canker of grapefruit. Image analysis ratings of the same leaves in different photographs correlated near perfectly ($r = 0.99$) while human ratings had lower correlations ($r = 0.89$-0.94). This trend was also seen in several other statistical measures of accuracy and precision. Image analysis preformed best relative to human raters at very high or very low levels of disease coverage. While human rater performance was not particularly low in this on study, the lack of precision and accuracy of visual ratings in Parker et al. (1995) and Poland and Nelson (2011) suggest that the value of image analysis can be even greater in some traits and can have a major impact on the results of experiments.

Skin Color Measurement

Some studies utilize only a subjective five-point scale for color intensity (Reeves 1988, Buhrig et al. 2015), but measurement of skin color for research purposes is most often performed using a colorimeter. These colorimeters typically report color using the Hunter LAB, CIE LAB, or Hue, Chroma, Lightness (HCL) color spaces, which are meant to simulate color as experienced by the human eye better than RGB (International Commission on Illumination 2019, Kuehni 2001). Image analysis has been used in the commercial context for sorting off-type or defective tubers (Tao et al. 1995, Patel et al. 2012), though not in the research context until more recently. Caraza□Harter and Endelman (2020) developed an image analysis method to extract HCL values from RGB images and collect hue, lightness, and chroma values of the tubers in an image. They

found high plot-based heritability (>0.8) for each color component as well as high collinearity between all three components. Image based color measurement could provide the advantage of highly accurate, quantitative measurement at a lower equipment and labor cost than colorimeters.

Shape Measurement

Measurement of tuber shape is often performed with either visual scales or using calipers to measure length to width ratio. Visual scales can be very simple, with only round and long categories, or include more classifications between these states (Prashar et al. 2014, Si et al. 2017, Van Eck et al. 1994). Calipers are able to give highly quantitative measurements of length to width ratio. However, using calipers is time consuming and doesn't capture other aspects of tuber shape such as eye depth and knobbing, which are important to potato marketability (Chung et al. 1988). Early attempts at automated shape measurement using machine vision were focused on the commercial grading context. These were focused on identifying off type tubers in a single variety and lacked the scope and accuracy needed for research purposes (Hassankhani and Navid 2012, Razmjooy et al. 2012, Tao et al. 1995, Zhou et al. 1998). Si et al. (2017) developed an image analysis program using images of tubers in a lightbox taken with an RGB camera. The program was created specifically for research and breeding, able to measure at high accuracies of at least 94%, with more oblong tubers performing better than round tubers. This analysis was extended to images of potatoes passing on a grader belt where they were able to achieve similar accuracies with most potatoes despite the speed of the grader (Si et al. 2018). Neilson et al. (2021) created a similar program to evaluate length to width ratio, as well as a trait they refer to as circularity, in order to better account for shape factors like eye depth. Both of these traits were found to be moderately heritable between plants grown in greenhouse and field conditions. These studies are promising for low cost, high throughput quantitative shape measurement, however factors like eye depth and knobbing may still require other forms of measurement.

Area Correlation with Weight

While weight is a trait that cannot be directly extracted from an image, a common approach is to estimate weight using area. Zhou et al. (1998) created an image analysis program to measure potato area, among many other traits, in a commercial context. They used area data along with grader yield data to determine the correlation between potato tuber area and weight, which they found to be very high ($r = 0.97$). Their study only included three, unspecified potato varieties, however a similar correlation ($r = 0.98$) in tubers from many genotypes has been observed (Neilson et al. 2021). While, to my knowledge, there are limited studies on the correlation between potato tuber area and weight, studies exist in other crops that may provide insight into how this trait may behave in potatoes. Jahns et al (2001) found a similarly very high weight to area correlation in tomato ($r=0.99$). However, a study that directly assessed two different varieties of mangoes found a significant difference in their area to weight correlations (Patel et al. 2020). Unequal correlations in area and weight between genotypes seem physically necessary due to differences in density and shape caused by genetic factors (Henninger et al. 2000). The degree of those differences may not be so severe as to make weight estimates based on area impractical in certain applications but should be considered before substituting these estimates for direct weight measurement.

Skin Set Measurement

Beyond basic skinning severity scales, more quantitative means of measuring skin maturity have been developed. The earliest means, which is widely deployed in studies of skinning or skin maturity (Lulai and Orr 1993, Boydston et al. 2018, Krupek et al. 2021), is a handheld torque meter designed to measure the force needed to pull the phellem layer off the tuber (Halderson and Henning 1993). A more recent method uses image analysis to determine the percentage of the tuber area that has been skinned in the process of harvesting and grading, which reduces some of the manual labor required in using a torque meter (Caraza-Harter and Endelman 2020). While these measurements may be similar, differences between potential skinning (as measured by pulling the skin) and realized skinning (as measured by skinned area) should be investigated.

Skin finish measurement

Given the lack of a clear definition of skin finish, skin finish scales appear to have little consistency between studies. Scales tend to go from smooth skin to russeted skin often with few or even no intermediate states (da Silva Pereira et al. 2021, De Jong 1981, McCord et al. 2011, Raimo et al. 2018). Some define the trait based on marketability, using ratings much like a merit scale (Clulow et al. 1995). Ginzberg et al. (2012) utilize a five-point russeting scale based on percentage of the surface area with russeted skin rather than the severity of the russeting.

Bruise measurement

Bruising has been primarily studied in the context of impact with machinery or when falling into storage bins. Damaged sustained from impact has led to the development of potato shaped accelerometers or accelerometers made to be inserted into potatoes for quality control and calibration of harvesting and processing equipment (Praeger et al. 2013). In order to cause bruising comparable to that suffered during harvest and processing for research purposes, potatoes are dropped from a fixed height, hit with a pendulum at a fixed speed, or skinned abrasively (Dean et al. 1993). Another means of testing that is less related to industry processes but results in more precise measurements is mashing and homogenizing potato samples, allowing them to oxidize for a fixed time, then using spectroscopy to determine the levels of pigment in tuber tissue (Lærke et al. 2002). This damage is sometimes evaluated qualitatively as a binary scale of bruising presence or absence, or with slightly more quantitative visual scales where darker and larger bruises increase severity (Dean et al. 1993, Muthukumarappan et al. 1994). More complex indexes that account for things such as the width and depth of bruises can also be used (Lærke et al. 2002). Hyperspectral imaging has been used by several studies to identify bruising (Ji et al. 2019, López-Maestresalas et al. 2016, Ye et al. 2018), but use of RGB imaging to measure bruising is more rare. Oppenheim et al. (2019) created a program to identify bruised tubers, along with other defects, but this only identified the presence of bruising and did not measure severity. Marique et al. (2005) classified bruised regions on RGB video of peeled potatoes but required prepping of photos to avoid misclassifications.

Scab measurement

The complex and variable symptoms of common scab can make the creation of a visual disease severity scale more difficult than one based only on infected surface area. Some visual scales ignore lesion types and focus only on percent coverage (Haynes et al. 2010), however most reported scales make some distinction between lesion types. A common distinction is that between more superficial "scab" lesions and deeper "pit" lesions. Some only make this distinction (Wilson 2001), but others even further divide lesions based on depth and overlap (Driscoll et al. 2009, Wanner and Haynes 2009, Haynes et al. 2010). When lesion types are scored separately, these scores can also be combined using an index (Wilson 2001). While these scales provide ways to break down the visual symptoms of common scab, they are all limited in precision (less than 10 categories) and subject to emphasis on what the creator of the scale finds to be the most important indicators of scab resistance with little scientific reasoning given for the weights assigned to different lesion types. Dacal-Nieto et al. (2011) utilized hyperspectral imaging to automate scab identification and Oppenheim et al. (2019) did the same with RGB images, but quantification of scab severity using image analysis appears to be unpublished.

Chip and fry color measurement

While the USDA Munsell color chart is still used to measure fry and chip color, there has been a widespread and long-term effort to provide quantitative measurements of chip and fry color. Spectroscopic analysis of acrylamide levels can provide quantitative data (Guo et al. 2019, Pedreschi et al. 2006a), but direct measurement of fry and chip color is also possible with image analysis. Early light-based methods that utilized the Hunter LAB scale with spectrophotometer devices required liquid from either fried chips or raw tubers absorbed onto filter paper, rather than direct observation of chips (Hyde and Shewfelt 1960, Chubey and Walkof 1968). Eventually, methods were developed for obtaining Hunter LAB scores from samples of crushed chips (Work et al. 1981). For French fries, only the outer surface is darkened, making methods that require crushing and flattening impractical. Therefore, in fries light reflectance percentages are used to

determine darkness (Panigrahi et al.,1996). Shock et al. (1994) compared reflectance values in French fries to the Munsell color chart and found the results to be highly correlative (r = 0.98), though the percentages offered by reflectance meters are more precise than the five-point visual scale, at least at a large enough sample size.

Image based evaluation of fry and chip color appears to have originated in evaluation of grayscale video frames. Grayscale video frames have been used to extract average brightness of chips, as well as to rate dark spots on chips (Scanlon et al. 1994, Segnini et al. 1999). These results were found to strongly correlate with subjective scale ratings. Pedreschi et al. (2006b) implemented a system to evaluate chips fried at different temperatures using still images from a color camera. While this system utilized a color RGB camera, images were converted to the CIE Lab color space and grayscaled for L, a, and b values before processing, making this process very similar to the previously mentioned studies with the additional collection of a and b values. Despite the development of several image based systems for chip evaluation, they have seen little adoption, perhaps due to their dependence on samples of only one chip limiting either phenotyping speed or reliability.

Machine Learning

Image analysis performed manually by researchers can quickly become prohibitively tedious work, preventing high-throughput use, along with suffering from many of the same rater bias and fatigue issues of visual scales (Marée et al. 2005). This can be avoided by programming image analysis software to automatically segment and measure the color or size of regions of an image. However, directly coding instructions for segmentation and evaluation of images can become difficult for objects of interest that lack a constant, unique color and shape that can be described through programming languages (Madabhushi & Lee 2016, Liakos et al. 2018). These issues can be delt with by further extending the automation of image analysis through machine learning.

Machine learning uses given data to identify patterns in that data and other similar data. This can be done through supervised or unsupervised classifiers. Supervised classifiers take training data that has been pre-classified and look for patterns in the classes that would allow determination of how to classify new sets of data. Unsupervised

classifiers have no training stage but rather take data sets and determine ways to divide the data into clusters (Liakos et al. 2018). Here I will be focusing on supervised classifiers.

Machine learning algorithms can be further divided based on what kind of models they form to classify data. Some of these include decision trees, Bayesian models, instance-based models, and neural networks (Breiman 2001, Fix and Hodges 1989, LeCun et al. 2015, Liakos et al. 2018). Decision trees will be of particular interest here, although I will cover literature utilizing other methods. Decision trees divide data into classes based on a series of binary choices based on a single value in the data. These decisions then branch off into further decisions until the tree reaches sufficient size to accurately classify data (Breiman 2001).

Machine Learning in potatoes

Use of machine learning in combination with image analysis has been performed previously in potatoes. Oppenheim et al. (2019) used deep convolutional neural networks to classify RGB images of potato tuber skin into categories of black scurf, silver scurf, common scab, black spot, and healthy categories. Whole images were classified rather than pixels to categorize tubers rather than measure severity. Tubers were placed correctly over 90% of the time except healthy tubers which were classified as silver scurf infected about 17% of the time. Noordam et al. (2005) used multiple classifiers to detect black rot, greening, and mechanical damage in samples of raw French fries using both RGB and multispectral images. They found the highest accuracy with the support vector machines classifier with hyperspectral images and with the k-nearest neighbor classifier using RGB images. Accuracies across the four russet varieties used were as high as a 10% range in accuracy across different clones. Razmjooy et al. (2012) examined tubers for general defects using RGB video and support vector machine classification intended to be implemented in a commercial tuber sorting context. Defective tubers were correctly identified 95% of the time.

Specific use of decision tree algorithms, in particular Random Forest (Breiman 2001), can be seen in some instances, usually along with other models. Dacal-Nieto et al. (2011) combined hyperspectral imaging with the Random Forest and Support Vector

Machines classifiers to measure common scab severity in terms of percent surface coverage. Cross validation showed an accuracy of 95% for Random Forrest and 97% for Support Vector Machines. Duarte-Carvajalino et al. (2018) used RGB images captured from a drone to measure late blight severity among potato plant canopies, using many different machine learning classifiers to determine severity. Of the classifiers used, convolutional neural networks and Random Forest had the best performance, with mean absolute errors of 12% and 13%, respectively. A variety of machine learning classifiers have been successfully implemented for different defects and diseases of potato in several contexts. Random Forest appears to perform similarly to other popular methods like Support Vector Machines and neural networks, but the performance of one classifier over another seems highly dependent on the task involved.

Genomic selection

A potential application for high quality, quantitative phenotype data is genotypic selection (Cabrera□Bosquet et al. 2012). Genomic selection has emerged as a promising means to increase genetic gain compared to traditional phenotypic or marker assisted selection in many crops. Markers have historically been used as an enhancement to phenotypic selection, providing utility for traits that take time to emerge or are difficult to measure (Bernardo and Yu 2007). While marker assisted selection has been available since even few genetic markers were available, genomic selection has only become feasible somewhat recently as more dense marker arrays have emerged in crops. Rather than relying on finding significant markers for traits of interest as is done in marker assisted selection, genomic selection utilizes every marker across the genome as random effects in a mixed model to determine an expected breeding value for traits of interest. This is of particular advantage over both phenotypic selection and marker assisted selection in low heritability traits controlled by many loci (Bernardo and Yu 2007). Genomic selection has gradually seen adoption in many diploid and allopolyploid crops including maize, soybean, wheat, and pea (Bernardo and Yu 2007, Jarquín et al. 2014, Pérez-Rodríguez et al. 2012, Lorenz et al. 2012, Annicchiarico et al. 2019). However, this is only recently becoming realized in autopolyploid crops such as potato.

Autopolyploids experience unique conditions that limit the applicability of genomic selection models created for diploids (Wu et al. 2001). A major difference is the possibility for five allelic dosages rather than the three possible in a diploid. These allelic states are potentially more difficult to distinguish from each other and allow for the possibility of more than two alleles per genotype. Inheritance of multiple chromosome sets from one parent also allows for more than two alleles to be present at one locus and for dominance effects to be heritable (Amadeu et al. 2020, Paterson 2005). Other complexities of autopolyploids include double reduction, which results in sister chromatids sorting to the same gamete during meiosis and means that even fully heterozygous genotypes can produce some homozygous gametes (Meirmans et al. 2018, Wu et al. 2001). These complexities in polyploid genomics require more steps be taken in quantitative genetic analyses, including genomic selection.

Distinguishing between five allele dosages required the development of additional genomic tools such as high read depth genotyping by sequencing (GBS) and single nucleotide polymorphism (SNP) arrays combined with mixture models to distinguish allele dosages (Voorrips et al. 2011, Uitdewilligen et al. 2013). In the genomic selection models developed from this genotype data, steps must be taken to account for factors including dominance, double reduction, and multivalent pairing (Wu et al. 2001). When combined, these additional measures and tools may enable genomic selection in polyploids comparable in utility to that already used in diploids.

Genomic Selection in potatoes

While genomic selection in potatoes has only recently become attainable, its potential benefits have led to quickly expanding interest and research. A crucial component of this research is what models should be used, and what data improves their reliability. Genomic best linear unbiased prediction (GBLUP), Bayesian least absolute shrinkage (LASSO), and reproducing kernel Hilbert spaces (RKHS) have been compared using the same population and traits and were found to have similar performance to each other (Habyarimana et al. 2017). While not necessarily proven to be substantially better at creating reliable models than other methods, GBLUP and its variants appear to have gained dominance in potato genomic selection (Endelman et al. 2018, Gemenet et al.

2020, Sood et al. 2020, Sverrisdóttir et al. 2017). Providing allele dosage data has improved models for yield, fry color, and specific gravity (Endelman et al. 2018). The partitioning of additive and non-additive genetic variance has had mixed results across studies, sometimes improving reliability but in other cases having no significant impact (Enciso-Rodriguez et al. 2018, Endelman et al. 2018, Gemenet et al. 2020). This may depend on the architecture of the trait being analyzed (Gemenet et al. 2020). Combining pedigrees with genotype data generally improves model reliability, depending on the trait and population studied (Enciso-Rodriguez et al. 2018, Sood et al. 2020). In some cases, models built from pedigree data without genotypes may even have performance equivalent to GBLUP models (Endelman et al. 2018). Overall, components to add to these models such as pedigree data and non-additive variance seem to differ in effectiveness depending on the traits and population involved but appear to be effective in at least some circumstances.

Other important factors to determine for potato genomic selection are necessary population sizes, markers, and read depth needed for useful models. Model accuracy can be inconsistent across training populations and models applied to different populations are often unreliable. However, reliability can be improved by using larger training populations or combining populations (Sverrisdóttir et al. 2017, Sverrisdóttir et al. 2018). Genotyping large populations can become prohibitively expensive, so ways of optimizing genotyping are needed for wide adoption of GS. Strategically reducing genotyping costs for GS can be accomplished by using principal component analysis on phenotype data to focus genotyping on individuals with multi trait variation and pruning SNPs based on linkage disequilibrium without loss in model quality (Selga et al. 2021). High read depth is critical for creating reliable models using GBS data, however costs can be reduced by using as few as 1000 SNPs without losing significant model quality (Gemenet et al. 2020). Increases to genotyping and model efficiency, using methods including those discussed here are needed for practical, widespread implementation of this technology.

Conclusion

Breeding for quality traits is an essential component of creating an improved, marketable potato. However, the large number of traits involved in quality requires collection of a lot of phenotype data. While the most common approach to this is the use of visual scales, these scales lack accuracy and precision due to broad categorical measurements and inconsistency between raters. Image analysis is a way to improve quality trait phenotype data. While successful adoption of image analysis has been seen in some quality traits, many breeders still primarily use visual scale measurements.

In cases of complex quality traits, machine learning can be used to identify and create trait data. This is a rapidly expanding area of research but remains even less tested and utilized than more basic forms of image analysis. An ideal application of more accurate and precise phenotype data in pursuit of genetic improvement of potato is genomic selection. Genomic selection has only recently become available in tetraploid potato due to the development of specialized tools to account for the complexities of polyploid genetics. While this is a recent development, potato genomic selection is also a rapidly expanding area of research likely to be utilized by breeders for clone and parent selection. Combined with the phenotyping technologies of image analysis and machine learning, this should provide selection models that can be used to accelerate progress in the difficult pursuit of potato quality trait breeding.

Chapter 2
TubAR: an R package for quantifying tuber shape and skin traits from images

Summary

Potato market value is heavily affected by tuber quality traits such as shape, color, and skinning. Despite this, potato breeders often rely on subjective scales that fail to precisely define many phenotypes. Individual human raters and the environments in which ratings are taken can bias visual quality ratings. Collecting quality trait data using machine vision allows for precise measurements that will remain reliable between raters and breeding programs. Here we present TubAR (Tuber Analysis in R), an image analysis program designed to collect data for multiple tuber quality traits at a low cost to breeders. To assess the efficacy of TubAR in comparison to visual scales, red potatoes were evaluated using both methods. Broad sense heritability was consistently higher for skinning, roundness, and length to width ratio using TubAR. TubAR collects essential data for fresh market potato breeding programs while maintaining efficiency by measuring multiple traits through one phenotyping protocol.

Introduction

Tuber quality traits, including shape, color, and skinning (excoriation), are key determinants for the market value of potatoes (Carputo et al. 2004). Therefore, they are important selection targets for potato breeders. These tuber quality traits are traditionally rated on subjective five or nine point visual scales (Buhrig et al. 2015, Prashar et al. 2014, Reeves 1988, Van Eck et al. 1994). While these scales simplify breeder note taking and allow for comparison across programs, they may not encompass all meaningful aspects of a trait. For instance, in the US National Chip Processing Trial, tuber shape is rated on a five-point scale from compressed to elongate as shown in Figure 2.1. The scale fails to account for other deviations from ovoid, such as shoulders/boxy shape, pointy ends, or lumpiness. Use of calipers to measure tuber length to width ratios can provide a more precise measurement of tuber shape, however this is labor intensive and does not capture other elements of tuber shape that are important to processors and consumers, such as knobbing which can lead to loss of yield in harvesting and processing (Chung et al. 1988).

Several sources of inconsistency arise in applying visual ratings. Results can vary across scientists due to individual differences in sensory perception (e.g. varying degrees of color perception capacity) and internalized definition of the target phenotype, as well as across time due to breeder experience level, fatigue, and differences in lighting during evaluation. In research involving many scientists over a large area, consistent, clear phenotyping standards are vital to ensuring high quality data and increasing the likelihood of meaningful results (Parker et al. 1995, Poland and Nelson 2011). Within-field or within-experiment check clones may be used to "visually calibrate" one's scoring before evaluating experimental entries. However, genotype by environmental interactions account for a portion of phenotypic variance and relying on generalizations about the standard cultivars' expected performance would result in a visual mis-calibration in some instances (Liu et al. 2019). Machine vision can improve consistency across programs and environments by eliminating differences due to rater bias, as well as measure traits of interest more accurately (Parker et al. 1995, Bock et al. 2008, Poland and Nelson 2011).

Machine vision is commonly utilized to assist in commercial quality sorting in a variety of crops including potato. Most machine vision platforms for potatoes are designed for industrial scale processors and implementing these systems in breeding programs may involve expensive equipment and might not provide the precision or consistency needed for research applications (Cubero et al. 2011, Moreda et al. 2012, Zhou et al. 1998). Specialized equipment for measurement of color (colorimeter) and skinning (torque meter) exist, but these are expensive and labor intensive as they are limited to the evaluation of one or very few traits. Accurate and precise image analysis of multiple tuber traits using standard digital camera images would therefore be advantageous to efficient and low-cost measurement of quantitative tuber quality traits.

Image analysis systems for potato tubers with accuracy appropriate for research and breeding program applications have been previously developed for both tuber shape and color. Si et al. (2017) used watershed segmentation to identify tubers in standard two-dimensional digital images and measured the length width ratio of the segmented tuber shapes. Accuracy of the ratios compared to caliper measurements was 96% for white potatoes and 94% for red potatoes, with the lower accuracy in red potatoes attributed to their generally more circular shape. Caraza-Harter and Endelman (2020) evaluated red potato skin color and skin set in digital images using the RGB Measure Plugin in ImageJ (Schneider 2012). Plot based heritability, using image analysis for phenotyping, was above 0.75 for the hue, chroma, and lightness measures of skin color, as well as for skinning.

While these image analysis methods provide accurate results at a low cost, high throughput phenotyping requires more automation and simplification of the analysis process. The tuber image analysis process could greatly benefit from having a single program and protocol to measure multiple quality traits at once. TubAR (Tuber Analysis in R) provides a simple and efficient method for phenotyping multiple tuber traits from light box images simultaneously.

Methods

Plant material and phenotyping

We collected a random sample of 10 USDA medium (2.5 – 3.25'') sized potatoes from each plot of a nitrogen trial held at the Sand Plain Research Farm in Becker, MN in the summers of 2018 and 2019 described in Stefaniak et al. (2021). Entries were advanced red potato selections from the University of Minnesota – Twin Cities breeding program and a selection of red potato cultivars favored by Minnesota growers. Plots were desiccated 90 days after planting (DAP) and mechanically harvested 104 DAP and tubers sized with a Kerian sizer (Kerian Machines Inc. Grafton, ND). Tuber shape was rated categorically as "boxy", "elongate", "oblong", "pear" or "round". Skinning was rated on a 0 to 5 scale (none to severe) and tuber color intensity was rated on a 0 to 5 scale (light to intense red color).

Additionally, we collected samples of red potato clones from the University of Minnesota breeding program that were grown over the 2019 and 2020 field seasons. In 2019 potatoes were grown at the North Central Research and Outreach Center in Grand Rapids, MN. Plots were desiccated 98 DAP and mechanically harvested 118 DAP. These tubers were not graded or rated using visual scales. In 2020 potatoes were grown at the Sand Plain Research Site. Plots were desiccated 90 DAP and mechanically harvested 104 DAP. Tubers were graded on an AgRay sorter and visually rated as above in 2020.

Following the methods described by Caraza-Harter and Endelman (2020), tubers were gently washed with water to remove soil and allowed to dry. For each plot, ten tubers were staged in an Ortery Photosimilie 200 software-controlled light box (Ortery Technologies Inc.) with front and rear lights turned on. One image was taken per plot. Photographs were taken with a Rebel T6i camera with a 24mm lens, ISO 100, 1/30 sec shutter speed and aperture f/5.6. A CameraTrax 24ColorCard (CameraTrax.com CT24-23-1315) was placed in the lower right corner of the image. Images were 6000 x 4000 pixels with resolution of 72 x 72. Images were saved in .jpg (lossy compression) format.

Image segmentation

Tubers were isolated from the image background through image segmentation. Images were read into the R statistical software (R Core Team 2018) environment with the package EBImage (Pau et al. 2010) and resized to reduce computation time with the function resize. Transformation from RGB to the CIELab color scale was done with the function convertColor (R Core Team 2018). We chose the CIELab color scale because it is the international standard for description of the color of objects because it is most similar to human perception (International Commission on Illumination 2019). Each component matrix of the color scales (R, G, B, L, a, and b) was inspected for capacity to create separation between tuber and background pixels. The resulting components of the image were subjected to a threshold to create a binary filter. Within-tuber gaps, due to lightly colored patches of skin or skinning, were filled with the function fillHull (Pau et al. 2010). Image segmentation was conducted with the function bwlabel (Pau et al. 2010). A minimum object size in pixels was enforced to remove objects resulting from shadows and small debris (e.g. tuber skin flakes and residual soil).

Color correction

We used three-dimensional thin-plate spline (TPS) to color correct images to minimize the effects of any changes in light levels. TPS for color correction warps (transforms) pixel color values of a conserved object in an image to a predetermined reference RGB color value, in this case, the center pixel of each color chip in the 24-chip color card. An interpolation function is then used to transform the color values of the other pixels in the image using the difference between the observed color of the color card pixels and the reference colors (Menesatti et al. 2012). The tps3d function from the Morpho package (Schlager 2017) was utilized to create the interpolation function and determine the corrected color values for each pixel.

Shape measures

Several measures were calculated on a per-tuber basis with the objective of identifying measures that correlate well with visually apparent aspects of tuber shape. The perimeter and area were recorded for each tuber as well as the convex hull of each

tuber. The convex hull is the tuber shape modified by removing indents, which are often created by micro-environmental factors. Maximum length was determined as the maximum of distances between all possible pairs of perimeter pixels. The minimum bounding box was fit for each tuber resulting in measures of length, width, and length-to-width ratio. Convex hull area and convex hull perimeter were used to calculate roundness. Roundness ranges from zero (for a straight line) to one (for a perfect circle) (Van der Werff and Van der Meer 2008).

$$\text{roundness} = \frac{\text{convex perimeter}^2}{4\pi \cdot \text{convex area}}$$

Color and skinning measures

On a per tuber basis, pixels were converted from RGB to the CIELab color scale with the function convertColor. We used the nlsLM function from the minpack.lm package (Elzhov et al. 2016) to fit a sigmoid curve with the formula:

$$y = \frac{a}{1 + e^{-c*(x-d)}} + e$$

in which y is the percent of the tuber pixels designated skinned at b (from Lab) threshold x. The parameters a, c, d, and e were initiated at the maximum y value, 1, the median x value, and minimum y value, respectively. The threshold for designating pixels skin (red) or skinned (flesh, white to yellow) was set at 1.5x the d value, whereby the algorithm responds to differences in tuber skin and flesh colors for each image. It is assumed that all tubers within an image comprise a sample from the same variety under the same treatment.

Excluding the portions identified as skinned, tuber skin color was quantified in terms of redness as the median a (from Lab) value for each tuber and skin lightness as the median L (from Lab) value for each tuber (Figure 2.2).

Heritability Calculations

Broad sense heritability for redness and skinning was calculated using data from visual ratings and then from image analysis data collected from the potato tubers. We used the lmer function from the lme4 R package (Bates et al. 2015) to create linear mixed-effect models from trait values, clone, nitrogen rate, and block among the nitrogen trial populations.

$$P_{ijk} = \mu + G_i + B_j + R_k + \varepsilon_{ijk}$$

Where P_{ijk} is the trait phenotype of clone i in block j at nitrogen rate k, μ is the intercept, G_i is the random effect of genotype, B_j is the fixed effect of block, R_k is the fixed effect of nitrogen rate, and ε_{ijk} is the random effect of the residual. For the breeding program data, field year was used instead of block and no fixed effect from nitrogen was present, making the linear mixed-effect model:

$$P_{ijk} = \mu + G_i + B_j + \varepsilon_{ij}$$

Variance components, V_G and V_E, were taken from the lmer function output and used to estimate broad sense entry-mean heritability.

Sample size variance calculations

In order to evaluate the optimal number of tubers per photo, we determined the mean standard error of standardized output values for each of the five phenotypes described above using 3 to 10 tuber samples, which is the practical limit of what can easily fit in the lightbox for most clones. Samples were taken from images of 224 clones harvested from the University of Minnesota breeding program in 2020. Smaller sample sizes were simulated from trait data of 10 samples by shortening the sample data lists. Because tuber placement within images was not ordered, these smaller tuber samples would still be random. For each trait, the standard error at a given tuber number was averaged across the 224 clones. These averages were plotted to evaluate gain in precision from each additional tuber included in an image.

Visual scale consistency test

We created a set of 50 red potato sample pictures, then created 2 more sets of the same sample pictures in randomized orders. The skin.all function in TubAR was used to remove the backgrounds of the photos to limit information about the clone genotype. Four members of the University of Minnesota Potato Breeding and Genetics Lab with prior experience in rating tubers using visual scales rated the pictures from each of the three sets using the skinning, color intensity, and shape visual scales previously discussed. We instructed participants to rate the pictures quickly as if the sample were passing by on a conveyor belt in order to simulate scoring of tubers during grading. Scores were then derandomized and the coefficient of determination (R^2) for each pair of sets scored by the same rater, as well as every pair of sets between each rater was determined using the *cor* function in R. The maximum, minimum, and average coefficient of determination was determined between scores of different sets from the same rater as well as between different raters.

Because the skinning trait rated by TubAR is theoretically comparable to that rated with the visual scale, the same set of 50 sample pictures was rated for skinning using TubAR. The ratings were converted from percentages to a three point scale, mimicking the ratings from human observers. This set of skinning values was compared to each of the sets rated by the human raters using the maximum, minimum, and average coefficient of determination as discussed above. The standard deviation for visual scores of each trait for each sample image were averaged for each trait to measure variation in rating across raters.

An additional set of 30 sample images of red potatoes collected from the USDA-ARS Potato Breeding and Genetics program in Prosser, WA using a Nikon D7100 DSLR camera and a HAVOX HPB-80D photo studio light box with a non-reflective black background. Images were evaluated by a scientist from the USDA and one from the University of Minnesota to determine difference in ratings between programs and demonstrate TubAR's extendibility to other lightbox set ups. Color intensity and shape were rated, while skinning was not due to low variance among the sample images. Trait scores were compared within one rater's scores and between raters based on R^2 values. Human color ratings were compared to lightness and redness values from TubAR and

human shape scores were compared to length to width ratio and roundness using R^2 values.

Time efficacy test

To quantify the difference in time to get trait data using TubAR versus previously available methods for quantitative tuber trait measurement, we timed the use of TubAR, calipers, and a colorimeter to collect data from 10 tubers.

For TubAR measurements, timing was started at the beginning of staging the tubers in the lightbox. Once a photo of 10 tubers was taken and saved, the timer was stopped. An R script was run on a personal computer (AMD Ryzen 5 3600 3.6 GHz processor, 16 GB of RAM, Microsoft Windows 10 Pro version 21H1) to collect skin and shape data from the photo, with the timer being restarted upon running the script, and stopped at the console printing trait values. To reflect the scalability of TubAR, a sample of 100 photos was processed with TubAR using the skin.all function using four processor cores, timing started upon running the R script and ended upon the creation of matrices of median skin and shape trait values.

For the alternative manual measurement method, ten tubers, calipers, a reflectance colorimeter (Photovolt Instruments Photoreflectometer 577PC), and a computer were assembled beforehand, and a timer was started once measurement began to be taken. Length to width measurements and reflectance values were taken for each tuber and recorded in a spreadsheet, with the timer being stopped after the last value was recorded. This was performed by three individuals with the average time being used in further calculations.

Total time to record data for one sample was directly timed for both methods, while time to collect data for 100 samples was based on multiplying the time to perform the manual component of each method, and in the case of TubAR adding the processing time for 100 photos.

Weight area correlation

A total of 337 potato breeding lines and named cultivars maintained by the USDA-ARS Potato Breeding and Genetics program in Prosser, WA and the U.S. Potato

Genebank in Sturgeon Bay, WI were planted as twice replicated, five-hill plots in the Pear Acres field site at Washington State University Irrigated Agricultural Research and Extension Center in Prosser, WA. Tubers from the plots were assessed for yield in order to compare tuber weight to tuber area. Tubers were rinsed by applying water to the samples for several minutes inside modified Kobalt 4-cu ft 0.5-HP cement mixers (Lowes LLC) lined with a ¼ inch yoga mat material (YogaDirect.com LLC). Measurements of sample yield were be taken using Ohaus Valor 7000 scale controlled by Python script run on a Raspberry Pi 3 computer. Those same samples were photographed and total tuber area in pixels was determined for each sample using TubAR. The cor function in R was used to determine the R^2 between the weight data and total area data from each sample.

Results

Parameters for TubAR use

Image size was reduced by a factor of four which was found to speed computation time while preserving the image sufficiently, as determined by visual inspection of the reduced images. Examples of the original and reduced images are shown in Figure 2.3.

To determine how the number of tubers per image affects the consistency of the measurements, we calculated standard errors for subsets of the tubers in each image. For all traits we observed a decrease in standard error with each additional tuber. However, the marginal effect of each additional tuber decreased as total tuber number increased (Figure 2.4).

Effectiveness as compared to visual ratings

We compared the performance of TubAR to visual ratings by looking at heritabilities and correlations. TubAR trait heritabilities were consistently higher than visual scores for the 2018 population (Table 2.1). In the 2019 nitrogen trial visual color intensity heritability was higher, with TubAR redness and lightness heritabilities being comparatively lower. TubAR trait heritabilities were recorded for the breeding program populations but could not be compared to visual scores because they were only scored

visually one year. We could not calculate heritability for skinning in the breeding population due to low variation across clones.

R-squared values between comparable visual and TubAR measurements ranged from 0.26 to 0.41 (Figure 2.5). We observed negative correlations for roundness and lightness because the scales for the TubAR measurements and the visual measurements run in opposite directions.

Judging the effectiveness of TubAR using correlations assumes the visual ratings are accurate. We had multiple raters rate the same images to determine consistency across raters. Coefficients of determination for human visual scores (Table 2.2) tended to be higher between ratings by the same individual than between individuals but ranged widely in every case (less than 0.2 to more than 0.6). Shape scores had the highest within and between correlations while color intensity had the lowest. The average R^2 between TubAR scores and human ratings was very similar to the between rater average (≈0.25). The average standard deviation of scores across the ratings for color intensity was 0.71, 0.76 for skinning, and 0.39 for shape.

In order to look at rater consistency in a different population and between different breeding programs, we also had a rater from the University of Minnesota and the USDA rate tubers from the USDA for color intensity and shape. Within rater R^2 values averaged 0.54 for color and 0.66 for shape. Between rater R^2 values averaged 0.12 for color and 0.60 for shape. TubAR values for lightness and redness were compared to color intensity. Length to width ratio and roundness were compared to shape. Color intensity had an average R^2 of 0.11 with redness and 0.41 with lightness. Shape rating had an average R^2 of 0.58 with length to width ratio and 0.65 with roundness.

Comparing tuber weight and tuber area as calculated by TubAR provides a measure of effectiveness that does not depend on the accuracy of visual ratings. The R^2 between tuber sample weights and total tuber photo area was 0.77.

Time investment

Measuring and recording color and length to width ratio for a single, ten-tuber sample took 5 minutes 40 seconds using TubAR, and 4 minutes 14 seconds using a calipers and photovolt. Taking a picture for TubAR took 2 minutes 23 seconds with the

remaining time being processing in R. Processing 100 samples in TubAR took 54 minutes 1 second. When the photo taking time is multiplied by 100 and added to TubAR processing time, the expected time to collect data for 100 samples in TubAR is 4 hours 52 minutes. The caliper and colorimeter time is multiplied by 100 to get the expected manual measurement time of 7 hours 3 minutes.

Discussion

TubAR was created to improve the precision and accuracy of tuber quality trait measurements over that of the visual scales often used by breeding programs to phenotype for selection. The more accurate the information used in breeder decisions, the more effective selection can be. This is particularly important as potato breeders begin implementing genomic selection cycles, where inaccurate predictions can drive populations in the wrong direction (Caruana et al. 2019, Enciso-Rodriguez et al., 2018, Endelman et al., 2018, Gemenet et al., 2020, Habyarimana et al., 2017, Selga et al., 2021, Sood et al., 2020, Stich & Van Inghelandt, 2018, Sverrisdóttir et al., 2017, Sverrisdóttir et al., 2018).

Accuracy and Precision

When visual ratings are used, clones are assigned a whole number value between one and five. TubAR increases the potential for precision by relying on a continuous numeric scale rather than a rating. Accuracy is more difficult to measure but from a breeding perspective, it is most important to be accurate about the genetic component of a trait. For traits measured using multiple methods in the same population, comparing heritability can provide information on the relative accuracy of measurement techniques (Caraza-Harter & Endelman, 2020). Because location and genotype remain the same across measures, any changes in heritability can be attributed to the accuracy of the phenotyping method. TubAR roundness, length to width ratio, and skinning consistently had higher heritability than the visual scale measurements, while redness, and lightness did not (Table 2.1).

It is important to note that redness and lightness are affected by tuber washing/drying. Visual inspection showed that images from 2018, where both TubAR

color measurements exhibited high heritability as compared to the visual scores, featured consistently dry tubers, whereas images from 2019, where heritability was dramatically lower, featured some wet and some dry tubers. Washing tubers before taking pictures improves the accuracy of color ratings but all tubers should be left to completely dry before being photographed. TubAR is only as good as the input images, so care must be taken for consistent images.

The increases we observed in heritability for many TubAR traits in comparison to visual scales could be attributed to either an improvement in our ability to measure traits or to refining the definition of the traits. Breaking down a trait like shape or color into its component pieces may better reflect the underlying factors. For example, lightness increases with length of time in storage while hue, or color family, remains unchanged (Caraza-Harter & Endelman, 2020). Similarly, nitrogen affects lightness but not redness (Jones et al., 2021) and heat affects the expression levels of anthocyanins but not which ones are expressed (Liu et al., 2019). Separating color into lightness and redness, may be allowing us to distinguish between the aspects of color more dependent on environment and those that are more heavily genetic.

A second method we explored for determining the accuracy of TubAR measurements was comparing the values to visual scores (Figure 2.5). However, the correlation between scores from different raters (Table 2.2) were so low as to bring into question the value of the visual ratings as a benchmark. The correlation between TubAR traits and visual ratings was consistent with the correlation between measurements from different raters. We observed a higher correlation when comparing area as determined from images to tuber weight. However, this correlation was lower than the extremely high correlation ($R^2 = 0.94$) reported by Zhou et al. (1998). The discrepancy could be accounted for by the fact that Zhou et al. only examined three potato cultivars while we measured diverse samples from a breeding program. Significant variation in area to weight correlation between cultivars has been found in mango (Patel et al. 2020). This is likely to be true for potatoes as well given genetic variation in shape and specific gravity both of which would influence the correlation between weight and area (Slater et al. 2014).

Practicality of Implementation

We were able to collect and process images for a very low material cost. While some labor cost is saved in the faster processing of many images, staging and taking photos of each set of tubers is still labor intensive. However, when taking measurements for many samples, time is ultimately saved compared to using calipers and a colorimeter to manually measure traits. Additionally, TubAR measures many more traits than the two measured by manual devices and adding trait measurements does not add to the time it takes to collect data. It should also be noted that much of the processing time using TubAR does not require a human to be actively involved, potentially lowering labor costs well beyond the time saved compared to manual measurement.

The time required in staging tubers can be balanced with the precision desired in an experiment. The standard errors seen in different sample sizes largely decrease with every additional tuber from three up to the maximum ten in each trait. While the standard errors continue to decrease, each additional tuber decreases the standard error less than the previous tuber. This creates a situation of diminishing returns where the gain in precision will eventually not make up for the extra time spent staging more tubers. This tradeoff is particularly important for larger market classes, such as russets, where fitting ten tubers in one image may not be practical for a lightbox, or when time and labor available for phenotyping is constrained.

Potential Benefits

Machine vision phenotyping systems such as TubAR can provide more accuracy from year to year within a breeding program if the same system is used to evaluate clones from generation to generation. They can also be used to increase accuracy between programs in situations such as national trials where multiple raters would introduce bias to visual scales, as seen by the low correlations between ratings of different scientists.

While TubAR's ability to score many traits from one image increases the speed that each of those traits can be scored compared to manual or visual scale measurements, staging tubers, tags, and color cards for each photo is still labor intensive and sets a practical limit on the amount of tubers that could be measured in a year by most breeding

or research programs. Modifying this program for use in a potato sorter may allow high-throughput phenotyping by eliminating the need to manually stage each photograph.

Our current version of TubAR focuses on quality traits crucial to fresh market red potato including skin color, shape, and skinning. There is potential for other traits of interest to be measured with TubAR through a similar image analysis which would increase the applicability of TubAR to other market classes. Traits of interest that may include russeting, greening, eye depth, and diseases including common scab, Rhizoctonia, and silver scurf. One benefit of TuBAR is that we can retain images and as we add functionality, we can measure additional traits in previous years' breeding populations. Only limited modifications may be necessary to use the program to measure the color of chips and fries or even the color and shape of other crops such as apples.

TubAR is available at https://github.com/shannonlabumn/TubAR. Instructions, a vignette, and sample data are available for download with the package.

Tables

Table 2.1 Broad-sense heritability of several tuber quality traits in red potatoes from three populations.

Trait	Nitrogen trial 2018	Nitrogen trial 2019	Breeding program 2019-2020
Skinning (Visual)	0.63	0.62	N/A
Skinning (TubAR)	0.68	0.68	N/A
Color Intensity (Visual)	0.59	0.71	N/A
Redness (TubAR)	0.63	0.14	0.33
Lightness (TubAR)	0.63	0.12	0.60
Shape (Visual)	0.36	0.41	N/A
Roundness (TubAR)	0.64	0.50	0.51
Length/Width Ratio (TubAR)	0.61	0.48	0.48

Table 2.2 Minimum, maximum, and average coefficients of determination (R^2) for color intensity, skinning, and shape measurements. "Within" values represent a single rater's consistency across three times rating the same picture set. "Between" values reflect consistency across raters. The only trait measured by visual rating and TubAR was skinning and final row reports R^2 values for comparisons between rater scores and TubAR scores.

Ratings	Maximum	Minimum	Average
Color intensity Between	0.61	0.00032	0.22
Color intensity Within	0.64	0.047	0.38
Skinning Between	0.69	0.010	0.25
Skinning Within	0.70	0.10	0.42
Shape Between	0.68	0.077	0.43
Shape Within	0.76	0.17	0.47
Skinning TubAR	0.39	0.074	0.26

Figures

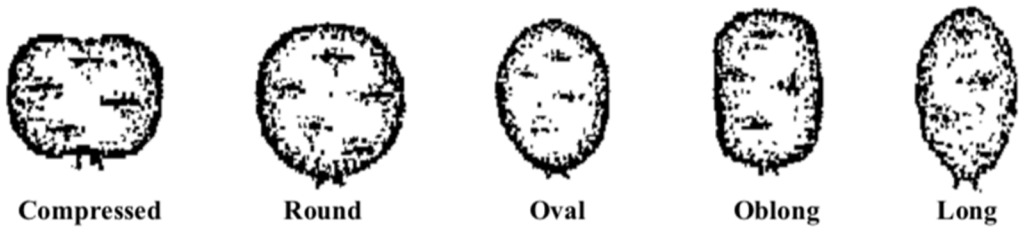

Figure 2.1 USDA form OMB NO 0581-0055 Tuber shape classification. (USDA Plant Variety Protection Office 2015)

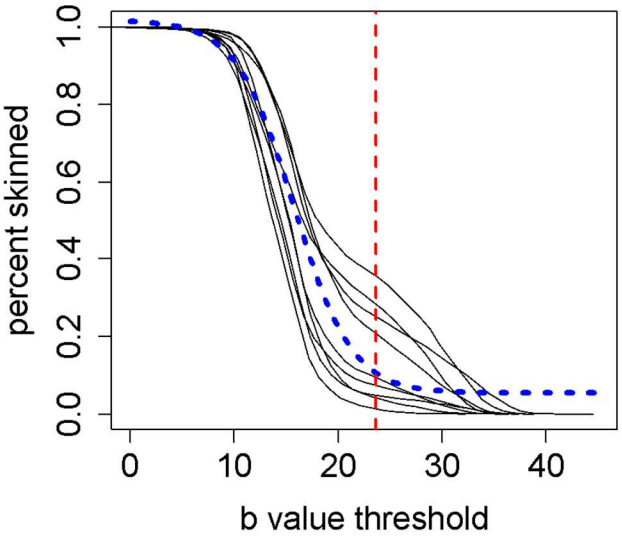

Figure 2.2 Percent of the tuber surface designated skinned relative to the Lab b value threshold within an image for each tuber (solid, black). The sigmoid curve fitted across tubers (short dashes, blue) was used to set the b threshold for the image (long dashes, red).

Figure 2.3 Image processing with TubAR. Original image (a, 6000 x 4000 pixels) versus resized image (b, 1500 x 1000 pixels) and indexed tubers with skinned area shown in gray (c).

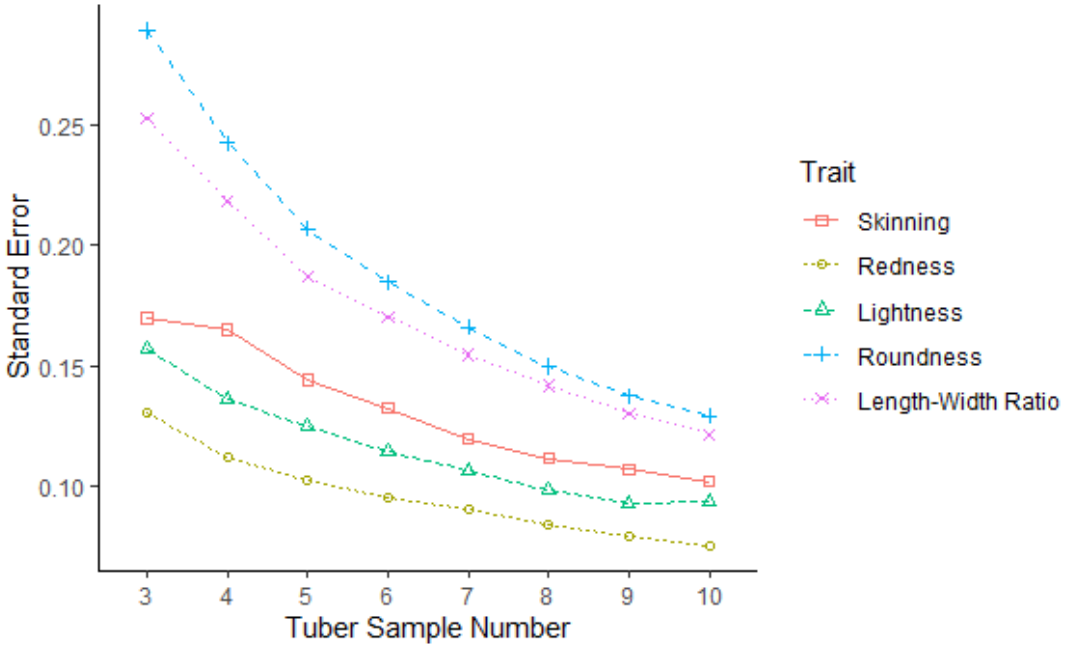

Figure 2.4 The mean standard errors of five standardized TubAR trait values given different numbers of sample tubers for 224 clones from the University of Minnesota breeding program.

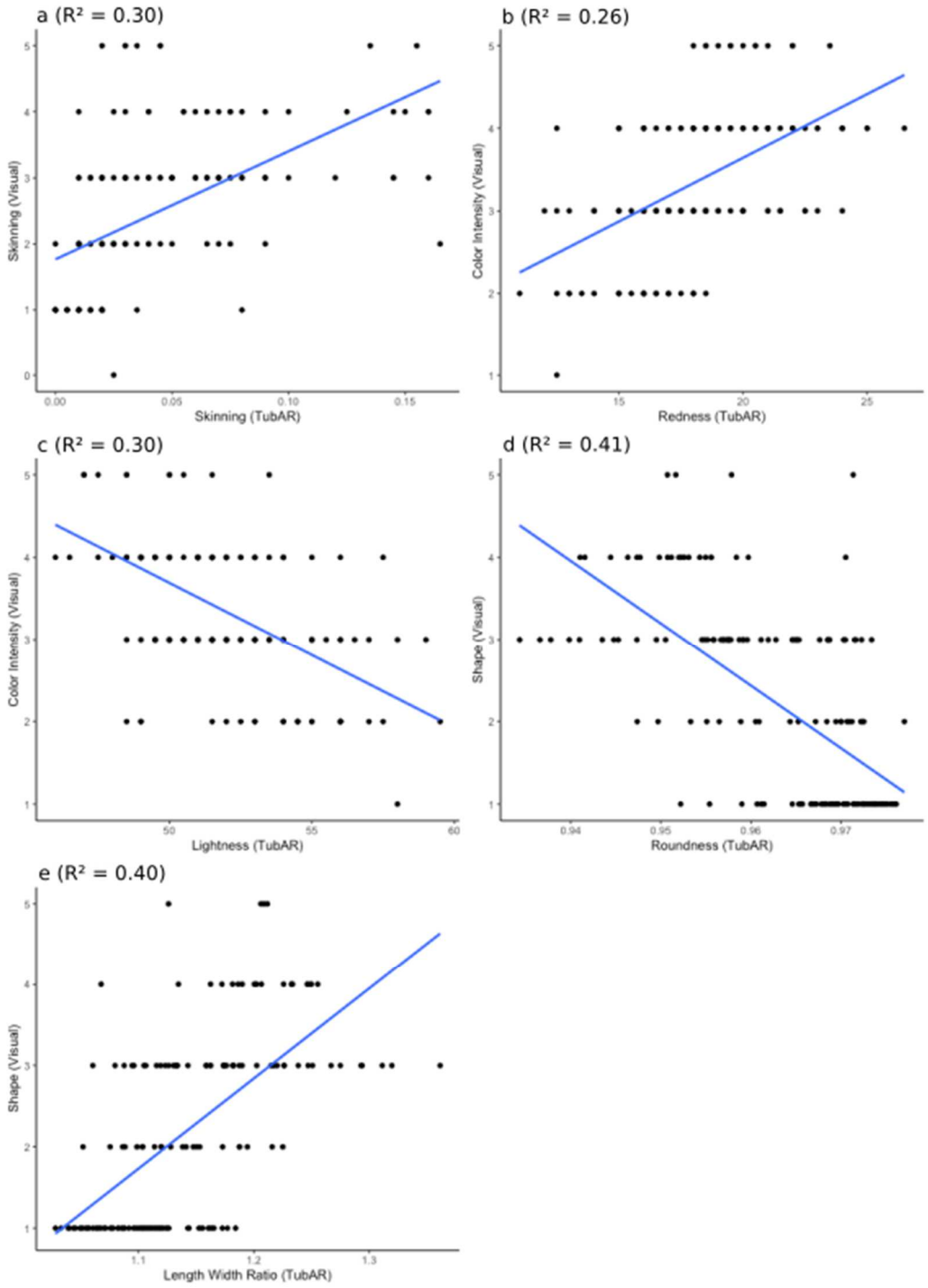

Figure 2.5 Scatterplots and regression lines of TubAR traits and visually rated traits. (a) Skinning percent determined by TubAR vs. skinning 1 to 5 visual rating, (b) Redness from TubAR vs. color intensity 1 to 5 visual rating, (c) Lightness from TubAR vs. color intensity 1 to 5 visual rating, (d) Roundness as measured by TubAR vs. shape visual rating, (e) Length Width Ratio taken from the length and width measured by TubAR vs. shape visual rating.

Chapter 3

Creation of machine learning models for identification and measurement of pressure bruise and skin finish

Summary

Tuber quality traits are important factors in breeding new potato varieties. While measurement of some quality traits can be automated by programming image analysis systems directly, others are difficult to define in ways that can be easily turned into computer code. However, leaving these traits to be measured using visual scales risks losses in precision and accuracy. Machine learning allows these traits to be identified by image analysis programs using example data, combining the simplicity of training a human rater with the accuracy, precision, and scalability of automated image analysis. Images were visually classified to create training data for pressure bruise and skin finish models which were used to create Random Forest models. These classifier models were used to create probability maps which were analyzed using the EBImage R package to get percent bruise coverage or the dominant skin finish type. These values were compared to manual scores of the same images. Model accuracy based on a test set made from the training data was 94.2% for the pressure bruise model, and 93.5% for the skin finish model. Model scores and manual scores had an R^2 of 0.90 for pressure bruise and 0.66 for skin finish. This procedure could be applied to many other tuber traits with only minimal changes and low equipment cost. Using image "patches" rather than individual pixels may be able to improve model accuracy while maintaining the potential for quantitative trait measurement.

Introduction

Tuber quality traits are of upmost importance in the marketability of potato varieties. These include a wide range of traits including tuber skin color, flesh color, shape, eye depth, skinning, skin finish, and bruising. While some of these traits are relatively simple to directly program image analysis software to identify and measure, other traits are more intricate and difficult to identify consistently using computer code. In cases dealing with more complicated traits, techniques used in simple image analysis procedures, such as color thresholding, would be insufficient and impractical for a programmer to implement directly. Despite the difficulty of explaining complicated traits through computer code, humans can often learn to identify these traits through subconscious learning processes when given examples of the trait (Jordan and Mitchell 2015, Sung 1996). Unfortunately, while human raters can identify traits, asking them to provide precise percentages or other numerical data can lead to inconsistency and rater bias which machine vision seeks to mitigate (Bock et al. 2008, Parker et al. 1995, Poland and Nelson 2011). Machine learning may provide a means to identify traits by examples like humans can while maintaining the objectivity provided by computers.

Machine learning uses example data to determine how to identify future information (Jordan & Mitchell, 2015). It has been applied to a rapidly widening variety of subjects, including agricultural image analysis. This includes field applications such as weed species identification and foliar disease identification (Binch and Fox 2017, Ferentinos 2018, Moshou et al. 2006). As well as post-harvest applications, such as fruit quality assessment (Dhiman et al. 2022). The cameras and platforms used in these scenarios vary from standard RGB digital camaras to hyperspectral cameras able to measure light frequencies beyond the perception of the human eye (Dacal-Nieto et al. 2011, Oppenheim et al. 2019, Razmjooy et al. 2012).

Image analysis of potato tubers, due to the below ground nature of tubers, is in the realm of post-harvest imaging. Studies of classification of tuber quality in RGB images have often focused on the context of defect detection at the whole tuber level (Oppenheim et al. 2019, Razmjooy et al. 2012). These have primarily been developed for commercial applications and do not offer the precision needed in breeding and research applications. Dacal-Nieto et al. (2011) developed a hyper spectral imaging system to

measure common scab severity in terms of percent surface coverage. This system requires a hyperspectral camera and mirror scanner to generate multiple hyperspectral images for a single tuber. They were able to produce a Random Forest model with 95% accuracy, but this system requires an expensive and complex equipment set up limited to evaluation of one tuber at a time. Models that could utilize images of multiple tubers taken with a commonplace RGB DSLR camera would be accessible to a wider range of research programs and require less data preparation.

Pressure bruise and skin finish are good candidates for the development of machine learning tools to automate their identification, due to their relative complexity and lack of a high throughput, objective means to measure these quality traits otherwise. Pressure bruise is an economically important, but understudied defect that occurs in tubers after long periods of storage. Over time, tubers near the bottom of storage piles will develop circular gray, round bruises at points of contact with other tubers (Schaper and Yaeger 1982, Lulai et al. 1996). This leads to moisture loss and lower economic value. Studies of pressure bruise resistance are lacking but could be made easier and more effective with an automated tool to measure pressure bruise. Skin finish is another trait of high economic importance which lacks extensive study (Jemison et al. 2008). Skin finish lacks precise definition in the literature but relates to the texture and quality of the tuber skin. When measured, raters tend to rely on highly subjective scales, sometimes difficult to distinguish from a general merit scale (Clulow et al. 1995, da Silva Pereira et al. 2021, De Jong 1981, McCord et al. 2011, Raimo et al. 2018). Machine based skin finish classification could provide much needed objectivity to the trait and allow better communication between scientists and breeding programs.

We present machine learning classifiers which accurately and precisely measure pressure bruise severity and skin finish using low cost, simple RGB images. We also created a simple and repeatable pipeline for the creation of models for other potato quality traits.

Methods

Plant material

Between October 2020 and March 2021, we stored chipping potatoes that were grown commercially near Grand Forks, North Dakota in storage totes (1000#, Macroplastic 32-S Pro-bin) modified with holes drilled into the base to ensure airflow. Samples of 5-10 tubers were stored in mesh bags and placed in the totes on top of bulk tubers to avoid sample contact with the tote base. We compressed the potatoes to increase the pressure exerted by contact between tubers. The storage totes received pressure from a ½ inch UHMW polyethylene plate exhorting 2.1 lb/in^2 of pressure, approximately the forces that would be exhorted by a 18 ft pile of tubers. Pressure was monitored every 48-72 hours. We stored chipping the potatoes at 46^oF based on industry standards. Temperature and humidity were controlled using a Techmark 755 Controller. We also received samples of pressure bruised red and yellow skinned from commercial storage facilities near Grand Forks.

For skin finish, potatoes from the University of Minnesota and North Dakota State University potato breeding programs were used. The potatoes from the University of Minnesota were grown at the Sand Plains Research Farm in Becker, Minnesota in the 2019, 2020, and 2021 field seasons. Potatoes from North Dakota State University were grown in Baker, Minnesota in the 2021 field season.

Imaging

Images were taken following the methodology of Caraza-Harter and Endleman (2020). For pressure bruise, we took pictures of samples of three to ten potatoes arranged in a Photosimilie 200 Lightbox (Ortery Technologies Inc.). Pictures were taken using a Canon Rebel T6i camera with a 18-55mm lens, ISO 100, 1/60 sec shutter speed and aperture f/5.6. All photos included a 3-color card and used a wooden stage painted blue (Windjammer #550B-5, Behr Paint Company) as a background.

For skin finish samples from the University of Minnesota, we took pictures of ten potatoes per sample using a Photosimilie 200 light box. Photographs were taken with a Rebel T6i camera with a 24mm lens, ISO 100, 1/30 sec shutter speed and aperture f/5.6.

We included a 24-color card (CameraTrax CT24-23-1315) and identification tag in all images with a smooth plastic white background under a translucent plexiglass stage. North Dakota State University used a Samitan 24”x24”x24” light box. Images were taken with a Canon EOS 800D camera, 18-55mm lens, ISO 100, 1/50 sec shutter speed, and aperture f/4.625. Images included an identification tag, 24-color card (CameraTrax CT24CC05000A), and a matte black background.

Model training

We used 50 photos of pressure bruised potatoes as the training data set for the pressure bruise model. The skin finish model used a sample of 80 photos, 40 from the University of Minnesota and 40 from NDSU chosen to represent the widest possible range of skin finishes and skin colors.

The ImageJ2/FIJI Trainable Weka Segmentation Plugin (v. 3.3.1) (Arganda-Carreras et al. 2017, Schindelin et al. 2012) was used to create training classification data. For the pressure bruise models, classes used were “Bruise” (Pressure bruise tissue), “Healthy” (Tuber tissue not affected by bruising, including tissue affected by other abiotic or biotic diseases), and “Background” (Areas of the photo that are not tuber tissue). For skin finish models, classes used were “Shiny” (smooth, shiny tuber tissue), “Dull” (Smooth, but not shiny tuber tissue), “Rough” (Exfoliating skin without a russeting pattern), “Russet” (thick skin with a net-like pattern), and Background (Areas of the photo that are not tuber tissue). See Figure 3.1 for examples of each skin finish class. Default training features were used for creating training data.

Models were created using the Knowledge Flow program included with Weka (v. 3.8.5) (Hall et al., 2009). Data was filtered to have an equal number of pixels of each class and in the case of pressure bruise, subsampled to a random subset of half the initial instances in order to increase the efficiency of model creation and not surpass the maximum Java heap size. Data was randomly split into 75% training data and 25% test data. The Random Forest (Breiman 2001) decision tree model was used with default settings. The Weka Classifier Evaluator was used to evaluate the model using the test set to determine accuracy, precision, recall, and F-measure (Bekkar et al. 2013).

Comparison to human measurements

We had one rater manually measure a subset of ten pressure bruise sample pictures for bruised area and total tuber area in pixels using the ImageJ2/FIJI Measure function. These values were compared to percentages of pixels classified with over 50% certainty as bruise by the pressure bruise model divided by tuber area as measured by the TubAR function shape.all.

For skin finish, one researcher rated a subset of 50 sample pictures categorically as Shiny, Dull, Rough, or Russet, with half marks allowed for samples that had a mixture of skin finish types. This same subset was processed using the skin finish model to create probability maps. Pixels from these maps classified to each skin finish type were counted using the EBImage package in R (Pau et al. 2010). The skin finish with the most area was considered the categorical rating provided by the model. Half scores were given in the case of samples with two classifications within 10,000 pixels of total area. These scores were converted to numerical scores along with the categorical human classifications and compared using the coefficient of determination to determine the percentage of images where these classifications aligned.

Results

Model data

The test set was composed of 25% of the training data for each model or 191071 pixels for pressure bruise and 194045 pixels for skin finish. Model accuracy was 94.2% for the pressure bruise model, and 93.5% for the skin finish model. Root mean squared error was 0.1818 for pressure bruise and 0.1703 for skin finish. The confusion matrices for pressure bruise (Table 3.1) and skin finish (Table 3.2) show how individual pixels were classified by the model compared to manual classification.

Additional model statistics for pressure and skin finish are provided in Table 3.3 and Table 3.4, respectively. Precision indicates what proportion of the instances that were assigned to a class are truly part of that class. Precision ranged from 0.902 for bruise to a rounded value of 1.000 for background for the pressure bruise model and from 0.887 for rough to 0.997 for background in the skin finish model. Recall indicates what proportion

of the members of a class were correctly assigned. For pressure bruise, recall ranged from 0.881 for healthy to 1.000 for background and for skin finish recall ranged from 0.892 for rough to 0.996 for background. The F-measure provides a combined precision and recall score. F-measures for pressure bruise ranged from 0.902 for healthy to 1.000 for background, and for skin finish ranged from 0.889 for rough to 0.997 for background.

Example probability maps created by the models in the Trainable Weka Segmentation plugin can be seen in Figure 3.2 for pressure bruise and Figure 3.3 for skin finish. These probability maps indicate the certainty of class assignment for each pixel using black for zero and white for one. Background classifications tend to very strict for both models, while tuber trait classifications tend to have some more pixels with medium gray values (indicating classification certainties further away from either 0 or 1).

Comparison to human measurements

In order to assess the reliability of the models for classification of whole sample images and confirm their usefulness in trait measurement, sample image classifications by the models were compared to human ratings. The bruised area of sample tubers over total tuber area as measured by manual image analysis and the Random Forest classifier had a coefficient of determination of $R^2 = 0.90$. Categorical skin finish ratings by a human rater and based on percent coverage as measured by the skin finish classifier had a coefficient of determination of $R^2 = 0.66$.

Discussion

Model accuracy on test set

Accuracy of the models in terms of agreement with manual pixel classifications was approximately 94% for both traits. This is a similar accuracy to previous studies of potato quality trait classification with various machine learning algorithms and imaging systems (Dacal-Nieto et al. 2011, Marino et al. 2019, Ming et al. 2018). Misclassifications rarely occurred between background pixels and tuber pixels, which would be expected given the sharp difference in color between background stages and tubers. When misclassifications of this type occurred, it tended to be on either very light

yellow or dark purple potato skin. Misclassifications in skin finish tended to be between adjacent skin finish types, e.g., shiny misclassified as dull, rather than more distant types, e.g., shiny misclassified as russet.

The machine learning models used individual pixel within manually classified regions as data instances, rather than whole images or regions of the image classified by the human rater. Therefore, this measure of accuracy is limited to those pixels which were manually classified and added to the training set. This could possibly lead to a bias away from ambiguous pixels which cannot be visually classified as those may have not been selected as examples for training data. Therefore, further assessment of the models based on whole images is required, as was carried out in the comparisons to human ratings.

Agreement with human measurements

Human measurements of traits provide a means of comparison to determine the reliability of whole image classification by the machine learning models. The pressure bruise model largely followed the manual bruise measurements ($R^2 = 0.90$). The percentages given by the model trended consistently higher than the human measurements. This may be due to shadows near the edges of tubers being classified as bruised tissue. However, this effect seems to be largely consistent from sample to sample resulting in the same rank ordering of sample bruise severity.

Skin finish categorical ratings based on human rating and the classifier model correlated only moderately ($R^2 = 0.66$). Some of this may be due to the complexity of the skin finish trait being difficult to capture within only four categories. Skin traits with little biological relation to each other may still be grouped into the same category, leading to ambiguity. The limitations of single pixel classification might also contribute to misclassifications. Pixel based classification has the advantage of allowing more quantitative measures of quality traits, such as percent surface area coverage. However, some skin traits characterized by patterns across the tuber, such as russeting, may be difficult to identify based on only a single pixel. Classification of small groups of pixels, or "patches," may allow these patterns to be better identified while allowing for quantitative measurements (Marino et al. 2019).

Random forest compared to other algorithms

The machine learning algorithm used in model creation can have a major impact on model performance (Singh et al. 2017). We chose Random Forest to create the classification models due to its use in past literature and relative high performance for its computational intensity as implemented in Weka. The computer hardware and software employed here was not adapted to make practical use of some other highly accurate algorithms common in the literature, such as Support Vector Machines (Boser et al. 1992) and convolutional neural networks (LeCun et al. 1989), which have both outperformed Random Forest in at least some contexts (Dacal-Nieto et al. 2011, Duarte-Carvajalino et al. 2018). Preliminary tests showed that the less computationally intensive J48 decision tree algorithm (Quinlan 1993) achieved results that might be sufficient for some situations, but with lower accuracy and much more "noise," or inaccurate shifts between classifications from pixel to pixel. Different hardware and software set ups may justify the use of other algorithms, however Random Forest seems sufficient for this application.

While Random Forest and Trainable Weka Segmentation can be utilized in an environment such as a personal computer, the models take time to load and use with images. Other algorithms with faster performance and similar accuracy would be desirable to reduce processing time, though preliminary tests of applicable algorithms available with the base Weka program did not present candidates that fit these criteria.

Expansion into other traits

Part of the goal in our creation of these models for pressure bruise was to create a pipeline that would allow for the efficient creation of classification algorithms for a wide variety of abiotic and biotic tuber diseases and skin traits. This process could be used with tubers exhibiting symptoms of diseases such as common scab and silver scurf with minimal changes. Similar methods could also be used for other types of plant tissue or other species. The integration of the Trainable Weka Segmentation interface allows for the creation of training data by individuals with little programming experience.

Integration into the TubAR package

Functions utilizing the machine learning methods detailed here could greatly expand the functionality of our tuber image analysis platform, TubAR. The simplicity of the Trainable Weka Segmentation could assist in the creation of many trait measurement functions, however integrating a Java-based program into an R package introduces complexities to creating a cohesive program. Currently, only the trait measurements after image classification can be performed in TubAR itself. Smooth communication between R and Java would be required for functions using this method. Alternatively, because the primary advantage to Trainable Weka Segmentation is in creation of model training data, we could import this data into R in a compatible manner to make the end-user functions purely R based.

Tables

Table 3.1 Confusion matrix for the pressure bruise model. Pixels are placed in rows according to classification by the model and columns according to manual classification.

	Background	Bruise	Healthy
Background	66679	5	8
Bruise	9	62679	4224
Healthy	16	6816	50635

Table 3.2 Confusion matrix for the skin finish model. Pixels are placed in rows according to classification by the model and columns according to manual classification.

	Background	Shiny	Dull	Rough	Russet
Background	38667	100	10	30	2
Shiny	37	36789	1470	314	199
Dull	42	1233	35133	1789	612
Rough	42	112	1375	34622	2658
Russet	2	22	188	2293	36304

Table 3.3 Precision, recall, and F-Measure statistics for the pressure bruise model.

Class	Precision	Recall	F-Measure
Background	1.000	1.000	1.000
Bruise	0.902	0.937	0.919
Healthy	0.923	0.881	0.902

Table 3.4 Precision, recall, and F-Measure statistics for the skin finish model.

Class	Precision	Recall	F-Measure
Background	0.997	0.996	0.997
Shiny	0.962	0.948	0.955
Dull	0.920	0.905	0.913
Rough	0.887	0.892	0.889
Russet	0.913	0.935	0.924

Figures

Figure 3.1 Skin finish category examples. From left to right: Shiny (smooth, shiny tuber tissue), Dull (Smooth, but not shiny tuber tissue), Rough (Exfoliating skin without a russeting pattern), Russet (thick skin with a net-like pattern).

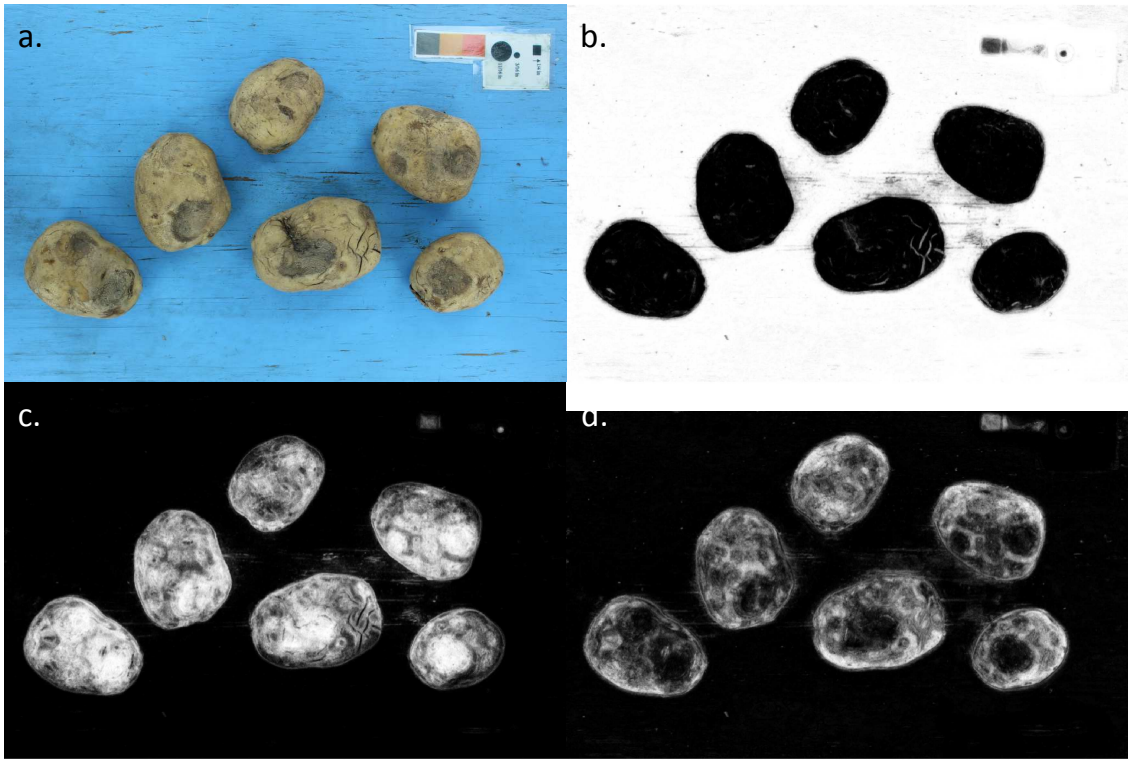

Figure 3.2 Probability maps generated by the pressure bruise classifier model. The pixels from the original image (a) were given a grayscale value to represent the likelihood of a pixel to be classified as background (b), bruise (c), or healthy (d).

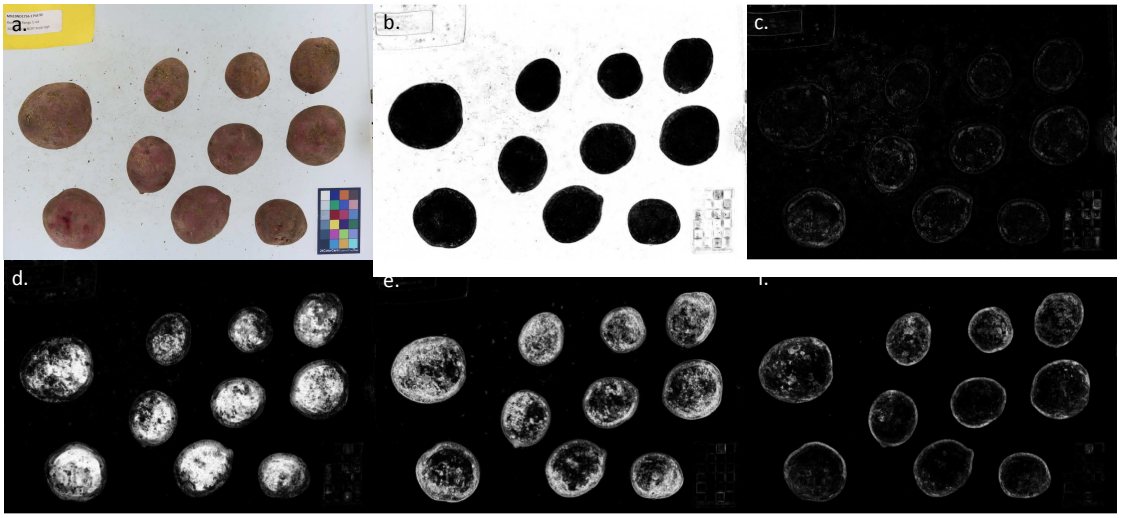

Figure 3.3 Probability maps generated by the skin finish classifier model. The pixels from the original image (a) were given a grayscale value to represent the likelihood of a pixel to be classified as background (b), shiny (c), dull (d), rough (e), or russet (f).

Bibliography

Agrios G. 2005. Plant pathology, 5th edn. Elsevier Academic, New York.

Amadeu, R. R., Ferrão, L. F. V., Oliveira, I. D. B., Benevenuto, J., Endelman, J. B., & Munoz, P. R. 2020. Impact of dominance effects on autotetraploid genomic prediction. Crop Science 60(2): 656-665.

Amrein, T. M., Limacher, A., Conde-Petit, B., Amadò, R., & Escher, F. 2006. Influence of thermal processing conditions on acrylamide generation and browning in a potato model system. Journal of agricultural and food chemistry 54(16): 5910-5916.

Annicchiarico, P., Nazzicari, N., Pecetti, L., Romani, M., & Russi, L. 2019. Pea genomic selection for Italian environments. BMC genomics, 20(1): 1-18.

Arganda-Carreras, I., Kaynig, V., Rueden, C., Eliceiri, K. W., Schindelin, J., Cardona, A., & Sebastian Seung, H. 2017. Trainable Weka Segmentation: a machine learning tool for microscopy pixel classification. Bioinformatics 33(15): 2424-2426.

Baritelle, A. L., & Hyde, G. M. 2003. Specific gravity and cultivar effects on potato tuber impact sensitivity. Postharvest Biology and Technology 29(3): 279-286.

Baritelle, A., Hyde, G., Thornton, R., & Bajema, R. 2000. A classification system for impact-related defects in potato tubers. American Journal of Potato Research 77(3): 143-148.

Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67(1): 1-48.

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. 2013. Evaluation measures for models assessment over imbalanced data sets. J Inf Eng Appl 3(10).

Bernardo, R., & Yu, J. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Science 47(3): 1082-1090.

Bethke, P. C., & Bussan, A. J. 2013. Acrylamide in processed potato products. American Journal of Potato Research 90(5): 403-424.

Binch, A., & Fox, C. W. 2017. Controlled comparison of machine vision algorithms for Rumex and Urtica detection in grassland. Computers and Electronics in Agriculture 140: 123-138.

Bock, C. H., Parker, P. E., Cook, A. Z., & Gottwald, T. R. 2008. Visual rating and the use of image analysis for assessing different symptoms of citrus canker on grapefruit leaves. Plant Disease 92(4): 530-541.

Bonar, N., Liney, M., Zhang, R., Austin, C., Dessoly, J., Davidson, D., ... & Hornyik, C. 2018. Potato miR828 is associated with purple tuber skin and flesh color. Frontiers in plant science 1742.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. 1992. A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory, 144-152.

Boydston, R. A., Navarre, D. A., Collins, H. P., & Chaves-Cordoba, B. 2018. The effect of vine kill method on vine kill, tuber skinning injury, tuber yield and size distribution, and tuber nutrients and phytonutrients in two potato cultivars grown for early potato production. American journal of potato research 95(1): 54-70.

Bradshaw, J. E., Hackett, C. A., Pande, B., Waugh, R., & Bryan, G. J. 2008. QTL mapping of yield, agronomic and quality traits in tetraploid potato (Solanum tuberosum subsp. tuberosum). Theoretical and Applied Genetics 116(2): 193-211.

Braun, S. 2013. Quantitative trait loci analysis and breeding for resistance to common scab in potato. Ph.D. Thesis, University of Wisconsin-Madison 133.

Braun, S., Gevens, A., Charkowski, A., Allen, C., & Jansky, S. 2017. Potato common scab: A review of the causal pathogens, management practices, varietal resistance screening methods, and host resistance. American Journal of Potato Research, 94(4): 283-296.

Breiman, L. 2001. Random forests. Machine learning 45(1): 5-32.

Brooke, R.C. 1996. Potato Bruising: How and Why Emphasizing Black Spot Bruise. Braun-Brumfield Inc, Ann Arbor, Michigan.

Brown, J., Caligari, P. D. S., Mackay, G. R., & Swan, G. E. L. 1984. The efficiency of seedling selection by visual preference in a potato breeding programme. The Journal of Agricultural Science 103(2): 339-346.

Buhrig, W., Thornton, M. K., Olsen, N., Morishita, D., & McIntosh, C. 2015. The influence of ethephon application timing and rate on plant growth, yield, tuber size distribution and skin color of red LaSoda potatoes. American journal of potato research 92(1): 100-108.

Byrne, S., Meade, F., Mesiti, F., Griffin, D., Kennedy, C., & Milbourne, D. 2020. Genome-wide association and genomic prediction for fry color in potato. Agronomy 10(1): 90.

Cabrera□Bosquet, L., Crossa, J., von Zitzewitz, J., Serret, M. D., & Luis Araus, J. 2012. High□throughput phenotyping and genomic selection: The frontiers of crop breeding converge F. Journal of integrative plant biology 54(5): 312-320.

Campos, H., & Ortiz, O. 2020. The potato crop: its agricultural, nutritional and social contribution to humankind. Springer Nature.

Caraza□Harter, M. V., & Endelman, J. B. 2020. Image□based phenotyping and genetic analysis of potato skin set and color. Crop Science 60(1): 202-210.

Caraza-Harter, M. V., & Endelman, J. B. 2022. The genetic architectures of vine and skin maturity in tetraploid potato. bioRxiv.

Carputo, D., Aversano, R., & Frusciante, L. 2004. Breeding potato for quality traits. Meeting of the Physiology Section of the European Association for Potato Research 684: 55-64.

Caruana, B.M, Pembleton, L.W, Constable, F, Rodoni, B, Slater, A.T, & Cogan, N.O.I. 2019. Validation of Genotyping by Sequencing Using Transcriptomics for Diversity and Application of Genomic Selection in Tetraploid Potato. Front Plant Sci 10.

Castleberry, H. C., & Jayanty, S. S. 2012. An experimental study of pressure flattening during long-term storage in four russet potato cultivars with differences in at-harvest tuber moisture loss. American journal of potato research 89(4): 269-276.

Castleberry, H. C., & Jayanty, S. S. 2017. Susceptibility to pressure flattening correlates with texture analysis of potato tubers. American Journal of Potato Research 94(5): 556-566.

Chen, X., Salamini, F., Gebhardt, C. 2001. A potato molecular-function map for carbohydrate metabolism and transport. Theor Appl Genet 102: 284–295.

Chubey, B. B., & Walkof, C. 1968. A rapid, non-destructive method for estimating chipping quality of potatoes. Canadian Journal of Plant Science 48(6): 636-637.

Chung, B., Armstrong, D., & Grice, S. 1988. Effect of irrigation frequency on the incidence of second growth of Russet Burbank potatoes in north-west Tasmania. The Journal of Agricultural Science 111(2): 399-402.

Clulow, S. A., McNicoll, J., & Bradshaw, J. E. 1995. Producing commercially attractive, uniform true potato seed progenies: the influence of breeding scheme and parental genotype. Theoretical and Applied Genetics 90(3): 519-525.

Cubero, S., Aleixos, N., Moltó, E., Gómez-Sanchis, J., & Blasco, J. 2011. Advances in machine vision applications for automatic inspection and quality evaluation of fruits and vegetables. Food and bioprocess technology 4(4): 487-504.

da Silva Pereira, G., Mollinari, M., Schumann, M. J., Clough, M. E., Zeng, Z. B., & Yencho, G. C. 2021. The recombination landscape and multiple QTL mapping in a Solanum tuberosum cv.'Atlantic'-derived F1 population. Heredity 126(5): 817-830.

Dacal-Nieto, A., Formella, A., Carrión, P., Vazquez-Fernandez, E., & Fernández-Delgado, M. 2011. Common scab detection on potatoes using an infrared hyperspectral imaging system. In International conference on image analysis and processing 303-312. Springer, Berlin, Heidelberg.

De Jong, H. 1981. Inheritance of russeting in cultivated diploid potatoes. Potato Research 24(3): 309-313.

De Jong, H., & Burns, V. J. 1993. Inheritance of tuber shape in cultivated diploid potatoes. American Potato Journal 70(3): 267-284.

De Wilde, T., De Meulenaer, B., Mestdagh, F., Govaert, Y., Vandeburie, S., Ooghe, W., ... & Verhé, R. 2005. Influence of storage practices on acrylamide formation during potato frying. Journal of Agricultural and Food Chemistry 53(16): 6550-6557.

Dean, B. B., Jackowiak, N., Nagle, M., Pavek, J., & Corsini, D. 1993. Blackspot pigment development of resistant and susceptibleSolanum tuberosum L. genotypes at harvest and during storage measured by three methods of evaluation. American Potato Journal 70(3): 201-217.

Dean, C. J., Knowles, L. O., & Richard Knowles, N. 2018. Auxin modulates gibberellin-induced effects on growth, yield, and raw product recovery for frozen processing in potato (Solanum tuberosum L.). American Journal of Potato Research 95(6): 622-641.

Dees, M. W., & Wanner, L. A. 2012. In search of better management of potato common scab. Potato research 55(3): 249-268.

Dhiman, B., Kumar, Y., & Kumar, M. 2022. Fruit quality evaluation using machine learning techniques: review, motivation and future perspectives. Multimedia Tools and Applications 1-23.

Douches, D. S., Maas, D., Jastrzebski, K., & Chase, R. W. 1996. Assessment of potato breeding progress in the USA over the last century. Crop Science 36(6): 1544-1552.

Driscoll, J., Coombs, J., Hammerschmidt, R., Kirk, W., Wanner, L., & Douches, D. 2009. Greenhouse and field nursery evaluation for potato common scab tolerance in a tetraploid population. American Journal of Potato Research 86(2): 96-101.

Duarte-Carvajalino, J. M., Alzate, D. F., Ramirez, A. A., Santa-Sepulveda, J. D., Fajardo-Rojas, A. E., & Soto-Suárez, M. 2018. Evaluating late blight severity in potato crops using unmanned aerial vehicles and machine learning algorithms. Remote Sensing 10(10): 1513.

Elzhov, T.V, Mullen, K.M, Spiess, A.N., and Bolker, B. 2016. minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds. R package version 1.2-1.

Enciso-Rodriguez, F, Douches, D, Lopez-Cruz, M, Coombs, J.J, & de los Campos, G. 2018. Genomic Selection for Late Blight and Common Scab Resistance in Tetraploid Potato (Solanum tuberosum). G3 Genes|Genomes|Genetics 8:2471–2481.

Endelman J.B, Schmitz Carley C.A, Bethke P.C, Coombs, J.J, Clough, M.E, da Silva, W.L, De Jong, W.S, Douches, D.S, Frederick, C.M, Haynes, K.G, Holm, D.G, Miller, J.C, Munoz, P.R, Navarro, F.M, Novy, R.G, Palta, J.P, Porter, G.A, Rak, K.T, Sathuvalli, V.R, Thompson, A.L & Yencho, G.C. 2018. Genetic Variance Partitioning and Genome-Wide Prediction with Allele Dosage Information in Autotetraploid Potato. Genetics 209:77–87.

FAO. 2021. World Food and Agriculture – Statistical Yearbook 2021. Rome.

Ferentinos, K. P. 2018. Deep learning models for plant disease detection and diagnosis. Computers and electronics in agriculture 145: 311-318.

Fix, E., & Hodges, J. L. 1989. Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique 57(3): 238-247.

Flinn, B., Rothwell, C., Griffiths, R., Lägue, M., DeKoeyer, D., Sardana, R., ... & Regan, S. 2005. Potato expressed sequence tag generation and analysis using standard and unique cDNA libraries. Plant molecular biology 59(3): 407-433.

Gemenet DC, Lindqvist-Kreuze H, De Boeck B, Pereira, G.S, Mollinari, M, Zeng, Z.B, Yencho, G.C, & Campos, H. 2020. Sequencing depth and genotype quality: accuracy and breeding operation considerations for genomic selection applications in autopolyploid crops. Theor Appl Genet 133: 3345–3363.

Ginzberg, I., Barel, G., Ophir, R., Tzin, E., Tanami, Z., Muddarangappa, T., ... & Fogelman, E. 2009. Transcriptomic profiling of heat-stress response in potato periderm. Journal of Experimental Botany 60(15): 4411-4421.

Ginzberg, I., Minz, D., Faingold, I., Soriano, S., Mints, M., Fogelman, E., ... & Yermiyahu, U. 2012. Calcium mitigated potato skin physiological disorder. American journal of potato research 89(5): 351-362.

Guo, K., Demory, B., Meah, S. Z., Zhai, T., Martinez, R., Islam, M. N., ... & Rodriguez-Saona, L. 2019. Non-destructive detection of acrylamide in potato fries with high-power supercontinuum lasers. In Fiber Lasers XVI: Technology and Systems 10897: 276-283. SPIE. San Francisco.

Habyarimana E, Parisi B, & Mandolino G. 2017. Genomic prediction for yields, processing and nutritional quality traits in cultivated potato (Solanum tuberosum L.). Plant Breeding, 136:245–252.

Halderson, J. L., & Henning, R. C. 1993. Measurements for determining potato tuber maturity. American Potato Journal, 70(2): 131-141.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. 2009. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter 11(1): 10-18.

Han, K. H., Sekikawa, M., Shimada, K. I., Hashimoto, M., Hashimoto, N., Noda, T., ... & Fukushima, M. 2006. Anthocyanin-rich purple potato flake extract has antioxidant capacity and improves antioxidant potential in rats. British Journal of Nutrition 96(6): 1125-1134.

Hassankhani, R., and H. Navid. 2012. Potato sorting based on size and color in machine vision system. Journal of Agricultural Science 4: 235–244.

Haynes, K. G., Wanner, L. A., Thill, C. A., Bradeen, J. M., Miller, J., Novy, R. G., ... & Vinyard, B. T. 2010. Common scab trials of potato varieties and advanced selections at three US locations. American Journal of Potato Research 87(3): 261-276.

Heineck, G. C., McNish, I. G., Jungers, J. M., Gilbert, E., & Watkins, E. 2019. Using R□ Based Image Analysis to Quantify Rusts on Perennial Ryegrass. The Plant Phenome Journal 2(1): 1-10.

Hosaka, K., Matsunaga, H., & Senda, K. 2000. Evaluation of several wild tuber-bearing Solanum species for scab resistance. American journal of potato research 77(1): 41-45.

Hyde, R. B., & Shewfelt, A. L. 1960. Measurement of chipping qualities in Manitoba-grown potatoes. Canadian Journal of Plant Science 40(4): 607-610.

International Commission on Illumination. 2019. Colorimetry – Part 4: CIE 1976 L*A*B* Colour space. International Commission on Illumination.

Jahns, G., Nielsen, H. M., & Paul, W. 2001. Measuring image analysis attributes and modelling fuzzy consumer aspects for tomato quality grading. Computers and Electronics in Agriculture 31(1): 17-29.

Jansen, G., & Flamme, W. 2006. Coloured potatoes (Solanum tuberosum L.)–anthocyanin content and tuber quality. Genetic Resources and Crop Evolution 53(7): 1321-1331.

Jarquín, D., Kocak, K., Posadas, L., Hyma, K., Jedlicka, J., Graef, G., & Lorenz, A. 2014. Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC genomics 15(1): 1-10.

Jemison Jr, J. M., Sexton, P., & Camire, M. E. 2008. Factors influencing consumer preference of fresh potato varieties in Maine. American Journal of Potato Research 85(2): 140-149.

Ji, Y., Sun, L., Li, Y., & Ye, D. 2019. Detection of bruised potatoes using hyperspectral imaging technique based on discrete wavelet transform. Infrared Physics & Technology 103: 103054.

Johansen, R. H., Farnsworth, B., Nelson, D. C., Secor, G. A., Gudmestad, N., & Orr, P. H. 1988. Russet Norkotah: A new russet-skinned potato cultivar with wide adaptation. American Potato Journal 65(10): 597-604.

Jones, C. R., Michaels, T. E., Schmitz Carley, C., Rosen, C. J., & Shannon, L. M. 2021. Nitrogen uptake and utilization in advanced fresh□market red potato breeding lines. Crop Science, 61(2): 878-895.

Jordan, M. I., & Mitchell, T. M. 2015. Machine learning: Trends, perspectives, and prospects. Science 349(6245): 255-260.

Jung, C. S., Griffiths, H. M., De Jong, D. M., Cheng, S., Bodis, M., Kim, T. S., & De Jong, W. S. 2009. The potato developer (D) locus encodes an R2R3 MYB transcription factor that regulates expression of multiple anthocyanin structural genes in tuber skin. Theoretical and Applied Genetics 120(1): 45-57.

Kaack, K., Larsen, E., & Thybo, A. K. 2002. The influence of mechanical impact and storage conditions on subsurface hardening in pre-peeled potatoes (Solanum tuberosum L.). Potato research 45(1): 1-8.

Kaur, A., Singh, N., & Ezekiel, R. 2008. Quality parameters of potato chips from different potato cultivars: Effect of prior storage and frying temperatures. International journal of food properties 11(4): 791-803.

Keijbets, M. J. H. 2008. Potato processing for the consumer: developments and future challenges. Potato Research 51(3): 271-281.

Kelderman, K. T. 2017. Pressure Flattening of Potatoes During Bulk Storage.

Keren-Keiserman, A., Baghel, R. S., Fogelman, E., Faingold, I., Zig, U., Yermiyahu, U., & Ginzberg, I. 2019. Effects of polyhalite fertilization on skin quality of potato tuber. Frontiers in plant science 1379.

Koch, M., Naumann, M., & Pawelzik, E. 2019. Cracking and fracture properties of potato (Solanum tuberosum L.) tubers and their relation to dry matter, starch, and mineral distribution. Journal of the Science of Food and Agriculture 99(6): 3149-3156.

Krupek, F. S., Dittmar, P. J., Sargent, S. A., Zotarelli, L., & Rowland, D. 2021. Impact of early potato desiccation method on crop growth, skinning injury, and storage quality maintenance. American Journal of Potato Research 98(3): 218-231.

Kuehni, R. G. 2001. Color space and its divisions. Color Research & Application 26(3): 209-222.

Lærke, P. E., Christiansen, J., & Veierskov, B. 2002. Colour of blackspot bruises in potato tubers during growth and storage compared to their discolouration potential. Postharvest biology and technology 26(1): 99-111.

LeCun, Y., Bengio, Y., & Hinton, G. 2015. Deep learning. Nature 521(7553): 436-444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. Neural computation 1(4): 541-551.

Li, L., Zhang, Q., & Huang, D. 2014. A review of imaging techniques for plant phenotyping. Sensors 14(11): 20078-20111.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. 2018. Machine learning in agriculture: A review. Sensors 18(8): 2674.

Lindqvist-Kreuze, H., Khan, A., Salas, E., Meiyalaghan, S., Thomson, S., Gomez, R., & Bonierbale, M. 2015. Tuber shape and eye depth variation in a diploid family of Andean potatoes. BMC genetics 16(1): 1-10.

Liu, Y., Lin-Wang, K., Espley, R. V., Wang, L., Li, Y., Liu, Z., Zhou, P., Zeng, L., Zhang, X., Zhang, J. & Allan, A. C. 2019. StMYB44 negatively regulates anthocyanin biosynthesis at high temperatures in tuber flesh of potato. Journal of experimental botany 70(15): 3809-3824.

López-Maestresalas, A., Keresztes, J. C., Goodarzi, M., Arazuri, S., Jarén, C., & Saeys, W. 2016. Non-destructive detection of blackspot in potatoes by Vis-NIR and SWIR hyperspectral imaging. Food control 70: 229-241.

Lorenz, A. J., Smith, K. P., & Jannink, J. L. 2012. Potential and optimization of genomic selection for Fusarium head blight resistance in six□row barley. Crop science 52(4): 1609-1621

Love, S. L., Werner, B. K., & Pavek, J. J. 1997. Selection for individual traits in the early generations of a potato breeding program dedicated to producing cultivars with tubers having long shape and russet skin. American Potato Journal 74(3): 199-213.

Lulai E., Freeman T.P. 2001. The importance of phellogen cells and their structural characteristics in susceptibility and resistance to excoriation in immature and mature potato tuber (Solanum tuberosum L.) periderm. Annuals of Botany 88: 555-561.

Lulai, E. C., & Corsini, D. L. 1998. Differential deposition of suberin phenolic and aliphatic domains and their roles in resistance to infection during potato tuber (Solanum tuberosumL.) wound-healing. Physiological and Molecular Plant Pathology 53(4): 209-222.

Lulai, E. C., & Orr, P. H. 1993. Determining the feasibility of measuring genotypic differences in skin-set. American Potato Journal, 70(8): 599-610.

Lulai, E. C., Glynn, M. T., & Orr, P. H. 1996. Cellular changes and physiological responses to tuber pressure-bruising. American potato journal, 73(5): 197-209.

Lynch, D. R., Kawchuk, L. M., Yada, R., & Armstrong, J. D. 2003. Inheritance of the response of fry color to low temperature storage. American journal of potato research 80(5): 341-344

Madabhushi, A., & Lee, G. 2016. Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical image analysis 33: 170-175.

Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., & Fritschi, F. B. 2020. Soybean yield prediction from UAV using multimodal data fusion and deep learning. Remote sensing of environment 237: 111599.

Marée, R., Geurts, P., Piater, J., & Wehenkel, L. 2005. Random subwindows for robust image classification. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1: 34-40. IEEE. San Diego, California.

Marino, S., Beauseroy, P., & Smolarz, A. 2019. Deep Learning-based Method for Classifying and Localizing Potato Blemishes. ICPRAM, 107-117.

Marique, T., Pennincx, S., & Kharoubi, A. 2005. Image segmentation and bruise identification on potatoes using a Kohonen's self□organizing map. Journal of food science 70(7): e415-e417.

McCord, P. H., Sosinski, B. R., Haynes, K. G., Clough, M. E., & Yencho, G. C. 2011. Linkage mapping and QTL analysis of agronomic traits in tetraploid potato (Solanum tuberosum subsp. tuberosum). Crop science 51(2): 771-785.

Meirmans, P. G., Liu, S., & van Tienderen, P. H. 2018. The analysis of polyploid genetic data. Journal of Heredity 109(3): 283-296.

Menesatti, P., Angelini, C., Pallottino, F., Antonucci, F., Aguzzi, J., & Costa, C. 2012. RGB color calibration for quantitative image analysis: The “3D Thin-Plate Spline” warping approach. Sensors 12(6): 7063-7079.

method for potato inspection using machine vision. Computers & Mathematics with Applications 63(1): 268-279.

Ming, W., Du, J., Shen, D., Zhang, Z., Li, X., Ma, J. R., Wang, F., & Ma, J. 2018. Visual detection of sprouting in potatoes using ensemble□based classifier. Journal of food process engineering 41(3): e12667.

Moreda, G. P., Muñoz, M. A., Ruiz-Altisent, M., & Perdigones, A. 2012. Shape determination of horticultural produce using two-dimensional computer vision–A review. Journal of Food Engineering 108(2): 245-261.

Moshou, D., Bravo, C., Wahlen, S., West, J., McCartney, A., De Baerdemaeker, J., & Ramon, H. 2006. Simultaneous identification of plant stresses and diseases in arable crops using proximal optical sensing and self-organising maps. Precision Agriculture 7(3): 149-164.

Muthukumarappan, K., Gunasekaran, S., Buelow, F. H., & Curwen, D. 1994. Investigations on potato storage management for control of pressure bruising. In American Society of Agricultural Engineers Meeting (USA).

Neilson, J. A., Smith, A. M., Mesina, L., Vivian, R., Smienk, S., & De Koyer, D. 2021. Potato Tuber Shape Phenotyping Using RGB Imaging. Agronomy 11(9): 1781.

Neubauer, J. D., Lulai, E. C., Thompson, A. L., Suttle, J. C., Bolton, M. D., & Campbell, L. G. 2013. Molecular and cytological aspects of native periderm maturation in potato tubers. Journal of Plant Physiology 170(4): 413-423.

Noordam, J. C., Otten, G. W., Timmermans, T. J., & van Zwol, B. H. 2000. High-speed potato grading and quality inspection based on a color vision system. In Machine Vision Applications in Industrial Inspection VIII 3966: 206-217). International Society for Optics and Photonics. San Jose, California.

Noordam, J. C., van den Broek, W. H., & Buydens, L. M. 2005. Detection and classification of latent defects and diseases on raw French fries with multispectral imaging. Journal of the Science of Food and Agriculture 85(13): 2249-2259.

Novy, R. G., Whitworth, J. L., Stark, J. C., Love, S. L., Corsini, D. L., Pavek, J. J., ... & Olsen, N. 2008. Premier Russet: A dual-purpose, potato cultivar with significant resistance to low temperature sweetening during long-term storage. American Journal of Potato Research 85(3): 198-209.

Oberholzer, M., Östreicher, M., Christen, H., & Brühlmann, M. 1996. Methods in quantitative image analysis. Histochemistry and cell biology 105(5): 333-355.

Oppenheim, D., Shani, G., Erlich, O., & Tsror, L. 2019. Using deep learning for image-based potato tuber disease detection. Phytopathology 109(6): 1083-1087.

Panigrahi, S., Wiesenborn, D., Schaper, L., & Bierwagen, G. 1996. Spectral reflectance properties of French fries. Applied Engineering in Agriculture 12(6): 721-724.

Parker, S. R., Shaw, M. W., & Royle, D. J. 1995. The reliability of visual estimates of disease severity on cereal leaves. Plant Pathology 44(5): 856-864.

Patel, K. K., Kar, A., & Khan, M. A. 2020. Development and an application of computer vision system for nondestructive physical characterization of mangoes. Agricultural Research, 9(1): 109-124.

Patel, K. K., Kar, A., Jha, S. N., & Khan, M. A. 2012. Machine vision system: a tool for quality inspection of food and agricultural products. Journal of food science and technology 49(2): 123-141.

Paterson, A. H. 2005. Polyploidy, evolutionary opportunity, and crop adaptation. Genetics of adaptation 191-196.

Pau, G., Fuchs, F., Sklyar, O., Boutros, M., and Huber, W. 2010. EBImage - an R package for image processing with applications to cellular phenotypes. Bioinformatics 26(7): 979-981.

Pavek, J. J., Corsini, D. L., Love, S. L., Hane, D. C., Holm, D. G., Iritani, W. M., ... & Thornton, R. E. 1992. Ranger Russet: A long russet potato variety for processing and fresh market with improved quality, disease resistance, and yield. American Potato Journal 69(8): 483-488.

Pedreschi, F., Kaack, K., & Granby, K. 2006a. Acrylamide content and color development in fried potato strips. Food Research International 39(1): 40-46.

Pedreschi, F., León, J., Mery, D., & Moyano, P. 2006b. Development of a computer vision system to measure the color of potato chips. Food Research International 39(10): 1092-1098.

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J. M., Crossa, J., Manès, Y., & Dreisigacker, S. 2012. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3: Genes| Genomes| Genetics 2(12): 1595-1605.

Poland, J. A., & Nelson, R. J. 2011. In the eye of the beholder: the effect of rater variability and different rating scales on QTL mapping. Phytopathology 101(2): 290-298.

Praeger, U., Surdilovic, J., Truppel, I., Herold, B., & Geyer, M. 2013. Comparison of electronic fruits for impact detection on a laboratory scale. Sensors 13(6): 7140-7155.

Prashar, A., Hornyik, C., Young, V., McLean, K., Sharma, S. K., Dale, M. F. B., & Bryan, G. J. 2014. Construction of a dense SNP map of a highly heterozygous diploid potato population and QTL analysis of tuber shape and eye depth. Theoretical and Applied Genetics 127(10): 2159-2171.

Quinlan, J.R. 1993. Programs for machine learning. Morgan Kaufmann, San Mateo, California.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/. Accessed 8 February 2022.

Raimo, F., Pentangelo, A., Pane, C., Parisi, B., & Mandolino, G. 2018. Relationships between internal brown spot and skin roughness in potato tubers under field conditions. Potato Research 61(4): 327-339.

Rak, K., Bethke, P. C., & Palta, J. P. 2017. QTL mapping of potato chip color and tuber traits within an autotetraploid family. Molecular breeding 37(2): 1-15.

Razmjooy, N., Mousavi, B. S., & Soleymani, F. 2012. A real-time mathematical computer method for potato inspection using machine vision. Computers & Mathematics with Applications 63(1): 268-279.

Reeves, A. F. 1988. Varietal differences in potato tuber greening. American Potato Journal 65(11): 651-658.

Richael, C. M. 2021. Development of the genetically modified Innate® potato. Plant Breeding Reviews 44: 57-78.

Roe, M. R., Carlson, J. L., McManimon, T. M., Hegeman, A. D., & Tong, C. 2014. Differential accumulation and degradation of anthocyanins in Red Norland periderm is dependent on soil type and tuber storage duration. American journal of potato research 91(6): 696-705.

Scanlon, M. G., Roller, R., Mazza, G., & Pritchard, M. K. 1994. Computerized video image analysis to quantify color of potato chips. American Potato Journal 71(11): 717-733.

Schaper, L. A., & Yaeger, E. C. 1982. Horizontal and vertical pressure patterns of stored potatoes. Transactions of the ASAE 25(3): 719-0724.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., ... & Cardona, A. 2012. Fiji: an open-source platform for biological-image analysis. Nature methods 9(7): 676-682.

Schlager, S. 2017. Morpho and Rvcg - Shape Analysis in R. In Statistical Shape and Deformation Analysis, ed. Zheng, G., Li, S., Szekely, G. 217-256. Academic Press.

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. 2012. NIH Image to ImageJ: 25 years of image analysis. Nature Methods 9(7): 671–675.

Segarra, J., Buchaillot, M. L., Araus, J. L., & Kefauver, S. C. 2020. Remote sensing for precision agriculture: Sentinel-2 improved features and applications. Agronomy 10(5): 641.

Segnini, S., Dejmek, P., & Öste, R. 1999. A low cost video technique for colour measurement of potato chips. LWT-Food Science and Technology 32(4): 216-222.

Selga C, Koc A, Chawade A, & Ortiz R. 2021. A Bioinformatics Pipeline to Identify a Subset of SNPs for Genomics-Assisted Potato Breeding. Plants 10:30.

Shetty, K. K., Dwelle, R. B., Fellman, J. K., & Patterson, M. E. 1991. Blackspot injury of film-wrapped potatoes in relation to bruising and respiratory gases. Potato research 34(3): 253-260.

Shock, C. C., Stieber, T. D., Zalewski, J. C., Eldredge, E. P., & Lewis, M. D. 1994. Potato tuber stem-end fry color determination. American potato journal 71(2): 77-88.

Si, Y., Sankaran, S., Knowles, N. R., & Pavek, M. J. 2017. Potato tuber length-width ratio assessment using image analysis. American Journal of Potato Research 94(1): 88-93.

Si, Y., Sankaran, S., Knowles, N. R., & Pavek, M. J. 2018. Image-based automated potato tuber shape evaluation. Journal of Food Measurement and Characterization 12(2): 702-709.

Siano, A., Kerckhoffs, L. H. J., Roskruge, N., & Sofkova-Bobcheva, S. 2018. Yield and tuber quality variability in commercial potato cultivars under abiotic stress in New Zealand. Massey University. Palmerston North, New Zealand.

Singh, A., Halgamuge, M. N., & Lakshmiganthan, R. 2017. Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms. International Journal of Advanced Computer Science and Applications 8(12).

Slater, A. T., Wilson, G. M., Cogan, N. O., Forster, J. W., & Hayes, B. J. 2014. Improving the analysis of low heritability complex traits for enhanced genetic gain in potato. Theoretical and applied genetics 127(4): 809-820.

Śliwka, J., Wasilewicz□Flis, I., Jakuczun, H., & Gebhardt, C. 2008. Tagging quantitative trait loci for dormancy, tuber shape, regularity of tuber shape, eye depth and flesh colour in diploid potato originated from six Solanum species. Plant Breeding 127(1): 49-55.

Sood S, Lin Z, Caruana B, Slater, A.T, & Daetwyler, H.D. 2020. Making the most of all data: Combining non-genotyped and genotyped potato individuals with HBLUP. The Plant Genome 13: e20056.

Sowokinos, J. R. 2001. Biochemical and molecular control of cold-induced sweetening in potatoes. American Journal of Potato Research 78(3): 221-236.

Stefaniak, T. R., Fitzcollins, S., Figueroa, R., Thompson, A. L., Schmitz Carley, C.A, & Shannon, L. M. 2021. Genotype and Variable Nitrogen Effects on Tuber Yield and Quality for Red Fresh Market Potatoes in Minnesota. Agronomy 11(2): 255.

Stich, B & Van Inghelandt, D. 2018. Prospects and Potential Uses of Genomic Prediction of Key Performance Traits in Tetraploid Potato. Front Plant Sci 9: 159.

Su, W. H., Zhang, J., Yang, C., Page, R., Szinyei, T., Hirsch, C. D., & Steffenson, B. J. 2020. Automatic evaluation of wheat resistance to fusarium head blight using dual mask-RCNN deep learning frameworks in computer vision. Remote sensing 13(1): 26.

Sung, K. K. 1996. Learning and example selection for object and pattern detection. Massachusetts Institute of Technology.

Sverrisdóttir, E, Byrne, S, Sundmark, E.H.R, Johnsen, H.O, Kirk, H.G, Asp, T, Janss, L, & Nielsen, K.L. 2017. Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing. Theor Appl Genet 130:2091–2108.

Sverrisdóttir, E., Sundmark, E. H. R., Johnsen, H. Ø., Kirk, H. G., Asp, T., Janss, L., ... & Nielsen, K. L. 2018. The value of expanding the training population to improve genomic selection models in tetraploid potato. Frontiers in plant science 9: 1118.

Tao, Y., Heinemann, P. H., Varghese, Z., Morrow, C. T., & Sommer Iii, H. J. 1995a. Machine vision for color inspection of potatoes and apples. Transactions of the ASAE 38(5): 1555-1561.

Tao, Y., Morrow, C. T., Heinemann, P. H., & Sommer III, H. J. 1995b. Fourier-based separation technique for shape grading of potatoes using machine vision. Transactions of the ASAE 38(3): 949-957.

Thornton, M. K., Lee, J., John, R., Olsen, N. L., & Navarre, D. A. 2013. Influence of growth regulators on plant growth, yield, and skin color of specialty potatoes. American journal of potato research 90(3): 271-283.

Topcu, Y., Sapkota, M., Illa-Berenguer, E., Nambeesan, S. U., & van der Knaap, E. 2021. Identification of blossom-end rot loci using joint QTL-seq and linkage-based QTL mapping in tomato. Theoretical and Applied Genetics 134(9): 2931-2945.

Uitdewilligen, J. G., Wolters, A. M. A., D'hoop, B. B., Borm, T. J., Visser, R. G., & Van Eck, H. J. 2013. A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PloS one 8(5): e62355.

Urbany, C., Colby, T., Stich, B., Schmidt, L., Schmidt, J., & Gebhardt, C. 2012. Analysis of natural variation of the potato tuber proteome reveals novel candidate genes for tuber bruising. Journal of Proteome Research 11(2): 703-716.

Urbany, C., Stich, B., Schmidt, L., Simon, L., Berding, H., Junghans, H., ... & Gebhardt, C. 2011. Association genetics in Solanum tuberosum provides new insights into potato tuber bruising and enzymatic tissue discoloration. BMC genomics 12(1): 1-14.

USDA Plant Variety Protection Office. 2015. USDA form OMB NO 0581-0055. USDA-AMS. https://www.ams.usda.gov/resources/st470-potato. Accessed 8 February 2022.

Van der Werff, H. M. A., & Van der Meer, F. D. 2008. Shape-based classification of spectrally identical objects. ISPRS Journal of Photogrammetry and Remote Sensing 63(2): 251-258.

Van Eck, H. J., Jacobs, J. M., Stam, P., Ton, J., Stiekema, W. J., & Jacobsen, E. 1994. Multiple alleles for tuber shape in diploid potato detected by qualitative and quantitative genetic analysis using RFLPs. Genetics 137(1): 303-309.

Voorrips, R. E., Gort, G., & Vosman, B. 2011. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC bioinformatics 12(1): 1-11.

Vulavala, V. K., Fogelman, E., Faigenboim, A., Shoseyov, O., & Ginzberg, I. 2019. The transcriptome of potato tuber phellogen reveals cellular functions of cork cambium and genes involved in periderm formation and maturation. Scientific reports 9(1): 1-14.

Wanner, L. A. 2006. A survey of genetic variation in Streptomyces isolates causing potato common scab in the United States. Phytopathology 96(12): 1363-1371.

Wanner, L. A., & Haynes, K. G. 2009. Aggressiveness of Streptomyces on four potato cultivars and implications for common scab resistance breeding. American Journal of Potato Research 86(5): 335-346.

Weiss, M., Jacob, F., & Duveiller, G. 2020. Remote sensing for agricultural applications: A meta-review. Remote Sensing of Environment 236: 111402.

Wilson, C. R. 2001. Variability within clones of potato cv. Russet Burbank to infection and severity of common scab disease of potato. Journal of Phytopathology 149(10): 625-628.

Work, T. M., Kezis, A. S., & True, R. H. 1981. Factors determining potato chipping quality 103. Life Sciences and Agriculture Experiment Station, University of Maine at Orono.

Wu, S. S., Wu, R., Ma, C. X., Zeng, Z. B., Yang, M. C., & Casella, G. 2001. A multivalent pairing model of linkage analysis in autotetraploids. Genetics 159(3): 1339-1350.

Ye, D., Sun, L., Tan, W., Che, W., & Yang, M. 2018. Detecting and classifying minor bruised potato based on hyperspectral imaging. Chemometrics and Intelligent Laboratory Systems 177: 129-139.

Yeatman, J. N., & Aulenbach, B. B. 1973. Relationship of instrumental measurements to visual impressions of potato chip color. Sensory Evaluation of Appearance of Materials 545: 128.

Zhang, Y., Jung, C. S., & De Jong, W. S. 2009. Genetic analysis of pigmented tuber flesh in potato. Theoretical and Applied Genetics 119(1): 143-150.

Zhou, L., Chalana, V., & Kim, Y. 1998. PC□based machine vision system for real□time computer□aided potato inspection. International journal of imaging systems and technology, 9(6): 423-433.

Zhu, X., Richael, C., Chamberlain, P., Busse, J. S., Bussan, A. J., Jiang, J., & Bethke, P. C. 2014. Vacuolar invertase gene silencing in potato (Solanum tuberosum L.) improves processing quality by decreasing the frequency of sugar-end defects. PloS one 9(4): e93381.

Zorrilla, C., Navarro, F., Vega, S., Bamberg, J., & Palta, J. 2014. Identification and selection for tuber calcium, internal quality and pitted scab in segregating ‘Atlantic’x ‘Superior’reciprocal tetraploid populations. American journal of potato research 91(6): 673-687.

Appendices

Appendix A. Weka Knowledge Flow scripts for Random Forest model creation and testing.

Figures A1 and A2 are screenshots of scripts for creation of Random Forest models created using Weka Knowledge Flow (Hall et al., 2009). These scripts required only modification of the input file sources to create both the pressure bruise and skin finish models.

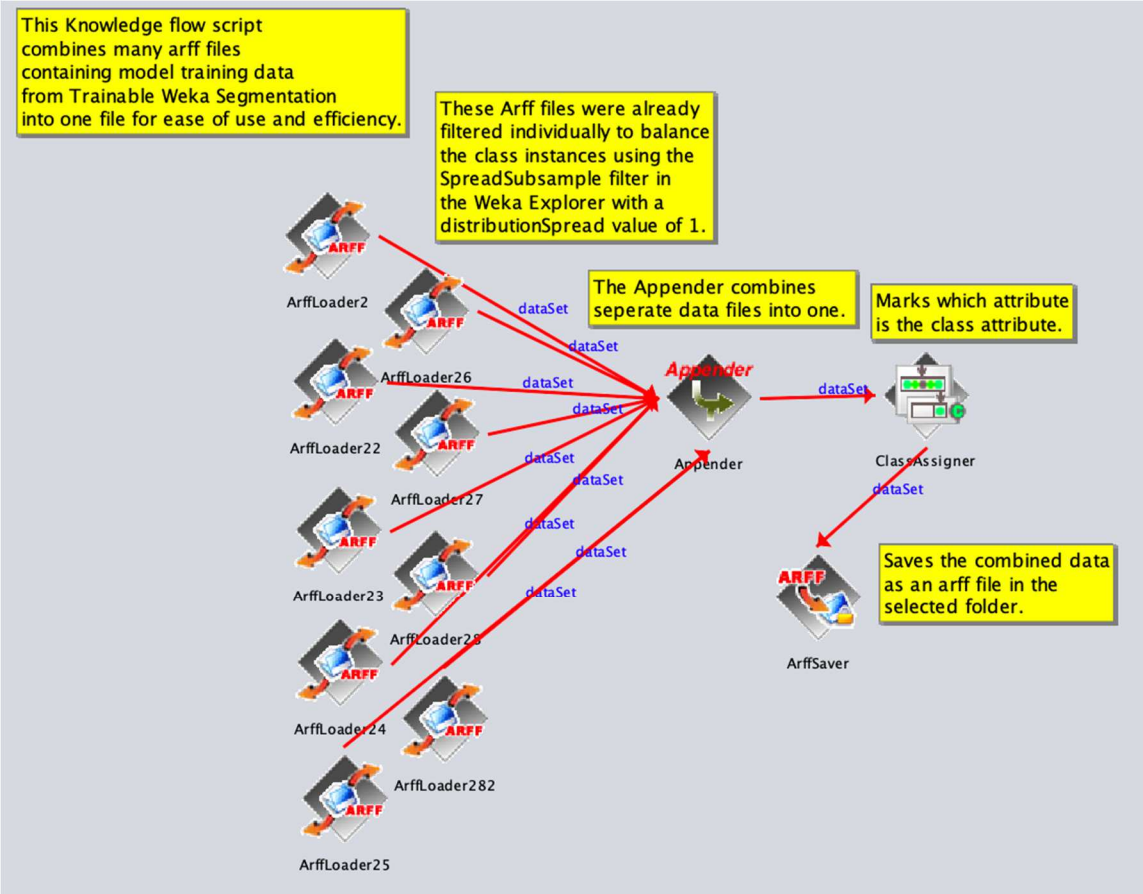

Figure A1. Screenshot of Weka Knowledge Flow script for merging image training data files to use in machine learning model creation into a single file.

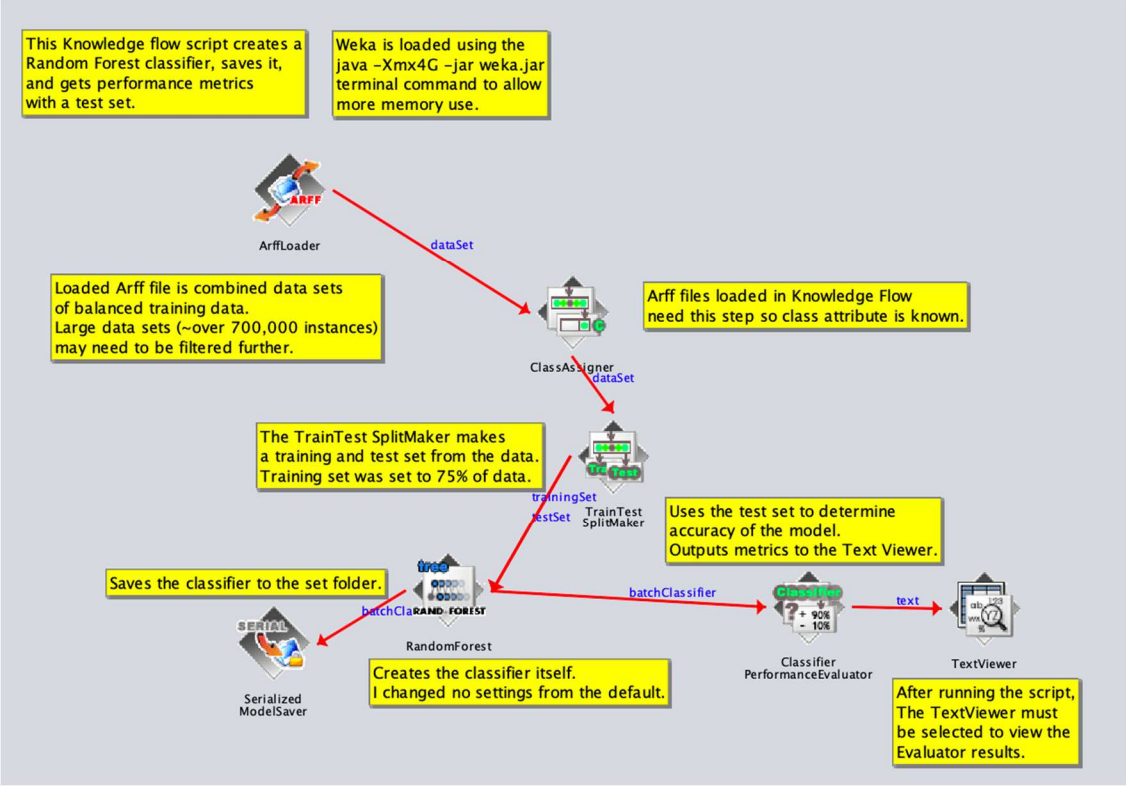

Figure A2. Screenshot of Weka Knowledge Flow script for creation of the Random Forest model using combined training data.

Appendix B. Retrospective on creation of an efficient, easy to learn, and easy to use program for quantitative analysis of complex potato tuber quality traits from RGB lightbox images.

Current approach:

The goal of this project was adding machine learning functionality to the existing TubAR package to measure quality traits including pressure bruise. No easy means for creating the necessary training data from image data in R presented itself in my early research, or at least not nearly as easy as the Trainable Weka Segmentation (TWS) plugin in FIJI/ImageJ2 (for simplicity I will call this application in any form ImageJ). However, TubAR was created in part to avoid scalability issues in ImageJ so going back to ImageJ did not seem like an option. Regardless, TWS seemed like the best way to create training data. The way TWS data was formatted seemed very difficult to move back to R and use effectively. The probability maps that could be created by models in TWS, however, could easily be analyzed in R using the EBImage package that TubAR already utilized. Given that classifier models could be created once manually in ImageJ or the Weka application then retrieved to make the probability maps, the R script would not have to be designed to make the model itself.

The focus then became how to operate TWS from R to get probability maps of desired image inputs using an existing classifier. Some packages for operating Java, Weka, or ImageJ from R exist but they seemed difficult to use compared to using the *system* function in R to control ImageJ from the OS terminal. Having R create an ImageJ script and tell the terminal to tell ImageJ to run said script seemed to be a promising solution. I was able to do this with an ImageJ macro script, but it opened all the normal windows in ImageJ that open when using the GUI manually, which seemed to be an extremely inelegant and resource inefficient solution. Headless mode in ImageJ is theoretically able to avoid this but I was unable to get TWS to work correctly in headless mode using the macro scripting language.

I later found a script written to create probability maps in TWS for a full folder of images without opening each image written in BeanShell (available at https://imagej.net/plugins/tws/scripting). Using this script in place of my macro script indeed avoided many windows from opening but still opened some ImageJ windows and worse yet required user input in a Java window. This seems avoidable but I lack the knowledge of Java or Beanshell to effectively alter this script to take the directories for the images from R.

To summarize, the ideal way the program, as currently imagined, would work is as follows:

1. The classifier model is created in Weka using data created in TWS. This doesn't involve R as this doesn't have to be scaled. Refer to Appendix C for creating the models.
2. An R function takes the location of the classifier and the images to be measured for a trait and inserts these into a boilerplate script that is sent to ImageJ using the OS terminal.
3. The script runs in ImageJ headless mode, not opening any windows and terminating itself once it's finished.

4. The EBImage package can then use the probability map TIFF files (either in the same folder as the originals or sent to a new folder) generated by TWS to measure things like percent of the tuber surface area affected by a trait or potentially things like how many pressure bruises are on the tubers in the image. This stage works more like how TubAR functions currently work.

Possible advances and alternatives:
Here is a list of possible future paths for the project in order of similarity to the current project and my opinion of the order in which they should be attempted:

1. Someone more knowledgeable about Java or Beanshell may be able to get the current system working much more cleanly and efficiently. If this is implemented, I would consider separating this project from TubAR as its own package.
2. Image analysis and machine learning tools exist in R that I have not fully explored and may allow either TWS to be avoided completely while creating good training data, or training data from TWS could be reformatted to be used to create models in R using a package such as randomForest.
3. PlantCV is often cited in the literature for similar tasks to this and seems to be worth further investigation. If we had not started with the TubAR project in R, I likely would have used Python for this reason. This is also the way that I know of to implement the image "patches" idea mentioned in Chapter 3.
4. Further investigation of ImageJ and discussion with scientists at the University of Wisconsin suggests that ImageJ scripts may be more capable of processing large scale image collections than was originally thought when the TubAR project was started. Running this program primarily from ImageJ seems possible, though Java still seems to present some issues with memory limitations and generally being harder to code in than R or Python.