Using a Genetic Algorithm with Histogram-Based Feature Selection in Hyperspectral Image Classification

Neil S. Walton
Gianforte School of Computing
Montana State University
Bozeman, Montana
neil.walton@student.montana.edu

John W. Sheppard
Gianforte School of Computing
Montana State University
Bozeman, Montana
john.sheppard@montana.edu

Joseph A. Shaw
Dept. Elec. & Comp. Engineering
Montana State University
Bozeman, Montana
joseph.shaw@montana.edu

ABSTRACT

Optical sensing has the potential to be an important tool in the automated monitoring of food quality. Specifically, hyperspectral imaging has enjoyed success in a variety of tasks ranging from plant species classification to ripeness evaluation in produce. Although effective, hyperspectral imaging is prohibitively expensive to deploy at scale in a retail setting. With this in mind, we develop a method to assist in designing a low-cost multispectral imager for produce monitoring by using a genetic algorithm (GA) that simultaneously selects a subset of informative wavelengths and identifies effective filter bandwidths for such an imager. Instead of selecting the single fittest member of the final population as our solution, we fit a univariate Gaussian mixture model to the histogram of the overall GA population, selecting the wavelengths associated with the peaks of the distributions as our solution. By evaluating the entire population, rather than a single solution, we are also able to specify filter bandwidths by calculating the standard deviations of the Gaussian distributions and computing the full-width at half-maximum values. In our experiments, we find that this novel histogram-based method for feature selection is effective when compared to both the standard GA and partial least squares discriminant analysis.

CCS CONCEPTS

• **Computing methodologies** → **Genetic algorithms**; • **Applied computing** → **Computer-aided design**;

KEYWORDS

Genetic algorithm, hyperspectral imaging, feature selection, histogram, produce monitoring

ACM Reference Format:
Neil S. Walton, John W. Sheppard, and Joseph A. Shaw. 2019. Using a Genetic Algorithm with Histogram-Based Feature Selection in Hyperspectral Image Classification. In *Genetic and Evolutionary Computation Conference (GECCO '19), July 13–17, 2019, Prague, Czech Republic.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3321707.3321748

GECCO '19, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6111-8/19/07…$15.00
https://doi.org/10.1145/3321707.3321748

1 INTRODUCTION

Every year in the United States, more than 43 billion pounds of fruits and vegetables are thrown away before they ever make it onto a plate [9]. This equates roughly to a 30 percent rate of loss for post-harvest produce, amounting to nearly 50 billion dollars a year in lost produce at the retail and consumer levels. These losses result from a variety of factors, including mechanical injury, bruising, sprout growth, rot, secondary infection, biological aging, and over-ripening [9, 15, 18]. These concrete characteristics all inform the more nebulous concept of overall produce quality.

This judgment of quality is innately subjective when the evaluation is carried out by a human; sensory preferences for color, texture, smell, and taste vary from person to person. Because of this variability, instrumental measurements are often preferred to sensory judgments when it comes to monitoring food quality. Methods such as mass spectrometry and high performance liquid chromatography are used in food monitoring, but both require sample destruction during analysis [16]. This means that only a representative sample of the produce is tested, which can give insights into the average quality of the produce being monitored, but it fails to capture the produce-specific characteristics necessary to perform tasks such as classification and sorting [1].

Several non-destructive techniques for food quality monitoring exist. One such method is hyperspectral imaging. Hyperspectral imaging combines the spatial information provided by conventional imaging and the spectral information captured by spectroscopy [16]. One advantage of hyperspectral images is that they contain a vast amount of information. The corresponding disadvantage of hyperspectral images is that they contain a vast amount of information. That is to say, the wealth of data provided by this technology can help lend valuable insight into a variety of problems; however, due to the curse of dimensionality, many standard processing techniques quickly become impractical. As a brief illustration, a single 1000×1000 pixel image taken by an imager with a 600 nm spectral range and a 2 nm spectral resolution results in a 300 million point data cube. Because of this, hyperspectral imaging has been a prime candidate for dimensionality reduction techniques.

In this study, we examine the effects of feature selection on hyperspectral image classification. We capture hyperspectral images of avocados and tomatoes and use the data to classify the produce as "fresh" versus "old", and also consider the highly cited Indian Pines dataset [6]. For hyperspectral imaging, the feature space from which a subset of features is selected comprises the set of wavelengths at which reflectance is measured by the imager. By selecting an informative subset of wavelengths, noise, redundant information, and the size of the data cube can all be reduced significantly.

The feature selection process can also assist in the design of cheaper multispectral imagers.

A large variety of feature selection techniques have been applied to hyperspectral data. A hybrid feature subset selection algorithm that combines weighted feature filtering and the Fuzzy Imperialist Competitive algorithm [26] has been used successfully to reduce the classification error of hydrothermal alteration minerals in hyperspectral datasets [31]. In another study, ground cover classification of hyperspectral images is improved by selecting features using simulated annealing to maximize a joint objective of feature relevance and overall classification accuracy [25]. On the same datasets used in [25], Feng *et al.* develop an unsupervised feature selection technique that improves classification error by optimizing the maximum information and minimum redundancy criterion via the clonal selection optimization algorithm (MIMR-CSA) [13].

While a large amount of active research investigates new advanced feature selection methods, and many other feature selection techniques (such as forward selection [2], backward elimination [23], and random forest-based methods [2, 10]) exist, in hyperspectral applications, two of the most commonly utilized feature selection techniques are partial least squares discriminant analysis (PLS-DA) [5] and genetic algorithms (GAs) [36]. PLS-DA has been adapted for feature selection by utilizing the coefficients produced by the PLS-DA method in order to rank the features by importance (i.e., from largest to smallest coefficient). The top k features are then selected for use in the analysis. PLS-DA has been used in application areas ranging from differentiating between fresh and frozen-to-thawed meat [4], to predicting the chemical composition of lamb [19], as well as many others [11, 30, 32, 35]. Likewise, in recent years, GAs have been widely used for feature selection in hyperspectral data analysis [11, 14, 20, 38].

In each of the aforementioned studies, the goals of dimensionality reduction are largely limited to reducing noise, eliminating redundant information, improving accuracy for a given prediction task, and reducing the size of the problem to be analyzed. These studies make the assumption that, in application, a hyperspectral imager will be used to capture the full spectral response at each pixel, then the selected wavelengths will be extracted and passed through the given prediction algorithm. However, hyperspectral imagers are prohibitively expensive for mass deployment in most retail settings, often costing tens of thousands of dollars per imager.

As the main contribution of this study, we propose a new feature selection technique based on the standard GA to assist in multispectral imager design. After the GA has satisfied its stopping criterion, instead of selecting the fittest member of the final population as the solution, we use a histogram-based approach that analyzes the overall population, in a method we call the **H**istogram **A**ssisted **G**enetic Algorithm for **R**eduction **i**n **D**imensionality (HAGRID). Not only does this method offer a new way of determining the solution for a GA, but it also allows for the analysis of the distribution of selected features, which, in the context of wavelength selection for hyperspectral data, allows for the determination of filter bandwidths for a multispectral imager.

The rest of the paper is organized as follows — section 2 gives an overview of hyperspectral and multispectral imaging, section 3 covers the formulation of the GA used in this paper, section 4 provides details for the HAGRID method, section 5 discusses the experimental setup and methods, section 6 provides the experimental results, and section 7 ends with conclusions.

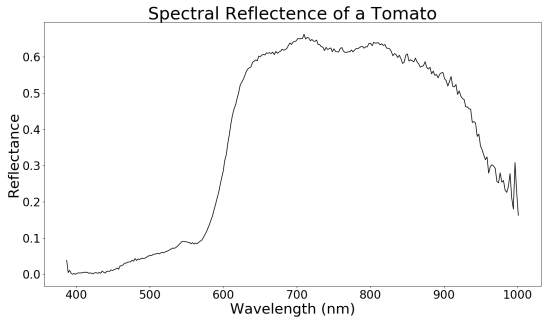


Figure 1: Sample spectral reflectance curve of a tomato.

2 HYPERSPECTRAL IMAGING

2.1 Overview

Hyperspectral imaging combines the two main components of conventional imaging and spectroscopy by capturing spatial and spectral information simultaneously [16]. The image produced by a hyperspectral imager can thus be thought of as a cube, consisting of two spatial dimensions and one spectral dimension. When incident light strikes an object, a percentage of that light is absorbed by the object, and a percentage is reflected off the surface [1]. When the percentages of light reflected at various wavelengths are measured, a spectral reflectance curve (Fig. 1) is produced. It is this spectral reflectance curve that defines the spectral dimension of a hyperspectral image. Hyperspectral imagers usually measure reflectance over a portion of the visible and near-infrared (NIR) spectrum, which covers wavelengths of light ranging from 400–2500 nanometers (nm).

Two main parameters inform the collection of spectral information for a hyperspectral imager. An imager has a spectral range and a spectral resolution. The spectral range dictates the range of wavelengths of light over which the imager is able to measure reflectance. The spectral resolution indicates the spacing between these measurements. For example, if an imager has a spectral range of 400-800 nm and a spectral resolution of 10 nm, the imager records the reflectance of light at 400 nm, 410 nm, all the way up to 800 nm, for each pixel in the spatial plane. It is worth noting that each reflectance measurement is centered around a wavelength determined by the spectral range and resolution, but the imager captures some response in a band around the wavelength center. As such, each individual reflectance measurement can be thought of as the integral of a Gaussian curve centered at a given wavelength, with spread proportional to the resolution of the imager.

2.2 Multispectral Imaging

Where hyperspectral imagers usually measure reflectance at hundreds of wavelengths of light, a multispectral imager takes these measurements at only a handful of wavelengths and therefore can

be a lot cheaper to manufacture and purchase. Multispectral imagers are also more flexible in terms of design and customization. They consist of a number of bandpass filters that each record the reflectance centered around a certain wavelength of light. Three main aspects of these filters can be customized – the number of filters included in the imager, the wavelength at which each filter is centered, and the bandwidth of each filter (where a larger bandwidth filter measures the reflectance over a larger range of wavelengths surrounding the filter center).

There are two main designs for multispectral imagers and both utilize bandpass filters. A bandpass filter allows for the transmission of light in a discrete spectral band [34]. These filters are centered at specific wavelengths of light and have fixed bandwidths. The first type of multispectral imager is known as a filter-wheel camera. This type of camera consists of a rotating wheel of bandpass filters that pass sequentially in front of the camera, allowing specific ranges of the spectrum to pass through to be measured by the camera [7]. The other main design utilizes multiple-bandpass filters. Instead of sequentially passing several filters in front of the camera, a multiple-bandpass filter comprises a single checkerboard pattern of microfilters. Each microfilter consists of a set configuration of bandpass filters, and these microfilters are tiled to create the larger multiple-bandpass filter. The amount of light transmitted through each bandpass filter in a given microfilter is measured and combined into a single pixel value, and these pixel values are combined across microfilters to create the entire multispectral image [34].

There can be a large amount of redundant information and noise present in a hyperspectral data cube. By intelligently selecting bandpass filters for a multispectral imager (either using domain knowledge or algorithmic feature selection), both the size of the data and the noise present in the data can be reduced greatly while still capturing the majority of the relevant information. Often, the wavelength centers for these filters are known *a priori* based on domain expert knowledge [21, 24]. Even so, algorithmic feature selection tends to do well in selecting relevant wavelength centers. Regardless of how the wavelengths are selected, the usual approach in designing a multispectral imager is to incorporate bandpass filters of standard width centered around these wavelengths (usually 10, 20, or 30 nm, though the bandwidths are customizable). While a large volume of literature explores methods for selecting the wavelength centers, very little work has been done in specifying the bandwidths of the filters algorithmically. Our proposed method seeks to accomplish both simultaneously.

2.3 Hyperspectral Produce Monitoring

Hyperspectral imaging has seen success in domains ranging from pharmaceuticals, to astronomy, to agriculture [16], but one prominent application areas is produce quality monitoring. A vast array of characteristics inform the concept of produce quality. Hyperspectral imaging has been able to help automate quality assurance, succeeding where manual inspections fail, reducing the processing time, and making the overall process cheaper for many quality monitoring tasks. While a comprehensive review of the various applications of hyperspectral imaging in produce monitoring is beyond the scope of this paper, the following studies offer a representative sample of the possibilities hyperspectral imaging offers.

In a 2006 study, Nicolai *et al.* were able to identify apple pit lesions that were invisible to the naked eye by applying PLS-DA to hyperspectral images of apples harvested from trees known to display bitter pit symptoms [27]. One interesting finding here is that the lesions could be identified with as few as two latent variables in the PLS model, indicating that a small portion of the spectrum can be sufficient to improve performance significantly for certain tasks. Serrant *et al.* were able to apply PLS-DA to hyperspectral images of grapevine leaves to identify *Peronospora* infection with a high degree of accuracy [33]. Polder *et al.* applied linear discriminant analysis (LDA) to hyperspectral images of tomatoes in order to assign the tomatoes to one of five ripeness stages [28]. The authors saw a significant improvement over the classification performance using RGB images, dropping the error rate from 51% to as low as 19% in some of their experiments. In a similar vein as [28], in this study, we investigate the impacts of feature selection on ripeness classification of avocados and tomatoes.

3 GENETIC ALGORITHM

In order to design a multispectral imager, (to borrow the phraseology of Michael Mahoney [22]) we need a set of wavelengths, not a set of eigenwavelengths. That is to say, we cannot design an imager that captures data for transformed subsets of wavelengths; an imager must measure reflectance at a subset of real wavelengths. As such, we must consider only feature selection techniques, rather than feature extraction techniques, when it comes to multispectral imager design. The genetic algorithm [17] is one such technique that can be utilized effectively for feature selection [36].

The individuals in our GA population are represented as integer arrays, where the integers represent a subset of indices corresponding to the wavelengths to be selected. We employ tournament selection, binonomial crossover, and generational replacement. For our mutation operator, if a gene (i.e., single wavelength index) is chosen for mutation, an integer is drawn randomly from the uniform distribution over $[-3, 3]$ and added to the index value. In this way, the mutation is restricted to adjacent wavelengths.

We use two different fitness functions for our experiments. Both use decision trees [29] to perform classification on the given datasets. For each member of the population, a decision tree is built using the subset of wavelengths represented by the individual. Ten-fold cross-validation is then performed on the given classification task using the decision tree, and the fitness score is the average classification accuracy attained across the ten folds.

In the first fitness function (*fitness1*), the fitness is simply equal to the classification accuracy obtained by the decision tree. In the second fitness function (*fitness2*), we make two alterations. The first is a dispersive force that adds a large penalty to solutions that select wavelengths within 20 nm of each other to encourage wavelength diversity. The second alteration aims to approximate an imager with a larger spectral resolution of 30 nm. To accomplish this, we bin the reflectance of wavelengths within 15 nm on either side of the selected wavelength center before feeding the data into the decision tree. The fitness is again equal to the classification accuracy attained by the decision tree.

4 HISTOGRAM-BASED APPROACH

4.1 Overview

Our proposed method changes only the determination of the solution after the GA has satisfied its stopping criterion and is therefore agnostic to the specific *selection*, *crossover*, *mutation*, and *replacement* operations used in the formulation of the GA. Instead of selecting the fittest individual from the final generation of the algorithm, we use a histogram-based approach that analyzes the overall population in order to determine the solution.

4.2 Population Clustering

Once the GA has terminated, we are left with a population of heterogeneous individuals. To produce the solution using HAGRID, instead of selecting the single fittest individual, we first produce a histogram of all of the wavelengths selected across every member of the population. Empirically, the distribution of the wavelengths roughly appears to follow a mixture of Gaussian distributions (Fig. 2a). However, the number of components present in the histogram (i.e., the number of individual Gaussian distributions that comprise the mixture model) does not necessarily equal the number of wavelengths to be selected.

For example, suppose we set the number of wavelengths to be selected to $k = 5$. The histogram of the entire population may have five distinct peaks, or it may have several more than five. The mismatch between these values is due to the existence of heterogeneous subpopulations that comprise the overall GA population (Fig. 2a). In order to identify subpopulations in the overall population, we use hierarchical agglomerative clustering (HAC) [12, 37] to partition the population into similar groups using the centroid linkage method. Once subpopulations have been identified, all but the subpopulation with the highest average fitness are discarded. In this way, we ensure the remaining population is homogeneous (in that the wavelengths are drawn from the same multimodal Gaussian distribution), and exclude the subpopulations with the worst performance (Fig. 2b,c).

Let us denote the population size as n. At each step in HAC, we calculate the pairwise distance between each of the clusters, which results in $O(n^2)$ distance calculations. In the case where all individuals are placed in a single cluster, we must run n iterations of HAC, resulting in an overall time complexity of $O(n^3)$. However, because n is usually relatively small for genetic algorithms, in practice, the clustering step is fast.

4.3 Fitting a Gaussian Mixture Model

Once the subpopulations have been identified and the subpopulation with the highest average fitness has been isolated, a Gaussian mixture model is fit to the histogram of the remaining subpopulation. As the name suggests, a Gaussian mixture model is a model consisting of several constituent Gaussian distributions that together comprise a multimodal Gaussian distribution. The parameters of the individual distributions (i.e. distribution mean and variance) are predicted using the Expectation-Maximization (EM) algorithm [37]. We assume the distributions in the mixture model each follow the univariate normal, given by:

$$f(x|\mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right\},$$

where μ_i and σ_i^2 are the mean and variance of the i^{th} distribution, respectively. If there are k components in the mixture model, then μ_i and σ_i^2 must be estimated for $i = 1, 2, ..., k$ [37].

To begin, all values of μ_i and σ_i^2 are initialized randomly. Then the algorithm iterates between the expectation step and the maximization step until convergence. This convergence is determined by the difference between parameter estimates in subsequent iterations falling below a threshold value. In the expectation step, the posterior probability of each data point being generated by each of the k distributions is calculated using the parameter estimates for μ_i and σ_i^2. In the maximization step, these posterior probabilities are used determine the maximum likelihood estimates of the parameters. After the estimates have converged, the parameters of each of the k distributions are returned.

Let $n' \leq n$ be the size of the selected subpopulation, t be the number of iterations for the EM algorithm, and k be the number of components in the Gaussian mixture. The time complexity of EM can thus be expressed as $O(tn'k)$. In practice, EM converges with small values of t. The value of k is usually small (3–10 in this study) and n' is bounded by the population size of the GA. In practice, the EM step of HAGRID is fast.

4.4 Selecting Features

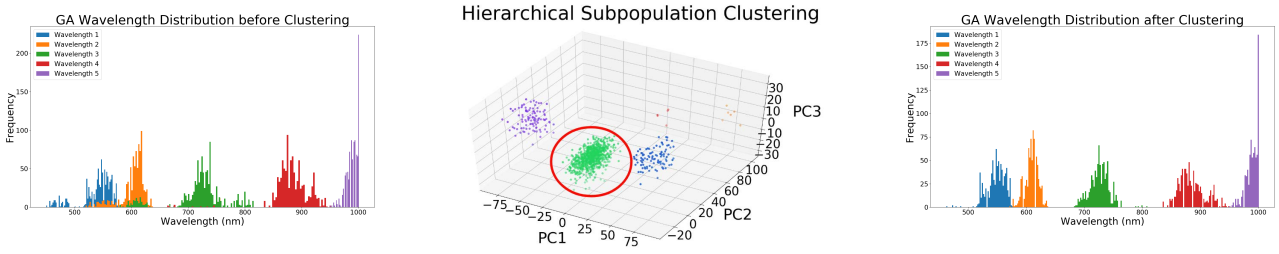
After the parameters of the Gaussian mixture model have been estimated, we can use those parameters to select the wavelength centers and filter bandwidths for the multispectral imager.

In order to select the wavelength centers for the multispectral imager, we select the estimated means from the output of the EM algorithm. These means correspond to the peaks of the individual Gaussian distributions that comprise the Gaussian mixture model. Here, the assumption is made that more informative wavelengths are selected a higher proportion of the time by members of the GA population, and therefore occur more frequently in the histogram of wavelengths.

We set the bandwidth of each filter based on the standard deviation (square root of the variance) of the Gaussian distribution associated with the wavelength center for that filter. The bandwidth of a filter is equal to the full-width at half-maximum (FWHM) of the corresponding Gaussian distribution of the filter. We set this value based on the definition of FWHM $= 2\sigma\sqrt{2\ln 2}$, where σ is the standard deviation. Here, the rationale is that the mean wavelength of a given Gaussian distribution is the most informative and most frequently selected wavelength, but the adjacent wavelengths are selected a relatively high proportion of the time as well, and are likely informative themselves. By setting the bandwidth of the filters based on the standard deviations of each Gaussian mixture component, we hope to capture most of the information across the most relevant wavelengths.

4.5 Transforming Data

Once the filter wavelength centers and bandwidths have been determined, we can transform the original hyperspectral data to mimic



(a) (b) (c)

Figure 2: Subpopulation clustering to select five wavelengths: a) overlapping heterogeneous subpopulations in the overall population of the GA; b) subpopulation clusters projected onto three dimensions using principal component analysis (the subpopulation with the highest average fitness is circled in red); c) histogram of the single selected subpopulation.

data that has been captured by a multispectral imager. First, we generate the Gaussian distributions determined by the predicted means and standard deviations. Next, we multiply these Gaussians by the original data to simulate the reflectance measurements taken by a multispectral imager. Third, we integrate under each Gaussian to produce a set of k discrete values, where k is the number of filters included in the imager. In any camera, standard color, multispectral, hyperspectral, or otherwise, even though the filters let in light over a band of wavelengths, the total amount of light is recorded as a single value for each filter, hence the integration step.

5 EXPERIMENTAL SETUP AND METHODS

5.1 Hyperspectral Imaging and Staging

For the collection of the hyperspectral images analyzed in this study, we use the Resonon Pika L hyperspectral imager. This imager has a spectral range of 387–1023 nm and a spectral resolution of roughly 2.1 nm, resulting in 300 spectral channels. The Pika L is a line-scan imager, meaning a horizontal sweep is made across the object to be imaged, and vertical slices of the image are successively combined into a single data cube.

Images produced in this way are stored as Band Interleaved by Line (BIL) files. All images are dark-corrected and calibrated to Spectralon reference panels. Spectralon is a specially designed reflective material that reflects nearly 100% of the incident light that strikes it. Because of this, the ratio of light reflected off the object of interest to the light reflected off the Spectralon panel approximates the percentage of total light reflected off the object.

The Pika L imager is placed on a rotational stage to allow for the sweep across the produce staging area. This staging area consists of a flat surface on which the produce is placed as well as a backdrop to block out external sources of light. Both the flat surface and backdrop are covered in a non-reflective paper to better control the source and direction of the illumination. The produce staging area is illuminated by two Westcott softbox studio lights. The entire hyperspectral imaging system (including the imager, lights, rotational stage, and produce staging area) can be viewed in Fig. 3.

5.2 Data

The avocado and tomato datasets used in this study are captured using the Resonon Pika L imager and the staging environment

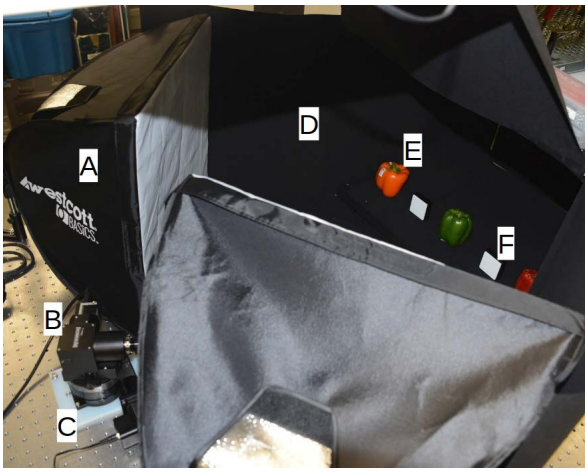


Figure 3: The hyperspectral imaging system and produce staging area. A) Lightboxes used for illumination. B) Resonon Pika L hyperspectral imager. C) Rotational stage. D) Non-reflective surface. E) Produce for imaging. F) Spectralon calibration panel.

described in Section 5.1. Images of avocados and tomatoes are taken once daily from the time of initial purchase until manually judged to be past the point of edibility. At the time of each image capture, each piece of produce in the image is labeled manually as either "fresh" or "old." To create the dataset used in the final experiments, 5×5 pixel patches are sampled repeatedly from different regions on each piece of produce. The average spectral response over these patches is then taken to smooth the spectral reflectance curve of each sample. Due to hardware limitations, a large amount of noise is present in the hyperspectral images past 1000 nm, so we exclude the last ten channels from the avocado and tomato datasets, resulting in each image containing 290 spectral channels. As a result of the exclusion of these ten wavelengths and the smoothing over the pixel patches, each data point is a one-dimensional array consisting of 290 reflectance measurements over the spectral range of 387–1001 nm. A description of the avocado and tomato datasets can be found in Table 1.

Dataset	Fruit	Samples/fruit	Total Examples
Avocado	18/10	20/36	720
Tomato	25/23	20/22	1006

Table 1: Sample sizes for produce data sets showing number of each fruit and number of samples per fruit for "fresh" and "old" respectively.

We also run experiments on the Indian Pines dataset [6]. The dataset consists of spectral reflectance curves for 16 classes of land cover, ranging from trees, to corn, to stone and steel. Each reflectance curve comprises 220 spectral channels with a spectral resolution of roughly 10 nm. As a preprocessing step, we remove all classes with fewer than 100 examples, leaving 12 classes and a total of 10,062 data points.

For all experiments, there are two main phases. In the first phase, feature selection is performed to select wavelengths and bandwidths. In the second phase, the performances of these feature subsets on the classification tasks (described in Section 5.3) are analyzed. In order to better comment on the generalizability of the feature selection methods and to disentangle the two phases, the datasets are divided into two portions, with one portion of the data being used for feature selection, and the other portion being used for classification evaluation.

5.3 Experimental Design

In our experiments, we aim to demonstrate two main objectives. First, we intend to show that our histogram-based feature selection approach is at least as good as existing methods. Second, we intend to demonstrate that filter bandwidth prediction is a viable use for the new method. In our experiments, the first of these objectives is evaluated using the classification accuracy attained on each dataset. We evaluate the second objective on the avocado and tomato datasets. For each dataset, a subset of the wavelengths is selected through various feature selection techniques, then the spectral responses at those wavelengths are fed into a feedforward neural network to perform the given classification. The various wavelength selection techniques and filter bandwidth settings are compared on the basis of classification accuracy. All experiments are run using 5×2 cross-validation and significance testing is performed using unpaired t-tests at the $\alpha = 0.05$ level.

As stated in Section 1, two of the most commonly used feature selection methods for produce monitoring applications are PLS-DA and the standard GA. As such, we compare our HAGRID method to both of these methods. Any feature selection method ought to outperform more simplistic wavelength choices, such as RGB and RGB+NIR, so for the sake of completeness, we include these in the comparison experiments as well[1]. Finally, we perform the given classification task using *all* available wavelengths.

For each algorithm (including the two fitness function variants), we select 3, 5, and 10 wavelengths for the avocado and tomato datasets. As RGB, RGB+NIR, and *all* wavelengths each contain set numbers of wavelengths, the number and values of wavelengths for these three subsets are not varied over experiments. This results in 18 total subsets of wavelengths to be compared for these two datasets. GA methods using the *fitness2* fitness function are denoted with an asterisk (i.e. GA* and HAGRID*); the absence of an asterisk denotes *fitness1* (i.e. GA and HAGRID). For the Indian Pines dataset, because the spectral resolution is already more coarse than the avocado and tomato datasets, we omit the use of *fitness2*. We then use the standard GA and HAGRID to select 3, 5, 10, and 15 wavelengths, and compare the accuracies to RGB, RGB+NIR, and all wavelengths.

For the filter bandwidth experiments on the avocado and tomato datasets, we do not find any other methods in the literature that specifically address the problem of algorithmically determining filter bandwidths for a multispectral imager. However, the bandwidths of RGB and NIR filters follow known Gaussian distributions and have known wavelength centers. In addition, standard filter bandwidths exist for custom wavelength-centered filters. In order to offer some comparison, we consider four alternatives to the HAGRID method. First, we compare known standard RGB wavelength centers and filter bandwidths[2]. Second, we compare known standard RGB+NIR wavelength centers and filter bandwidths. Third, we select wavelength centers using the standard GA and set the bandwidths to 20 nm, which is a standard filter size, commonly available at the retail level. Fourth, we select wavelength centers using HAGRID, and again set the bandwidths to 20 nm.

For each of the GA methods and fitness functions, we again select 3, 5, and 10 wavelength centers, while RGB and RGB+NIR remain constant across experiments, resulting in 20 total wavelength center and filter bandwidth combinations. Once the wavelength centers and filter bandwidths are determined, the data in the avocado and tomato datasets are transformed, as described in Section 4.5. From there, the transformed data are fed into a feedforward neural network to classify "fresh" versus "old."

5.4 Parameter Settings

For our experiments, the population size for all GA variants is set to 1,000 and the algorithms are run for 300 generations. The population size is set to a relatively large value to ensure we have a large enough number of individuals to produce a histogram that can be analyzed meaningfully. The crossover rate, mutation rate, and tournament size are tuned using a grid search. The crossover rate takes on a values from the set {0.1, 0.25, 0.35, 0.5}. The mutation rate is chosen from {0.05, 0.10}, which is relatively high to encourage diversity in the population. Finally, we consider tournament sizes of 3 and 5. We also perform a basic grid search to tune parameters for the feedforward neural networks. For all networks, we use a single hidden layer, the Adam optimizer, rectified linear units (ReLU), and a softmax classifier for the output layer. Between different experiments, the learning rate varies between 0.0005 and 0.05, while the number of nodes in the hidden layer varies between 5 and 10.

[1]The wavelengths for red, green, blue, and NIR light used in these experiments are 619 nm, 527 nm, 454 nm, and 857 nm, respectively.

[2]The RGB wavelength centers and bandwidths are derived from the known Gaussian fits for a standard Nikon camera.

		Avocado		Tomato		Indian Pines	
Method	# of Wavelengths	Accuracy	Standard Deviation	Accuracy	Standard Deviation	Accuracy	Standard Deviation
RGB	3	81.001%	1.31%	65.211%	1.46%	59.829%	1.10%
RGB+NIR	4	82.945%	2.80%	74.232%	3.93%	69.214%	1.22%
All	290/290/220	84.667%	1.27%	78.014%	2.24%	89.186%	0.98%
PLS-DA	3	80.165%	3.10%	63.06%	4.18%	-	-
PLS-DA	5	80.501%	3.01%	62.704%	1.40%	-	-
PLS-DA	10	80.555%	3.29%	62.465%	1.81%	-	-
GA	3	84.888%	2.70%	77.177%	1.38%	73.986%	0.84%
GA	5	82.388%	1.81%	75.943%	1.67%	78.991%	1.32%
GA	10	84.889%	2.55%	76.501%	1.19%	81.624%	2.39%
GA	15	-	-	-	-	83.391%	1.15%
HAGRID	3	81.166%	2.74%	73.640%	1.75%	76.108%	0.70%
HAGRID	5	82.554%	1.76%	77.021%	2.51%	77.041%	0.60%
HAGRID	10	84.723%	1.88%	76.500%	1.53%	82.331%	1.40%
HAGRID	15	-	-	-	-	**84.154**%	1.19%
GA*	3	81.388%	1.22%	76.262%	2.78%	-	-
GA*	5	85.112%	2.12%	76.224%	1.52%	-	-
GA*	10	83.945%	2.72%	75.748%	1.97%	-	-
HAGRID*	3	81.944%	2.24%	75.666%	2.14%	-	-
HAGRID*	5	**86.776**%	2.10%	78.407%	2.17%	-	-
HAGRID*	10	84.000%	2.45%	**79.005**%	2.11%	-	-

Table 2: Classification accuracy using RGB, RGB+NIR, all wavelengths, and feature selection. The best accuracy for each dataset is shown in bold. Asterisks denote the use of *fitness2*.

6 RESULTS

6.1 Feature Selection

Results for the feature selection experiments are summarized in Table 2. For both the avocado and tomato datasets, PLS-DA performs the worst across the board. This is not surprising, as it does not take into account variable interaction when performing feature selection. The highest accuracy for the avocado dataset (86.776%) is obtained by HAGRID* with five wavelengths (denoted HAGRID*/5). Further, HAGRID*/5 performs significantly better than RGB ($p < 0.0001$), RGB+NIR ($p = 0.0042$), all wavelengths ($p = 0.0192$), and the best PLS-DA result ($p = 0.0001$) at the $\alpha = 0.05$ level. It is worth noting that for the avocado dataset, all GA and HAGRID methods are able to classify the data at least as well as when utilizing all wavelengths. For the tomato dataset, HAGRID*/10 yields the highest accuracy (79.005%), which is significantly better than RGB ($p < 0.0001$), RGB+NIR ($p = 0.0049$), the best PLS-DA solution ($p < 0.0001$), and the best standard GA solution (GA/3, $p = 0.0430$).

Since HAGRID changes only how the solution is selected from the final population, one complete run of the GA is used for the corresponding GA and HAGRID results. For example, only one run of the GA is required to provide results for GA*/3 and HAGRID*/3. In this scenario, the GA is run using *fitness2*, the fittest member of the population is selected for GA*/3, and the same population is used for the histogram approach of HAGRID*/3.

For the avocado dataset, HAGRID outperforms the standard GA for four of the six head-to-head comparisons; although, the difference between the methods is not statistically significant in any of these cases. For the tomato dataset, HAGRID outperforms the standard GA in three of six experiments with HAGRID being significantly better than its GA counterpart for HAGRID*/10 ($p = 0.0422$) and HAGRID*/5 ($p = 0.0237$).

For the Indian Pines dataset, no feature selection method considered is able to match the accuracy attained using all 220 available wavelengths. Of the feature selection methods, HAGRID/15 achieves the highest overall accuracy of 84.154%, which is significantly higher than all other methods besides GA/15 ($p = 0.1836$). In three of the four head-to-head experiments, HAGRID outperforms its GA counterpart, but the difference is significant only in the case of GA/3 and HAGRID/3 ($p < 0.0001$).

The new HAGRID method has been shown to perform at least as well as the standard GA, but also has the benefit of estimating the bandpass filter bandwidths. Another possible benefit includes allowing for uncertainty quantification.

6.2 Bandwidth Prediction

Results for the filter bandwidth experiments are summarized in Table 3. The PLS-DA method is omitted from this section due to its poor performance in the feature selection experiments. For this section, let "H" denote the histogram-based bandwidths and "S" denote standard 20 nm bandwidths. For both datasets, the simulated RGB and RGB+NIR filters tend to perform the worst overall. For the avocado dataset, the best solution is found by HAGRID*/10/H, which achieves an accuracy of 85.889%. The best non-histogram bandwidth approach for the avocado dataset is GA*/5/S, which achieves a classification accuracy of 85.723%. Although HAGRID*/10/H and GA*/5/S are not statistically significantly different from each other ($p = 0.8790$), they both are significantly better

			Avocado		Tomato	
Method	Filter Bandwidth	# of Wavelengths	Accuracy	Standard Deviation	Accuracy	Standard Deviation
RGB	Known	3	80.388%	2.33%	67.275%	2.75%
RGB+NIR	Known	4	82.334%	2.13%	74.314%	2.75%
GA	20 nm	3	84.222%	1.22%	77.335%	2.08%
GA	20 nm	5	83.276%	2.43%	76.339%	2.07%
GA	20 nm	10	84.945%	2.44%	75.703%	2.07%
HAGRID	20 nm	3	81.112%	2.74%	74.791%	1.35%
HAGRID	20 nm	5	82.945%	3.00%	**78.766**%	2.25%
HAGRID	20 nm	10	84.444%	1.94%	78.608%	1.76%
GA*	20 nm	3	83.168%	1.88%	76.227%	2.40%
GA*	20 nm	5	85.723%	2.98%	76.223%	1.69%
GA*	20 nm	10	84.610%	1.88%	76.621%	1.31%
HAGRID*	20 nm	3	84.722%	2.41%	73.241%	1.47%
HAGRID*	20 nm	5	85.500%	2.10%	77.614%	1.72%
HAGRID*	20 nm	10	85.055%	2.35%	76.581%	1.49%
HAGRID	Histogram	3	82.945%	2.17%	74.197%	3.08%
HAGRID	Histogram	5	83.334%	2.21%	76.264%	2.07%
HAGRID	Histogram	10	85.721%	1.97%	77.016%	2.84%
HAGRID*	Histogram	3	83.223%	1.72%	73.397%	1.72%
HAGRID*	Histogram	5	85.612%	1.68%	77.773%	1.95%
HAGRID*	Histogram	10	**85.889**%	1.22%	76.740%	1.24%

Table 3: Classification accuracy using various wavelength centers and simulated filter bandwidths. The best results for each dataset are shown in bold. Asterisks denote the use of *fitness2*.

than RGB ($p < 0.0001$ and $p = 0.0005$, respectively) and RGB+NIR ($p = 0.0004$ and $p = 0.0126$, respectively). Note that for both *fitness1* and *fitness2*, the HAGRID/H method achieves the highest accuracy.

For the tomato dataset, the best histogram bandwidth determination is achieved by HAGRID*/5/H, with 77.773% accuracy. However, for this dataset, the histogram-based determination is outperformed by both HAGRID/5/S and HAGRID/10/S, with the former achieving the highest overall classification accuracy of 78.766%. Again, the difference between HAGRID*/5/H and HAGRID/5/S is not statistically significant ($p = 0.3312$), but both significantly outperform RGB ($p < 0.0001$ in both cases) and RGB+NIR ($p = 0.0066$ and $p = 0.0015$, respectively).

7 CONCLUSIONS

In the majority of head-to-head comparisons for the wavelength selection experiments, the HAGRID method outperforms its corresponding standard GA formulation. The filter bandwidth experiments are a little more varied, with the histogram determination of bandwidths performing the best for the avocado dataset, but second best for the tomato dataset. Overall, the fact that in all five experiments, HAGRID produces the best overall result is encouraging.

The most computationally intensive portion of a genetic algorithm is the iteration through the generations, not the selection of the solution from the final population. Since HAGRID is simply a new way of selecting the solution from this final population, it can be utilized in tandem with the standard selection of the fittest individual without adding much overhead, and the two methods can then be compared for the selection of the best solution. Specifically, the complexity of HAGRID is $O(n^3 + tn'k)$, but the values of each term are usually small, leading to fast runtimes in practice.

While here HAGRID is applied to multispectral imager design, there is no reason why it cannot be extended to other feature selection problems where the input space is continuous. The method may also have extensions to optimization problems where the variable to be optimized is continuous. As mentioned in Section 6.1, the fact that HAGRID considers a distribution of solutions, rather than a single solution opens a number of possibilities, including uncertainty quantification and other statistical evaluations.

There are several directions for future work. In general, the manual classification of produce is subjective, which introduces a fair amount of noise into the data. One way of reducing this noise would be to use a tool such as a penetrometer, which measures the force required to dent or penetrate a surface. Penetrometer readings could be taken for produce at various ages, and the learning target would then be predicting these readings based on hyperspectral data, making the classification much more objective. We would also like to further investigate other methods for feature selection in the context of hyperspectral image data, including random forests [8] and layer-wise relevance propagation [3]. Both methods can be used to derive scores of importance for individual features, which could be leveraged for effective wavelength selection. Another area of interest is the fitness functions utilized in the process. Parameters such as filter prices could be included in the fitness function to optimize the cost/performance trade-off inherent in imager design.

REFERENCES

[1] Judith A Abbott. 1999. Quality measurement of fruits and vegetables. *Postharvest biology and technology* 15, 3 (1999), 207–225.

[2] Elfatih M Abdel-Rahman, Fethi B Ahmed, and Riyad Ismail. 2013. Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. *International Journal of Remote Sensing* 34, 2 (2013), 712–728.

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015).

[4] Douglas F Barbin, Da-Wen Sun, and Chao Su. 2013. NIR hyperspectral imaging as non-destructive evaluation tool for the recognition of fresh and frozen–thawed porcine longissimus dorsi muscles. *Innovative Food Science & Emerging Technologies* 18 (2013), 226–236.

[5] Matthew Barker and William Rayens. 2003. Partial least squares for discrimination. *Journal of Chemometrics: A Journal of the Chemometrics Society* 17, 3 (2003), 166–173.

[6] Marion F. Baumgardner, Larry L. Biehl, and David A. Landgrebe. 2015. 220 Band AVIRIS Hyperspectral Image Data Set: June 12, 1992 Indian Pine Test Site 3. https://purr.purdue.edu/publications/1947/1. (Sep 2015). https://doi.org/doi:/10.4231/R7RX991C

[7] Johannes Brauers, Nils Schulte, and Til Aach. 2008. Multispectral filter-wheel cameras: Geometric distortion model and compensation algorithms. *IEEE transactions on image processing* 17, 12 (2008), 2368–2380.

[8] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[9] Jean C Buzby, Hodan Farah-Wells, and Jeffrey Hyman. 2014. The estimated amount, value, and calories of postharvest food losses at the retail and consumer levels in the United States. *USDA-ERS Economic Information Bulletin* 121 (2014).

[10] Jonathan Cheung-Wai Chan and Desiré Paelinckx. 2008. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment* 112, 6 (2008), 2999–3011.

[11] Jun-Hu Cheng, Da-Wen Sun, and Hongbin Pu. 2016. Combining the genetic algorithm and successive projection algorithm for the selection of feature wavelengths to evaluate exudative characteristics in frozen–thawed fish muscle. *Food chemistry* 197 (2016), 855–863.

[12] William HE Day and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification* 1, 1 (1984), 7–24.

[13] Jie Feng, Licheng Jiao, Fang Liu, Tao Sun, and Xiangrong Zhang. 2016. Unsupervised feature selection based on maximum information and minimum redundancy for hyperspectral images. *Pattern Recognition* 51 (2016), 295–309.

[14] Yao-Ze Feng, Gamal ElMasry, Da-Wen Sun, Amalia GM Scannell, Des Walsh, and Noha Morcy. 2013. Near-infrared hyperspectral imaging and partial least squares regression for rapid and reagentless determination of Enterobacteriaceae on chicken fillets. *Food Chemistry* 138, 2-3 (2013), 1829–1836.

[15] TM Gajanana, D Sreenivasa Murthy, and M Sudha. 2011. Post harvest losses in fruits and vegetables in South India–A review of concepts and quantification of losses. *Indian Food Packer* 65 (2011), 178–187.

[16] AA Gowen, CPo O'Donnell, PJ Cullen, G Downey, and JM Frias. 2007. Hyperspectral imaging–an emerging process analytical tool for food quality and safety control. *Trends in food science & technology* 18, 12 (2007), 590–598.

[17] John Henry Holland. 1992. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

[18] Adel A Kader. 2004. Increasing food availability by reducing postharvest losses of fresh produce. In *International Postharvest Symposium 682*. 2169–2176.

[19] Mohammed Kamruzzaman, Gamal ElMasry, Da-Wen Sun, and Paul Allen. 2012. Non-destructive prediction and visualization of chemical composition in lamb meat using NIR hyperspectral imaging and multivariate regression. *Innovative Food Science & Emerging Technologies* 16 (2012), 218–226.

[20] Shijin Li, Hao Wu, Dingsheng Wan, and Jiali Zhu. 2011. An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine. *Knowledge-Based Systems* 24, 1 (2011), 40–48.

[21] Renfu Lu. 2004. Multispectral imaging for predicting firmness and soluble solids content of apple fruit. *Postharvest Biology and Technology* 31, 2 (2004), 147–157.

[22] Michael W Mahoney and Petros Drineas. 2009. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences* (2009), 697–702.

[23] Kezhi Z Mao. 2004. Orthogonal forward selection and backward elimination algorithms for feature subset selection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 34, 1 (2004), 629–634.

[24] Scott A Mathews. 2008. Design and fabrication of a low-cost, multispectral imaging system. *Applied optics* 47, 28 (2008), F71–F76.

[25] Seyyid Ahmed Medjahed and Mohammed Ouali. 2018. Band selection based on optimization approach for hyperspectral image classification. *The Egyptian Journal of Remote Sensing and Space Science* (2018).

[26] Mostafa Moradkhani, Ali Amiri, Mohsen Javaherian, and Hossein Safari. 2015. A hybrid algorithm for feature subset selection in high-dimensional datasets using FICA and IWSSr algorithm. *Applied Soft Computing* 35 (2015), 123–135.

[27] Bart M Nicolai, Elmi Lötze, Ann Peirs, Nico Scheerlinck, and Karen I Theron. 2006. Non-destructive measurement of bitter pit in apple fruit using NIR hyperspectral imaging. *Postharvest biology and technology* 40, 1 (2006), 1–6.

[28] Gerrit Polder, Gerie WAM van der Heijden, and IT Young. 2002. Spectral image analysis for measuring ripeness of tomatoes. *Transactions of the ASAE* 45, 4 (2002), 1155.

[29] J. R. Quinlan. 1986. Induction of decision trees. *Machine Learning* 1, 1 (1 March 1986), 81–106.

[30] P Rajkumar, N Wang, G EImasry, GSV Raghavan, and Y Gariepy. 2012. Studies on banana fruit quality and maturity stages using hyperspectral imaging. *Journal of Food Engineering* 108, 1 (2012), 194–200.

[31] Amir Salimi, Mansour Ziaii, Ali Amiri, Mahdieh Hosseinjani Zadeh, Sadegh Karimpouli, and Mostafa Moradkhani. 2018. Using a Feature Subset Selection method and Support Vector Machine to address curse of dimensionality and redundancy in Hyperion hyperspectral data classification. *The Egyptian Journal of Remote Sensing and Space Science* 21, 1 (2018), 27–36.

[32] Innapa Saranwong, Jinda Sornsrivichai, and Sumio Kawano. 2001. Improvement of PLS calibration for Brix value and dry matter of mango using information from MLR calibration. *Journal of Near Infrared Spectroscopy* 9, 4 (2001), 287–295.

[33] S Serranti, G Bonifazi, V Luciani, and L D'Aniello. 2017. Classification of Peronospora infected grapevine leaves with the use of hyperspectral imaging analysis. In *Sensing for Agriculture and Food Quality and Safety IX*, Vol. 10217. International Society for Optics and Photonics.

[34] George Themelis, Jung Sun Yoo, and Vasilis Ntziachristos. 2008. Multispectral imaging using multiple-bandpass filters. *Optics letters* 33, 9 (2008), 1023–1025.

[35] Di Wu and Da-Wen Sun. 2013. Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A Review Part I: Fundamentals. *Innovative Food Science & Emerging Technologies* 19 (2013), 1–14.

[36] Jihoon Yang and Vasant Honavar. 1998. Feature subset selection using a genetic algorithm. In *Feature extraction, construction and selection*. Springer, 117–136.

[37] Mohammed J Zaki, Wagner Meira Jr, and Wagner Meira. 2014. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.

[38] Li Zhuo, Jing Zheng, Xia Li, Fang Wang, Bin Ai, and Junping Qian. 2008. A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine. In *Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Classification of Remote Sensing Images*, Vol. 7147. International Society for Optics and Photonics.