Addis Ababa University
School of Graduate Studies

Faculty of Computer and Mathematical Sciences
Department of Computer Science

Raw Quality Value Classification of Ethiopian Coffee Using Image Processing Techniques: In the case of Wollega region

By
Asma Redi Baleker

A Thesis Submitted to:

The School of Graduate Studies of Addis Ababa University

In Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science

November, 2011
Addis Ababa

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF COMPUTER SCIENCES

Raw Quality Value Classification of Ethiopian Coffee Using Image Processing Techniques: In the case of Wollega region

By

Asma Redi Baleker

MEMBERS OF THE EXAMINATION BOARD:

Name	Signature
1. Dr. Sebsibe Hailemariam	_____
2. Dr. Fitsum Admasu	_____
3. _____	_____

Acknowledgement

My profound sincere thanks go to my advisor Dr. Sebsebe H/Mariam who efficiently guided me in the thesis work. I forward my heartfelt appreciations for his motivation, regular guidance, supervision, excellent suggestions and comments in the course of proposal development, data collection, data analysis and the overall thesis work. He has invested his precious time, energy and scientific ideas and knowledge in all phases of this research.

Next comes my special heart-felt appreciations and thanks to my husband, Yishak Sahle, without whose support starting from initiation of this research, the work could not have turned up real. The motivation, encouragement and experience I obtained from him gave me the strength not only for this work but also for my future endeavors.

Many thanks and appreciations go to all ECX staff and especially the coffee and tea quality control & liquoring center team, for their continuous and meaningful support in data collection.

I would like to thank my lovely daughter Daania and my dear Sons Luliad and Dagim for their patience and support in the course of the research, where their love and spirit refreshed and strengthened me during the accomplishment of this intensive research work.

Last, but not least, my profound gratefulness to my father and mother for their assistance, love and encouragement starting from my childhood up to the current level of work. Thanks and appreciations to my brothers and my sisters.

Table of Contents

List of Tables

List of Figures

Abstract

Development of an automated computer vision system aiming in the establishment of technological and innovative approaches towards sample coffee bean raw quality value classification by extracting the relevant coffee bean features is the focal issue of this exploratory research. Of paramount significance in this regard is addressing the identified problems of the tedious and inefficient manual grading and sorting mechanisms of one of the most important agricultural products in Ethiopia, coffee. Prevalent sorting and classification approaches are characterized by subjective assessments of the features and nature of this huge economy representing crop, thereby influencing quality control and productivity aspects of the product. The major objective of the research spans extraction and selection of the important coffee bean morphological and color features that are useful for the purpose of classification of the raw quality grade level of sample coffee beans by designing, analyzing and testing a digital image processing model.

The automated raw quality value classification experimentation comprised the analysis of images of washed coffee beans of varying grades from Wollega region, using major attributes of morphological structures (shape and size), and color features. Grades 2 – 9 of the coffee beans were available, providing a total of 27 samples, which yielded 324 sample images after a series of re-sampling measures of same into 12 sub-samples. The overall image processing work to develop models and depict trends for an efficient raw quality value classification involved sequential phases of image acquisition, image enhancement and segmentation, feature extraction, attribute selection, classification and performance evaluation.

The Naïve Bayes, C4.5 and Artificial neural networks (ANN) were implemented for such classification purposes. A combined morphological and color features aggregate function dataset was used to develop the base model, though model attempts with separate features were conducted. Feed-forward multilayer perceptrons with two hidden layer and back-propagation algorithms are used in the ANN classifiers.

Discretization of the raw quality value in to three interval classes was done to improve the performance of the model. 75% split evaluation technique was implemented for the Naïve Bayes and ANN classifiers as 10-foldcross validation evaluation techniques implemented in C4.5. Naïve Bayes classifier yielded higher model performance (82.72% correctly classified), followed by C4.5 (82.09%) and the ANN classifier (80.25%). Model robustness and sensitivity was analyzed by using perturbation analysis involving manipulations of model evaluation techniques and dataset characters. Alteration of number of beans in discretization and the use of different number of hidden layers constitute the trial modeling in this regard. Classification model was also run with various combinations of features of the coffee beans as listed with the attribute selection feature of Weka tool, where the final selection of the 21 features was done at a maximal model performance level for the Naïve Bayes and ANN classification approachs. C4.5 selected 10 features as it has its own attribute selection characteristics.

An additional simulation was done with regression analysis for the sake of evaluation and trends analysis of the model outputs. A higher relation was resulted from this statistical approach between the raw quality values and the mentioned coffee bean features, supporting suitability and accuracy of dataset for classification in this research.

1. Introduction

1.1. Background

Coffee is a beverage obtained from cherry, the fruit of coffee plant. The coffee plant refers to several species of the genus *coffea* of the **madder** (common name for family *Rubicaecea*) family, which is actually a tropical evergreen shrub that has the potential to grow to a height of 100 feets. *Coffee Arabica* and *Coffee Robusta* are the two most commonly cultivated species of coffee plants with a huge economic significance. Arabica accounts for about 70 percent of the world's coffee production, as Robusta coffee represents about 30 percent of the world's market [4]. Ethiopia produces only Arabica coffee, which is widely believed to have originated in the country. Arabica coffee is a wild crop that grows in the forests of the south-western parts of the country, which contain an important source of genetic resources for the world coffee industry [28].

Coffee, known to be indigenous to Ethiopia, is the major export commodity of the country generating a meaningful income and taking the lion’s-share of the GDP of the country [38]. In 1998, coffee exports were responsible for about two-thirds of the foreign exchange earnings of the Ethiopian economy. Coffee is the major cash crop of the country and is frequently claimed to provide a livelihood for about 25 per cent of the country's population. Of the total annual production of around 250,000 tones, about half is consumed domestically [36].

A globally recognized excellent quality and flavor characterize the coffee grown in Ethiopia. Ethiopia is the producer for diversified types and grades of coffee from its different regions including Sidama, Wollega, Yirgacheffe, Jimma, Harar and several other places.

Today Ethiopia stands the biggest coffee producer and exporter country in Africa and amongst the leading in the world. Ethiopia is probably the oldest exporter of coffee in the world ranking sixth largest coffee producer after Brazil, Colombia, Vietnam, Indonesia and India, and the seventh largest exporter worldwide, in 2005, when exports were

recorded to amount 2.43 million bags, comprising 2.82 percent of world trade in coffee beans (ICO statistical database). The bulk of current Ethiopian exports dominantly reach Japan, Germany and Saudi Arabia [28].

A significant seasonal intra-annual variability exists in the price of coffee. Factors reflecting domestic supply and the periodic trends of the global coffee demand and supply are attributable to these fluctuations. In addition, the variations can be seen between different varieties and grades of coffee [4].

Coffee production for international market passes through several processes in order to be competitive at the world market. Due to this, the government has given serious monitoring and care to preserve the inherent coffee quality characteristics to satisfy customers' preferences [20]. Accordingly, every arrival coffee produced has to go through the monitoring of Ethiopian Commodity Exchange (ECX) to certify that the supplied coffee has met the minimum requirement of national standard for domestic and international markets. ECX offers an integrated warehouse system from the receipt of coffee on the basis of industry accepted grades and standards for each traded coffee by type to the ultimate delivery. Arrival coffee is deposited in warehouses operated by ECX in major surplus regions of the country.

Coffee-grading and quality control are very useful procedures in encouraging, as well as enforcing good-quality coffee production and provision. Besides, such activities assist encourage and ensure dependable and competent exporters, paving the ground for creating lasting business collaboration with overseas clients. Sorting and grading serve as a device for controlling the quality of an agricultural commodity so that buyer and seller can do business without personally examining every lot sold.

The term coffee grading exhibits varied and obscure set of terms at the various coffee-growing countries, and few are distinguished by logical clarity [36]. In Ethiopia, coffee grading is conducted through the combination of two methods. The grading is done on the basis of points assigned to the sample for its Raw Quality (measured using physical appearance of the coffee beans) and Liquor Value (cup test), being 40% and 60%

respectively. Raw Quality entails attributes of shape & make (15%), Color (15%) & odor (10%), whereas Cup Cleanness (15%), Acidity (15%), Body (15%) and Flavor/ Character (15%) constitute Liquor Value assessment. Outcomes of these parameters will then help determine the grade, the task being handled by coffee and tea quality control & liquoring center of ECX (Ethiopian Commodity Exchange) [4].

ECX conducts the manual grading system by taking a representative sample of 3kg coffee beans per each arriving truck from suppliers, of which 300g from each sample is used for raw evaluation analysis. The rest part of each sample will be classified for roasting, reference and for clients' display. The classified samples from each truck will be tagged uniformly for future identification and processing.

For a coffee stock to be marketable, standardized moisture content and coffee bean size are a pre-requisite. Before the grading process starts the coffee bean moisture level should not exceed the preset maximum moisture level, 11.5%. In addition, the coffee beans screen size should not be less than the standard Ethiopian coffee screen size, 14 units. If the sample coffee beans do not meet the above two conditions, the coffee will be rejected.

A minimum of three experts should participate in the grading activity for both the raw quality and liquor value grading decisions. These experts work independently at the same sample coffee beans and finally put their results together for final evaluation of their points for each attribute. If there is a difference on one of the points, they will convince one another and elaborate their reason to let their respective points come up to uniform final decision. The final grade will be given by an overall agreement of all the experts. Most of the time the manual grading process is amply reliable and to a good deal consistent.

There is no education or training at higher educational institution or school level in the country for producing experts in grading of coffee. Mostly, plant science professionals are preferred for this process. Such demands of the experts are fulfilled by providing on-

the job training at the center which could extend from three months to three years, by experienced experts of the center.

Few experts and a number of technical assistants participate in the classification, sorting, grading and evaluation of the coffee bean samples at the ECX, Addis Ababa branch. This number of staff for such a big task is insufficient, in particular at the time when the centre is decentralizing these practices of grading and classification to various regions of the country.

1.2 Statement of the problem

Despite the fact that coffee is one of the major agricultural products in Ethiopia, sorting and grading of this product is accomplished using traditional and manual procedures and approaches. This method employs visual and manual methods of inspection of the major entities used, including appearance, texture, shape, size and color of coffee beans, exposing the quality assessment to inconsistent results and subjectivism [11]. In addition, the tedious and time-taking human operator inspection activity for grading this high value product is very expensive, less efficient and less effective generating less descriptive and biased data information for quality control and other innovative improvement activities.

Lack of a specialized specific field of study and qualification at country level for sorting, grading and classification of this item represents an important drawback which affects the reliability, efficiency and consistency of the practice. The cost incurred to fulfill this gap at various scales of trainings to generate capable experts is also significant. This will be a serious problem when observed from the perspective of extending and decentralizing the classification and grading activities to many other regions of the country.

As a result, replacement of the human operator systems with the consistent, non-destructive, superior speed, precise and cost effective automated system of coffee quality classification and determination is necessary for such commercial products that generate a huge amount of income to the country.

The automated computer vision classification and grading system enables eliminating possible and potential human error and bias in the process. The application of this

objective inspection technique has expanded into many and diverse industries for food and agriculture to assist the inspection and grading of various fruits and vegetables in a non-destructive method. Its speed and accuracy satisfy the ever-increasing production and quality requirements, thereby promoting the development and expansion of totally automated processes. In addition, it has been successfully applied in the analysis of grain characteristics and evaluation of food crops [39].

This research project has tried to employ relevant and applicable image analysis, processing and classification techniques and models, with the aim of introducing and achieving the mentioned benefits of technological approaches for coffee bean raw quality value classification and determination. The new technologies of image analysis and machine vision have not been fully explored in the development of automated machine in agricultural and food industries. This calls for launching exploratory researches for applying, evaluating and developing these emerging technologies to assist the betterment of quality control and productivity issues of the sectors.

1.3 Motivation

Technological advancement is gradually finding applications in the agricultural and food industries, in response to the recurrent chronic global challenges, i.e., meeting the needs of the ever-increasing population. Efforts are being geared towards the replacement of human operator with automated systems, as human operations are inconsistent, costly and/or less efficient at various aspects of human endeavors. Automation involves the accomplishment of an action aiming to control a process at optimum efficiency as controlled by a system that operates using instructions that have been programmed into it or response to some activities [25]. Automated systems in most cases are faster and more precise [34].

Advances in computer technology have produced a surge of interest in image analysis during the last decade and the potential of this technique for the guidance or control of agricultural and food processes have been recognized. Series of studies have been conducted in recent years to investigate the application of computer vision technology to

sorting and grading of agricultural products. Labor intensive manual processes that are less efficient, accurate and effective induce the growing need for automation in the food industries of the developing countries [34].

In Ethiopia, technologies of image analysis or computer vision have not been explored in a significant manner in the development of automation in agricultural and food industries. Particularly, Ethiopian coffee quality inspection is based on traditional ways of classification and grading system [20]. Therefore, the implementation of imaging technology in the sector will have a paramount importance to facilitate commercial activities by increasing efficiency, to sustain dependability of customer preferences and to promote the market.

1.4. Objectives of the study

1.4.1. Main Objective

The main objective of this study is to identify coffee bean features which enable to measure the raw quality grade level of coffee beans by designing, analyzing and testing a digital image analysis and processing model, based on the coffee bean morphological and color features.

1.4.2 Specific Objectives

The specific objectives of this research are formulated as:

- Studying and understanding the determinant parameters/factors of physical characteristics of coffee which are used to evaluate the raw quality value of a given coffee sample.
- Studying the existing statistical calculations/computational tools that are used to determine the grade level of a given coffee sample in the manual coffee bean grading procedures.

- Extracting and analyzing coffee bean features that are useful to classify and determine the raw quality value of a representative sample, using image analysis and processing methods.

- Developing a suitable model for the classification and determination of the raw quality value of sample coffee, using results of the study and analysis of digital coffee bean images, by employing various classifiers

- Measuring the performance of the developed classification models from the prototype implementation or experimentation.

- Conducting comparative analysis for the accuracies of predictions of coffee quality by a human expert and the proposed computer vision system.

1.5. Scope and limitations of the study

The research work is based on washed coffee produced in the production year 2009/10, from Wollega region. Grades 2-9 were under consideration for the purpose of the research as the only samples during the data collection phase. It is on the basis of this dataset that the physical property (raw value) of Ethiopian coffee bean be studied for modeling classification and raw quality value computation systems using computer vision systems.

Wollega coffee is chosen, due to the decentralization measures taken within the ECX, restricting the Addis Ababa centre to deal with products of the Wollega.

The nature of liquor quality value evaluation techniques, comprising the measurement of parameters like odor, cup cleanness, acidity, and flavor, limit the modeling tasks to utilize only the raw quality value of the coffee grading system. It is however clear that a combined and integrated assessment of both the attributes of raw quality and liquor quality values could have generated a more meaningful and reliable grade determination model.

The available smaller number of dataset can also be mentioned as an important limitation in the research, thereby restricting the models being used and the performances of the utilized models.

1.6. Methodological approach

The computer vision based quality determination and evaluation comprised the analysis of images of coffee beans of varying grades from Wollega region to identify and evaluate the different parameters that are necessary to accomplish the research. The major attributes of focus with this regard are morphological structures (shape and size), and color features, whose differential existence and nature laid the ground for understanding and determining coffee quality, manifested through raw quality value in this particular research. Statistical computations and relevant analyses and processing techniques then enabled the final decision on the classification and determination of the various coffee packages under surveillance in an efficient and more consistent manner with meaningful temporal speed. The details of the methodological approach in this research are given below.

1.6.1. Data Collection and sampling technique

All sample coffee beans that are used for this research were obtained from coffee and tea quality control & liquoring center at Addis Ababa, a wing of the ECX accomplishing the mentioned grading tasks. These samples used are certified coffee beans by domain experts of the center.

A meaningful representative amount of washed coffee bean samples were taken from the various grade levels of coffee from Wellega region, as sorted by the centre. Coffee bean samples from other regions with their respective grades could not be accessed due to the decentralization of the center into different regions, where the Addis Ababa center deals with processing of only the Wellega coffee. In addition, samples from washed coffee were the only available resources for the purpose of the research project since these were the only available items under production and processing during the sample collection period. Necessary attributes and statistical datasets regarding the various samples have

been recorded from the tagged coffee bean packages, processed by the ECX experts during the manual grading activities.

Grades 2 – 9 of the Wellega coffee were used for the purpose of this research, where samples ranging from 2 to 4 were available for gathering from each grade, summing up to 27 coffee bean samples, each weighing 300gms. Clarity and visibility of each coffee bean and sound spacing between each coffee bean was an important task persuaded while taking the images for further processing and analysis purposes. This forced re-sampling of each of the 27 samples into 12 sub-samples of 25gms. weight, providing a total of 324 coffee bean images. All the sampled coffee beans of the region were harvested during the 2009/2010 production year. Table 1.1 shows detailed information about the collected coffee bean samples. A summarized description of same is also shown in Table1.2.

1.6.2. Digitizing sample coffee bean images

The level and quality of illumination affect digitizing activities using computer vision systems as with the human eye [25]. The performance of the illumination system greatly influences the quality of an image and plays an important role in the overall efficiency and accuracy of the system, underlining the need for manipulation of the illumination system specifications like type, angle and the use of constant light [19]. The aim is to provide the digitizing system with uniform lightning or balanced illumination. Adjustments of the imaging environment with the provision of a suitably uniform light and prohibition of the interference of external lighting sources assisted attainment of a uniform and balanced illumination for capturing the sample coffee bean images.

The data acquisition system in this research paid due concern with this regard to generate clear, unbiased and simplified digital coffee bean sample database for further analysis and processing. White background, with perpendicular and fixed orientation of imaging with the beans suitably spaced for the sake of ease of segmentation activities comprises the major adjustments of the data retrieval phase. The images captured likewise using a digital camera were then transferred into a computer, displayed on a screen and stored on the hard disk in JPEG format as digital color images.

1.6.3 Feature Extraction

Automated computer systems for classification, sorting and grading of agricultural products demand the extraction of relevant features that characterize the items under study. This research involved the extraction of morphological and color features from digitized images of sampled coffee beans to generate a useful input database for raw quality value classification. Color features of the sample coffee beans were extracted from segmented coffee bean images resulting from histogram tresholding. Morphological features were extracted from the binary images produced by histogram tresholding of the gray scale images of the original coffee bean color images.

1.6.4. Model selection

Developing the raw quality value classification demands suitable and applicable selection of models to run, compute and analyze the empirical dataset generated through image processing and analysis approaches. Artificial Neural Network, Naïve Baye's, and C4.5 classification model were employed to carry out the intended tasks of developing the raw quality value classifier.

The Naive Bayes classifier is also found an important classification approach that requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification [42]. Neural networks are data driven self-adaptive methods in that they can adjust themselves to the data without any explicit specification of functional or distributional form for the underlying model. They are also recognized as universal functional approximators, in that the neural networks can provide a projection of any function with an arbitrary accuracy [45]. In addition, C4.5 classifier was found to be important for the classification problem. It creates a decision tree based on the attribute values of the available training data in order to classify a new item, by identifying the attribute that discriminates the various instances most clearly. Possibility of higher information gain is raised as a consequence from the feature that tells most about the data instances.

1.6.5. Tools Selection

All sample coffee bean images were acquired using SONY color digital camera model number DSC-W220 with specification of 12.1 mega pixels, 2.7'' LCD screen, Carl Zeiss Vario Tessar lens with wide-angle lens of 30mm, optical zoom 4x, full HD 1080. The camera was mounted over the illumination chamber on a stand which provided easy vertical movement.

Selection of the relevant image processing, classification and regression tools is vital. One important tool with this regard was the ImageJ program which runs for processing and analysis of coffee bean images, particularly for activities of pre-processing, segmentation, and analysis and feature extraction. This windows platform public domain Java image processing program, inspired by National Institute of Health [8], is open source software. It was designed with an open architecture in-order to provide extensibility via Java plugins. It assists to display, edit, process, and analyze coffee bean images in the process of coffee classification and grading.

The outcomes of extracted features of coffee bean images is represented using the Microsoft Excel application program. The same application was used to drive new features that represent the dataset and have more discriminative power.

The neural network classification tool, the NeuroSolustions for excel version 6.0, and Weka (Waikato Environment and Knowledge Analysis) 3.6.4 were used for modeling the classifier. Weka 3.6.4 is used to implement the Naïve Bayes and C4.5 classification model. J48 is weka implementation of C4.5. Weka is a machine learning software, written in Java, and developed at the University of Wiakato in New Zealand. It is open source software.

1.6.6. Evaluation Technique

Each model was evaluated by running a test dataset on the classifier built using the training dataset. The model performance of the classifiers was returned as an output that contains performance matrices and percentage accuracy measures for each class, further summarized into a confusion matrix. Confusion matrix is a kind of a contingency table,

used to drive true positives, true negatives, false positives and false negatives indicating the correct/incorrect allotment of samples into their respective classes.

1.7 Layout of the thesis

General description for the state-of-affairs in coffee grading and sorting, research problem and motivation and fundamental methodological approaches to achieve the sought-after objectives in the image processing for coffee bean raw quality value classification is provided in this chapter one of the thesis report. The remaining parts are organized in additional four chapters. Brief discussions about machine vision systems, existing sorting and grading approaches, image processing techniques, feature extraction and classification approaches are explained in chapter two of the report. In addition, important related works are revised in this chapter.

The scientific and technological state-of-arts persuaded in the research are clearly explained in chapter three of the paper, with due focus on the setup and overall design of all sequential phases in the automated raw quality value classification experimentation process. Basic procedures in developing, calibrating, validating and evaluating classification models are also elaborated in this chapter. Chapter four of this report provides the details of the implementation of the various stages in the raw quality value classification model with the associated achievements and discussions. Finally, conclusions and recommendations in the perspective of this research are given in chapter five.

2. Literature review

2.1. Machine Vision System

A machine vision system, which can be described as the technological integration of a camera and a computer, provides an alternative to the manual inspection of biological products [15], whereby the automation contributes to reducing operating costs and increasing product value and quality.

Manual inspections of products using features that correlate with quality (like fruit size, color and weight) are being replaced by machine vision systems in different industries with their acceptance widening in recent years [16].

A wide amplitude of inspection, including defect detection, dimensional measurement, orientation detection, grading, sorting and counting, could be conducted with such automated techniques. Machine vision incorporates several advantages over the conventional methods of inspection. Capability of being compatible with other on-line processing tasks, taking dimensional measurements more accurately and consistently than a human being, and provision of measure of color and morphology of an item objectively than subjectively could be some of these. The absence of physical contact involved makes this method more hygienic and the possibility of damage during inspection to fragile biological products is very low [25].

Additional benefit of machine vision systems is the non-destructive and undisturbing manner in which information is attained [44], in addition to its attractive feature in that it can be used to create a permanent record of any measurement at any point in time [39], [24].

2.2. Coffee processing and grading

Coffee processing involves all the activities for converting raw coffee fruit into commodity green coffee. The cherry has the fruit or pulp which will later be removed leaving the seed or the coffee bean [4]. Various methods are employed for processing

coffee, each of which with a significant effect on the flavor of roasted and brewed coffee. The most commonly applied methods of processing, after having the coffee package harvested and picked, are wet process and dry process.

The wet process involves removal of the fruit covering the seeds/beans before drying. Coffee processed by the wet method is called wet processed or washed coffee [4]. The wet method requires the use of specific equipments and substantial quantities of water. After the green coffee is picked, the coffee is sorted by immersion in water. Bad or unripe fruit will float and the good ripe fruit will sink. The skin of the cherry and some of the pulp is removed by pressing the fruit by machine in water through a sieve screen. The beans could still contain a significant amount of the pulp clinging to it, demanding further cleaning.

Dry process, also known as unwashed or natural coffee, is the oldest method of processing coffee. The entire cherry after harvest is first cleaned and then sun-dried by placing on tables or thin layers on patios. The drying operation is the most important stage of the process, since it affects the final quality of the green coffee. A coffee that has been over dried will become brittle and produce too many defective broken beans. Insufficient drying also will facilitate the susceptibility of the coffee to fungal and bacterial attacks due to the inherent higher moisture content that is a conducive environment for the proliferation of such damaging agents.

There is no international standardization of coffee quality, as coffee is graded by a set of characteristics peculiar to each producing country. A sample of beans is taken from a bag, judged according to the standards of a particular country, and the sack of beans from which the sample was taken is given a quality rating, good or bad, depending on the outcome of the assessment. Due to the variation in classifications of grades, quality standardization and the descriptive terminologies used amongst producing countries, interpretation of the quality of coffee is difficult, demanding realization of the grading systems for each specific country. Starting from size determination, a standardized screen

size is used by each specific producing country for determining the real size of coffee beans, as large or small.

Raw quality evaluation and liquor value analysis are the two major components of coffee quality inspection in Ethiopia, the weight of each for the total grading being 40% for the former and 60% for the later. These two methods are universally accepted systems in both coffee producing and consuming countries tailored to the quality control system of respective countries.

Moisture content and bean size are the parameters that are preliminarily tested in the coffee grading process. The upper limit of moisture content is 11.5%. The lower limit of screen size for Ethiopian coffee bean is 14 units, where 1 unit is 1/64 of inch [20]. For values lower than this, the coffee is considered as inferior quality and no further analysis of grading proceeds. With regard to elevated moisture than the upper limit, further reprocessing is recommended to minimize the moisture content.

Visual inspection and numerous manual techniques are used to assess the physical properties of coffee in raw quality analysis. The most important physical properties for this purpose are shape, size, color, uniformity or irregularity and defect count of the coffee bean [4], [20].

Chemical properties of coffee are investigated with the other component of grading, liquor value analysis, which relies on human sense of test to identify and classify coffee. The parameters in this case are acidity, body and flavor. Acidity is a primary coffee taste created as the acids in the coffee combine with the sugars to increase the overall sweetness of the coffee. Body refers to the texture and sensation of coffee in the mouth; for example, light or heavy feelings of the coffee. Flavor is an aroma or the smell perception of the elements present in roasted coffee. Each of these parameters account for 20% of the total 60% weight for the liquor value analysis.

Coffee bean features

The most common features important for grading coffee comprise appearance (bean size, uniformity, color), number of defective beans per sample, cup quality, which includes flavor and body, and the extent to which beans are well and evenly roasted.

Coffee beans are examined closely for several different traits. Similarity in shape and uniformity in size are very important in terms of affecting the evenness of roasting [18]. Roasting coffee beans of varying shape and size results in uneven roasting since smaller beans tend to roast differently than larger sized beans, leading to the browning and popping of some individual beans before others.

The color of beans is also of paramount importance to coffee graders as uneven coloring reflects dissimilarity in the rate of drying. It also suggests that the beans were mixed from a variety of cultivators, implying inconsistent flavor and roasting. It is important that beans are separated geographically and by cultivator for a quality product. Separate harvest, processing and drying are preconditions for a good brew. Graders also smell the coffee beans, as a good bean releases a fresh aroma. Inadequate processing adds a musty or smoky hint to the aroma, which is an undesired quality in a cup of coffee [18].

Coffee grading process using raw quality analysis in the manual system utilizes mostly morphological structures like size, shape and color features of the coffee beans. Generally, there are 14 parameters used in the manual system, being:

- Pod: coffee bean with its seed coat not removed.
- Immature: a coffee bean harvested as a pre-mature crop
- Insect damage : one affected by insects
- Broken : with some part of the coffee bean broken or removed
- Foxy : brownish colored coffee beans due to some damages
- Green : a juvenile and immature coffee bean
- White : an important parameter reflecting improper storage
- Black : dark colored coffee beans due to over-drying and a total damage

- Soiled : beans containing soil cover
- Spongy : a parameter showing moldy beans due to wet storage and damage
- Stone
- Stick
- Wanza
- Earth

Refer to external materials mixed with the coffee beans and considered as defects.

Each of the above mentioned coffee bean parameters possess pre-determined importance values for the purpose of raw quality value determination, depending on the effect they exert on the nature and quality of the sample coffee. With this, the existing ones of the mentioned parameters will be identified and grouped into their respective types for a given sample coffee bean. Each parameter type will then be counted and multiplied with its pre-determined importance value. The value of each will then be summed up and converted to 40% to determine the raw quality value of that particular sample.

With this regard, morphological structures and color features of coffee bean are of paramount importance for manual coffee raw quality grading purposes. As a result, it is clear that those parameters that are correlated to this manual parameterization needs to be employed in automating the coffee raw quality grading system.

2.3. Digital Image Processing

An image may be defined as a two-dimensional function, **f(x, y)**, where **x** and **y** are spatial (plane) coordinates, and the amplitude of **f** at any pair of coordinates **(x, y)** is called the intensity of the image at that point. When **x**, **y**, and the amplitude values of **f** are all finite, discrete quantities, we call the image a digital image. In a most generalized way, a digital image is an array of numbers depicting spatial distribution of a certain field parameters (such as reflectivity of electromagnetic radiation, emissivity, temperature or some geophysical or topographical elevation) [32].

A digital image is composed of a finite number of elements, each of which having a particular location and value. These elements are referred to as picture elements, image

elements, pels, and pixels. Pixel is the term most widely used to denote the elements of a digital image [32].

Digital image consists of discrete picture elements called pixels. Associated with each pixel is a number represented as DN (Digital Number) that depicts the average radiance of relatively small area within a scene, with DN values normally ranging of from 0 to 255 in 3 elements or 0 to 2^{24}-1 in single numbering. The size of this area effects the reproduction of details within the scene. As the pixel size is reduced more scene detail is preserved in digital representation.

The field of digital image processing refers to processing digital images by means of a digital computer. Digital image processing involves efficient techniques of data acquisition and retrieval through sound image representation, display, pre-processing and segmentation approaches.

2.3.1. Image Representation and Display

The overall objective of all the ways to acquire images is to generate digital images from sensed data. The output of most sensors is a continuous voltage wave whose amplitude and spatial behavior are related to the physical phenomenon being sensed. Conversion and representation of the continuous sensed data into digital form is required to create a digital image. Sampling and quantization models are important tools by providing systematic procedures for such conversion and representation tasks [20].

An image may be continuous with respect to the x- and y-coordinates, and also in amplitude. To convert it to digital form, the function needs to be sampled in both coordinates and in amplitude. Digitizing the coordinate values is called sampling, providing the set of pixels. Digitizing the amplitude values, which is the gray level, is called quantization [43]. Quantized images are commonly represented as sets of pixels encoding color/brightness information in matrix form. Figure 2.1 models the broad route in the transformation process employing sampling and quantization.

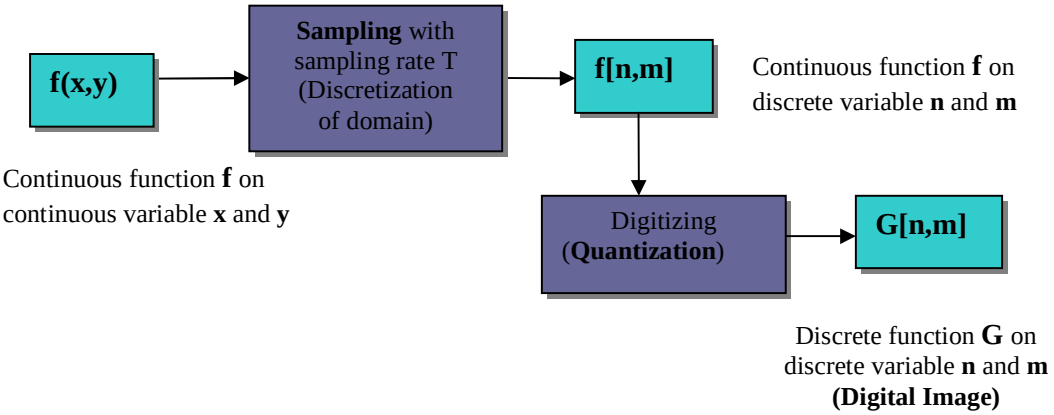


Figure 2.1: Sampling and Quantization Processes

2.3.2. Image pre-processing

Image pre-processing refers to the various initial image enhancing techniques of the captured raw images. The images captured should first be transferred onto a computer and be converted to a digital image. Digital images, though displayed on the screen as pictures, are digits readable by the computer and are converted to tiny dots or picture elements representing the real objects [25]. A series of image pre-processing techniques precede the analysis and processing activities, to enhance image quality and avoid distortion, for a valuable information retrieval [19].

Uncontrolled information during the capture of images is also checked by noise removal procedures of image pre-processing [25]. Electronic noise and noise resulting from the data transfer process from the camera to the computer can be removed using different types of filtering techniques presently in use.

Image pre-processing is the term for operations on images at the lowest level of abstraction. These operations do not increase image information content but they decrease it if entropy is an information measure [29]. The aim of pre-processing is an improvement of the image data that suppresses undesired distortions or enhances some image features relevant for further processing and analysis task.

2.3.3. Image Segmentation

Image segmentation is an essential component of image analysis technique that determines the quality of the final outputs. Image segmentation enables to discriminate objects from background, into non-overlapping sets. It also subdivides an image into its constituent parts or objects [37]. The level to which this subdivision is carried out depends on the problem being viewed. Segmentation involves partitioning of an image into a set of homogeneous and meaningful regions, such that the pixels in each partitioned region possess an identical set of properties or attributes. These sets of properties of the image may include gray levels, contrast, spectral values, or textural properties. Three popular image segmentation techniques: thresholding, edge-based, and region-based techniques are described next as discussed in [29].

A) Thresholding

The operation of separating objects and background into non-overlapping sets involves employment of a simple but effective tool for image segmentation, thresholding [1], [40]. Thresholding is used in characterizing image regions based on constant reflectivity or light absorption of their surface. This shows the fact that regions with similar features are characterized and extracted together.

B) Edge detection

Edge detection is a fundamental tool used in most image processing applications to obtain information from the frames as a predecessor step to feature extraction and object segmentation. This process detects outlines of an object and boundaries between objects and the background in the image. An edge-detection filter can also be used to improve the appearance of blurred image. Edge detection is more common for detecting discontinuities in gray level than detecting isolated points and thin lines, as isolated points and thin lines do not occur frequently in most practical images [37]. There are different methods of edge detection techniques including Sobel Operators, Roberts Cross Edge Detector and Canny Edge Detector Technique.

The Sobel operator performs a 2-D spatial gradient measurement on an image. Typically it is used to find the approximate absolute gradient magnitude at each point in an input grayscale image. The Sobel edge detector uses a pair of 3x3 convolution masks (figure 2.2), one estimating the gradient in the x-direction (columns) and the other estimating the gradient in the y-direction (rows).

-1	o	1
-2	0	2
-1	0	1

1	2	1
0	0	0
-1	-2	1

Figure 2.2: The Sobel convolution masks

The Roberts Cross operator performs a simple, quick to compute, 2-D spatial gradient measurement on an image. It thus highlights regions of high spatial frequency which often correspond to edges. In its most common usage, the input to the operator is a grayscale image, as is the output. Pixel values at each point in the output represent the estimated absolute magnitude of the spatial gradient of the input image at that point. Figure 2.3 shows Roberts cross convolution mask.

1	0
0	-1

0	+1
-1	0

Figure.2.3.Roberts cross convolution mask

Canny technique is very important method to find edges of an image and the critical value for threshold, after isolating noises from the image, with this noise detection having no adverse effect on the features of the edges in the image [37].

C) Region Based

Region based segmentation involves the grouping together and extraction of similar pixels to form regions representing single objects within the image. In this process the other regions are deleted leaving only the feature of interest.

The image is partitioned into connected regions by grouping neighboring pixels of similar intensity levels [10], in the beginning. Adjacent regions are then merged under some criterion involving homogeneity within resulting segments, inhomogeneity across neighboring segments or sharpness of region boundaries.

2.4. Feature extraction

Retrieval and measurement of agricultural product features like size, color, shape, position and contour measurement via edge detection and linking, and textural measurements on regions for better classification refers to *feature extraction*. Focus is on the enhancement of indexing and retrieval by using various methods of capturing visual content of images. The issue of choosing the right features worthy of being extracted is guided by a number of concerns including their possession of ample information about the image, ease in computation and rapid retrieval and consistence with the human perceptual characteristics [33].

Feature extraction focuses on a set of specifically known features characterizing the application domain, probably with some consideration for non-overlapping or uncorrelated features [19]. As a formative procedure of various attributes and properties associated with regions or objects, it operates mainly on abstracted image information obtained through segmentation [9]. The image objects could be measured and described based on their features and characteristics after proper completion of the image segmentation of the external grading system process of the samples.

2.5. Classification approaches

Recognition of the characteristics of objects in an image from a specific set of measured values of features of the object facilitates the stratification of an image into various classes with similar features. This is the core business of design of classifiers, which utilizes specified features of an object as its inputs, thereby generating a classification label or value depicting the correct class allotment of the object. A set of objects with predefined labels or values- the learning set-serve as the immediate basis for this mechanism.

Though various techniques of classification mechanisms are applied for the mentioned tasks of object recognition, this paper tried to deal with the Naïve Bayes, C4.5 and the Artificial Neural Network classifiers, being described as in the following sections.

2.5.1. Naïve Bayes Classifier

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even though these features depend on the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods [23].

Bayes rule helps as the foundation for designing learning algorithms or function approximators. As the interest is to learn some target function describing X independent and Y dependent variables that can be explained as $f : X \rightarrow Y$, or equivalently, $P(Y|X)$, the training data is used to learn estimates of $P(X|Y)$ and $P(Y)$. New independent attributes can then be classified using these estimated probability distributions, plus Bayes rule. This type of classifier is called a generative classifier, as it allows view of the distribution $P(X|Y)$ as describing how to generate random instances X conditioned on the target attribute Y [21].

Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood.

The Naive Bayes algorithm is based on the Bayes rule that assumes the attributes $X_1 \ldots X_n$ are all conditionally independent of one another, given Y (the dependent variable). This

assumption enables to dramatically simplify the representation of P(X|Y) and the problem of estimating it from the training data. Consider, for example, the case where X = (X_1,X_2). In this case

$$P(X|Y) = P(X_1,X_2|Y)$$
$$= P(X_1|X_2,Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

Where the second line follows from a general property of probabilities, and the third line follows directly from our above definition of conditional independence. More generally, when X contains n attributes which are conditionally independent of one another given Y, we have

$$P(X_1 \ldots X_n|Y) = \prod_{i=1}^{n} P(X_i|Y) \quad (1)$$

Notice that when Y and the Xi are Boolean variables, we need only 2n parameters to define $P(X_i = x_{ik}|Y = y\ j)$ for the necessary i,j,k. This is a dramatic reduction compared to the $2(2^n-1)$ parameters needed to characterize P(X|Y) if we make no conditional independence assumption.

Let us now derive the Naive Bayes algorithm, assuming in general that Y is any discrete-valued variable, and the attributes $X_1 \ldots X_n$ are any discrete or real valued attributes. Our goal is to train a classifier that will output the probability distribution over possible values of Y, for each new instance X that we ask it to classify. The expression for the probability that Y will take on its k^{th} possible value, according to Bayes rule, is

$$P(Y = y_k|X_1 \ldots X_n) = \frac{P(Y = y_k)P(X_1 \ldots X_n|Y = y_k)}{\sum j\ P(Y = y_j)P(X_{1\ldots}X_n|Y = y_j)}$$

where the sum is taken over all possible values y j of Y. Now, assuming the Xi are conditionally independent given Y, we can use equation (1) to rewrite this as

$$P(Y = y_k|X_1 \ldots X_n) = \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum j\ P(Y = y_j)\ \prod_i P(X_i|Y = y_j)} \quad (2)$$

Equation (2) is the fundamental equation for the Naive Bayes classifier. Given a new instance $X^{new} = (X_1 \ldots X_n)$, this equation shows how to calculate the probability that Y will take on any given value, given the observed attribute values of X^{new} and given the distributions P(Y) and $P(X_i|Y)$ estimated from the training data. If we are interested only in the most probable value of Y, then we have the Naive Bayes classification rule:

$$Y \leftarrow \underset{y_k}{\operatorname{argmax}} = \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum j\, P(Y = y_j)\prod_i P(X_i|Y = y_j)}$$

this simplifies to the following (because the denominator does not depend on yk).

$$Y \leftarrow \underset{y_k}{\operatorname{argmax}} = P(Y = y_k)\prod_i P(X_i|Y = y_k) \qquad (3)$$

An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

2.5.2. C4.5 classifier

C4.5 is developed by Ross Quinlan [31]**,** is an algorithm used to generate a decision tree that can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier**.**

C4.5 builds decision trees from a set of training data, using the concept of information gain ratio [13]. The training data is a set $S = s_1,s_2,...$ of already classified samples. Each sample $s_i = x_1,x_2,...$ is a vector where $x_1,x_2,...$ represent attributes or features of the sample. The training data is augmented with a vector $C = c_1,c_2,...$ where $c_1,c_2,...$ represent the class to which each sample belongs.

At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an

attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub-lists.

This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

This algorithm is useful to create decision trees with the following advantages:

- Handling both continuous and discrete attributes - In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.
- Handling training data with missing attribute values - C4.5 allows attribute values to be marked as ? for missing. Missing attribute values are simply not used in gain and entropy calculations.
- Handling attributes with differing costs.
- Pruning trees after creation - C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

Some premises guide this algorithm, as is mentioned below [12]:

- if all cases are of the same class, the tree is a leaf and so the leaf is returned labeled with this class;
- for each attribute, calculate the potential information provided by a test on the attribute (based on the probabilities of each case having a particular value for the attribute). Also calculate the gain in information that would result from a test on the attribute (based on

the probabilities of each case with a particular value for the attribute being of .a particular class);

• depending on the current selection criterion, find the best attribute to branch on.

This process uses the "Entropy", i.e. a measure of the disorder of the data. The Entropy of $\vec{y}$ is calculated by:

$$\text{Entropy}(\vec{y}) = -\sum_{j=1}^{n} \frac{|y_j|}{|\vec{y}|} \log \frac{|y_j|}{|\vec{y}|}$$

iterating over all possible values of $\vec{y}$. The conditional Entropy is

$$\text{Entropy}(j/\vec{y}) = \frac{|y_j|}{|\vec{y}|} \log \frac{|y_j|}{|\vec{y}|}$$

and finally, we define Gain by

$$\text{Gain}(\vec{y}, j) = \text{Entropy}(\vec{y} - \text{Entropy}(j|\vec{y}))$$

The aim is to maximize the Gain, dividing by overall entropy due to split argument $\vec{y}$ by value j.

Prunning is also an important step to the result because of the outliers. All data sets contain a little subset of instances that are not well-defined, and differ from the other ones on its neighborhood. After the complete creation of the tree, that must classify all the instances in the training set, it is pruned. This is to reduce classification errors, caused by specialization in the training set; this is done to make the tree more general.

2.5.3. Artificial Neural Network

Networks assist to model a wide range of phenomenon in various disciplines including physics, computer science, biochemistry, mathematics and telecommunications [6], as the constituent parts of most of their systems could be seen as a network of say proteins, computers, communities, etc. Networks vary enormously in type, though all are characterized by two broad components, a set of nodes, and connections between these nodes. The nodes can be seen as computational units. They receive inputs, and process them to obtain an output.

One type of network sees the nodes as ‘artificial neurons’. These are called artificial neural networks (ANNs). An artificial neuron is a computational model inspired in the natural neurons. Natural neurons receive signals through synapses located on the dendrites or membrane of the neuron. When the signals received are strong enough (surpass a certain threshold), the neuron is activated and emits a signal through the axon. This signal might be sent to another synapse, and might activate other neurons.

Artificial neural networks (ANN) are among the newest signal-processing technologies serving two important functions: pattern classifiers and non-linear adaptive filters [26]. ANNs can identify and learn correlated patterns between input datasets and corresponding target values. After training, ANNs can be used to predict the outcome of new independent input data [8].

The non-linear nature of the neural network processing elements provides the system with lots of flexibility to achieve practically any desired input/output map, i.e., some Artificial Neural Networks are universal mappers [6]. A value describing style in neural computation is represented in figure 2.4.

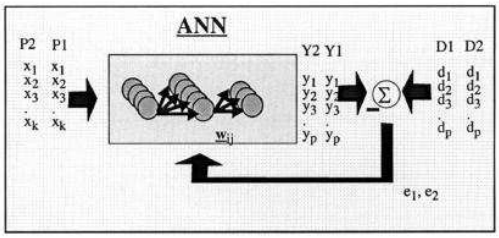


Figure 2.4: The style of neural computation [6].

This model is a variety of supervised training where an input is presented to the neural network providing a corresponding desired or target response as an output. The discrepancy between the desired response and the system output comprises an error, where this information is fed back to the system to adjust the system parameters in a systematic fashion as asserted by the learning rule. The process is repeated until the performance is acceptable.

In artificial neural networks, the designer chooses the architectural setup, topology of a network, the performance function, the learning rule, and the criterion to stop the training phase. ANN-based computations are recognized for their advantages in development time, resources and provision of much better performance under difficulties than other technologies. A back-draw of this technology includes the difficulty to introduce ‘a priori’ information into the design as the system automatically adjusts the parameters. In addition, when the system does not work properly it is also hard to incrementally refine the solution.

2.5.2.1. Neural Network topologies

Two most widely used topologies of ANN technology are described in detail below.

1. **Feed-forward neural networks**: allow signals to travel one way only; from input to output [8]. The data processing can extend over multiple (layers of) units, but no feedback connections are present, that is, connections extending from outputs of units to inputs of units in the same layer or previous layers [26].
2. **Feed-Back/Recurrent neural network:** can have signals traveling in both directions by introducing loops in the network [8]. Contrary to feed-forward networks, the dynamical properties of the network are important. In some cases, the activation values of the units undergo a relaxation process such that the neural network will evolve to a stable state in which these activations do not change anymore [6]. In other applications, the changes of the activation values of the output neurons are significant, such that the dynamical behavior constitutes the output of the neural network.

2.5.2.2. Training of artificial neural networks

A **neural network** has to be configured so that the application of a set of inputs produces (either 'direct' or via a relaxation process) the desired set of outputs. Various methods to set the strengths of the connections exist. One way is to set the weights explicitly, using a priori knowledge. Another way is to **'train' the neural network** by feeding it teaching patterns and letting it change its weights according to some learning rule. The learning situations can be categorized in to three distinct sorts. These are:

- **Supervised learning** or **Associative learning** in which the network is trained by providing it with input and matching output patterns. These input-output pairs can be provided by an external teacher, or by the system which contains the neural network (self-supervised) as it is represented in figure 2.7.

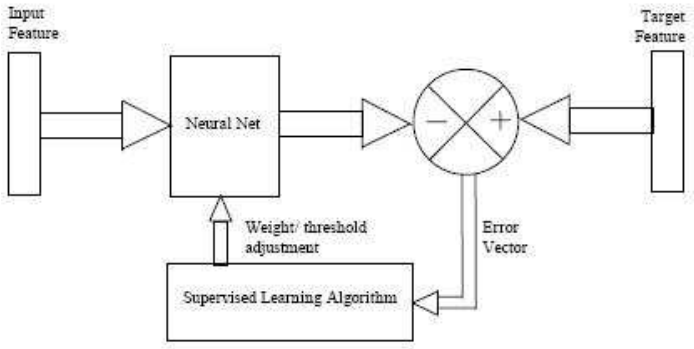


Figure 2.5: Training of artificial neural networks using supervised learning.

- **Unsupervised learning** or **Self-organization** in which an (output) unit is trained to respond to clusters of pattern within the input. In this paradigm the system is supposed to discover statistically salient features of the input population. Unlike the supervised learning paradigm, there is no a priori set of categories into which the patterns are to be classified; rather the system must develop its own representation of the input stimuli.
- **Reinforcement Learning** This type of learning may be considered as an intermediate form of the above two types of learning. Here the learning machine does some action on the environment and gets a feedback response from the environment. The learning system grades its action as good (rewarding) or bad (punishable) based on the environmental response and accordingly adjusts its parameters. Generally, parameter adjustment is continued until an equilibrium state occurs, following which there will be no more changes in its parameters. The self-organizing neural learning may be categorized under this type of learning.

2.6.2.3. The Backpropagation Algorithm

The backpropagation algorithm is used in layered feed-forward ANNs [6]. This means that the artificial neurons are organized in layers, and send their signals “forward”, and then the errors are propagated backwards. The network receives inputs by neurons in the input layer, and the output of the network is given by the neurons on an output layer. There may be one or more intermediate hidden layers. The backpropagation algorithm uses supervised learning, which means that the algorithm with examples of the inputs and outputs to be computed by the network will be provided. This enables calculation of the error (difference between actual and expected results). The idea of the backpropagation algorithm is to reduce this error, until the ANN learns the training data. The training begins with random weights, and the goal is to adjust them so that the error will be minimal.

2.5.2.4. Advantages and disadvantages of ANN

Although computing these days is truly advanced, there are certain tasks that a program made for a common microprocessor is unable to perform; though a software implementation of a neural network can be made with their advantages and disadvantages [6].

2.6. Related works

Over the past decade, advances in hardware and software for digital image processing have motivated several studies on the development of systems, to evaluate the quality of diverse agricultural products. The majority of these studies are focused on the application of computer vision system to agricultural products quality inspection and grading. Computer vision based inspection and grading of apple, oranges, strawberries, nuts, tomato, mushrooms, wheat, corn and rice are examples.

The study of apples using computer vision has attracted much interest and can reflect the progress of computer vision technology for fruit inspection. An image processing algorithm based on Fourier expansion was developed to characterize objectively the

apple shape, assisting to identify different phenotypes [30]. In this research, it was shown that four images per apple were needed to quantify the average shape of a randomly chosen apple. It was found that this profile analysis can be used to characterize existing shape descriptor lists. The researchers used Fourier analysis of apple peripheries as quality inspection/classification technique. This methodology gave an insight into the ways in which external product features affect the human perception of quality. This research revealed the fact that the involvement of more product properties in the classification and their more complex nature increases the error of human classification.

An experimental investigation of the use of computer vision in sorting fresh strawberries, based on size and shape, showed a result that the developed system was able to sort the 600 strawberries tested with an accuracy of 94-98% into three grades based on shape and five grades on size [22].

Subsequent researches employing machine vision to identify different varieties of wheat and to discriminate wheat from non-wheat components revealed the fact that wheat classification methods could be improved by combining computer vision analysis and hardness analysis [44]. Twenty three morphological features were used for the discriminant analysis of different cereal grains using machine vision. Classification accuracies of 98, 91, 97, 100 and 91% were recorded for CWRS (Canada Western Red Spring) wheat, CWAD (Canada Western Amber Durum) wheat, barley, oats and rye, respectively. The relationship between color and texture features of wheat samples to scab infection rate was studied using a neural network method [16].

In order to preserve corn quality, it is important to understand its physical properties and assess potential mechanical damages so as to design optimum handling and storage equipments [27]. Measurements of kernel length, width and projected area independent of kernel orientation have been performed using machine vision. The algorithm accuracy was between 0.86 and 0.89 measured by the correlation coefficient between predicted results and actual sieving for a 500g sample. The processing time of the size-grading program was reported as being between 0.66 and 0.74 second per kernel.

As rice is one of the leading food crops of the world, its quality evaluation is of importance to ensure it remains appealing to consumers. A digital image analysis method was developed for measuring the degree of milling of rice, whose comparison with the conventional chemical analysis provided a coefficient of determination of $R^2 = 0.9819$ for the 680 samples tested [14]. In China, an image analysis technique was developed recently to identify rice seed varieties using a neural network model for pattern classification. It used MATHLAB 6.5 programming language to extract color and morphological features of individual seeds. From color features of the mean and variance of RGB components were calculated. Six varieties (ey795, syz3, xs11, xy5968, xy9308, z903) rice seeds, which are widely planted in Zhejiang Province of china, were considered for the research work. The experimentation result indicated that the classification accuracies are 90%, 88%, 95%, 82%, 74%, 80% for ey7954, syz3, xs11, xy5968, xy9308, z903 respectively.

A digital image analysis technique based on morphological and color features was developed to classify different varieties of Ethiopian coffee based on their growing region [20]. For the classification analysis, ten morphological and six color features were extracted from each coffee bean image. The processing type of coffee (washed or unwashed) has been also predefined during the analysis. He also compared classification approaches of Naïve Bayes and Neural Network classifiers on each classification parameter, i.e., morphology, color and combination of the two. To evaluate the classification accuracy, from the total of 4844 datasets, 80% were used for training and the remaining 20% for testing. The classification system was supervised at the corresponding predefined classes of growing regions. Accordingly, it was found that the classification performance of neural networks classifier was better than Naïve Bayes classifier. It was also described that the discrimination power of morphology features was better than color features; however combined use of both morphology and color features resulted increased classification accuracy. The best classification accuracies (80.7%, 72.6%, 56.8%, 96.77%, 95.42% and 69.9% for Bale, Harar, Jimma, Limu, Sidamo and Welega, respectively) were obtained using neural networks when both morphology and color features were used together. The overall classification accuracy was 77.4%.

It can be concluded from the above researches that morphological structures and color are the most viable features used in computer vision systems for inspection and grading of agricultural products. As a classification technique, Artificial Neural Network is the most appropriate technique for classification, inspection and grading of agricultural products, especially coffee.

3. Design of the model

3.1. Introduction

The inspection and grading of coffee products should be understood as a complex and systematic series of a process whereby numerous phases starting from region identification to the final grading of the coffee beans are accomplished objectively and rationally.

The initial stage in the grading process is preparation of the raw coffee bean samples. These raw coffee products might come from different coffee producing regions, each with its own quality and characteristic features. This demands for classification of the arrival coffee beans into their respective source region. The so far mentioned morphological and color features are the parameters useful to carry out this sorting by region. Such region classifier models were developed previously for six coffee producing regions of Ethiopia by Habtamu, 2008, yielding a classification accuracy of 77.4%.

As shown in the general procedures of coffee processing (figure 3.1), region classification is followed by the activity of prediction of the raw and liquor quality values of the coffee bean samples. A series of team-based cup-tests of roasted sample coffee beans provides an averaged liquor quality value. The contribution of these values to the prediction of the coffee bean grade comprises 40% for raw quality and 60% for liquor quality values. The final grading of the sample coffee beans for each sample of a given region will then be decided by summing up their respective values for the liquor and raw qualities accordingly.

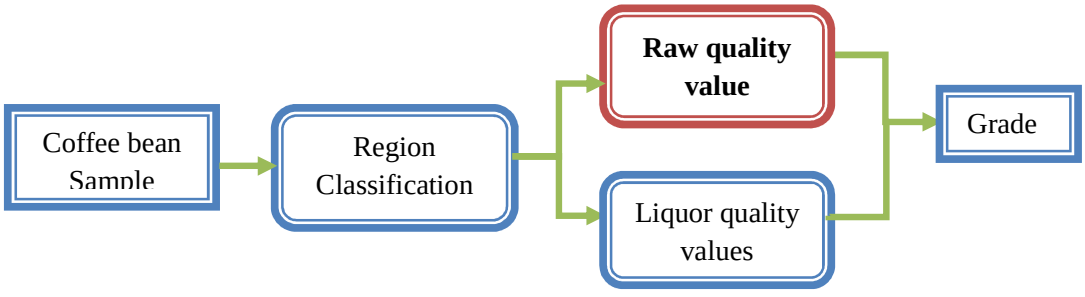


Figure 3.1 General procedures of the coffee grading process.

Unlike the liquor quality value of coffee beans which is dependent on subjective decisions, the raw quality value can be computed objectively and scientifically for a meaningful sorting, classification and grading purposes of coffee beans and other agricultural products. This requires the development of an automated system with minimum human error and bias in the grading process.

With this regard, this research aims to model an automated computer vision system which assists to determine the raw quality value of sample coffee beans by using image processing techniques. Training of the learning model by feeding data on pre-known raw quality values and selected coffee bean features comprises the core task of this work, whereby further modeling and evaluation of the unknown raw quality values for certain sample coffee beans is made possible using the selected feature vector of that sample as an input into the simulators.

The learning model development is made to the overall combined morphological and color feature attributes of the sampled coffee beans. Performance evaluation of the simulated models is produced as an output from the classifiers, depicting performance matrices, statistic values, percent correctly/incorrectly classified and confusion matrices. Further evaluation and analysis of the classifiers could also be conducted by running regression models, generating a statistical insight on the relation between the raw quality values and the coffee bean attributes used in the modeling process.

3.2. Coffee raw quality value classification

The problem of classification is concerned with the construction of a procedure that can be applied to differentiate items, in which each new item must be assigned to one of a set of pre-defined characteristic classes on the basis of observed attributes or features.

Accordingly, image analysis or computer vision based techniques were employed to determine the raw quality value of Ethiopian coffee by assisting characterize and formulate distinct pre-defined classes that served as the basis for assigning the new samples beans into their respective fit classes. The pre-defined classes depend on the values of the morphological and color features computed from coffee bean images.

The captured images for the purpose of this research were transferred into a computer and pre-processed for activities like noise reduction, to enhance the images for accurate use in extracting the necessary features using image analysis and processing techniques.

Further computations were made on the coffee bean features to generate simplified and representative data, which promoted further development of the final model. Generation of aggregate values and model based feature selection were important steps to proceed with the model simulation. Appropriate clustering and classification approaches were employed to classify a given sample to the fitting category using the generated data as input. Figure 3.2. depicts the various phases employed to develop a simulation model that assist determination of raw quality value.

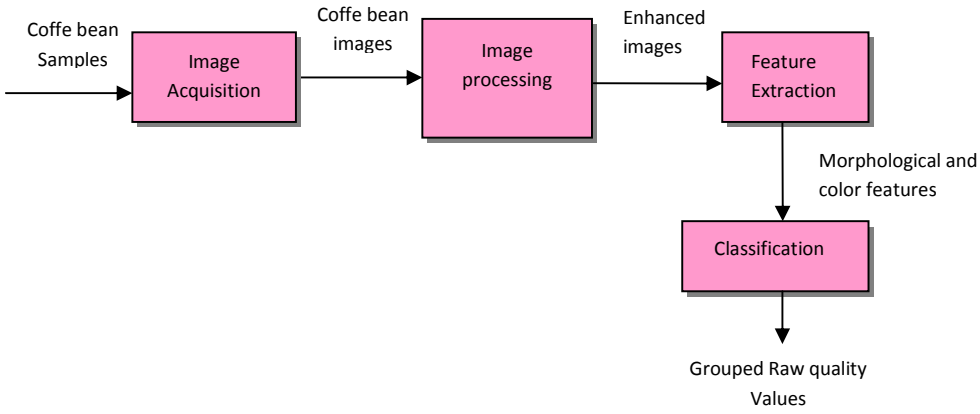


Figure 3.2: Raw quality value classification process

3.3. Image acquisition

A series of computational processes and mechanisms were accomplished repeatedly in the image capturing stage, till the most appropriate and suitable status was achieved finally. Illumination, background, coffee bean spacing, distance between the sensor and the scene, camera adjustment and manageable sample size selection were the most important issues worthy of manipulations for acquiring clear images with less noise in this controlled experiment.

All coffee bean sample pictures were taken from a fixed height (0.5 m) oriented in a perpendicular manner directly above the sample coffee beans to retain uniformity between all the image pictures of the samples (Figure 3.3). The camera was mounted on a stand with the mentioned elevation of 0.5m above the beans for the sake of enabling simple movement vertically and to avoid blurred pictures. A white background table was used to arrange the sample beans captured for clarity and unbiased pictures. The coffee beans were well spread out up to a maximal spacing that the fixed image capturing area allowed. This was meant to avoid any coffee bean over-lapping or touching particles, that could constraint further processing and analysis of the images, like in segmentation phase.

To obtain uniform lighting or balanced illumination, an incandescent lamp whose light source was 100W with a rated voltage of 220V was used in all experiments. The images were taken at a resolution of 1632 x 1224 pixels.

All the captured images were then transferred into computer and used as input data by the image processing software to carry out the necessary processes and analysis. Figure 3.4. shows a sample coffee bean image captured under the described environment and ready for application in further automated image processing.

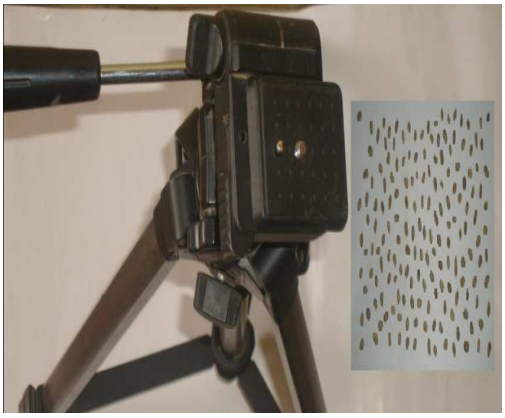


Figure 3.3. The image capturing environment.

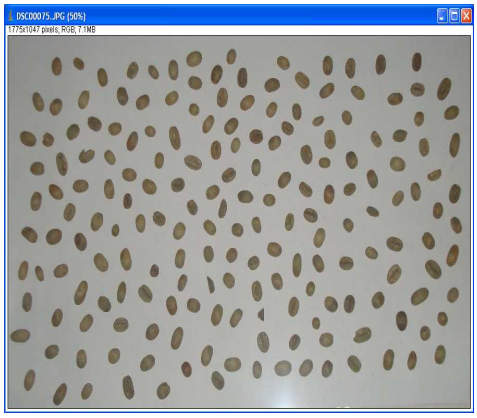


Figure 3.4. A sample coffee bean image

3.4. Coffee beans Image Processing

The "act of examining images for the purpose of identifying objects and judging their significance" can represent a meaningful definition for Image processing. This technique utilizes sensed data, the logical processes of its detection, identification, classification, measurement and evaluation assisting in understanding the significance of physical and cultural objects, their patterns and spatial relationships [32]. Likewise, this classification system of coffee products is primarily comprised of a combination of hardware and software, to capture images of the coffee bean samples at the simulated sample station and perform image processing with a high performance computer.

Major steps of image representation, pre-processing, segmentation and extracting necessary information comprise this image processing whereby images are handled in mechanisms ranging from efficient ways for transferring these images to computers to enhanced images using suitable media for the next development technologies.

All the collected images were displayed on a computer screen and stored in JPEG (Joint Photographers Expert Groups) format on a hard disk. Succeeding this, the images were further pre-processed to enhance the retrieval of accurate information.

Back ground subtraction was the first process conducted to avoid blurs, light distortions and other noises that could be formed due to illumination effects and some external objects on the background. Conversion of the RGB images to 8-bit gray scale image and histogram thresholding for the extraction of morphological features, and histogram thresholding for extracting color features from the thresholded images followed the background subtraction task in the original images. Conversion to gray scale images of the RGB images supports the production of binary images for the sake of extraction of morphological features.

Grayscale refers to a range of shades of gray without apparent color. Grayscale images are also called monochromatic, denoting the absence of any chromatic variation (i.e., no

color). Grayscale images are often the result of measuring the intensity of light at each pixel in a single band of the electromagnetic spectrum.

Histogram thresholding technique was applied for the sake of segmenting the images with the constituents partitioned into homogenous attributes. The upper and lower limits of thresholding used were 0 and 215, respectively. Extraction of morphological features was made possible from the binary images, which are the products of the thresholding of the gray scale images. A binary image represents a digital image with only two possible values for each pixel. Typically the two colors used for a binary image are black and white, though any two colors can be used. The color used for the object(s) in the image is the foreground color while the rest of the image is the background color.

Some coffee beans on many of the produced binary images were seen to contain holes (Figure 4.1. (d)), which might be the result of defects on the surface of the coffee or due to over drying of the coffee beans in the coffee processing phase. These holes affect the computation of some features like coffee bean area, which might affect the performance of the next process, feature extraction. ImageJ noise removal and hole filling feature assisted the removal and filling of the mentioned holes. Regardless of the enhancement measures taken in background subtraction, there appeared to exist some outliers on some coffee bean images. Unclear backgrounds of the coffee bean images resulting from inconsistent image acquisition environment could be the causes for these. The outliers tend to magnify the error rate as the system analyses them as part of the coffee beans. The outlier removal feature of the program ImageJ was used to remove the mentioned outliers from the coffee bean images.

3.5. Feature extraction

Processing and extraction of a meaningful set of empirical dataset of feature attributes from the pre-processed coffee bean images is an important start to carry out the computer-assisted coffee bean raw quality value computation tasks. The production of a set of known features, characteristic for the application domain, probably with some consideration for non-overlapping or uncorrelated features comprises the process of feature extraction. The collections of extracted attributes represent a particular feature,

and a vector of such a feature is called a pattern. Features are used as inputs to the algorithms for classifying the objects into different categories. Pattern recognition can be done by analyzing the morphology (shape and size), color, texture (spatial distribution of color), or a combination of these features of the images [25].

Particle analyzer method of ImageJ was used to extract morphological and color features of the sample coffee beans from the previously processed and analyzed binary and tresholded images. ImageJ conducts the calculation of the features for each coffee bean from the region of interest with in the concern image by giving a unique label for each bean. A total of 324 images of coffee beans are used for the extraction of both color and morphological features. These features were then stored in excel for further processing and analysis activities.

3.5.1. Morphological features

The most common measurements that are made on objects were those that describe shape. Shape features are physical dimensional measures that characterize the appearance of an object. Area, perimeter, major and minor axes lengths, and aspect ratio are some of the most commonly measured morphological features. Morphological features are widely used in automated grading, sorting and detection of objects in industry [20] [25] [19].

Here is provided a description of the morphological features extracted from each coffee bean image.

- **Area (A) -** The area A of the kernel is measured as the number of pixels in the polygon.
- **Perimeter (P) -** The perimeter P is the mathematical sum of the Euclidean distances between all the successive pairs of pixels around the circumference of the kernel.
- **Major axis length (MA) -** The length of the major axis is the longest line that can be drawn through the object.
- **Minor axis length (MI)** – The length of the major axis is the longest line that can be drawn through the object perpendicular to the major axis.

- **Aspect Ratio (AR) (Elongation)** - The elongation ratio of the length of the minor axis to the length of the major axis. This is given as:

$$AR = \frac{MA}{MI}$$

- **Circularity (Cr)-** This morphological attribute of the coffee beans is given by:

$$Cr = (4\pi A)/P^2$$

- **Roundness (R) –** This attribute is described as:

$$R = 4A/\pi MA^2$$

- **Feret Diameter (FD)** - This is the diameter of a circle having the same area as the object and is computed as:

$$FD = [(4A)/\pi]^{1/2}$$

3.5.2. Color features

Color is an important and the most straight-forward feature that humans perceive when viewing an image. Human vision system is more sensitive to color information than gray levels so color is the first candidate used for feature extraction [17]. The most commonly used color feature model in image processing is based on the primary spectral components of red (R), green (G) and blue (B). Color features of an object are extracted by examining the R, G and B levels of each pixel within the object's boundary. The histogram of these pixels shows the brightness distribution found in the object [25].

The three common perceptual descriptors of a light sensation in relation with RGB color are Brightness (B), Saturation (S) and hue (H) [20] [24]. Hence, the color features are extracted by computing the mean values of RGBs and HSBs of coffee bean images. Hue of an image is excluded from the features list in this experimentation due to its generation of complex and unclear image display (Figure 3.5.), making difficult the extraction of feature values from that particular image. Computation of the mean values for each component of these color spaces needs to split each component to separate image stacks. The RGB and HSB stack splitting assignments were done with the respective RGB and HSB stack splitting features of ImageJ (Figures 3.5. and 3.6.).

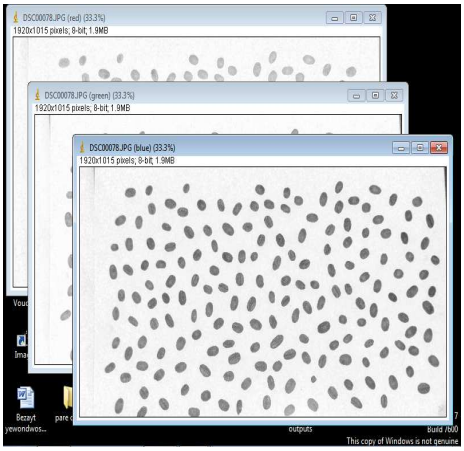


Figure 3.5. RGB stacks

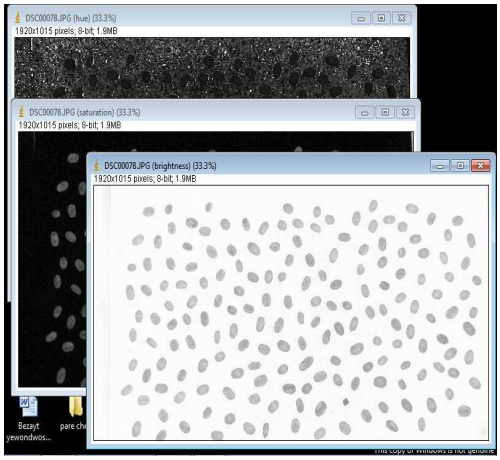


Figure 3.6. HSB Stacks

3.6 Aggregate features generation

The over all effect of the morphological and color features of coffee beans is an important determinant of the raw quality value of the samples. This demands the generation of an aggregate dataset that represents the cumulative effect of all the morphological and color feature values of the beans. Excel spreadsheet enabled computation of the aggregate values for each of the extracted color and morphological feature values of the sample coffee beans. These aggregate values include the maximum, minimum, average, variance and standard deviation statistic values for the morphological features and mean values for the color features.

3.7. Classification model

The overall processing and analysis of the sample coffee bean images targets the production and provision of an empirical dataset for statistical computations of raw quality value determination and classification in different suitable and applicable tools.

Classification was conducted in Weka and neurosolutions tools by employing the Naïve Bayes, C4.5, and artificial neural networks classifiers. Combined attributes of morphological and color features were used as input patterns to build all the classification models. Classifier-specific attribute selection techniques characterize the simulation processes in all tools, where model performances are realized to be sensitive to differential attribute perturbations. Selection of a set of appropriate input feature variables

is an important issue while trying to develop classification models by employing the most appropriate respective classifiers. The purpose of feature variable selection is to find the smallest set of features that can result in satisfactory model performance.

Naïve Bayes and C4.5 Classifiers

Weka classification tool was used to implement Naïve Bayes and C4.5 classifiers using the available dataset in this research. The Naïve Bayes classifier which highly depends on "independence assumptions'' is believed to be more suitable mainly when dealing with small number of samples. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix [47].

These classifiers demand the class, i.e., the dependent variable (the raw quality value in this case) as the last attribute in the rows of the dataset of selected attributes for modeling. Pre-processing of the dataset with discretization was necessary for a better performance of the simulation, with bins number of 15. Discretization, as a pre-processing step, was done to partition each numeric feature into a finite set of adjacent distinct intervals/items. A good discretization algorithm should not only characterize the original data to produce a concise summarization, but also help the classification performance [46]. In addition, discretization of the dependent variable, the raw quality value, was accomplished with bins number of 3, on the basis of the nature of the dataset. The kernel estimator is turned on in this classifier for the same reasons of improved performance.

Attribute selection feature of Weka is used to rank the attributes based on their importance on the classification performance of the Naïve Bayes classifier. The best attributes were then selected by running the model with the top prioritized ones part by part until the combination with the highest performance was achieved.

10-fold cross validation and different percent split evaluation techniques were tried, with 75% percent split yielding the most appropriate performance. As a result, the overall dataset was partitioned into a ratio of 3:1 to the training set and the test set, respectively,

to develop and evaluate the learning model. This was made possible by using the percent split option of the Weka explorer.

The selected attributes and the raw quality value as the last attribute in the rows were used to simulate the model using the classify button of the Weka explorer window, by manipulating the percent split field to 75%, indicating the training portion of the dataset of the total working data. The raw quality value represents the dependent variable, serving as the foundation for the sound approximation of the dataset to its fitting raw quality value class. An evaluation simulation of the performance of the model was conducted using the remaining 25% of the dataset for test set at the same simulation as the learning model.

C4.5 classifier, of all the input attributes, selects those attributes of the data that most effectively discriminate the dataset into its appropriate simulated category to build the learning and evaluation models. Stratified 10-fold cross validation technique was the most appropriate performance evaluation technique used to build and evaluate the simulated model.

In general, both models function the statistical analysis of deviations and distributions of the attributes by utilizing a predefined portion of the empirical dataset for model building (Figure 3.7.), and the residual predefined portion for the testing simulation of the model built (Figure 3.9.), thereby providing the classifier. Model outputs provide various statistic values, mean square error, percent corrects and percent incorrect and the confusion matrix, for evaluation, prediction and inference of model performances.

A separate simulation of classification models was run for morphological and color features of the coffee beans using similar procedures and tools as the combined ones. The aim was to generate an insight and conduct comparative analysis about the model outputs under such alternative attributes, in addition to their combined function.

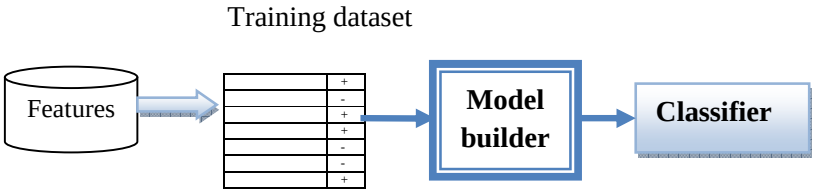


Figure 3.7. Learning Model for the Naïve Bayes and C4.5 classifiers.

Artificial Neural Network classification

The application of the artificial neural networks (ANN) comprised an important component in this computer-assisted raw quality value simulation assignment. The flexible learning algorithm, diverse network topology, fast learning capability and high error tolerance makes the neural networks powerful analytical tools. The neurosolutions for excel version 6.0 tool assists this artificial neural network classification task, with the dataset modified to the neurosolutions format.

Conversion of the numeric raw quality values to nominal scale was made possible with the associated discretization of the working raw quality values as high, medium and low, representing the data preparation approach for the simulation. These nominal values were used as an output column label name in the excel spreadsheet together with the associated input attribute values. The values for these nominal raw quality value output columns was then filled with the use of binary numbers, reflecting the presence or absence of the specific nominal value that represents the specific set of record in the actual dataset.

Sensitivity and robust model performance trials with differential input of sets of attributes iteratively in the model assisted final selection of attributes for the base model. Setting the input range (independent variable) and the desired output range (dependent variable) was an important procedure while working with neurosolutions for excel. Stratified random allocation of the dataset was made to training set, testing set and the cross-validation set, from the inherent three strata of the raw quality values. Likewise, 65% of the dataset as assigned to the training set, 10% to the cross-validation and 25% to the test set. Training data is the portion of the data employed to actually train the network. This is normally the largest portion of the data. Cross Validation data was used to intermittently

validate the training. Cross-validation is very useful tool for preventing over-training of the model. It used to stop the network training when the network starts to specialize too much on the training data. Test data was used to validate the results of a trained network. The analytical computations of dispersions and distributions of the coffee bean feature attributes are conducted in ANN by using the training set which is iteratively evaluated with the cross validation set to build the model. This will finally generate the classifier after being evaluated for its performance with the predefined test set (Figure 3.8.).

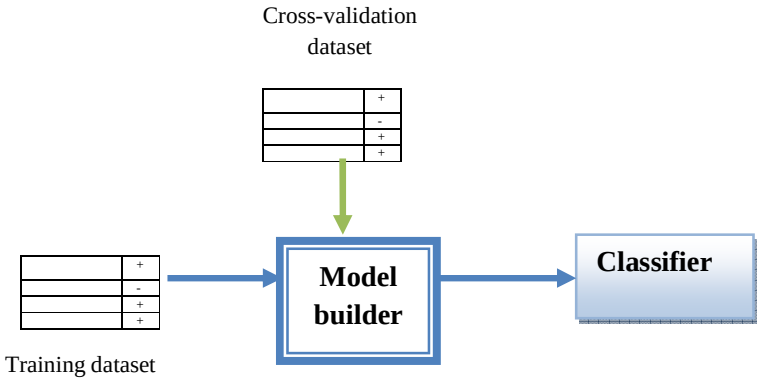


Figure 3.8. Learning Model design for Artificial Neural Networks Classifier.

A supervised feed forward multiple layer preceptors (MLP), a universal pattern classifier allowing the discriminant functions to take any shape, assisted to model the classifier with 2 hidden layers. Back propagation learning rule was incorporated to calculate the shares of the errors in model building and to modify connection weight. MLP is also suitable as the desired response of the outputs is known beforehand. It is one of the most commonly implemented neural network topologies.

The trained results are automatically tested for the neural networks, providing a summary of the network performances in the model as an output. The training and testing simulations in this classifier focus on confusion matrices, percent correct and performance matrices.

3.8 Regression Analysis

An additional evaluative and exploratory analysis was conducted by using the regression field of neurosolutions for excel to the overall working dataset that is used as an input to develop the base classifier. Regression is an important statistical method describing the relation between the dependent and independent variables that are utilized to build the classification model, thereby promoting an objective evaluation of the classifier itself. This regression of the dataset over the dependent variable, i.e., the raw quality values, assisted to compute the coefficient of correlation with the help of which can be predicted and projected the performances and evaluations of the simulated learning models in describing the variations of the raw quality values due to the mentioned attributes of the coffee beans.

Similar features of the ANN classifier as the classification task were applied for establishing the regression model in the mentioned classifier. Model outputs comprised correlation coefficient, mean square error and mean, maximum and minimum absolute errors.

4. Implementation and discussion:Raw quality value classification

Models must have precise structural and behavioral abstractions in order to be correctly simulated. The sought-after objective of this research, classification and determination of the raw quality values of unknown coffee bean samples on the basis of an automated learning model, was tried to be achieved by utilizing the approaches, procedures and tools as has been mentioned in the design part of this research.

Though differential outcomes and results could occur depending on the types of the techniques employed, an objective evaluation of performances was analyzed on numerous methods of statistical computational tools to make decisions on the best practicable and applicable ones. A detail provision of the outcomes and discussions of the dataset analysis, processing and simulation tasks with this regard will be the concern of the current part. The first two sub-sections will try to provide and discuss on the achievements of the sample coffee bean image analysis activities for the sake of an improved extraction of the required feature attributes. Next to this comes a sub-section telling the detailed story of the aggregate function development for further modeling tasks. Detailed provision of results and associated discussions with regard to the classification and evaluation aspects of the research will be the issue of the last two sub-sections. Focal issue in the last sub-section will be analysis of regression model evaluating the relations between the variables used as input for the classification purpose.

4.1. Image analysis and feature extraction results

ImageJ program assisted the processing, analysis and feature extraction activities of all the captured coffee bean images. Enhanced and segmented coffee bean images were pioneer outputs of the research, whereby same were used as inputs for the succeeding phase of feature extraction in the program.

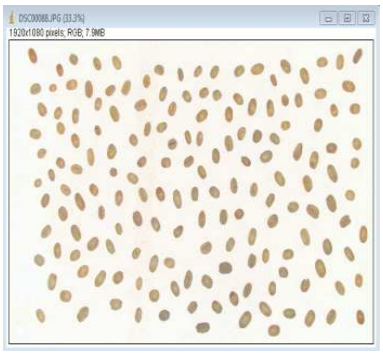
4.1.1. Enhanced and segmented images for data mining

A series of image processing techniques were used prior to extraction of coffee bean features from a given image. The developed output from such image enhancement, segmentation and advanced enhancement of all the coffee bean sample images is represented in figure 4.1a-f.

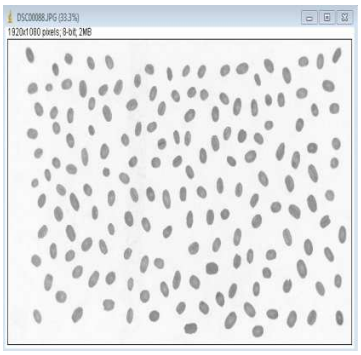
Consideration of morphological features calls for the image enhancement and segmentation procedures shown in figure 4.1. (a) - (e), as the extraction of color features was assisted only by conducting background subtraction and image tresholding.

Examination and manipulation of the images by minimizing noises from the background and by improving the clarity and type of the images made possible the succeeding acquisition of necessary features of the captured images for further character modeling and evaluation using computer assisted technologies. Improvement of the clarity of the beans with respect to the background, meaningfully subdivided individuals with similar attributes and geometrically well represented beans are seen as series of image development procedures, Figure 4.1.(a) –(e).

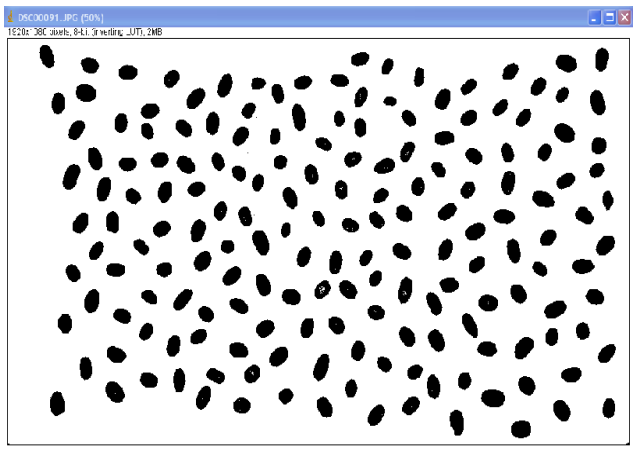
(a) Original image



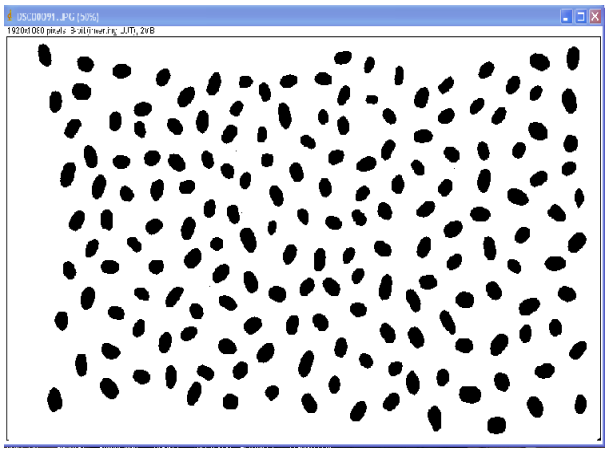
(b) Background subtracted image



(c) 8-bit grayscale image



(d) The binary image



(e). A binary image with the holes being filled

Figure 4.1. A representation of an original coffee bean image (a) which was then background subtracted (b) and then converted to gray scale image(c), a binary/segmented image (d), and holes filled image (e).

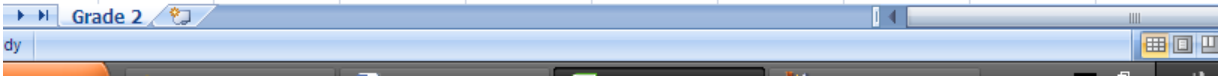
4.1.2. Extracted coffee bean features

Enhanced images assisted the generation and computation of the important and inherent characteristic features of these agricultural products. Morphological and color features of the coffee beans were of focal concern in this research and were extracted accordingly to generate numeric values for raw quality value determination modeling. A sample extracted morphological feature of a coffee bean sample is shown in Table 4.1.

A significant variability can be observed between the various values of the coffee bean morphological features, thereby influencing further simulation and inferences about the products. This implies for the fact that statistical transformations and variations evaluations are concern-worthy issues with regard to optimizing feature characterization and selection for further analysis and inferences. On the contrary, almost the entire sample beans possess higher similarity for their certain features.

Table 4.1. Sample morphological features

A	B	C	D	E	F	G	H	I	J
No	Label	Area	Perim.	Major	Minor	Circ.	Feret	AR	Round
1	2:DSC00087	340	69.598	23.819	18.174	0.882	24.739	1.311	0.763
2	2:DSC00087	401	78.669	30.23	16.889	0.814	31.241	1.79	0.559
3	2:DSC00087	539	90.669	33.296	20.612	0.824	33.734	1.615	0.619
4	2:DSC00087	361	70.426	24.812	18.525	0.915	25.554	1.339	0.747
5	2:DSC00087	280	64.184	21.209	16.809	0.854	23.022	1.262	0.793
6	2:DSC00087	412	76.184	26.797	19.576	0.892	27.857	1.369	0.731
7	2:DSC00087	407	75.598	25.476	20.341	0.895	26.249	1.252	0.798
8	2:DSC00087	403	74.527	25.959	19.766	0.912	26.249	1.313	0.761
9	2:DSC00087	472	82.326	30.542	19.676	0.875	31.064	1.552	0.644
10	2:DSC00087	371	72.426	24.95	18.933	0.889	25.612	1.318	0.759
11	2:DSC00087	519	85.012	29.976	22.045	0.902	30.676	1.36	0.735
12	2:DSC00087	444	78.184	27.843	20.304	0.913	28.46	1.371	0.729
13	2:DSC00087	270	69.113	28.192	12.194	0.71	28.443	2.312	0.433
14	2:DSC00087	349	69.598	24.033	18.489	0.905	24.739	1.3	0.769
15	2:DSC00087	430	77.598	27.549	19.874	0.897	27.857	1.386	0.721
16	2:DSC00087	544	90.083	33.302	20.799	0.842	33.287	1.601	0.625
17	2:DSC00087	619	95.64	35.314	22.318	0.85	35	1.582	0.632
18	2:DSC00087	417	78.669	29.058	18.272	0.847	30	1.59	0.629
19	2:DSC00087	500	86.426	32.584	19.538	0.841	32.985	1.668	0.6



The color attributes of all the sampled items were extracted in a similar manner and made available for further applications. A sample extracted color feature of a coffee bean sample is shown in Table 4.2.

Table 4.2. A sample color feature of a sample coffee bean.

	A	B	C	D	E
1	Blue	Green	Red	Brightness	Saturation
2	66.406	85.941	97.509	125.562	84.638
3	67.517	89.337	101.376	97.604	88.593
4	70.6	93.28	109.424	101.468	93.011
5	71.585	92.142	104.745	109.501	84.237
6	68.217	90.37	107.294	104.873	96.371
7	70.051	91.106	103.229	107.368	77.36
8	76.797	96.952	108.133	103.452	88.407
9	68.386	86.817	96.723	108.302	80.331
10	65.391	87.008	96.204	96.902	87.439
11	58.097	73.499	88.221	96.411	96.049
12	65.116	86.913	102.87	88.444	96.547
13	71.824	91.263	103.084	102.933	81.217
14	65.725	79.269	88.419	103.201	77.209
15	68.89	89.352	99.789	88.912	83.233
16	64.836	86.897	99.182	99.849	95.448
17	68.775	90.235	102.102	99.37	87.364

RawVAlue_9

Ready

Almost similar measures of central tendency values are observed for various samples, of their respective features, representing a kind of reduced variability.

4.2. Features Analysis: Derivation of aggregate values for modeling

Generation of an aggregate data that represents the cumulative effect of all features of the coffee beans can be mentioned as the preparation of the working dataset for the rest of the data mining activities.

Computation of an aggregate function for the measured morphological and color features of the samples was the first core business in the research under consideration. The aggregate values dealt with in this respect include the maximum, minimum, average,

variance and standard deviation values for each of the morphological features, being computed for all the sample coffee beans. The grand mean value of all the attributes of color features was computed for each of the samples and one dataset file is created. These aggregate functions are calculated from the 324 sampled coffee bean images. These samples were captured from eight different types of grades, each grade having a sample number extending from 2 – 4. A summary of the aggregate dataset for these morphological features is shown in table 4.3.

Table 4.3. Sample aggregate values from the morphological features.

	A	B	C	D	E	F	G	H	I	J	K	L
1	minArea	minPer	minMaj	minMin	minCir	minFer	minAR	minRoun	maxArea	maxPer	maxMaj	maxMin
2	873	109.397	38.302	25.026	0.658	39.217	1.092	0.417	2569	207.723	79.829	46.534
3	643	102.912	32.562	23.001	0.664	37.121	1.034	0.462	2495	199.137	77.458	45.583
4	418	77.497	26.648	19.972	0.759	28.16	1.111	0.492	2635	204.309	75.318	46.827
5	709	101.983	35.045	25.759	0.704	36.878	1.095	0.442	2743	211.765	82.47	47.397
6	637	93.64	29.531	27.465	0.742	31.321	1.075	0.484	2574	203.865	80.982	47.904
7	949	120.125	38.628	22.673	0.672	39.812	1.145	0.425	2475	196.267	75.037	46.246
8	497	98.811	34.054	16.34	0.64	34.785	1.124	0.422	2707	208.167	82.052	49.632
9	787	114.468	40.217	21.944	0.648	42.426	1.107	0.421	2935	214.995	82.82	47.21
10	960	114.711	39.047	29.139	0.677	39.408	1.112	0.454	2880	209.238	81.616	50.421
11	59	34.042	11.07	6.786	0.617	12.806	1.114	0.413	2683	207.823	78.84	46.587
12	700	106.225	37.915	23.507	0.683	39.925	1.12	0.457	3030	211.823	76.749	53.413
13	766	110.468	35.855	18.337	0.529	36.892	1.109	0.345	2621	204.267	81.068	47.537
14	899	112.225	39.313	24.314	0.669	40.497	1.167	0.423	2699	206.35	81.656	47.081
15	904	114.225	40.408	28.251	0.728	41.437	1.154	0.516	2942	209.279	77.695	50.199
16	914	123.64	41.163	21.701	0.629	42.19	1.15	0.353	2605	200.551	75.655	48.259
17	800	112.953	34.133	25.847	0.632	39.319	1.144	0.465	2724	214.652	84.778	46.261
18	824	110.326	40.455	25.934	0.744	40.792	1.064	0.485	2873	205.924	76.722	48.848
19	811	107.983	35.284	22.309	0.616	38.328	1.074	0.36	2820	211.823	80.395	49.089
20	582	100.811	34.222	20.037	0.571	36.056	1.051	0.347	2994	213.581	78.734	50.472
21	604	102.669	32.967	19.61	0.688	34.205	1.101	0.436	2610	210.35	82.578	47.772

Maj | Min | Cir | Fer | Sheet1 | AR | Sheet2 | Round

Ready

4.3. Classification simulation: Outputs and trends of raw quality values

Corresponding model results and trends for two classification tools, Weka and Neurosolutions, to implement the Naïve Bayes, C4.5 and Artificial Neural Network classifiers, will be provided in this part of the thesis report. A combined morphological and color features dataset was used in this experimentation.

Model robustness and sensitivity was experimented by altering model evaluating techniques, discritization bins and dataset characteristics. Development of model for separate color and morphological features was also another sensitivity and robustness trial on the base model. 10-fold cross validation and different percent split values for the

training and test dataset comprise the model evaluating technique trials. Dataset characteristics were perturbed for the number and type of the aggregate feature values.

The accuracy and efficiency of models relies on the procedures of setting up the model initialization and parameter selection. The overall automated system modeling activity is conducted by using attributes that were either selected by the model itself or by those selected based on their suitability to the particular model.

4.3.1. Naïve Bayes classification

The attribute selection feature of Weka explorer assisted the statistical selection of the morphological and color features for simulation purposes using information gain attribute evaluator with the ranker search method. Overall features of 45 were available in the initial working dataset, out of which only the first 25 were prioritized and selected for the modeling purpose, on the basis of performance evaluations with varying numbers of the selected features. Average and standard deviations of all the morphological features, minimum values of area, perimeter, major axis and minor axis, and all the color features were amongst the highly prioritized and selected features.

All the extracted features of coffee beans are numeric values, where the use of these values in Naïve Bayes classifier needs conversion of these continuous numeric values to discrete values. The preprocessing facility of Weka unsupervised filtering technique was used to discretize the dataset into distinct intervals. Accordingly all input attributes were discretized into 15 bins whereas raw quality value attribute in to 3 bins, as mentioned in chapter three of this paper.

The raw quality value discretization in the model generated an output of groups of characteristic coffee bean samples possessing strong similarity with that of the characteristic distribution of the raw quality values and manually obtained grades of the working dataset. This is believed to pave the way for a strong and promising classification and discrimination capability of the model. Such conclusions on the final model parameters and dataset natures for the simulation purpose were however attained through a series of alternative attempts to evaluate output and model sensitivity (Table

4.4). Shaded part in the table shows parameterization route of the base model using Naïve Bayes classifier.

Table 4.4. Model and dataset parameter perturbation attempts for robust model performance in Naïve bayes.

Model	**Attributes**	**Evaluation technique**	**Performace**
Naïve Bayes	All	75% percent split	74.07%
		10-fold cross validation	73.15% (**Worst**)
	Average and standard deviation of Morphological features and color features	75% percent split	80.24%
		10-fold cross validation	75.93%
	Minimum, Average, standard deviation of Morphological features and color features	75% percent split	81.48%
		10-fold cross validation	76.85%
	MinArea, MinPer, MinMaj, MinMin, Average, and Standard deviation of all morphological features and color features	75 percent split	82.72% **(Best)**
		10-fold cross validation	76.54%

Model building and evaluation

Selection of 75% of the data set was used for training the classification learning model and the remaining 25% for testing to evaluate the performance of the model. Weka screen shot of outcomes of the model is shown below in figure 4.3. The output contains different statistic values, indicating the performance status of the model. 82.72% (Figure 4.2) of the test data set was classified correctly, with the moderately higher kappa statistic value of 0.73 strengthening this result. Kappa is a chance-corrected measure of agreement between the classifications and the true classes, and a kappa of 1 indicates perfect agreement, whereas a kappa of 0 indicates agreement equivalent to chance.

The model also yielded a mean absolute error value (0.14) which is lower than the root mean squared error (0.33) reflecting good model outputs [7]. The difference between these two statistic values is also small, underlining the robustness of the classification model.

Better performance can be concluded for higher true-positive-rate, lower false-positive rate and north-west side of ROC space, i.e., approaching to a value of 1[5]. A ROC area,

representing plot of the true positive rate against the false positive rate, of above 0.9 is modeled for the simulation dataset, depicting the higher discrimination measurement ability of the model. In most situations the discrimination ability of the forecast is not really considered useful in practice unless the ROC area is greater than 0.7. A higher rate of the true positive and lower false positive values characterize the model results.

The confusion matrix in the form of a contingency table shows higher correct allotments to those with lower raw quality value and tends to decrease as the raw quality value increases. This can be linked to the differential existence of diverse and numerous defects at the lower raw quality value items, making it suitable and clear for discrimination into their relevant classes. It can be however seen from the table that an acceptably higher performance is yielded by the model.

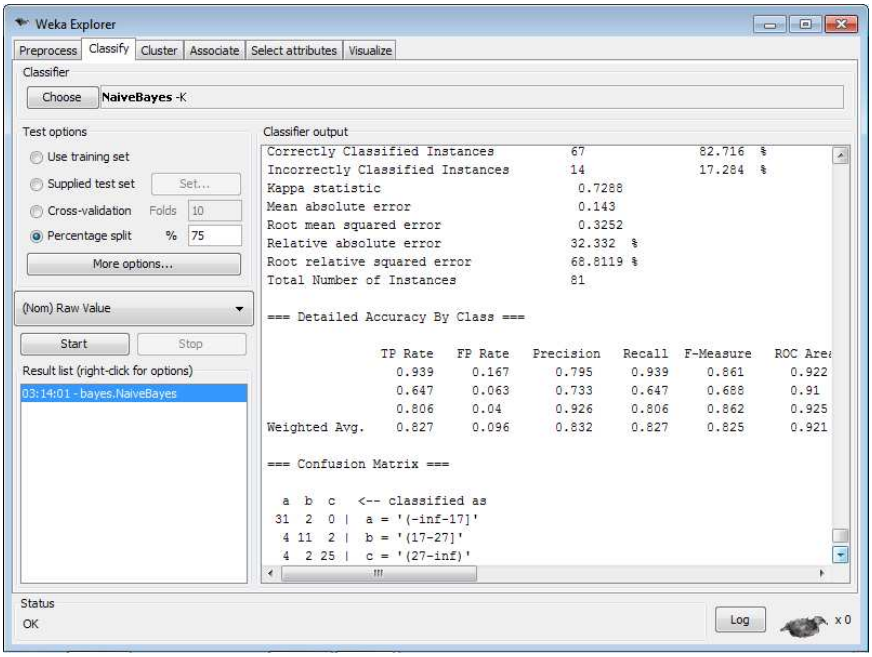


Figure 4.2. Screen shot of Weka environment for model simulation and evaluation.

The model was also run separately for color and morphological features of the concern sample dataset, yielding respective results of 56% and 79% for the percent correctly classified instances. This implied that the morphological features are stronger in discriminating the various classes for the sample coffee beans than the color features.

This achievement can also be linked and supported with the fact that the attribute selection feature of Weka has prioritized these morphological features of the dataset than the color counterparts.

An overall higher performances and lower discrepancies and errors were statistically computed for the models run with the combined features of color and morphological attributes of coffee beans (Table 4.5.).

Table 4.5: Summary of models evaluation using Naïve Bayes classifier.

Modeled statistic values	Using morphological features	Using color features	Using combined features
Correctly classified	79.01%	56.79%	82.72%
Kappa statistics	0.66	0.35	0.73
Mean absolute error	0.16	0.31	0.14
Root mean squared error	0.36	0.44	0.33
Relative absolute error	36.81%	70.65	32.33%
Root relative squared error	75.65%	93.20%	68.81%

4.3.2. C4.5 classification

The simulation with the C4.5 classifier involved similar approaches in discretization of dataset as the Naïve Bayes. Likewise, 15 bins discretization for all independent attributes and 3 bins discretization for the dependent attribute, raw quality value were persuaded. This classifier utilizes those attributes with higher discrimination power for the purpose of classification.

Model building and evaluation

Nine attributes, Minimum Area, Average of Perimeter, Major, Minor, Circularity, Roundness, Standard deviation of Circularity, and the Color Features, i.e., Blue, Red, and

Brightness, were selected by the classifier for building the model. Sensitivity and perturbation analysis for robustness of model performance were run for some attributes based on Weka prioritization for Naïve Bayes classifier. Differential application of evaluation techniques, including 10 fold cross-validation and 75% splits, assisted the perturbation analysis with varying attribute attempts. 10 fold cross-validation evaluation technique with the model selected attributes, shaded route (Table 4.6.), yielded higher performance.

Table 4.6. Sensitivity analysis with certain attribute and evaluation technique attempts to evaluate model performance.

Model	**Attributes**	**Evaluation technique**	**Performace**
C4.5	All	75% percent split	72.84% **(Worst)**
	MinArea, Average of Perimeter, Major, Minor, Circularity, round. Standard deviation of Circularity Color Features: Blue, Red, and Brightness	75% percent split	80.25%
		10-fold cross validation	82.09% **(Best)**
	MinArea, MinPer, MinMaj, MinMin, Average, and Standard deviation of all morphological features and color features	75 percent split	74.07%
		10-fold cross validation	79.32%

C4.5 with the base dataset of combined morphological and color features provided for 82.09% of the test data set to be correctly classified, with a kappa static value of 0.7 (Figure 4.3). A lower mean absolute error (0.16) than the root mean squared error value (0.31) is in compliance with model output.

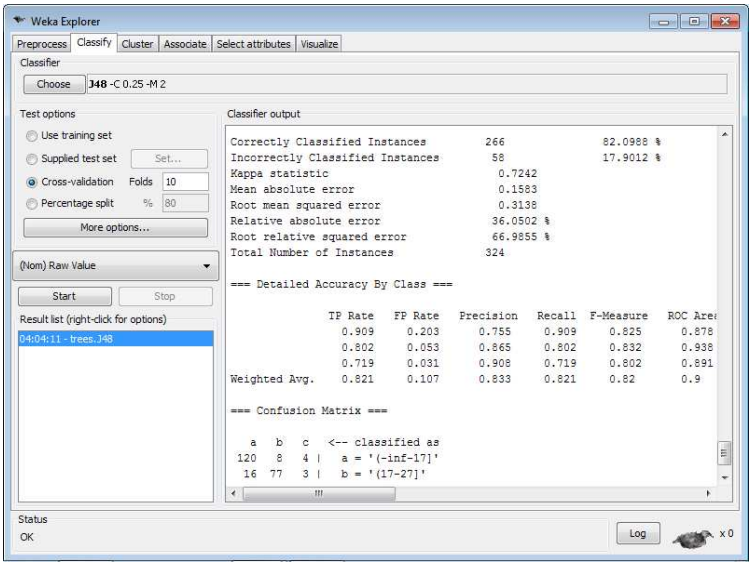


Figure 4.3. Screen shot of Weka environment for model simulation and evaluation with C4.5.

Similar trend of incremental performance with decreasing raw quality value was modeled as in the Naïve Bayes classifier. A separate run for color and morphological features yielded respective results of 71.91% and 63.27% for the percent correctly classified instances. It was underlined, as in the Naïve Bayes classifier, in the outputs for the morphological features to be stronger in discriminating the various raw quality value classes of coffee beans than the counterpart color features. It should however be noted that the combined features of color and morphological attributes modeled for an overall higher performances and lower discrepancies and errors (Table 4.7.).

Table 4.7: Summary of model evaluation for color and morphological features

Statistical measures/Model	Using morphological features	Using color features	Using combined features
Correctly classified	71.91%	63.27%	82.09%
Kappa statistics	0.56	0.43	0.72
Mean absolute error	0.23	0.28	0.15
Root mean squared error	0.37	0.41	0.31
Relative absolute error	53.39%	64.58%	36.05%
Root relative squared error	78.96%	87.75%	66.98%

4.3.3. Artificial neural network (ANN) classification

Discretization of raw quality values in to three was conducted manually by assigning three nominal values as high (H), medium (M), and low (L), which is further converted into binary valued attribute. Selection of ANN parameters was made after building the model with the different attributes and finally attributes with high performance were accepted. Higher performance was exercised with those attributes which yielded higher performance using the Naïve Bayes classifier (Table 4.8).

Table 4.8. Some of the trials made to select attributes with high performance for ANN classifier.

Model	Selected Attributes	Performance
Artificial Neural Networks	All	74.07%
	Minimum, Average, Std.Deviation, and Variance of Morphological features and all color features	62.96% **(Worst)**
	Minimum, Average, and Std.Deviation of Morphological features and all color features	64.20%
	Average and Std.Deviation of Morphological features and all color features	65.43%
	Minimum of Area, Perimeter, Major Axis, and Minor Axis; Average and Std.Deviation of Morphological features and all color features (Attributes which are used for Naïve Bayes)	80.25% **(Best)**
	MinArea, Average of Perimeter, Major, Minor, Circularity, round. Standard deviation of Circularity, Color Features: Blue, Red, and Brightness (Attributes which are used for C4.5)	67.90%

Model Training and evaluation

The training and testing of the classification model using ANN is simulated together by specifying the portion of the training rows and testing rows in the data set. In addition, some part of the rows should be specified for cross validation dataset to intermittently validate the model training. Figure 4.4 shows the environment in neurosolutions to build and test a classification model. Generalized feed forward multilayer perceptron topology with two hidden layers was used to simulate the model. There were four layers in this respect, an input layer consisting of 25 nodes for morphological and color features, the two hidden layers, and an output layer consisting of 3 nodes representing the nominal values of raw quality value.

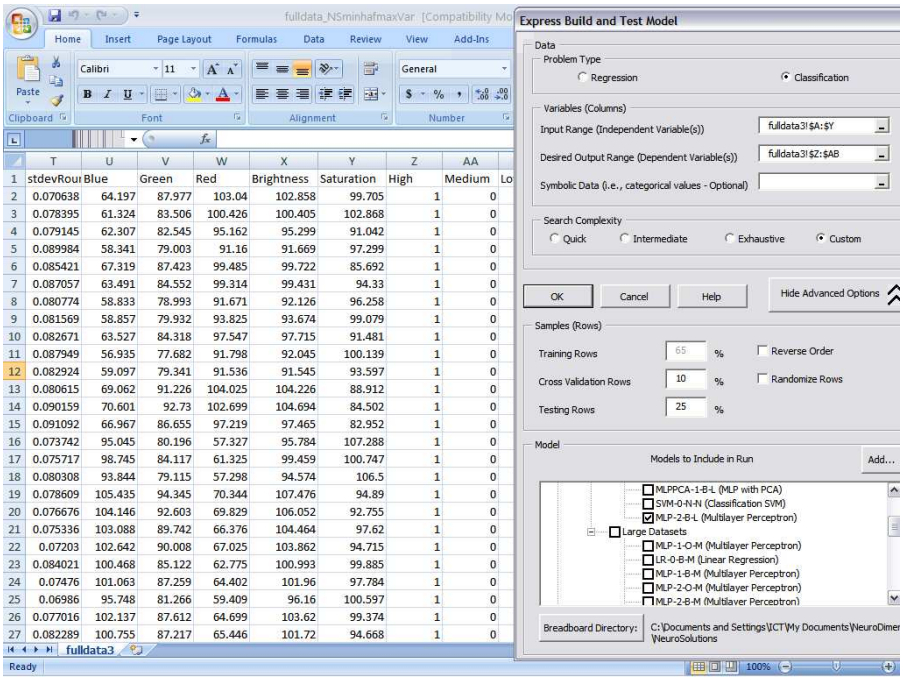


Figure 4.4. Neurosolution screen shot for data preparation and topology selection.

The simulation was conducted on the combined color and morphological features of sample coffee beans, though separate trials of the model were done for each of the attributes for evaluative purposes.

Classifier outputs for the three modeling dataset partitions (Table 4.9) yielded meaningfully lower values of mean square errors and higher values for correlation coefficients. 80.25% of the samples are classified correctly with regard to their real raw quality value group.

Table 4.9. Modeled statistic values for the training, cross-validation and testing results.

	Training	Cross Val.	Testing
# of Rows	211	32	81
MSE	0.00619	0.09438	0.09008
Correlation (r)	0.98549	0.79805	0.80426
# Correct	208	28	65
# Incorrect	3	4	16
% Correct	98.58%	87.50%	80.25%

A comparative analysis of the observed and expected raw quality values (Table 4.10.) of the evaluation model using ANN reveals higher classifier performance, ranging from very high to moderate discrimination trends. Almost all medium class testing datasets were assigned to their exact class.

Table 4.10. Confusion Matrix for the testing output of ANN model.

Actual	Target		
	High	Medium	Low
High	23	7	0
Medium	0	16	1
Low	1	7	26

Simulation with the use of morphological features modeled for 64.20% of the dataset to be correctly classified, which is higher than the simulation result while using color features only, 62.96%. Lower value of mean square error for simulation of morphological features than color features was modeled. In addition higher correlation coefficients for morphological features were found as compared to the color features. A summarized output description table is shown below (Table 4.11).

Table 4.11. Summarized results of modeling for the alternative features

	Morphology features	Color Features	Combined Features
MSE	0.205731469	0.327782871	0.090077588
Correlation (r)	0.60325102	0.144257092	0.804257115
# Correct	52	51	65
# Incorrect	29	30	16
% Correct	64.20%	62.96%	80.25%

4.3.4. Summary of classification performance

Statistical computations dealing with coffee bean raw quality value classification yielded almost similar model outputs for the various classification tools employed (Table 4.12), with the Naïve Bayes yielding higher values than all the others. The higher model

performace value in Naïve Bayes is attributed to suitability 0f the classifier under smaller number of datasets. This together with the resulting higher model statistic values and lower mean and relative error rates [7] for performance evaluation depict the suitability and possibility of implementing the automated classification system using similar dataset orientation and tools. Better performances could even be achieved through further in-depth implementations involving larger samples.

Table 4.12. Performance of the model in different classifiers.

Model	Combined Feature Performance (% correctly classified)	Morphology features	Color Features
C4.5	82.09	71.91%	63.27%
Artificial neural networks	80.25	64.20%	62.96%
Naïve Bayes	82.72	79.01%	56.79%

It becomes obvious that the two different coffee bean features when used in combination and separately in the classification tools yielded various model outputs. The combined feature classification models provide the higher model performance values, succeeded with those of the morphological features only (Table 4.12). [20] has also concluded similar achievements in another coffee bean classification experiment.

Robustness and sensitivity of the model was also analyzed with manipulation of selected and applicable model evaluation techniques. Discretization trials with varying number of bins comprised the final selection of 15 bins that provides better performance of all classifier models in Weka classification tool. The trial was made by increasing and decreasing the bins value until the optimum performance was obtained. Multiperceptron neural network models contain three or more layers, the input, output and one or more hidden layers while mapping a specific set of dataset input into an output model. The differential existence of one or two hidden layers in this model was simulated for the focal dataset and the respective performances compared. Despite the fact that one hidden layer is the most commonly applied one in many experiments, higher performance was

recorded in the current experiment with the use of two hidden layers (80.09% correctly classified) in the ANN model than one hidden layer (64.2% correctly classified).

Discritization of the raw quality values into three nominal classes assisted in achieving higher model performances for all the classification approaches. Discritization, by transforming continuous valued features into discrete intervals, makes classification performance more effective and higher [6].

Weka classification tools yielded higher classifier performance for dataset with lower raw quality values, as the artificial neural networks provided higher performance values for those of medium raw quality value individuals. This is in compliance with the manual raw quality value grading systems that depend on visual evaluations and classify items on the basis of visible defects, these being clear and abundant on the lower raw quality valued samples.

4.4. Regression analysis for classification variables

Regression trends were analyzed statistically to explore the relations between the raw quality values and the coffee bean morphological and color features for the sampled working dataset of varying grades. Such a regression is intended to assist the evaluative role of the experiment to the performance and robustness of the classification simulations by describing the relation between the classification variables. Such experimentation was done by running the regression analysis model in the neurosolutions for excel application. The dataset that was used for the classification model building was used as an input for this regression analysis too, to generate a deeper insight and inferences about the relationships between the concern dependent and independent variables. This enabled a more explicit statistical evaluation and trend analysis on the classifier developed model to measure and classify raw quality value of an incoming unknown sample entity.

With this regard, the regression analysis in neurosolution for excel yielded a more or less higher coefficient of correlation between the related variables for raw quality value determination in the performance evaluation model (0.86). The lower mean square error values of the model built also indicate higher accuracy of the regression model for theses variables.

Table 4.13. Model evaluation using regression analysis in neurosolution for excel.

Static Measures	Training	Cross Val.	Testing
# of Rows	194	49	81
MSE	0.002903	0.060498	0.062248
Correlation (r)	0.993049	0.847846	0.857992
Mean Absolute Error (MAE)	0.026983	0.083155	0.109878

Lower mean square errors in all cases than the mean absolute error underline model accuracy with minimum error factors.

Similar regression analysis was conducted on separate morphological and color features of the sampled coffee bean dataset, with the former providing better relational trends between the variables (0.63) than the later attribute (0.25). This result is in compliance with the yielded classification outputs for morphological features than the color features.

5. Conclusion and Recommendation

Automated sorting and classification systems for agricultural products are proven to be less costly, efficient and non-destructive. Application of this technology makes effective quality control and inspection aspects for such economically important commodities. With this regard, this research has focused on using image processing techniques and approaches to classify raw quality value of sample coffee beans by employing Naïve Bayes, C4.5 and artificial neural networks classification approaches. The achievements obtained in this research work indicate the possibility of applying classification of raw quality value using computer vision system.

Morphological and color features were the attributes extracted from the sample coffee bean images of various grade levels and used for the classification purpose. Image pre-processing techniques on the original images enhanced the harvesting of such important features in a reliable and effective manner.

The classification models built with all the classifier tools were evaluated for almost a homogenous performance, with the Naïve Bayes yielding the highest model performance (82.72% correctly classified), and the C4.5(82.09%) and artificial neural networks(80.25%) succeeding respectively. Combined aggregate feature values of both the morphological and color features were used to build and evaluate the base raw quality value classification model with all classifiers used in this research.

Sensitivity of the classifier model was however attempted by running the classification model for the separate aggregate feature values, i.e., morphological and color features. The separate trials produced higher performances in all classifiers for the morphological features than the color features, reflecting the suitability of the morphological features for grading and sorting the coffee bean samples than the color features. It should however be underlined that the highest overall performances and hence suitability for classification purposes in this research is concluded for the combined morphological and color features.

Discritization of the dataset into raw quality value intervals is concluded to be an important tool contributing to improved model performance. Likewise, discretization of

the raw quality values into three bins in classification generated higher performance evaluation outputs amongst a set of alternative discretization options.

Regression model for the relation between the raw quality values and the combined aggregate feature values of the sample coffee beans simulated higher correlation coefficients (0.86), implying the suitability and applicability of the classification achievements for the working dataset in this experimentation.

An important recommendation to forward from this research could be the launching of developing a working base classification model for raw quality value classification purposes by utilizing larger number of dataset from each grade level of coffee bean sample. This could also be supported by another future research that aims at prediction of actual raw quality values by utilizing a meaningfully larger dataset. Important is due consideration to the number and type of samples and to the data acquisition environment to effectively achieve the intended end-goals of developing an applicable automated computer vision system for this purpose.

This automated technique might also be a potential approach in Ethiopia assisting quality control and grading/sorting activities of other important agricultural products like fruits and cereals.

References

[1] Arifin, A.Z. Asa, A. 2006. *Image segmentation by histogram thresholding using hierarchical cluster* analysis, Pattern Recognition Letters 27, 1515–1521.

[2] Brosnan, T. and D.-W. Sun, 2004. "*Improving quality inspection of food products by computer vision*—a review." Journal of Food Engineering 61: 3-16.

[3] Carlos G., *2003. Artificial Neural Networks for Beginner*, Cornell University Library.

[4] Ethiopian Commodity Exchange, http://www.ecx.com.et/commodities.aspx, visited on September 1, 2009.

[5] Fawcett, T., 2004. *Evaluating Performance*, "ROC Graphs: Notes and Practical Considerations for Researchers", HP Laboratories, MS 1143, 1501 Page Mill Road, Palo Alto.

[6] Hacibeyoglu, M., Arslan, A., Kahramanli, S, 2011. *Improving Classification Accuracy with Discretization on Datasets Including Continuous Valued Features*, World Academy of Science, Engineering and Technology 78.

[7] Hyndman, R. and Koehler, A., 2006. *Another Look at Measures of Forecast Accuracy*. International Journal of Forecasting 22: 679-688.

[8] Jha G.K., *Artificial Neural Networks*, Indian Agricultural Institute, PUSA, New Delhi-110 012.

[9] Jusoh, N.A. and Zain, Jasni M. and Ismail N., Kamariah N. and Abd Kadir, Asmawaty T., 2007. *Comparison between Techniques in Feature Extraction.* In: ICEEI2007, Bandung, Indonesia.

[10] Karoui I., Boucher,R.F., J.-M., Augustin, J.-M. *Region-Based Image Segmentation Using Texture Statistics And Level-Set Methods*, GET, ENST Bretagne, CNRS TAMCIC, CS 83818 - 29238 Brest Cedex, France.

[11] Kavdir, I. and Guyer, D. E., 2003. *Apple Grading Using Fuzzy Logic*, Journal of Agricultural Engineering Research. 27:375-382.

[12] Korting, T.S., *C4.5 algorithm and Multivariate Decision Trees*, Image Processing Division, National Institute for Space Research – INPE, SP, Brazil.

[13] Kotsiantis, S.B., 2007. *Supervised Machine Learning: A Review of Classification Techniques*, Informatica 31, 249-268.

[14] Liu, W., Tao, Y., Siebenmorgen, T.J., Chen, H., 1997. *Digital image analysis method for rapid measurement of rice degree of milling*. In: 1997 ASAE Annual International Meeting Technical Papers, Paper No. 973028.

[15] Li, Q., Wang M. And Gu, W., 2002. *Computer vision based system for apple surface defect Detection*, Computers and Electronics in Agriculture 36 215-223.

[16] Majumdar, S., Jayas, D.S., Bulley, N.R., 1997. *Classification of cereal grains using machine vision1: Morphological features*. In: 1997 ASAE Annual International Meeting Technical Papers, Paper No. 973101.

[17] Malamasa, E. N., Petrakis, G. M., Zervakis, M., Petit, L., Legat, J., 2003. "*A survey on industrial vision systems, applications and tools*." Image and Vision Computing 21: 171–188.

[18] Maxwell C.N., 2007. *The Coffee Grading Process From Bean to Brew*, Ezine articles.

[19] Meftah, S. A., Abdul Rashid, M.S., 2011. *Recent Methods and Techniques of External Grading Systems for Agricultural Crops Quality Inspection* – Review, International Journal of Food Engineering, Volume 7, Issue 3.

[20] Minase H., 2008. *Image analysis for Ethiopian coffee classification*, Addis Ababa University department of Computer Science.

[21] Mitchell, T.M., 2010. Machine Learning, *Generative and Discriminative Classifiers:Naïve Bayes and Logistic Regression*, McGraw Hill.

.

[22] Nagata, M., Cao, Q., Bato, P.M., Shrestha, B.P., Kinoshita, O., 1997. *Basic study on strawberry sorting system in Japan*. In: 1997 ASAE Annual International Meeting Technical Papers, Paper No. 973095, ASAE.

[23] *Naive Bayes Classifier Introductory Overview*, StatSoft electronic statistics textbook, http://www.statsoft.com/, visited on May 11-2010.

[24] Narendra, V.G., Hareesh, K.S., 2010. *Quality Inspection and Grading of Agricultural and Food Products by Computer Vision*- A Review, Manipal Institute of Technology Manipal, Karnataka, India, *International Journal of Computer Applications* (0975 – 8887) Volume 2 – No.1.

[25] Narendra, V.G. and Hareesh, K.S., 2010. *Prospects of Computer Vision Automated Grading and Sorting Systems in Agricultural and Food Products for Quality Evaluation*, International Journal of Computer Applications (0975 – 8887), Volume 1 – No. 4.

[26] Neural networks: *A requirement for intelligent systems*, visited on May 11-2010.

[27] Ni, B., Paulsen, M.R., Reid, J.F., 1997. *Size grading of corn kernels with machine vision.* In: 1997 ASAE Annual International Meeting Technical Papers , Paper No. 973046.

[28] Nicolas, P., 2007. *Ethiopia's Coffee Sector: A Bitter or Better Future?*, Journal of Agrarian Change, Vol. 7 No. 2, pp. 225–263.

[29] Olga, M., 2009. *Image Pre-Processing Tool, Kragujevac J. Math.* 32, 97-107.

[30] Paulus, I., Schrevens, E., 1999. *Shape characterisation of new apple cultivars by Fourier expansion of digital images*. Journal of Agricultural Engineering Research 72, 113_ 118.

[31] Quinlan, J. R., 1993. *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers.

[32] Rafael, C.G. and Richard, E., 2002. Woods: *Digital Image Processing*, Second Edition, Pearson Education.

[33] Raicu, D.S., 2004. *Image Feature Extraction*, Visual Computing Workshop: Image Processing, DePaul University.

[34] Raji, A.O. and Alamutu, A.O., 2005. *Prospects of Computer Vision Automated Sorting Systems in Agricultural Process Operations in Nigeria*. Agricultural Engineering International: The CIGR Journal of Scientific Research and Development Vol. VII.

[35] Roseleena, J., Nursuriati, J., Ahmed, J. and Low, C.Y., 2011. *Assessment of palm oil fresh fruit bunches using photogrammetric grading system*, International Food Research Journal 18(3): 959-965.

[36] Roy, L., 2001. *The Ethiopian coffee filiere & its institutions: cui bono?*, Sheffield Hallam University, UK.

[37] Salem, S.A., 2010. *Image Segmentation by Using Edge Detection*, International Journal on Computer Science and Engineering Vol. 02, No. 03, 804-807.

[38] Surendra, K. and Ann, G., 2000. *ICO/CFC Study of Marketing and Trading Polices and Systems in Selected Coffee producing countries:* Ethiopia Country Profile.

[39] Tadhg, B. and Da-Wen, S., 2002. *Inspection and grading of agricultural and food products by computer vision systems,* a review, Computers and Electronics in Agriculture, 36: 193_ 213.

[40] Tobias, O.J., Seara, R., 2002. *Image Segmentation by Histogram Thresholding Using Fuzzy Sets,* IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 11, NO. 12.

[41] Visen1, N.S., Paliwal, J., Jayas, D.S., and White, N.D.G., 2004. *Image analysis of bulk grain samples using neural networks,* Department of Biosystems Engineering, University of Manitoba, Canada R3T 2M9.

[42] Wikipedia-Naive-Bayes-Classifier, **http://en.wikipedia.org/wiki/Naive_Bayes_classifier**, Visited on May 11-2010.

[43] Yang, C.C., Prasher, S.O., Landry, J.-A., Ramaswamy, H.S. and Ditommaso, A., 2000. *Application of artificial neural networks in image recognition and classification of crop and weeds.*

[44] Zayas, I.Y., Martin, C.R., Steele, J.L., Katsevich, A., 1996. *Wheat classification using image analysis and crush force parameters*. Transactions of the ASAE 39 (6), 2199_ 2204.

[45] Zhang, G.P., 2000. *Neural Networks for Classification: A Survey*, IEEE Transactions on systems, MAN AND CYBERNETICS-PART C: Applications and eviews, VOL. 30, NO. 4.

[46] Zhu Q., 2009. *Effective Supervised Discretization for Classification based on Correlation Maximization*, Coral Gables, FL 33124, USA.

[47] Harry Zhang, 2004. *The Optimality of Naive Bayes*, FLAIRS2004 conference

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualifications except as specified.

Asma Redi

_____

October 21, 2011

This thesis has been submitted for examination with our approval as university advisors.

Dr. Sebsibe H/Mariam _____ _____