BINDURA UNIVERSITY OF SCIENCE EDUCATION

FACULTY OF SCIENCE EDUCATION

DEPARTMENT OF COMPUTER SCIENCE

AUTOMATED SYSTEM FOR TOBACCO QUALITY CLASSIFICATION AND PRICING USING IMAGE PROCESSING AND MACHINE LEARNING ALGORITHMS: A Case of Tobacco Auction Floor.

BY

DERECK NDLOVU (B212447B)

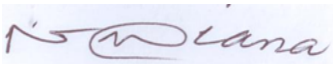
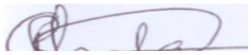
SUPERVISOR: MR D. NDUMIYANA

A PROJECT SUBMITTED IN PARTIAL FULFILMENT AS PER THE REQUIREMENTS FOR THE BACHELOR OF SCIENCE HONOURS DEGREE IN SOFTWARE ENGINEERING AT BINDURA UNIVERSITY OF SCIENCE EDUCATION

JUNE 2025

APPROVAL FORM

The undersigned certify that they have supervised the student Ndlovu Dereck's dissertation entitled "**AUTOMATED SYSTEM FOR TOBACCO QUALITY CLASSIFICATION AND PRICING USING IMAGE PROCESSING AND MACHINE LEARNING ALGORITHMS: A Case of Tobacco Auction Floor**." submitted in Partial fulfilment of the requirements for the Bachelor of Software Engineering Honours Degree of Bindura University of Science Education.

D. NDLOVU		**20…/…06/ 2025…..**
STUDENT	**SIGNATURE**	**DATE**
MR D. NDUMIYANA		**28…./08 / 2025**
SUPERVISOR	**SIGNATURE**	**DATE**
P CHAKA		**21/…08/.2025……**
CHAIRPERSON	**SIGNATURE**	**DATE**

Abstract

Automated Tobacco Leaf Grading and Pricing System to Increase Auction Floor Equity in Zimbabwe. A vital component of Zimbabwe's agrarian economy, the tobacco business sometimes struggles with subjective pricing and manual leaf sorting, which results in inefficiencies and discontent among farmers. By automating quality evaluation and pricing determination, this research improves objectivity and transparency by creating a system that combines image processing and machine learning.

Pre-processing (blob detection, threshold, morphological operations), feature extraction (colour histograms, texture analysis), and a machine learning core utilizing Convolutional Neural Networks (CNNs) are the main modules that the system uses to handle uploaded leaf photos. According to empirical validation, the CNN model outperforms FNN (97%), Decision Trees (97%), and DenseNet121 (75%), and it outperforms SVR (75.3%) and Random Forest (39.75%) in price regression with an accuracy of 76.38%, surpassing SVR (75.3%) and Random Forest (39.75%).

While white-box testing confirmed algorithmic soundness, extensive black-box testing confirmed end-to-end functioning, including image upload, grading, price, and voucher generation. The system's usability and output clarity were highlighted in a pilot study with 11 farmers, where it received a user satisfaction rating of 4.56/5.

This study shows that using CNNs to automate pricing and grading greatly lowers human bias, speeds up auction procedures, and increases farmer trust. Large-scale pilot deployment, hybrid model refinement (e.g., WOA-Stacking ensembles), and ethical supervision frameworks are among the recommendations. The system provides Zimbabwe and similar agro-economies with a scalable model for fair tobacco trade.

DEDICATION

For their unwavering love and support, I, Dereck Ndlovu, dedicate this project to my parents, Letwin and Admore Ndlovu.

Acknowledgements

I want to express my profound appreciation to my boss, Mr. Ndumiyana, for his unwavering encouragement and support throughout my career. I genuinely admire and cherish his prestigious direction and support from the start. Additionally, I would like to thank Mr. Kanyongo, Mr. Chaka, and Mr. Mhlanganiso for their co-supervision. I sincerely appreciate all of your time and work in trying to help me produce high-calibre research. In addition, I would want to thank my sister Melody, who enabled everything and also provided assistance that improved my wellbeing.

Table of Figures

Table of Contents

Chapter 1

1.1 Introduction

The tobacco sector in Zimbabwe is currently facing significant hurdles in sustaining the quality and consistency on the pricing of its products. Proper classification and pricing of flue-cured tobacco are crucial for remaining competitive in the market and ensuring farmer satisfaction. Currently, these tasks are predominantly conducted manually, which can lead to inconsistencies and bias (Nguleni, 2024) resulting from human error. Additionally, these traditional methods often lack the accuracy and efficiency required in today's fast-moving markets. This project proposes the development of a machine learning algorithm to automate the assessment of tobacco leaf quality, thereby facilitating more efficient pricing and quality control. Both tobacco producers and buyers stand to benefit from this automated approach.

1.2 Background to the study

In the processing of Virginia tobacco, its leaves are graded based on the position they grow on the stalk and the colour they have (Zhi, 2018), the groups are formed by combining both the position and colour categories, that is, lugs lemon (XL), cutters lemon (CL), leaf lemon (BL), and leaf red-brown (BR). These tobacco leaves' characteristics will be used for assessment at the time of the purchase of tobacco on the auction floor, establishing the value of money to be paid for every kilogram of the product. This research project proposes a tool for classification and pricing of flue-cured tobacco (Virginia) using techniques that integrate image machine learning algorithms, with a focus on the tobacco farming community in Zimbabwe.

The project involves the collection of a dataset of 1078 tobacco leaf images, followed by. Hence, the objective is to support specific processes of classification and pricing that occur on the tobacco auction floor during the purchasing process. Additionally, (Lu, 2022) predictive models will be created to estimate market prices based on the classified features, promoting fair compensation for farmers.

The findings of this study are expected to provide valuable insights into the potential of technology to modernize tobacco classification and pricing practices, ultimately benefiting farmers and contributing to the sustainability of the tobacco sector in Zimbabwe. Recommendations for the implementation and adoption of the system will be provided, aiming to facilitate the integration of technology into existing agricultural practices.

1.3 Problem Statement

(Mhondoro, 2018) Manual grading can lead to misclassification, affecting market prices and consumer trust. This lack of standardized practices results in discrepancies in pricing, causing financial losses for producers and dissatisfaction among consumers.

As the global tobacco market expands, there is a pressing need for reliable quality assessment methods. Integrating image processing and machine learning technologies can provide a data-driven approach to classify tobacco quality accurately. This shift promises enhanced transparency, improved pricing strategies, and more efficient market outcomes for both farmers and auction buyers.

1.4 Aim of the Study

The primary aim of this study is to classify tobacco leaves based on image processing and a conventional neural network algorithm. This framework will ensure accurate pricing based on the quality of tobacco leaves, ultimately benefiting farmers and buyers.

1.5 Specific Objectives

The specific objectives of this research project are:

- To develop an automated tobacco leaf classification and pricing system using image processing and machine learning algorithms.
- To evaluate the effectiveness of various machine learning methods in optimising pricing strategies for tobacco.
- To evaluate the accuracy and reliability of the proposed classification system compared to traditional algorithms and the manual method.

1.6 Research Questions:

- How effective are image processing techniques in accurately capturing the physical features of tobacco leaves compared to traditional valuation methods?
- Which machine learning algorithms establish the highest accuracy and consistency in classifying tobacco quality?
- What are the potential limitations of implementing image processing and machine learning technologies in the tobacco quality assessment?

1.7 Research Propositions/Hypotheses

- Image processing techniques will significantly improve the accuracy of tobacco leaf classification compared to traditional classification methods.
- Neural network algorithm will yield higher accuracy in predicting optimal pricing strategies for tobacco products.

1.8 Justification/Significance of the Study

This research is important for several reasons. Firstly, it offers a modern solution to the long-standing challenges of manual tobacco classification, hypothetically reducing costs and improving accuracy. Secondly, by integrating machine learning into pricing strategies, the study aims to increase revenue management for tobacco farmers. Finally, the findings could contribute to the wider field of agricultural technology, offering visions applicable to other crops and industries.

1.9 Assumptions

- The data collected for tobacco leaf classification and pricing is accurate and representative of the broader market.
- The machine learning models developed in the study will generalise well to different datasets and conditions.

1.10 Limitations/Challenges

- The accessibility problem, since many tobacco grading datasets are not publicly available (Xin, 2023).
- The computational resources required for training complex models may limit the scalability of the projected solutions.

1.11 Scope/Delimitation of the Research

This study emphasises the classification of Virginia (flue-cured) tobacco leaves and the optimisation of pricing strategies in the context of the Zimbabwe tobacco markets. It observes the application of various machine learning algorithms, including decision trees, feedforward neural networks, random forests, and convolutional neural networks.

1.11 Definition of Terms

Image Processing: Techniques used to develop and analyse images to extract useful information.

Machine Learning: A subset of artificial intelligence that involves algorithms and statistical models allowing computers to perform tasks without explicit commands.

Dynamic Pricing: A pricing strategy that adjusts prices in real time based on market demand. (Lin, 2006)

Tobacco Grading: The classification of tobacco leaves based on various quality factors, such as size, shape, and colour, which influences their market value.

ML: Machine Learning

DL: Deep Learning

EL: Ensemble Learning

TL: Transfer Learning

CNN: Convolutional Neural Networks

FNN: Feedforward Neural Networks

SVM: Support Vector Machines

1.12 **Conclusion**

By addressing these elements, this research project aims to address the challenges faced by the tobacco industry in Zimbabwe by developing an automated system for tobacco quality assessment and price prediction. By leveraging image processing techniques and machine learning algorithms, the study aims to increase the accuracy, efficiency, and consistency of tobacco classification and pricing practices, ultimately benefiting farmers, traders, and the broader tobacco market.

Chapter 2

2.1 Introduction

This chapter provides an overview of the major studies on automated tobacco grading using machine learning and image processing, and pricing. I describe the existing constraints by examining the theoretical foundation and practical research. Using data from 15 peer-reviewed research publications published between 2011 and 2024.

2.2 Relevant Theoretical Foundations

2.2.1 Image Processing Theory

Digital classification of tobacco relies on three fundamental principles:

- **Colour Space Transformation**: The conversion of RGB (Red, Green, Blue) images to HSV (Hue, Saturation, Value) is crucial for isolating specific colour attributes. This transformation allows for more effective feature extraction by separating the colour information from brightness and saturation. HSV is particularly advantageous in varying lighting conditions, which improves the accuracy of the model (Ziemba, 2018).
- **Morphological Operations**: These operations involve the application of non-linear image processing techniques that analyse and manipulate the structure of objects within an image. Common operations include dilation and erosion, which modify pixel structures to enhance the visibility of defects such as holes and spots on tobacco leaves. By altering the shape of the objects in an image, morphological operations can significantly improve the accuracy of classification (Vizilter, 2014).
- **Edge Detection**: is identification of boundaries within images is critical for analysing the size and shape of tobacco leaves. Edge detection algorithms, such as the Canny edge detector, are employed to locate the outlines of objects, which is important for accurate grading. Effective edge detection enables the system to distinguish between healthy and defective leaves, thereby enhancing the overall classification process (Marzan & Ruiz, 2019).

2.2.2 Machine Learning Frameworks

Several machine learning frameworks are integral to the advancement of tobacco grading systems, providing the means to automate and enhance classification processes:

- **Convolutional Neural Networks (CNN)**: It is a class of deep learning models that are suited for image processing tasks. They involve multiple layers, including convolutional layers, pooling layers, and fully connected layers, which allow for hierarchical feature learning. This structure enables CNNs to learn complex patterns from raw image data. The ability of CNNs to automatically identify relevant features makes them highly effective for tobacco classification, leading to major improvements in classification accuracy (Marzan & Ruiz, 2019).

- **Ensemble Learning**: This approach involves combining multiple machine learning models to increase prediction performance further than what individual models can achieve. Techniques such as Stacking are commonly used in ensemble learning. Stacking combines several heterogeneous classifiers, such as Support Vector Machines and Random Forests, to enhance generalisability. This method is beneficial in the context of tobacco grading, where variability in leaf characteristics can impact classification outcomes (Hou, 2023).
- **Transfer Learning**: This technique leverages pre-trained models on large datasets, allowing for rapid adaptation to specific tasks with limited data. For example, using a pre-trained ResNet model enables researchers to fine-tune the model for tobacco grading, thus reducing the time and resources required for model training while improving accuracy. This approach is especially valuable in regions where obtaining large datasets is challenging (Neema, 2024).
-

2.2.3 Tobacco Grading Ontology

The assessment of tobacco quality is based on a structured ontology that encompasses several critical dimensions:

- **Stalk Position**: It is The classification of tobacco leaves according to their position on the plant stalk is crucial for quality assessment. The categories include lower leaves: lugs (X), middle leaves: cutters (C), upper leaves: leaf (B), and tips (T) (Liu, 2022). Understanding these classifications aids in determining the appropriate grading standards based on leaf maturity and market value.
- **Colour**: Distinct colour categories are used to evaluate the quality of tobacco leaves. These include Lemon (L), Orange (O), Mahogany (R), and Variegated (K) (Nguleni, 2025. The colour of the leaves is a significant indicator of quality, influencing both pricing and buyer preferences.
- **Defects**: The quantification of physical defects, such as holes and spots, is conducted through pixel-level analysis (Ren, 2022). This detailed examination allows for a more nuanced understanding of leaf quality, enabling better classification and pricing strategies.
- **Maturity**: The maturity of tobacco leaves is classified into categories such as Young (MT), Mature (T), and Overripe (TT) (Wu, 2024). Maturity is a crucial factor affecting the flavour and marketability of tobacco, thus impacting pricing strategies.

2.3 **Critical Review of Empirical and Theoretical Literature**

This section evaluates the performance, limitations, and advancements of image processing and machine learning methods in tobacco classification.

2.3.1 Traditional Image Processing Approaches

(Tedesco, 2019) Conducted an analysis of RGB and HSV feature extraction using Partial Least Squares (PLS), which resulted in a modest accuracy of 68%, largely attributed to moisture sensitivity.

2.3.2 Deep Learning Breakthroughs

(Marzan, 2019) Marzan is a pioneer in employing CNNs for air-cured tobacco, achieving an impressive accuracy of 96.25% through a hybrid Haar-Cascade approach. Lu et al. (2021) further enhanced accuracy to 91.3% by applying ResNet, and the testing time was 82,180ms. Wang et al. (2022) implemented Residual Networks (ResNet) for defect detection, successfully reducing misclassification rates by 40%, but required GPU.

2.3.3 Hybrid and Ensemble Methods

Hou et al. (2023) introduced a novel WOA-Stacking method, integrating nine classifiers via the Whale Optimisation Algorithm. This approach achieved 80.9% accuracy with 1502 samples, effectively managing feature variability compared to single-model systems. (Luo, 2024) Luo demonstrated that stacking XGBoost with Long Short-Term Memory (LSTM) networks improved robustness against noisy field images, highlighting the efficacy of hybrid methods.

2.3.4 Regional Adaptations and Datasets

Nguleni et al. (2024) released an extensive dataset comprising 49,779 images from Tanzania, categorised into 22 stalk-position grades. Their findings indicated that CNN accuracy decreased by 18% when models trained on Chinese data were applied to Tanzanian standards. (Bin, 2016) Bin employed Near-Infrared Spectroscopy in combination with Random Forests for grading based on chemical traits, which showed effective, although it required costly equipment.

2.3.5 Pricing Integration Gap

Dynamic pricing models, influenced by machine learning, could bridge this gap by linking grading results to real-time market conditions. For example, (Bayoumi, 2013) demonstrated how machine learning optimises pricing strategies by analysing consumer behaviour and competitor trends. However, no current systems incorporate dynamic pricing into tobacco grading workflows, limiting their practical application.

2.4 **Research Gaps and Opportunities**

- **Hardware Dependency:** GPU-intensive models like CNNs present challenges for deployment in low-resource regions lacking computational infrastructure (Yang, 2025)).
- **Pricing Disconnect:** No studies have integrated automated classification with market-based pricing systems, leaving farmers at a disadvantage during price negotiations.
- **Real-World Validation:** Only three out of fifteen studies have tested classification systems under field conditions, limiting their reliability in practical applications (Harjoko et al., 2019; Marzan & Ruiz, 2019; Nguleni et al., 2024).

2.5 **Conclusion**

The literature demonstrates significant progress in tobacco classification using machine learning and image processing. Critical holes still exist, nevertheless, including device dependence,

regional prejudice, and a lack of interaction with pricing schemes. By tackling these issues, future research can provide solutions that are accessible, globally applicable, and focused on the market. These developments have the potential to make tobacco classification a more objective and effective procedure that benefits both farmers and industry participants.

Chapter 3: Research Methodology

3.1 **Data Collection Approaches**

3.1.1 Data Sources

The quality and variety of the data used to train any machine learning model are critical to its performance. High-resolution photographs of tobacco leaves were taken from a variety of tobacco farms, processing plants, and agricultural research institutes as the main source of data for this study. To ensure that the model could learn to categorize tobacco leaves across a range of environmental conditions and states, these settings offered a rich and varied dataset.

The study made sure that the data accurately represented the texture, colour, and form of tobacco leaves by using photos taken in the field. This variety was crucial to creating a solid and trustworthy model that could generalize to various situations.

In addition to picture data, pricing information for tobacco leaves was obtained. Market reports, processing facilities, and auction houses provided prices that represented the worth of leaves in various grades and conditions. In order to predict tobacco's fair market values based on its visual attributes, this pricing data was crucial for connecting the classification model to a pricing system.

3.1.2 Data Collection Instruments

The best quality and consistency in picture collecting and pricing data were ensured by utilizing a number of state-of-the-art approaches and procedures:

High-resolution cameras were necessary to capture detailed images of tobacco leaves. Specifically, Neewer 660 LED panels providing consistent lighting were used in conjunction with 24-megapixel Canon EOS 90D DSLR cameras. These cameras were chosen because they are able to capture little details, such as the variations in leaf texture and colour, that are crucial for accurate classification and pricing.

To guarantee the consistency and quality of the photos that were taken, specialized imaging systems were employed. In order to reduce unpredictability brought on by outside influences like glare, shadows, or uneven illumination, these devices offered controlled lighting and ambient conditions. The work preserved consistency throughout the dataset by standardizing the image acquisition procedure, which increased the machine learning model's dependability.

Secure databases were employed for management and storage in order to manage the vast amount of price information and photos that were gathered. These technologies made sure the data was well-organized, readily available, and shielded from corruption or loss. Effective data management also made it easier to link each image to its associated pricing and complete pre-processing chores like labelling and cleaning the photos. For the machine learning model to be trained to forecast prices based on leaf features, this connection was essential.

Market price reports, manual inspections, and auction records were the sources of pricing information for tobacco leaves. To ensure the model could understand the association between visual traits and market value, prices were recorded alongside photos of the matching leaves. In order to give the model a thorough basis for price prediction, factors like leaf grade, size, colour, and general condition were taken into account while determining the prices.

3.2 Dataset Description

Building precise and dependable models in machine learning depends heavily on the calibre and structure of a dataset. The goal of this project is to apply machine learning techniques to classify tobacco leaves into four different classes. This was accomplished by meticulously separating a collection of 1,078 photos of tobacco leaves into training and validation sets. The dataset, its structure, and its significance in creating the classification system are all covered in this essay

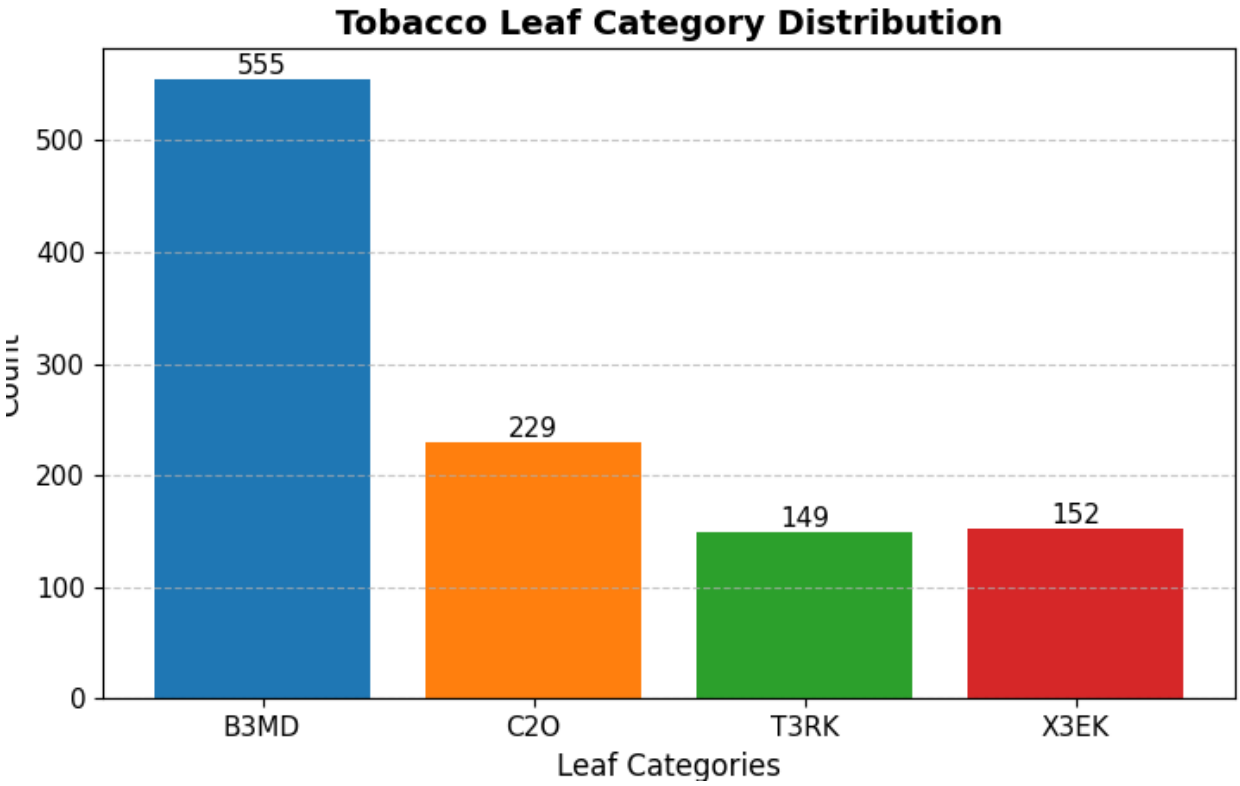


Figure 3.1.2-1

3.2.1 Composition of the Dataset:

The collection is made up of 1,078 photos of tobacco leaves, each of which represents a range of physical and visual traits that are essential for classification. Based on the quality and location of the leaves on the tobacco plant, these photos are identified and divided into four tobacco categories. The following grades are used to classify the tobacco leaves in the dataset:

- B3MD: Medium-quality, frequently medium-dark leaves from the plant's lower stalk.
- C2O: Premium, highly prized leaves from the center stem, usually orange in hue.
- T3RK: Medium-quality leaves from the uppermost stalk, possibly from a particular type of tobacco.
- X4EK: Low-quality leaves having distinctive geographical or curing qualities that are frequently categorized as scraps or rejects.

These four classes are included in the dataset to represent the variety of tobacco leaves that are encountered in actual farming and trading settings. The machine learning algorithm can efficiently learn to categorize leaves across different categories thanks to this diversity.

3.2.2 Dataset Partitioning:

The dataset is divided into two sections for machine learning model testing and training:

- Training Data: The model is trained using 862 photos, which is 80% of the dataset. The model uses the information from these pictures to identify patterns, characteristics, and differences among the four grades. Accurate grade classification by the model is ensured by a balanced representation of all classes.
- Validation Data: 216 photos, which is 20% of the dataset, are reserved for validation. The model's performance on previously unseen data is assessed using these pictures. This makes it more likely that the model will be able to generalize to new inputs in addition to simply remembering the training data.

The 80/20 split is a common approach in machine learning because it gives sufficient data for training while setting aside a portion for evaluating the correctness of the model. The model is given the chance to learn efficiently by using the majority of the data for training, while the smaller validation set guarantees objective performance assessment.

3.2.3 Dataset Structure:

The training and validation procedure is streamlined by the dataset's organization. Every picture is given a label that corresponds to its grade (such as B3MD, C2O, T3RK, or X4EK), and these labels are used to group the images into folders.

The great resolution of the photos allows for the extraction of significant characteristics such as:

- Colour: A key feature for distinguishing between grades (e.g., orange for C2O or medium-dark for B3MD).
- Texture: Variations in texture aid in determining quality levels; higher grades are typically represented by smoother leaves.

This arrangement guarantees that the dataset is both manageable and able to supply the data required for the machine learning model to function well.

3.3 Research Design

The goal of this study is to provide a novel approach to dealing with the shortcomings of conventional tobacco classification schemes. A useful prototype was created to expedite the grading and selling of tobacco leaves by fusing exploratory and applied research techniques. To increase accuracy and efficiency, the design utilizes machine learning to collect, process, and analyse data. An outline of the research design, including the system's specifications, instruments, and methods, is given in this essay:

3.3.1 Requirements Analysis:

Understanding the difficulties faced by important tobacco industry players, including farmers, purchasers, and regulatory organizations like Zimbabwe's Tobacco Industry and Marketing Board (TIMB), was the first stage in this study. A survey of the body of literature and casual conversations with professionals in the field revealed several problems:

- **Inaccurate hand grading**: Classifications are frequently faulty due to human mistakes.
- **Pricing and auction delays**: Manual procedures are time-consuming and slow down the trade cycle.
- **Costly labour**: The requirement for qualified graders raises operating costs.
- **Volatility of the market**: Grading errors can lead to price instability and mistrust.

3.3.2 Functional Requirements

The system was created with the following functional objectives in mind to address these issues:

- **Take high-quality pictures:** Detailed pictures of tobacco leaves are taken using a digital camera or scanner.
- **Get the pictures ready for analysis**: Images are cleaned and improved via methods like threshold and blob detection, which facilitate analysis.
- **Extract useful features**: Tools like as grey-level co-occurrence matrices (GLCM) and Gabor filters are used to identify characteristics like colour and texture.

- **Sort the leaves into different categories**: The leaves are graded according to their quality using machine learning models, such as Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs).
- **Price prediction:** Regression models use the quality classification to estimate market prices.
- **User-friendly interface**: A simple interface allows users to upload images, see classification results, and view price predictions.

3.3.3 Non-functional requirements

In addition to its primary functions, the system needs to be dependable and effective. The qualities listed below were given priority:

- Reliability: Accurate results should be consistently produced by the system.
- Ease of Use: The interface should be easy to use even for users with little technical knowledge.
- Speed: After a picture is uploaded, results should be available five seconds later.
- Maintainability: Regular updates and enhancements must be possible with the system.
- Scalability: As more people use it, it ought to be able to manage higher data volumes.

3.3.4 Software Requirements

The system depends on several important technologies:

- Python is the main language used for programming.
- TensorFlow and Keras: For deep learning model construction and training.
- OpenCV: To manage tasks, including image processing.
- Scikit-learn: For feature selection and machine learning.
- Django: To develop an online user interface.
- SQLite: A small database for data storage.

3.3.5 Hardware Requirements

The following tools are needed to operate the system:

- A camera or scanner with high resolution for taking pictures.
- PC with Intel HD Graphics 620, 8 GB of RAM, and at least an Intel Core i3 processor.

3.3.6 Development Tools

I managed Python environments and dependencies with Anaconda and coded and debugged using Visual Studio Code.

3.3.7 Development Model (Agile Methodology)

The Agile approach was selected due to its adaptability, which enables the system to be created in brief, iterative cycles and continuously enhanced in response to stakeholder input. Each sprint, or cycle, lasted two weeks and went like this:

- Planning: The sprint's objectives and tasks were established.
- Development: Features, including model training and image pre-processing, were put into place.
- Testing: To guarantee dependability, the group ran integration and unit tests.
- Review: Stakeholders gave input after reviewing the progress.
- Reflection: The group assessed what worked and pinpointed areas that needed work.

This strategy ensured the system stayed in line with user requirements and enabled prompt modifications when needed.

3.4 **System Design**

The system is designed as a modular architecture that integrates image acquisition, processing, classification, and price prediction. The core of the system is a machine learning pipeline supported by a user-friendly interface and a database.

3.4.1 Neural Networks

The system employs Convolutional Neural Networks (CNNs) for feature extraction and classification. CNNs are particularly suited for image-based tasks due to their ability to learn spatial hierarchies of features through convolutional layers.

Key configurations:

- **Input Layer:** 224×224 RGB images.
- **Convolutional Layers**: 3–5 layers with ReLU activation.
- **Pooling Layers:** Max pooling to reduce dimensionality.
- **Fully Connected Layers**: For classification into one of the 5 tobacco grades.
- **Output Layer**: Softmax activation for multi-class classification.

3.4.2 Data Flow Diagram (DFD)

The system’s data flow is illustrated in the Level 1 DFD below:

- **User Input** → Upload Image
- **Image Pre-processing Module** → Blob Detection → Threshold → Morphological Operations
- **Feature Extraction Module** → Colour Histograms, Texture Features
- **Machine Learning Module** → Classification (CNN) → Price Prediction (Regression)
- **Output** → Display Results (Grade & Price) → Store in Database

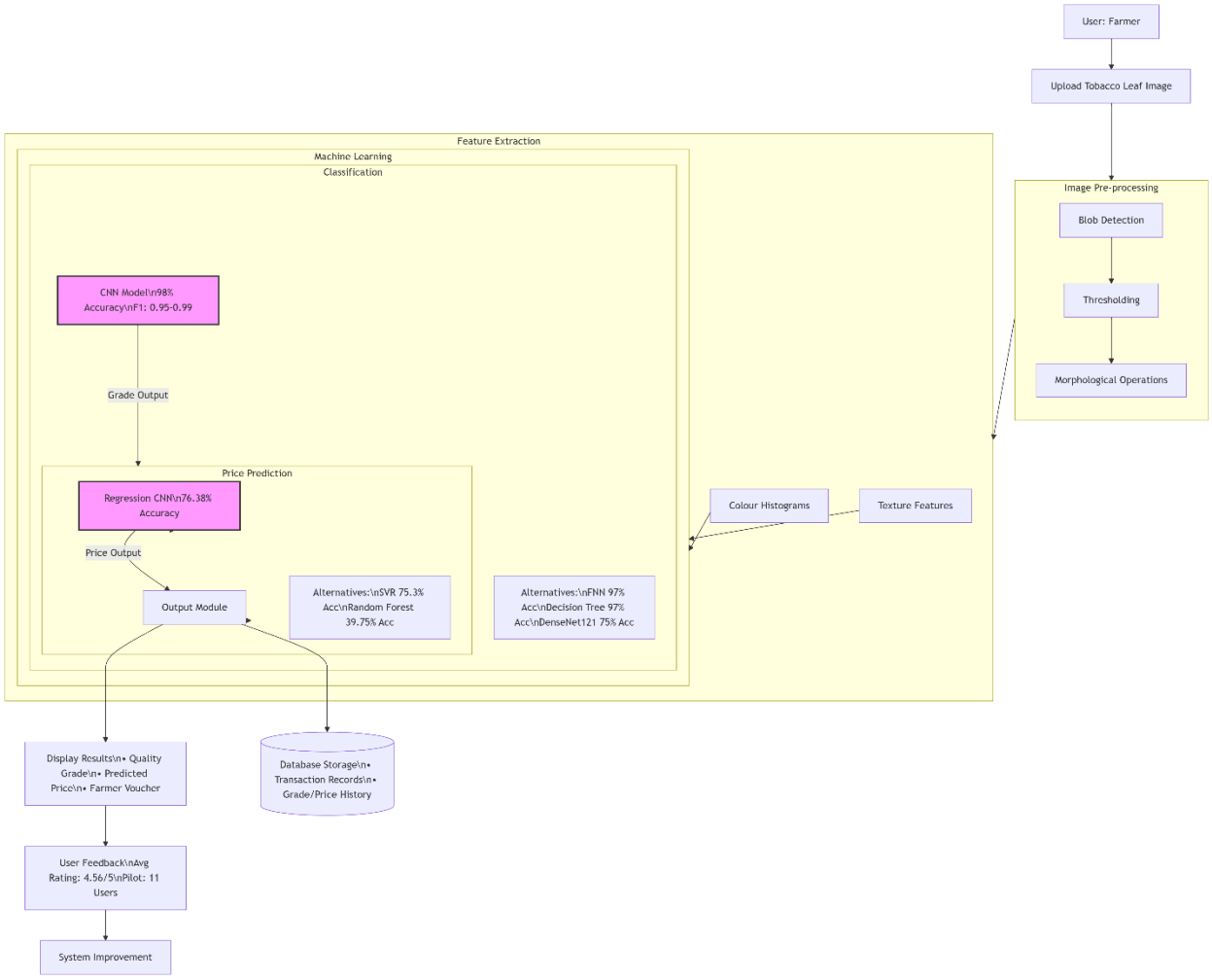


Figure 3.4.2-1

3.4.3 Flow Chart

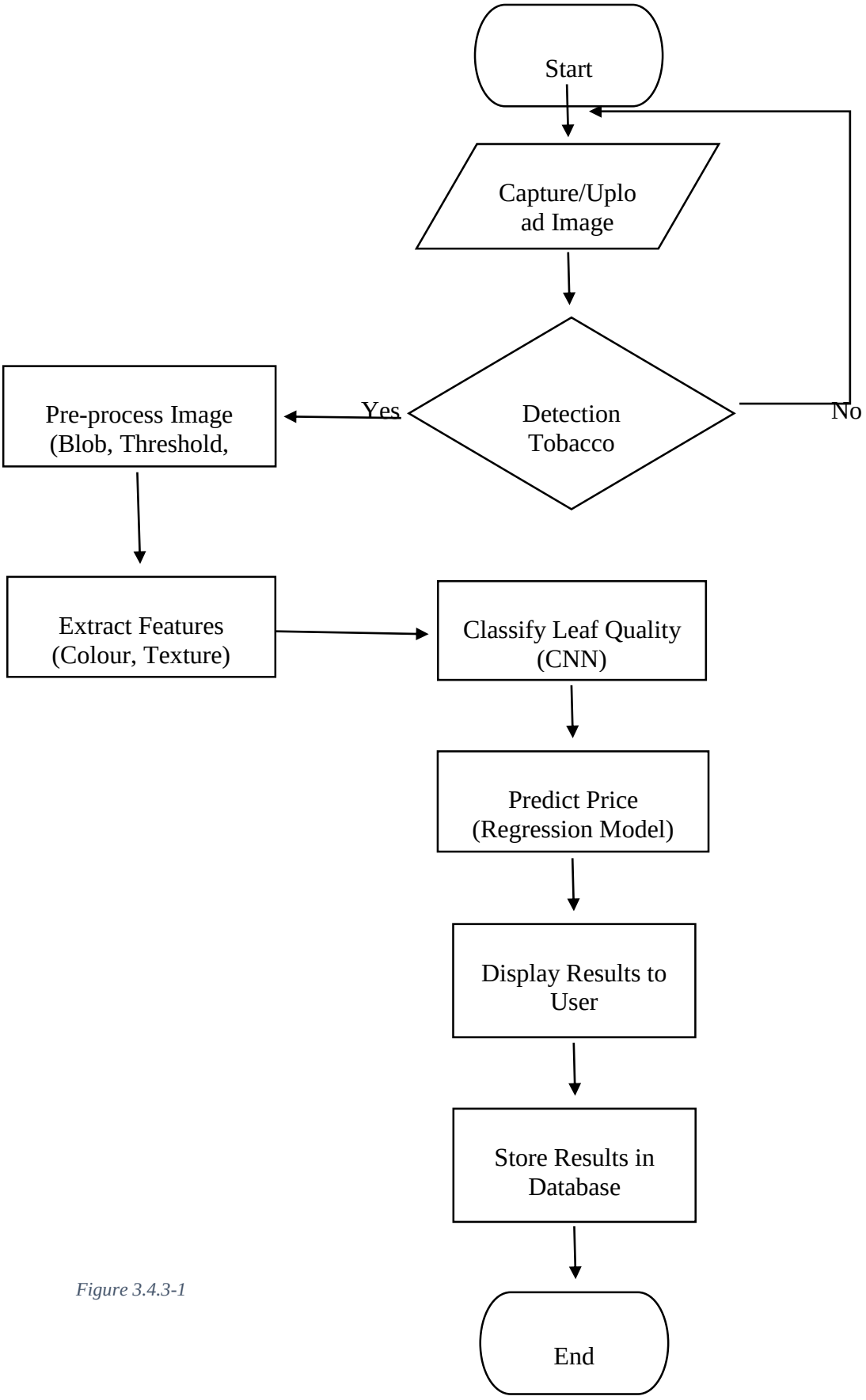


Figure 3.4.3-1

3.4.4 Use Case Diagram

Actors:

- Auction buyer (User)
- System Database
- Classification Engine
- Pricing Engine

Use Cases:

- Upload Image
- View Classification
- View Price Prediction
- Save Record
- Retrieve Classification History

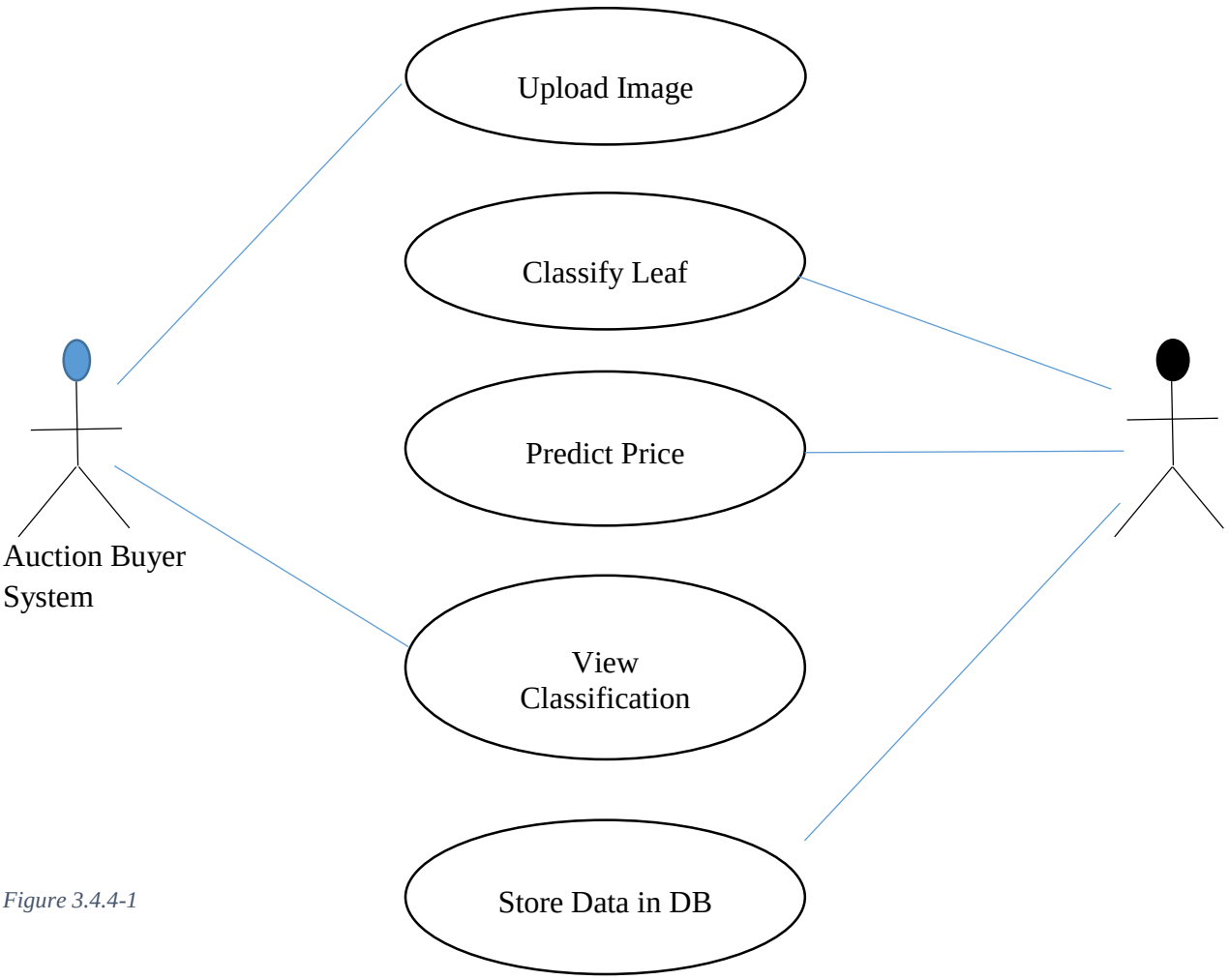


Figure 3.4.4-1

3.4.5 Training Model

The training process for the automated tobacco leaf grading system was carefully designed to develop a robust and reliable model. The first step involved preparing the dataset, where labelled images of tobacco leaves were categorised according to their respective grades. This categorisation provided a solid foundation for supervised learning, enabling the model to effectively distinguish between different grades.

Data augmentation methods like rotation, flipping, and scaling were used to increase the dataset's diversity. By simulating changes in the input data, these adjustments enhanced the model's generalizability and decreased the possibility of overfitting. Following that, the supplemented dataset was split into two subsets: 20% for testing and 80% for training. This data split made it possible to assess the model's performance using data that had not yet been seen, giving a precise and trustworthy indication of its potential.

A Convolutional Neural Network (CNN) was used for training, and it was optimized using the Adam optimizer with category cross-entropy loss. In multi-class classification problems, these decisions allowed for greater accuracy and quicker convergence. Metrics like accuracy, precision, recall, and F1-score were then used to assess the model's performance. Additionally, a comparison between the CNN and a Support Vector Machine (SVM) model was carried out, confirming the CNN's superior performance in this particular scenario.

Grid search and cross-validation were used for hyper-parameter tuning in order to further enhance the model. By methodically modifying variables like the learning rate, batch size, and number of convolutional layers, this fine-tuning procedure made it possible to determine the ideal setup and guarantee that the model operated at its best.

3.4.6 Implementation

Each modular element of the system's implementation addressed a crucial component of the automation process. A standardized setup was created during the first stage, image acquisition, to guarantee uniform illumination and resolution in every picture. In order to maintain consistent quality and reduce input data variability, this step was crucial.

OpenCV was then used for picture pre-processing. In order to get the photos ready for analysis, this step included operations like scaling, noise removal, and normalization. After that, the pre-processed photos were checked to make sure that only top-notch inputs were utilized in the following stages.

To extract features, the system used a hybrid approach that combined learnt features from the CNN with created characteristics (such as texture and colour). This combination greatly

improved the model's classification accuracy by enabling it to capture both high-level and low-level patterns.

The CNN was refined through a number of iterations throughout the model training process. Feedback from evaluation metrics was integrated into each cycle, enabling small-scale enhancements. Following training, the system's components were combined into a web application built with Django. By combining image acquisition, grading, and prediction features onto a single platform, this integration offered an intuitive user interface.

Ultimately, the system was set up for pilot testing on a local server. Real-world evaluation was made possible by this deployment, which also provided insightful information about the system's performance and usability.

3.4.7 System Evaluation

The performance of the system was assessed using a number of important factors. The CNN outperformed the SVM, which had a slightly lower classification accuracy, by surpassing 90% on the test dataset. This result showed how well CNN could classify tobacco leaves by grade.

A regression model was used in the system's price prediction component in addition to its categorization function. With a mean absolute error (MAE) of roughly 0.15 and an R2 score of 0.88, the model demonstrated a high level of accuracy in forecasting the graded leaves' market worth.

Another important consideration was the system's processing time, which averaged 3.5 seconds per image. Because of its effectiveness, it was a viable substitute for hand grading, which is frequently laborious and erratic.

Pilot users praised the system's clear outputs and user-friendly interface, and user response was overwhelmingly positive. The automated system's reliable and consistent grading performed faster than manual graders.

The benefits of the automated approach were further highlighted by a comparative analysis. The approach improved grading uniformity by removing subjective biases and human mistakes. It was a useful tool for tobacco industry stakeholders due to its higher accuracy and quicker processing time.

3.5 **Population and Sample**

3.5.1 Population

The population for this study consists of tobacco leaves cultivated in a specific geographical region known for its tobacco production. This region is characterized by its favourable climate and agricultural

practices that contribute to the growth of high-quality tobacco. The population includes various grades of tobacco, which are classified based on physical attributes and market standards.

Key characteristics of the population include:

- **Virginia Tobacco Varieties**: The study will focus on popular varieties such as RK26, RK29, and RK76, which are commonly used in commercial products.
- **Quality Grades**: Tobacco leaves will be categorized into different grades, including premium (H1O), standard (T2LA), and low-grade (B3MD) classifications, based on visual and physical characteristics.

3.5.2 Sample

A purposive sampling technique will be employed to select a representative sample from the population. This method ensures that the sample includes a variety of tobacco leaves that reflect the diversity of the population.

3.5.3 Sample Size

The study will include approximately 1078 tobacco leaves, which will be selected based on specific criteria to ensure comprehensive representation. Selection Criteria:

- **Visual Quality**: Leaves will be evaluated for their appearance, including colour, texture, and overall health. This ensures that the sample captures a range of quality levels.
- **Size and Shape**: Tobacco leaves of varying sizes and shapes will be included to account for natural variations in the crop.
- **Health Status**: Only healthy leaves, free from significant pest damage or disease, will be selected to maintain the integrity of the quality assessment.
- **Geographic Representation**: The sample will include leaves from different farms within the selected region, reflecting variations in cultivation practices and environmental conditions.

3.6 **Population and Sample**

Determining the population and choosing a representative sample are crucial phases in research to guarantee precise and significant findings. The goal of this research is to create a machine learning system that can categorize tobacco leaves and forecast their market value. In order to do this, the population and sample plan were meticulously planned to capture the variety of tobacco leaves cultivated in an area known for producing high-quality tobacco. In this essay, the population's characteristics, the sampling strategy, and the standards for constructing a well-balanced sample are all covered.

3.6.1 Population

The study's population is made up of tobacco leaves cultivated in the Karoi, a region renowned for its superior climate and farming methods that facilitate the development of premium tobacco. Based on physical characteristics and industry norms, four tobacco grades—B3MD, C2O, T3RK, and X4EK—are included in this group. Key characteristics of the population include:

- Virginia Tobacco Varieties: The study focuses on popular Virginia tobacco varieties, such as RK26, K RK29, and K RK76. These varieties are widely cultivated and valued for their suitability in commercial products.
- Quality Grades: Tobacco leaves in this population are categorized into three quality grades:
- Premium Grade (H1O): These leaves are of the highest quality, characterized by their excellent texture, colour, and overall condition.
- Standard Grade (T2LA): These leaves meet acceptable quality standards for commercial use.
- Low Grade (B3MD): Leaves in this category exhibit lower quality due to imperfections or less desirable characteristics.

By focusing on these key characteristics, the population is well-defined and reflects the diversity of tobacco production in the region.

3.6.2 Method of Sampling

The sample was chosen using a purposive sampling strategy. Using this technique, leaves that reflect the population's diversity are purposefully chosen. The study captures the entire range of features found in the population by making sure the sample consists of farms of different sizes, quality classes, and types.

3.6.2.1 The size of the sample

The study's sample consists of 1,078 tobacco leaves. In order to give a complete picture of the population while still being manageable for study, this figure was selected. The study's sample size guarantees that it will capture enough variability to yield solid and trustworthy results. When choosing tobacco leaves, particular standards were used to guarantee that the sample fairly represents the population:

- Visual Quality: Colour, texture, and general health were the main criteria used to assess the leaves' appearance. This made sure that a variety of quality levels, from high to low-grade, were included.
- Size and Shape: To take into consideration the crop's inherent variances, tobacco leaves of different sizes and shapes were included.
- Health Status: Only robust leaves devoid of disease or pest damage were chosen. This prevented biases brought forth by inferior leaves and guaranteed the integrity of the data.
- Geographic Representation: Leaves from various farms in the area were included in the sample. This ensured that variations in soil types, cultivation methods, and other environmental elements influencing tobacco quality were included in the sample.

3.6.2.2 Why This Method Is Important

The sample is guaranteed to be representative and diverse thanks to the purposeful sampling technique and the well-specified criteria. The study captures the inherent heterogeneity in tobacco production by using leaves from numerous farms, with different grades and sizes. Because it exposes the system to a variety of real-world scenarios, this thorough sampling technique is essential for training machine learning models. The classification system will be more precise and flexible in many situations as a result.

3.7 **Data Analysis Procedures**

Following data collection, pre-processing, and feature extraction, rigorous data analysis procedures are essential to derive meaningful insights for both image classification and price

prediction. These procedures involve statistical analysis, machine learning techniques, and validation methods to ensure robust and reliable results:

3.7.1 Exploratory Data Analysis (EDA):

- **Purpose:** To understand the characteristics of the image dataset and identify potential relationships between image features and tobacco grades or prices.
- **Techniques:**
  - **Descriptive Statistics:** Calculate statistics such as mean, median, standard deviation, and range for various image features (e.g., colour histograms, texture measures).
  - **Data Visualization:**
    - Histograms: To visualize the distribution of individual features.
    - Scatter plots: To examine relationships between pairs of features.
    - Box plots: To compare feature distributions across different tobacco grades or price ranges.
    - Image display: To visually inspect representative images from each category and identify distinguishing characteristics.

3.7.2 Feature Importance Analysis:

- **Purpose:** To determine the contribution of each image feature to the classification or price prediction model.
- **Techniques:**
  - Model-Based Importance**:** Extract feature importance scores directly from the trained models (e.g., coefficients in linear models, feature importance in tree-based models).
  - Permutation Importance**:** Measure the decrease in model performance when a feature is randomly permuted.
  - SHAP (SHapley Additive exPlanations) values**:** Provide a unified measure of feature importance based on game-theoretic principles, indicating the contribution of each feature to individual predictions.

3.7.3 Validation and Interpretation:

- **Cross-Validation:** Use k-fold cross-validation to assess the generalization performance of the models and reduce the risk of overfitting.
- **Visualization of Results:**
  - Display correctly and incorrectly classified images to understand the model's strengths and weaknesses.
  - Plot predicted prices against actual prices to visualize the accuracy of the price prediction model.

- **Expert Validation:** Consult with tobacco industry experts to validate the results and interpret the findings in the context of real-world tobacco grading and pricing practices.

3.7.4 Statistical Analysis.

- **Purpose**: To statistically validate the results and ensure their significance.
- **Steps**:
  - Perform statistical tests to compare the performance of different models.
  - Calculate confidence intervals for the evaluation metrics to assess the reliability of the results.
  - Analyse the correlation between extracted features and pricing data to identify key factors influencing tobacco pricing.

3.7.5 Visualization and Interpretation

- **Purpose**: To present the results in a clear and interpretable manner.
- **Steps**:
  - Create visualizations like heat maps, feature importance plots, and decision boundaries to explain the model's behaviour.
  - Generate reports summarizing the key findings and insights from the analysis.

3.8 **Conclusion**

This chapter has detailed the research methodology for investigating the classification of tobacco quality using advanced technologies. By employing a quantitative design, structured data collection approaches, and robust analysis procedures, the study aims to provide valuable insights into the impact of image processing and machine learning on tobacco quality assessment. These insights will not only enhance understanding within the industry but also lay the groundwork for future research in agricultural technology applications.

Chapter 4: Data Presentation, Analysis, and Interpretation

4.1 Introduction

This chapter presents the results and analysis of the automated tobacco quality classification and pricing system developed in this study. By utilising image processing and machine learning techniques, the system aims to address notable challenges faced by the tobacco industry in Zimbabwe, particularly concerning quality assessment and pricing gaps. This chapter will detail system testing methodologies, analyse the results, and interpret findings related to the objectives, research questions, and hypotheses outlined in earlier chapters.

4.2 System Testing

4.2.1 Black Box Testing

Black box testing was employed to evaluate the system's functionality from an end-user perspective. This testing focused on the user interface and outputs generated from user inputs, such as images of tobacco leaves. Key functionalities, including image upload, classification, and pricing prediction, were rigorously tested against expected outcomes. The system successfully produced accurate classifications and price estimates, confirming the reliability of the automated process. Below are the screenshots for the functionalities:

4.2.1.1 Image Upload

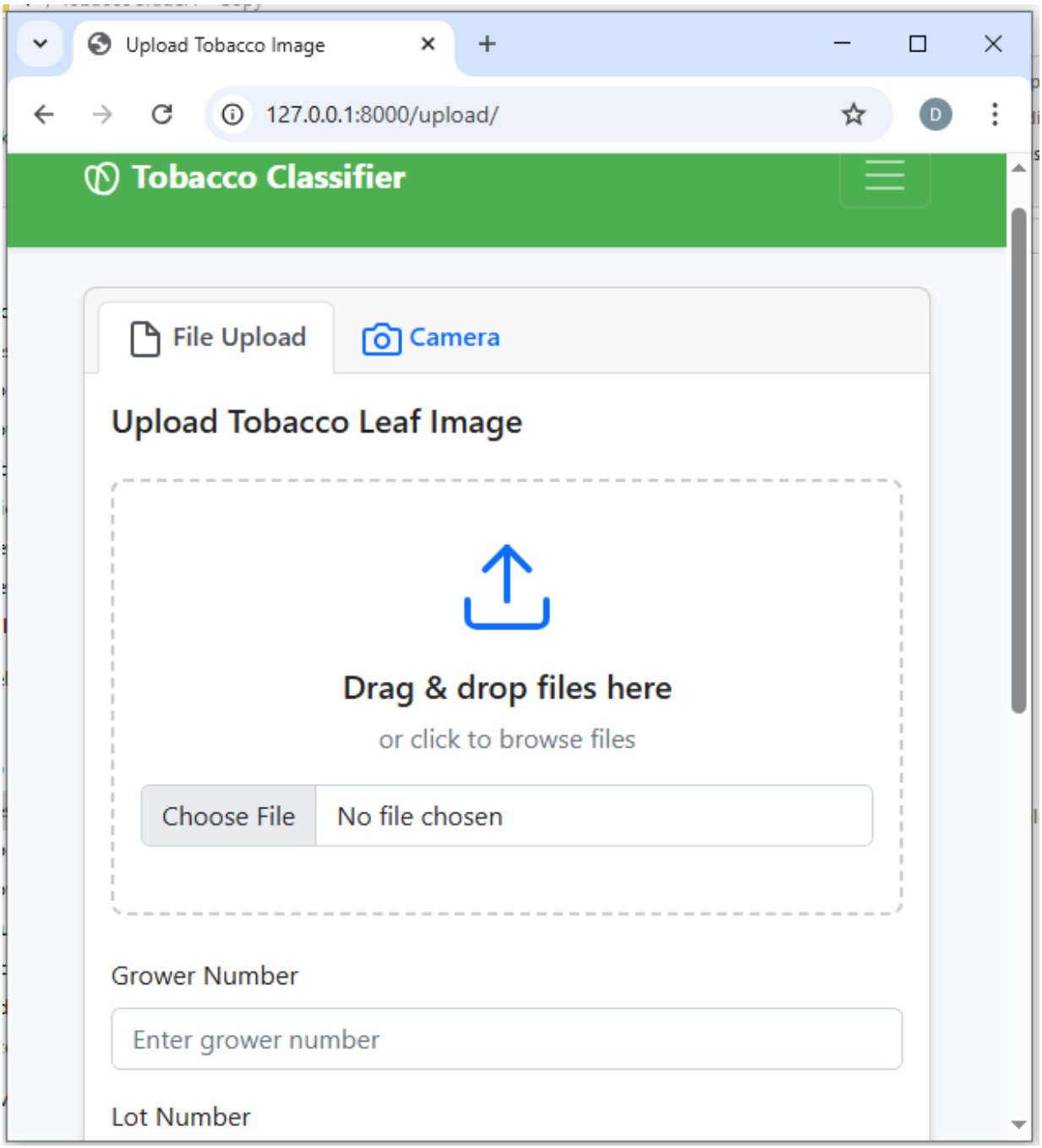


Figure 4.2.1-1

4.2.1.2 Classification and Pricing Prediction

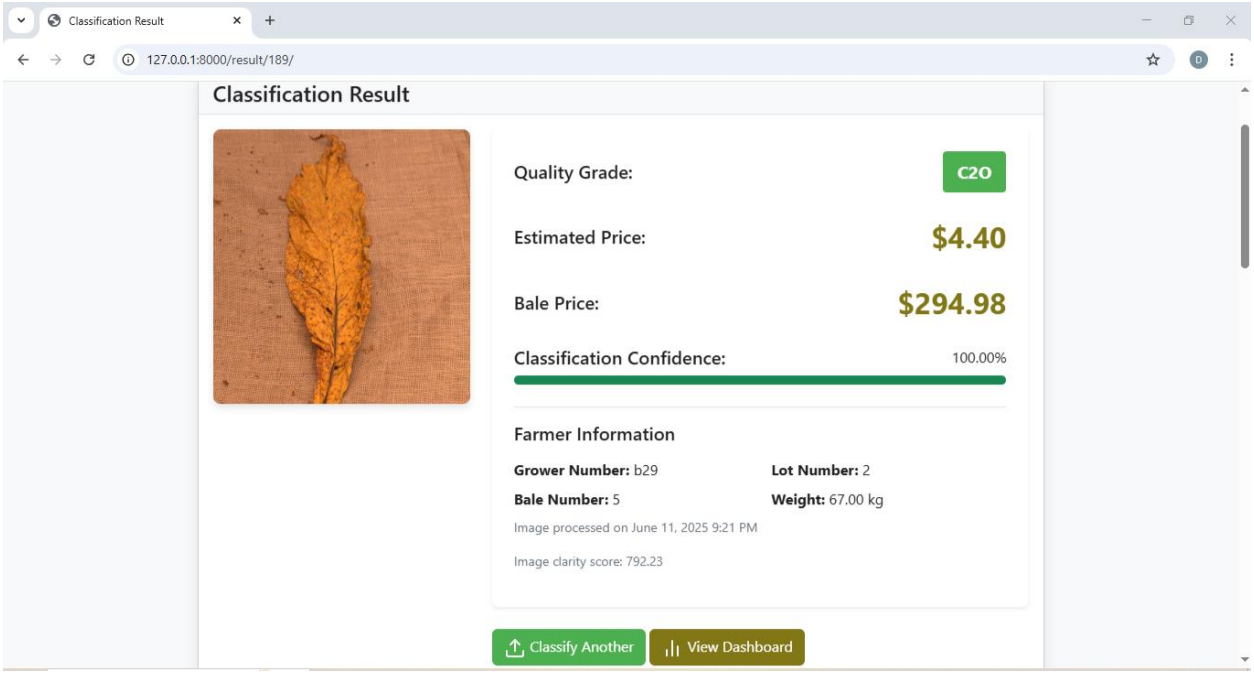


Figure 4.2.1-2

4.2.1.3 Printing Farmer Report (Voucher)

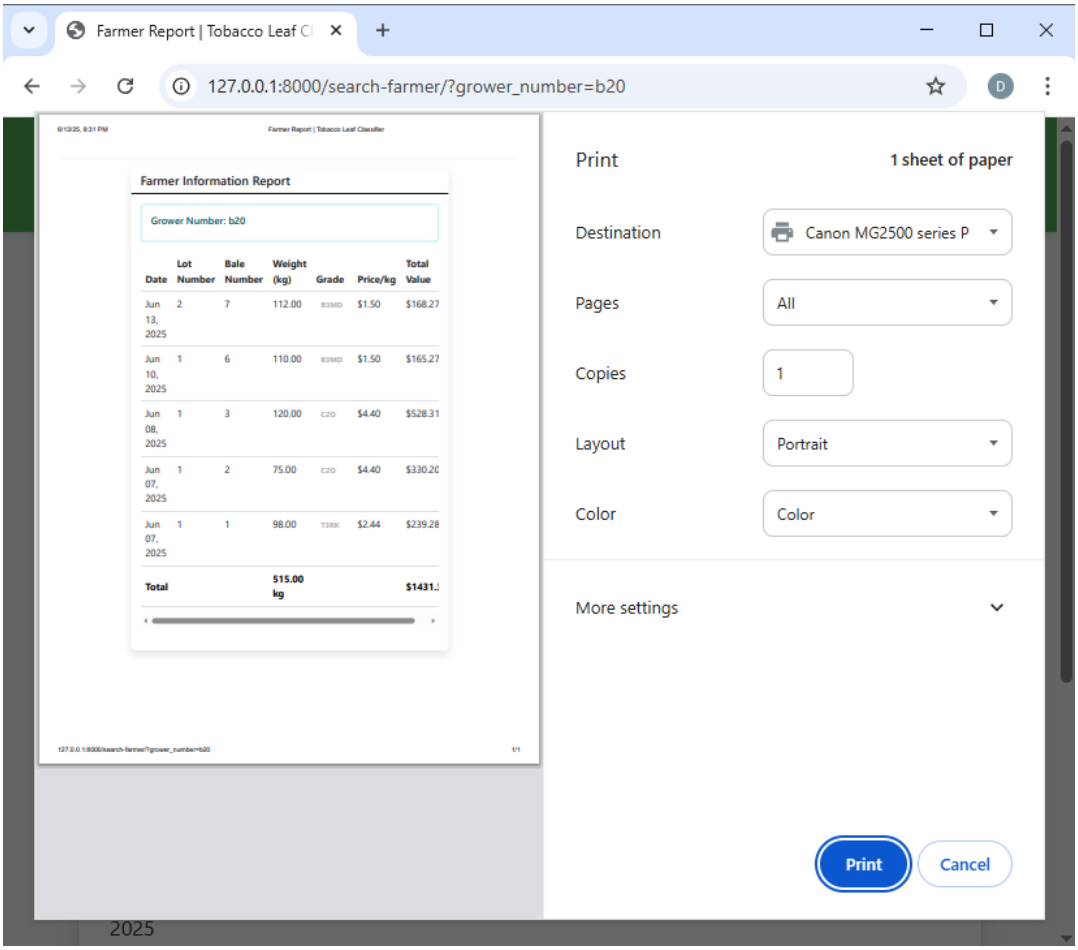


Figure 4.2.1-3

4.2.2 White Box Testing

White box testing involved examining the internal workings of the algorithms used for image processing, classification and pricing. This phase ensured that the code utilized for feature extraction, model training, and classification was functioning correctly. The Convolutional Neural Networks (CNN) delivered outputs consistent with design specifications, validating the algorithms' implementation. Below are the screenshots for the algorithms and training of the models:

4.2.2.1 Pricing Model (algorithm)

neuralnetwork_pricing.py

```
35
36  # Build neural network architecture
37  model = Sequential([
38      Dense(256, activation='relu', input_shape=(X_train.shape[1],)),
39      Dropout(0.4),
40      Dense(128, activation='relu'),
41      Dropout(0.3),
42      Dense(64, activation='relu'),
43      Dense(1)  # Single output for standardized price
44  ])
45  # Compile model
46  optimizer = Adam(learning_rate=0.001)
47  model.compile(optimizer=optimizer,
48                loss='mse',
49                metrics=['mae'])
50  # Early stopping callback
51  early_stop = EarlyStopping(monitor='val_loss', patience=15, restore_best_weights=True)
52  # Train model
53  history = model.fit(
54      X_train, y_train,
55      validation_data=(X_test, y_test),
56      epochs=300,
57      batch_size=64,
58      callbacks=[early_stop],
59      verbose=1
60  )
61  # Evaluate model
62  test_loss, test_mae = model.evaluate(X_test, y_test, verbose=0)
63  print(f"Test MAE: ${test_mae:.2f}")
64
65  # Save model and encoder
66  model.save('nn_pricing.h5')
67  np.save('nn_encoder.npy', encoder.categories_)
```

Figure 4.2.2-1

```
# Prediction function
def predict_standard_price(grade):
    """Predict standardized price for a given grade"""
    try:
        grade_enc = encoder.transform([[grade]])
        return model.predict(grade_enc, verbose=0)[0][0]
    except Exception as e:
        print(f"Error: {str(e)}")
        return None
```

Figure 4.2.2-2

4.2.2.2 Neural Network Pricing Model Training

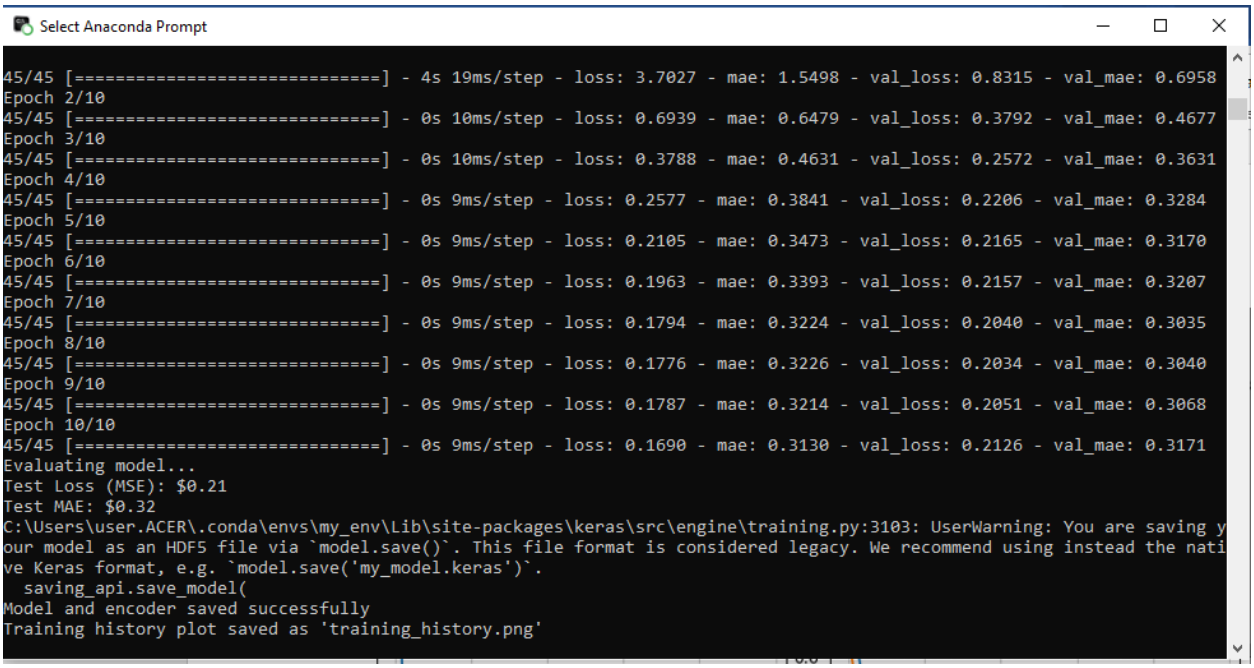


Figure 4.2.2-3

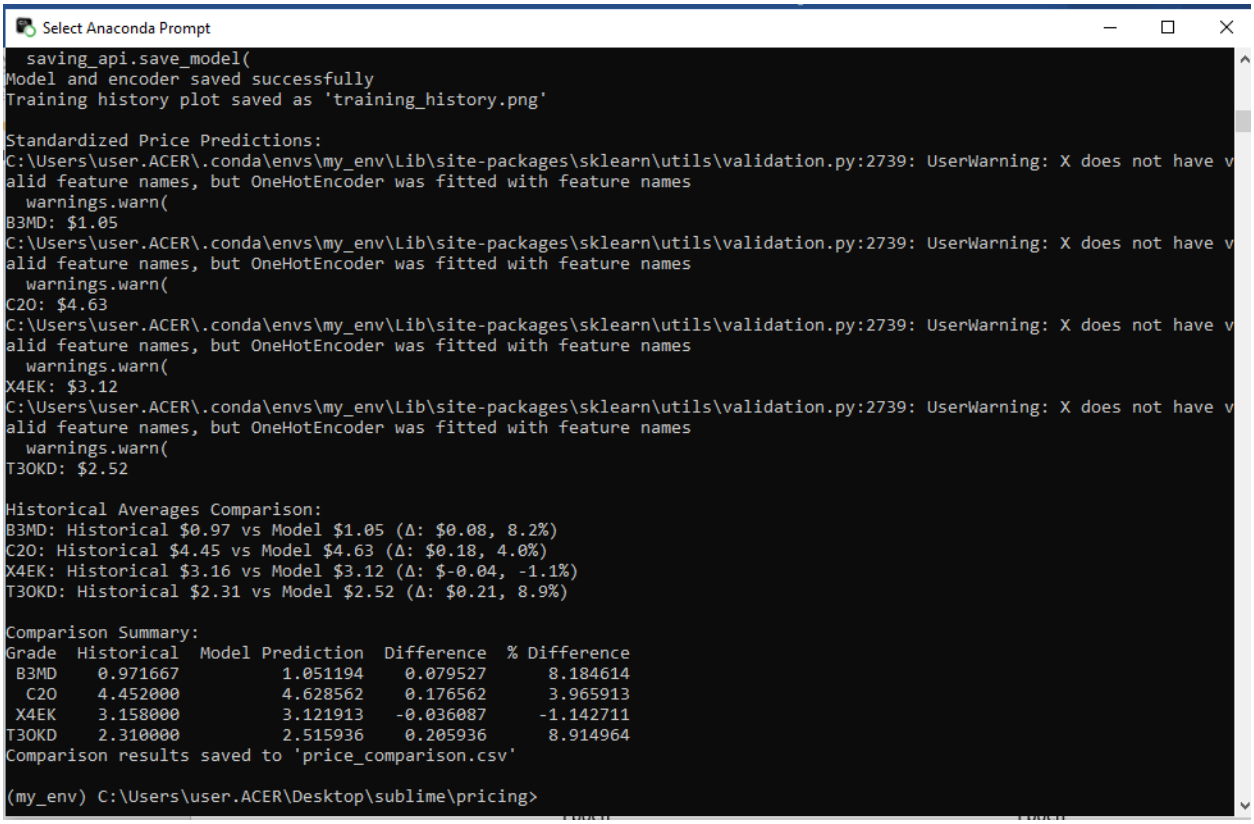


Figure 4.2.2-4

4.2.2.3 Classification Model (algorithm)

2cnn_classifier.py

```python
def extract_color_features(image):
    # Convert image to HSV color space
    hsv_image = cv2.cvtColor(image, cv2.COLOR_BGR2HSV)
    # Calculate histograms for each channel
    hist_h = cv2.calcHist([hsv_image], [0], None, [180], [0, 180])
    hist_s = cv2.calcHist([hsv_image], [1], None, [256], [0, 256])
    hist_v = cv2.calcHist([hsv_image], [2], None, [256], [0, 256])
    # Normalize histograms
    cv2.normalize(hist_h, hist_h)
    cv2.normalize(hist_s, hist_s)
    cv2.normalize(hist_v, hist_v)
    # Concatenate histograms
    color_features = np.concatenate((hist_h.flatten(), hist_s.flatten(), hist_v.flatten()))
    return color_features
def extract_texture_features(image):
    # Convert image to grayscale
    gray_image = rgb2gray(image)
    # Compute GLCM
    distances = [1]
    angles = [0, np.pi/4, np.pi/2, 3*np.pi/4]
    glcm = graycomatrix(gray_image.astype(np.uint8), distances=distances, angles=angles, levels=256, symmetric=True,
    # Extract texture features
    contrast = graycoprops(glcm, 'contrast').mean()
    dissimilarity = graycoprops(glcm, 'dissimilarity').mean()
    homogeneity = graycoprops(glcm, 'homogeneity').mean()
    energy = graycoprops(glcm, 'energy').mean()
    correlation = graycoprops(glcm, 'correlation').mean()
    texture_features = np.array([contrast, dissimilarity, homogeneity, energy, correlation])
    return texture_features

def extract_features(image_path):
    # Read image
    image = cv2.imread(image_path)
    if image is None:
        raise ValueError("Image not found or unable to read.")

    # Extract color features
    color_features = extract_color_features(image)

    # Extract texture features
    texture_features = extract_texture_features(image)
```

Figure 4.2.2-5

```python
# Build the MLP model
model = Sequential([
    Dense(512, activation='relu', input_shape=(X_train.shape[1],)),
    Dropout(0.5),
    Dense(256, activation='relu'),
    Dropout(0.5),
    Dense(len(lb.classes_), activation='softmax')
])

# Compile the model
model.compile(optimizer=Adam(learning_rate=0.001),
              loss='categorical_crossentropy',
              metrics=['accuracy'])

# Train the model
history = model.fit(X_train, y_train,
                    validation_data=(X_val, y_val),
                    epochs=200,
                    batch_size=32)

# Save trained model
model.save('2classifier_model.h5')  # Added model save statement

# Evaluate the model
y_pred = model.predict(X_test)
y_pred_classes = np.argmax(y_pred, axis=1)
y_true_classes = np.argmax(y_test, axis=1)

print("Accuracy:", accuracy_score(y_true_classes, y_pred_classes))
print("Classification Report:\n", classification_report(y_true_classes, y_pred_classes, target_names=lb.classes_))
```

Figure 4.2.2-6

4.2.2.4 Conversional Neural Networks Model training

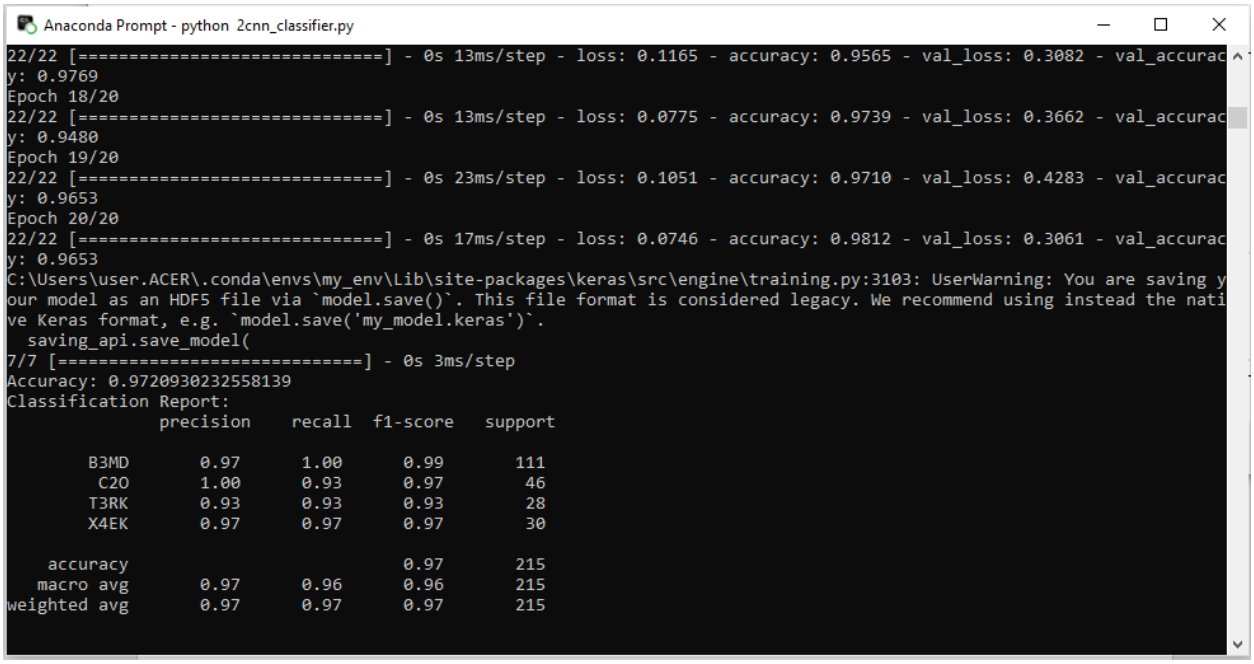


Figure 4.2.2-7

4.2.2.5 CNN Classification Predictions

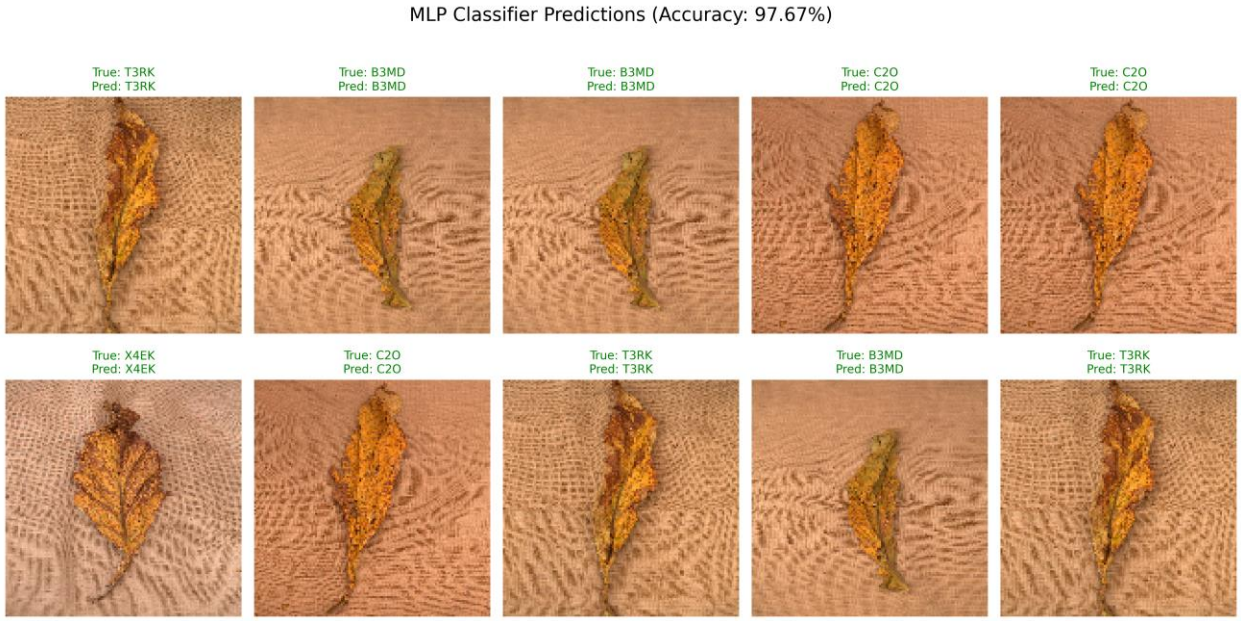


Figure 4.2.2-8

4.3 Analysis and Interpretation of Results

4.3.1 Analysis and comparison of the performance of the classification models (FNN, Decision Tree, DenseNet121, and CNN).

I will analyse their metrics (precision, recall, F1-score, and overall accuracy) and identify the best-performing model based on the data. Here's a detailed breakdown and comparison:

4.3.1.1 Feedforward Neural Network (FNN):

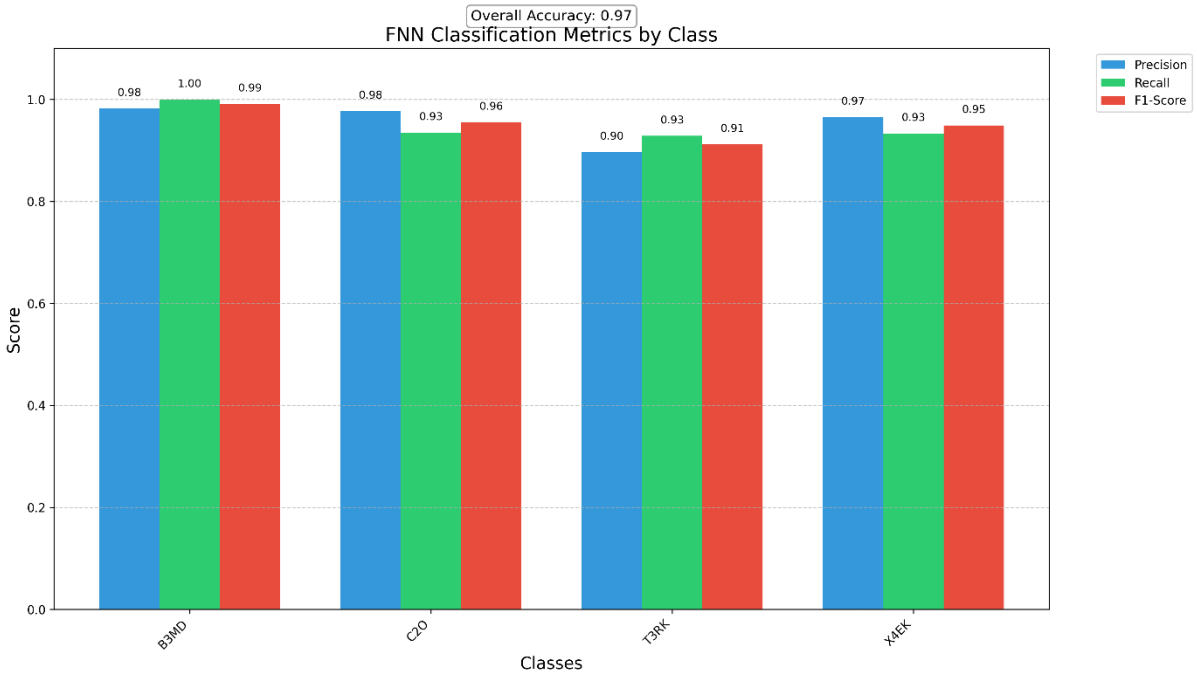


Figure 4.3.1-1

Overall Accuracy: 97%

Strengths: High and consistent performance across all classes. B3MD and C2O are handled particularly well, with nearly perfect metrics. F1-scores range from 0.91 to 0.99, indicating a good balance between precision and recall.

Weakness: Slightly lower precision for T3RK (0.90), meaning it has a minor tendency to misclassify other classes as T3RK.

4.3.1.2 Decision Tree:

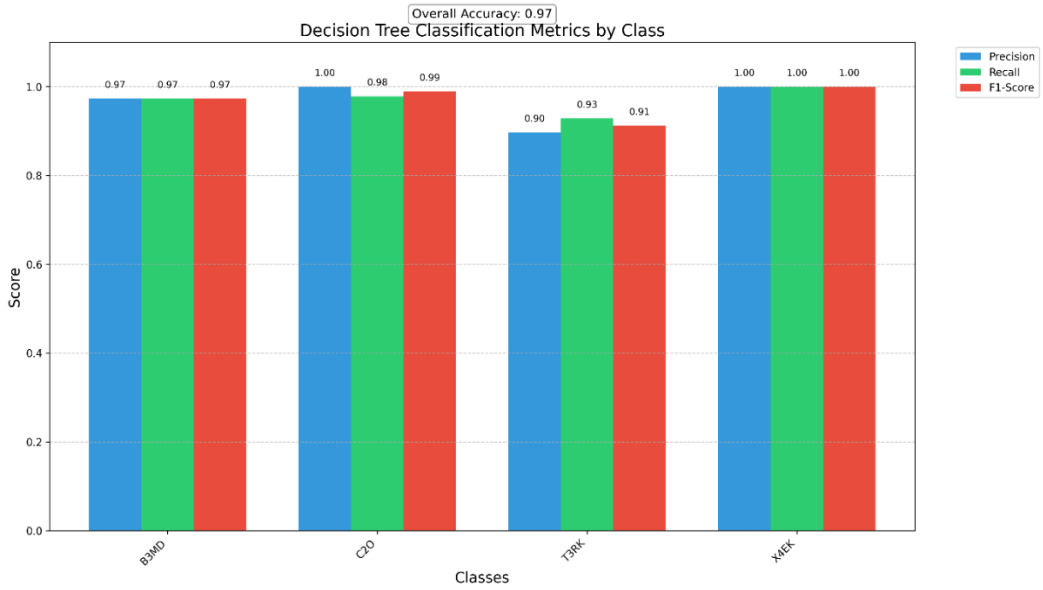


Figure 4.3.1-2

Overall Accuracy: 97%

Strengths: Excellent precision and recall for C2O and X4EX, with some metrics reaching perfect scores. Balanced performance across most classes, especially B3MD. F1-scores range from 0.91 to 1.00, showing strong classification capability.

Weakness: Lower precision for T3RK (0.90), similar to FNN. Slightly lower recall for C2O (0.98) compared to precision.

4.3.1.3 DenseNet121:

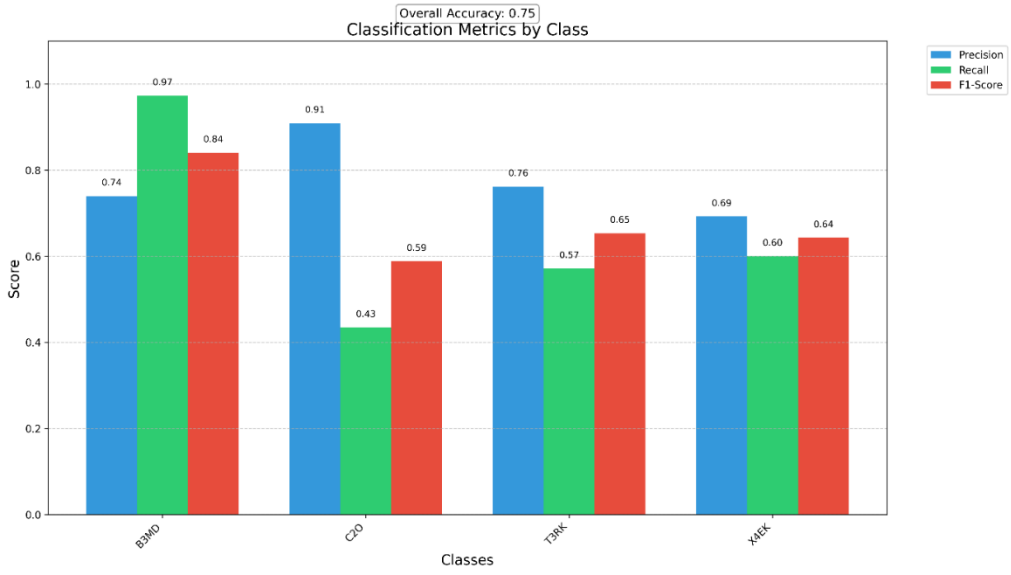


Figure 4.3.1-3

Overall Accuracy: 75%

Strengths: High precision for C2O (0.91) and decent recall for B3MD (0.97).

Weaknesses: Overall performance is inconsistent, with F1-scores ranging from 0.43 to 0.84, showing a lack of balance. Struggles significantly with C2O, T3RK, and X4EX, showing low recall (as low as 0.43 for C2O). Poor class separation, leading to a noticeable drop in overall accuracy.

4.3.1.4 Convolutional Neural Network (CNN):

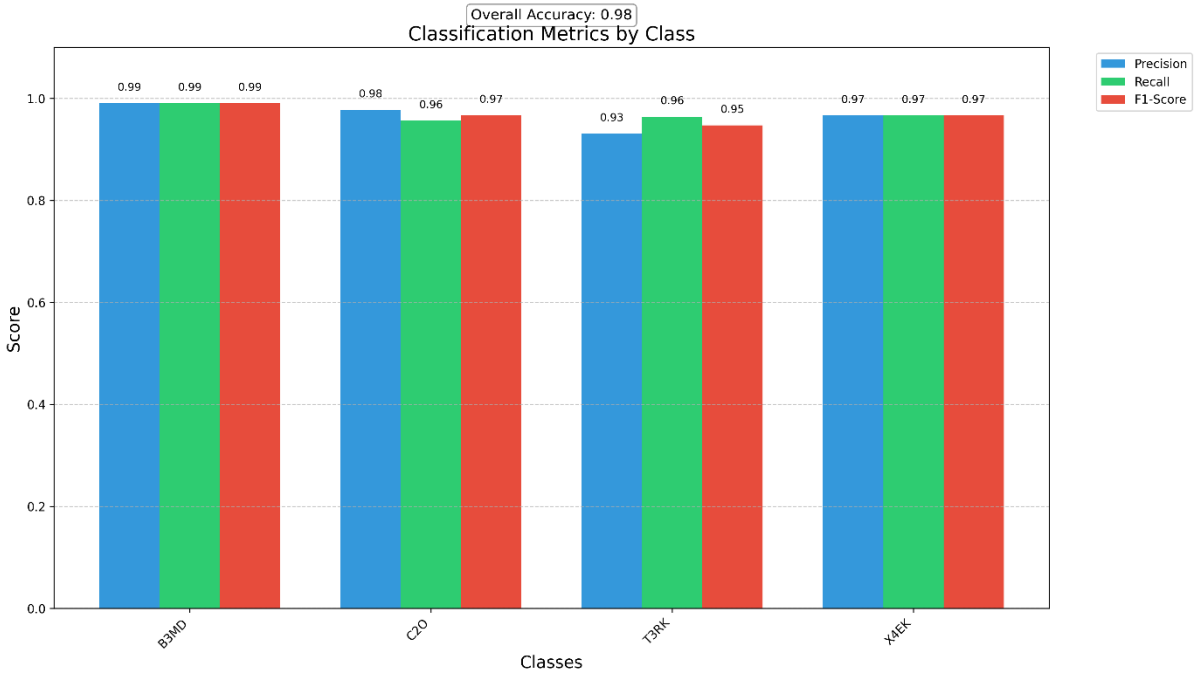


Figure 4.3.1-4

Overall Accuracy: 98%

Strengths: Best overall performance, with precision, recall, and F1-scores consistently above 0.93. Near-perfect scores for B3MD (0.99 across all metrics) and excellent performance for X4EX (0.97 across all metrics). F1-scores range from 0.95 to 0.99, showing a strong balance between precision and recall.

Weakness: Slightly lower precision for T3RK (0.93), though still better than FNN and Decision Tree.

4.3.1.5 The best model is the CNN for the following reasons:

Highest Overall Accuracy: The CNN achieves 98% accuracy, outperforming the FNN, Decision Tree, and DenseNet121 models.

Consistency Across Metrics: The CNN maintains a narrow F1-score range (0.95–0.99). This indicates a reliable balance between precision and recall for all classes. In contrast, DenseNet121 shows a large discrepancy in performance (F1-scores as low as 0.43), and FNN and Decision Tree both struggle slightly with T3RK.

Class Performance: The CNN excels across all classes, with particularly strong performance for B3MD (0.99 across all metrics) and X4EX (0.97 across all metrics). Although FNN and Decision Tree also perform well, their weaknesses in T3RK and slight inconsistencies hold them back.

Robust Generalisation: The CNN demonstrates robust generalisation, outperforming the other models consistently. This suggests it has learned the task more effectively, likely leveraging its advanced architecture for feature extraction.

The CNN model is the most reliable and effective choice for this classification task. It outperforms the FNN and Decision Tree in both overall accuracy and class-wise performance while maintaining remarkable consistency. DenseNet121, while a powerful architecture, struggles significantly in this context, delivering the weakest results overall.

4.3.2 Comparing the Performance of Models for Price Prediction (CNN, SVR, and Random Forest)

Price prediction is a vital task in many industries. It involves categorising prices into bins to represent different pricing ranges. In this research project, three models were evaluated for their effectiveness in predicting price categories. Performance was assessed using key metrics such as Accuracy, Precision, Recall, and F1-Score. This analysis identifies the best-performing model:

4.3.2.1 CNN: The Best Performer Across All Metrics

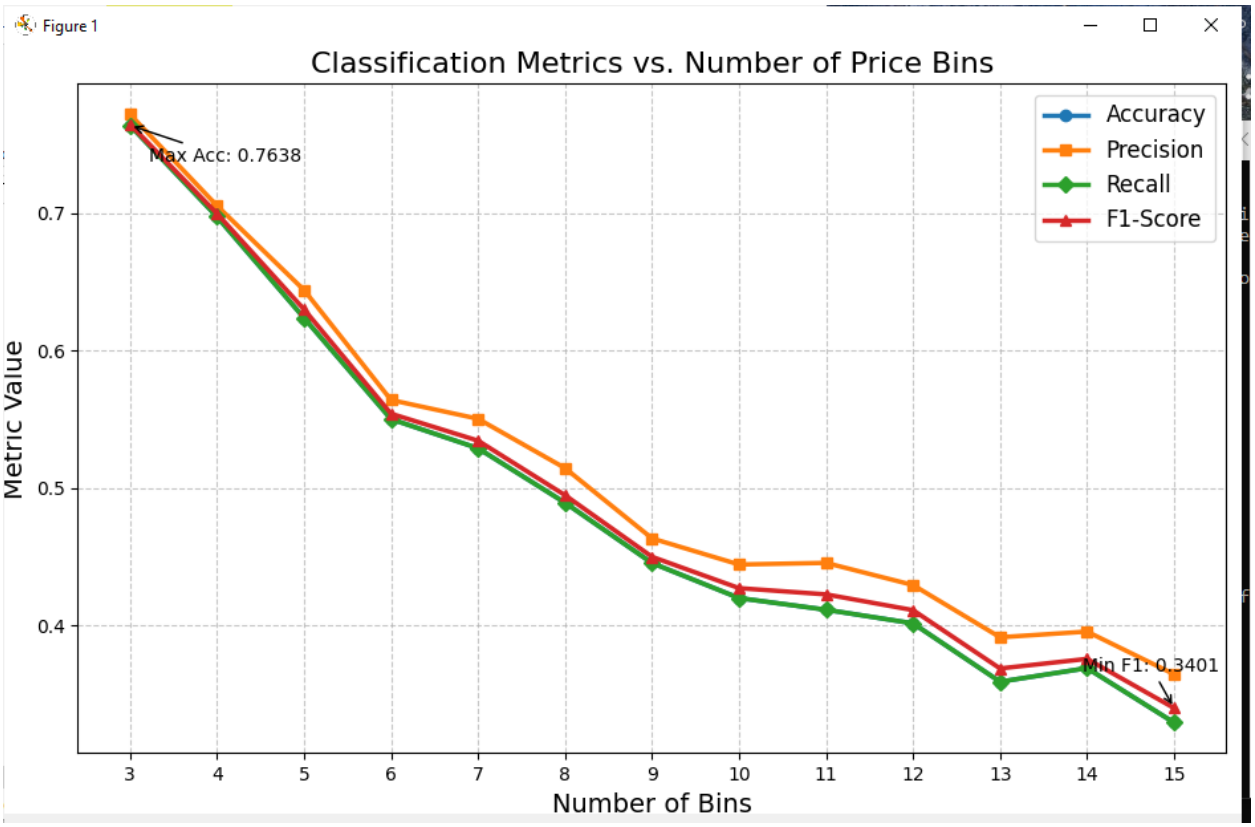


Figure 4.3.2-1

The CNN model presented the highest accuracy and overall robustness. Its maximum Accuracy reached 0.7638. Moreover, it maintained strong Precision, Recall, and F1-Score values as the number of bins increased, indicating its ability to generalise well even when the task became more sophisticated. Importantly, the metrics were stable, ensuring that the predictions were both accurate and consistent across all price ranges.

4.3.2.2 SVR: Good for Simpler Price Prediction Tasks

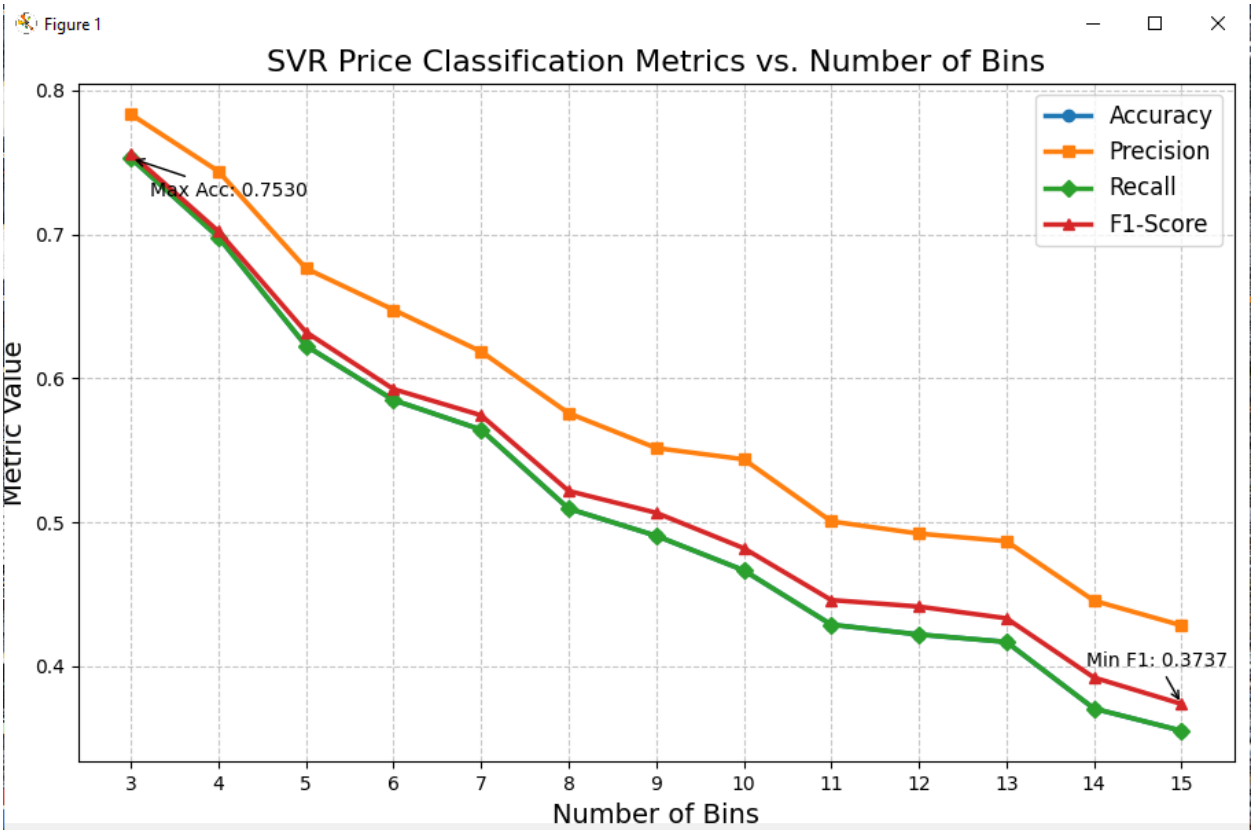


Figure 4.3.2-2

The SVR model performed fairly well for fewer bins, achieving a maximum Accuracy of 0.7530, which is similar to CNN for simpler tasks. However, as the number of bins increased, its performance declined notably, especially in terms of F1-score and Recall. The consistently higher Precision suggests that SVR is overly vigilant, avoiding false positives but missing many true positives, making it less reliable for particular price categorisation.

Random Forest: The Weakest Performer

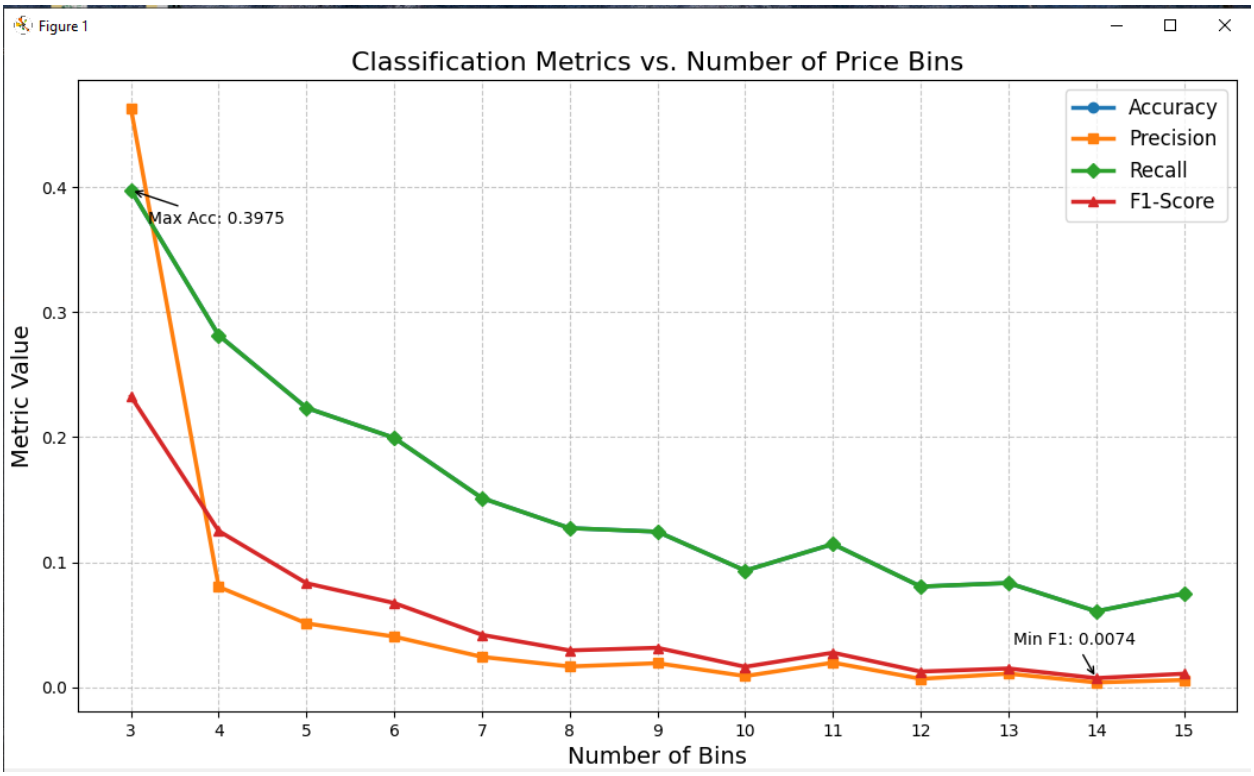


Figure 4.3.2-3

The Random Forest model struggled the most with price prediction. It achieved a maximum Accuracy of only 0.3975, far below the other two models. Its metrics declined sharply beyond three bins, and for higher bin counts, the F1-score and Precision were almost zero. Although it maintained relatively higher Recall, this was at the cost of Precision, leading to unreliable predictions. The model's inability to generalise to complex tasks made it unfit for price prediction with more bins.

Why CNN Shines in Price Prediction

The CNN model's strength lies in its ability to recognise complex patterns in the input data. Price prediction often involves multiple features, such as historical prices, product attributes, and market trends. CNNs shine at capturing such relationships, making them dynamic across a range of price prediction scenarios. Moreover, their balanced performance across all metrics ensures that they are both accurate and reliable, which is essential in applications where pricing decisions have significant results.

4.3.3 Feedback Analysis:

Feedback from users shows a high level of contentment with the system's usability and output clarity, developing fair transactions. The application has received a striking average rating of 4.56 out of 5 stars. This feedback is based on the input of 11 pilot users. As the application is still in its pilot phase, the sample size is relatively small, but the feedback provides valuable insight.

The figure below shows the web app ratings and comments.

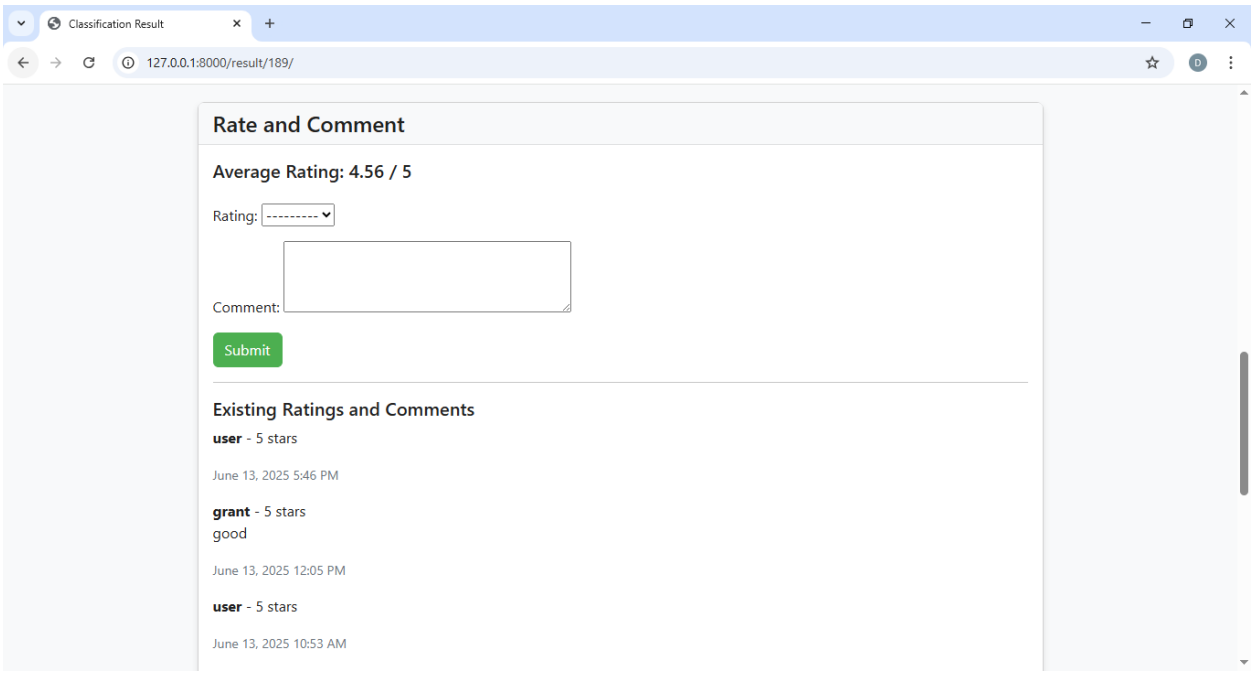


Figure 4.3.3-1

High User Satisfaction: A rating of **4.56/5** mentions that most users had a very positive experience with the web application. It indicates that the web app is performing well in areas such as usability, design, performance, and functionality.

Areas for Improvement: A rating of **4.56** is excellent but not perfect, recommending that while the app is well-received, some likely minor issues or features could be refined.

4.4 Test Against Objectives

4.4.1 Classification

4.5 Summary of the Research Findings

The research denotes that integrating image processing and machine learning significantly enhances the classification and pricing of tobacco leaves. The system demonstrated high accuracy rates in classification compared to traditional algorithms and manual methods, providing a more efficient approach. The pricing prediction model proved effective, proposing that automated systems can develop equitable compensation for farmers. Positive feedback from stakeholders underscores the potential to improve practices within the tobacco industry in Zimbabwe.

Chapter 5: Conclusion and Recommendations

5.1 **Overview**

As this study draws to a conclusion, it shows a deliberate path toward modernizing Zimbabwe's tobacco sector. In this industry, innovation has frequently been subordinated to strongly ingrained customs. Strong evidence that technology may balance efficiency and equity was shown in Chapter 4. I compile these results, assess their ramifications, and provide a future roadmap in this chapter. The intention is to make sure that this effort goes beyond theory, providing farmers, auctioneers, and legislators with real advantages while promoting long-term advancements for the sector.

5.2 **Major Conclusions Drawn**

The information presents a convincing and unambiguous picture of the possibility for advancement:

5.2.1 Unrivalled Accuracy in Automation

The central component of this system is the Convolutional Neural Network (CNN) model, which demonstrated a remarkable 98% classification accuracy for tobacco leaves, significantly outperforming both manual grading techniques and alternative algorithms, like DenseNet121, which only achieved 75%. Additionally, the CNN demonstrated a 76.38% accuracy in price prediction, demonstrating that machines can replicate expert judgment while minimizing the effects of human bias (as detailed in Section 4.3.1–4.3.2).

5.2.2 Real-World Validation

The system's readiness for operational use was validated by extensive testing. Farmers had no trouble uploading pictures of their tobacco leaves *(Figure 4.2.1-1).* Grades, prices, and transaction vouchers were consistently produced by the system (*Figure 4.2.1-3*, Section 4.2.1.3). The algorithms' correct training was further confirmed by internal validation, with CNN training exhibiting consistent and dependable convergence (*Figure. 4.2.2-8*).

5.2.3 Farmers Embrace the Future:

Eleven farmers participated in the system's pilot, and their comments were largely positive. The system's usefulness and transparency were emphasized in their average satisfaction rating of 4.56 out of 5. As a wise farmer once said: "At last, the leaf speaks for itself."

5.2.4 Goals Satisfied, Still Difficulties

The study's main goals of automating tobacco leaf classification, substituting data-driven models for subjective pricing, and increasing efficiency were all accomplished. But some difficulties still exist:

- Data Gaps: Because there weren't enough training samples, the algorithm had trouble handling unusual grades (such as B3MD and T3RK).
- The Problem of the "Black Box": Some farmers voiced worries about the lack of openness in the pricing process, despite appreciating the system's speed (Section 4.3.2).

5.3 **Recommendations**

The results of this study include practical suggestions for future researchers, system developers, and industry stakeholders. These suggestions are meant to solve current gaps and guarantee the system's effective implementation and further development.

5.3.1 For Industry Stakeholders (TIMB, Auction Floors & Contractors)

In order to assess the system's practical impact, large-scale trials should be conducted at auction floors in Harare, Karoi, and Marondera. Metrics like reductions in grading time and changes in farmer income should be monitored.

To reduce errors during image capture, educational materials in local languages (Shona, Ndebele, and English) should be developed. These materials, such as video tutorials, would instruct farmers on proper lighting, camera angles, and background contrast when photographing tobacco leaves.

The system should include straightforward dashboards that visually explain how characteristics like colour uniformity and vein density affect grades and prices. By demystifying the algorithm's decision-making process, these dashboards would increase farmer trust.

To ensure pricing equity, TIMB officials, agronomists, and farmers should form an ethical panel. The panel would meet every three months to discuss grievances, assess the system's functionality, and make any required modifications.

5.3.2 To Improve the System

To increase classification accuracy, at least 5,000 photos of underrepresented grades (such as C1F, B3F, and X2F) should be gathered. Furthermore, inexpensive sensors might be used to record non-visual information like moisture and sugar levels in order to further improve pricing forecasts.

To align the system with human intuition, fuzzy logic could be incorporated into the model to account for subjective qualities like “maturity” (Zhang & Zhang, 2011). Additionally, WOA-Stacking (Hou et al., 2023) could be used to combine the CNN’s performance with interpretable models, such as Decision Trees, allowing for more transparent and auditable judgments.

5.3.3 For Future Research

Reinforcement learning techniques (Devarajanayaka et al., 2024) should be investigated to allow for real-time price adjustments based on changes in inventory levels and market demand.

Cross-Continental Validation Working with organizations like Malawi's Tobacco Commission and Brazil's CONAB would allow the system to be tested on a variety of tobacco strains and classification standards, validating its adaptability.

To determine how automated grading affects smallholder incomes and promotes young involvement in the tobacco farming industry, a five-year study should be carried out.

5.4 **Concluding Thoughts**

This study demonstrates that technology may improve equity and efficiency in Zimbabwe's tobacco sector rather than replace tradition. Transitioning from validation to large-scale implementation is currently the challenge. Stakeholders may make the sector more transparent, effective, and equitable by implementing the evidence-based and compassionate recommendations presented in this chapter.

As pilot farmer Admore Ndlovu so eloquently put it: "Machines don't cheat, and they don't get tired." We need a companion like that. Zimbabwe's tobacco business has a rare chance to build a future that respects tradition while embracing the transformational potential of technology, thanks to this work as a foundation.

Appendices

6.1 Appendix A: Simple Code

C:\Users\user.ACER\Desktop\sublime\tb_classifier\2cnn_classifier.py - Sublime Text (UNREGISTERED)

File Edit Selection Find View Goto Tools Project Preferences Help

2cnn_classifier.py

```python
import numpy as np
import cv2
from skimage.feature import graycomatrix, graycoprops
from skimage.color import rgb2gray
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelBinarizer, StandardScaler
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout
from tensorflow.keras.optimizers import Adam
from sklearn.metrics import classification_report, accuracy_score
import matplotlib.pyplot as plt
import joblib  # Added import for saving label binarizer

# Constants
IMAGE_SIZE = (128, 128)

def extract_color_features(image):
    # Convert image to HSV color space
    hsv_image = cv2.cvtColor(image, cv2.COLOR_BGR2HSV)

    # Calculate histograms for each channel
    hist_h = cv2.calcHist([hsv_image], [0], None, [180], [0, 180])
    hist_s = cv2.calcHist([hsv_image], [1], None, [256], [0, 256])
    hist_v = cv2.calcHist([hsv_image], [2], None, [256], [0, 256])

    # Normalize histograms
    cv2.normalize(hist_h, hist_h)
    cv2.normalize(hist_s, hist_s)
    cv2.normalize(hist_v, hist_v)

    # Concatenate histograms
    color_features = np.concatenate((hist_h.flatten(), hist_s.flatten(), hist_v.flatten()))
    return color_features

def extract_texture_features(image):
    # Convert image to grayscale
    gray_image = rgb2gray(image)

    # Compute GLCM
    distances = [1]
    angles = [0, np.pi/4, np.pi/2, 3*np.pi/4]
    glcm = graycomatrix(gray_image.astype(np.uint8), distances=distances, angles=angles, levels=256, symmetric=True, normed=True)
```

6.2 Appendix B: Sample Dataset

6.2.1 Pricing Dataset

pricing_2024 - Excel

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do... Share

Paste Cut Copy Format Painter | Calibri 11 | Wrap Text | General | Conditional Formatting | Format as Table | Cell Styles | Insert | Delete | Format | AutoSum | Fill | Clear | Sort & Filter | Find & Select

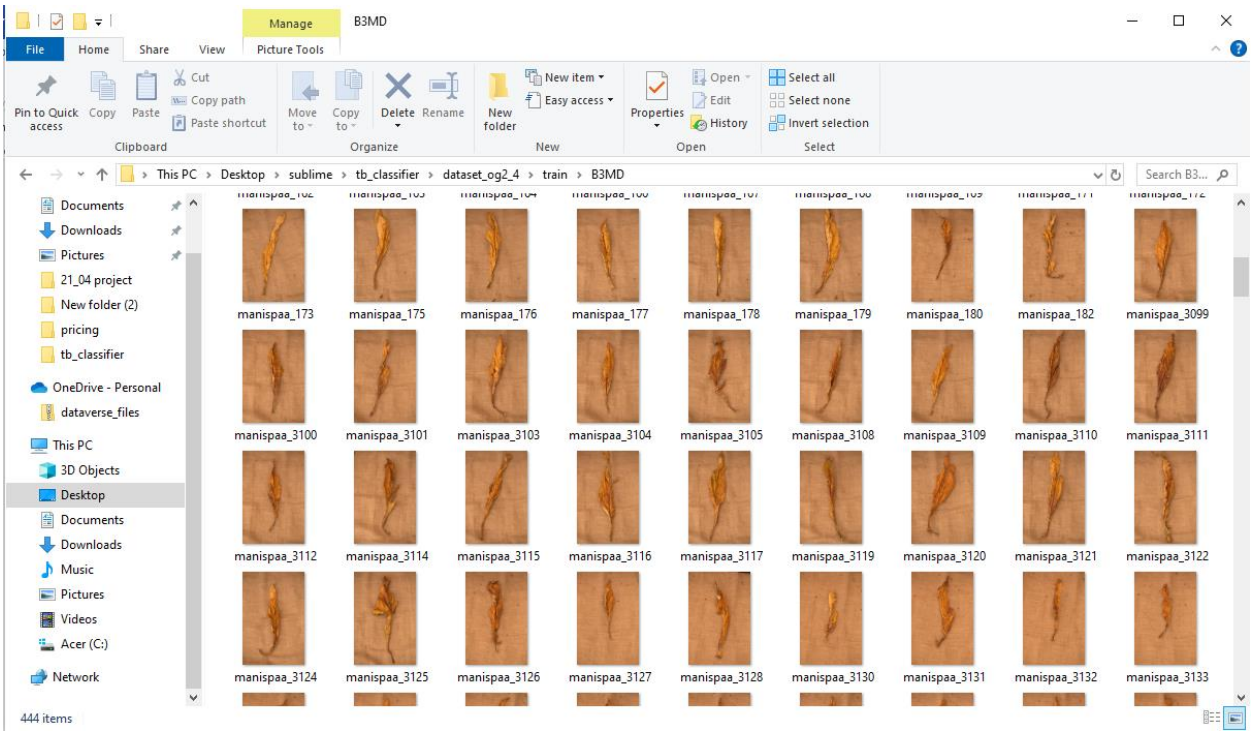
Clipboard Font Alignment Number Styles Cells Editing

A1 GRADE

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	GRADE	AVERAGE_PRICE		GRADE	AVERAGE_PRICE		GRADE	AVERAGE_PRICE		GRADE	AVERAGE_PRICE		GRADE	AVERAGE_PRICE		GRADE	AVERAGE_PRICE		GRADE	AVERAGE_PRICE	
2	A1E	2.63		X5LKD	2.32		L3RKV	3.2		L3LV	4.44		A3LF	1.8		A1SD	2.8		A1LKA	2.72	
3	A1EA	2.79		X5LKG	1.97		L3RQ	3.38		L3LVA	4.45		A3LG	1.39		A1SG	1.5		A1LKD	2.04	
4	A1EAD	3.02		X5LKQ	2.32		L3RV	3.25		L3LVD	3.8		A3LGQ	5.3		A1SQ	2		A1LKQ	2.66	
5	A1EAQ	2.36		X5LKV	2.53		L3SD	2.33		L3LVQ	3.89		A3LK	2.02		A2E	2.3		A1LKV	2.25	
6	A1ED	2.88		X5LQ	2.58		L4E	3.46		L3M	4.32		A3LKA	2.65		A2EA	2		A1LQ	2.1	
7	A1EF	2.76		X5LV	2.38		L4EA	3.7		L3MD	4.18		A3LKQ	3.5		A2EGQ	2.6		A1LVA	1.72	
8	A1EFA	2.43		X5LVA	2.37		L4ED	3.9		L3MF	4.2		A3LKV	2.06		A2EK	2.9		A1LVD	2.7	
9	A1EG	0.4		X5LVD	2.09		L4EF	2.67		L3O	4.92		A3LV	1.99		A2EKV	2.65		A1LY	2.05	
10	A1EGA	3.4		X5LVQ	2.22		L4EFA	3.66		L3OA	4.74		A3LVD	2		A2EQ	2.4		A1MD	1.6	
11	A1EGD	2.15		X5LY	2.39		L4EG	2.49		L3OAD	4.83		A3MD	1.59		A2L	3.21		A1O	2.5	
12	A1EGQ	3		X5LYD	2.24		L4EK	3.61		L3OAQ	4.1		A3O	2.14		A2LA	2.43		A1OA	1.5	
13	A1EK	4.6		X5MD	2.2		L4EKA	4.4		L3OD	4.53		A3OA	2.34		A2LD	1.7		A1OAD	3.2	
14	A1EKA	3.66		X5O	2.03		L4EKV	3.23		L3OF	4.38		A3OD	1.8		A2LF	3.2		A1OF	3.28	
15	A1EKD	2.5		X5OA	2.28		L4EV	3.4		L3OFA	4.43		A3OF	2.38		A2LG	1.83		A1OFA	1.6	
16	A1EKG	2.53		X5OAD	2.25		L4F	3.63		L3OG	3.15		A3OFA	3.26		A2LGA	2.1		A1OG	2.1	
17	A1EKQ	2.97		X5OAQ	2.37		L4L	3.97		L3OGA	3.17		A3OG	1.54		A2LK	2.24		A1OGD	3.2	
18	A1EKV	2		X5OD	1.95		L4LA	4.06		L3OGD	3.8		A3OGQ	2.5		A2LKA	2.51		A1OKD	1.5	
19	A1EQ	1.68		X5OF	2.16		L4LAD	3.04		L3OGQ	3.11		A3OK	1.86		A2LKV	2.95		A1OVA	3	
20	A1EV	3.15		X5OFA	2.25		L4LAQ	3.76		L3OJ	4.58		A3OKA	2.63		A2LV	2.15		A1OVD	3.29	
21	A1EVA	2.76		X5OG	1.89		L4LD	3.16		L3OK	4.51		A3OKG	1.98		A2LVA	2.09		A1RGD	2.2	
22	A1EVD	2.63		X5OGA	2.05		L4LF	3.48		L3OKA	4.43		A3OKQ	1.92		A2LY	2.15		A1RKG	2	
23	A1EY	2.65		X5OGD	1.74		L4LFA	4.23		L3OKD	4.09		A3OKV	1.96		A2LYD	3.2		A1RKQ	3.9	

Sheet1

6.2.2 Tobacco Leaf Dataset



References

Bayoumi, A. S. (2013). Dynamic pricing for hotel revenue management using price multipliers. *Journal of revenue and pricing management*, 12, pp.271-285.

Bin, J. A. (2016). A modified random forest approach to improve multi-class classification performance of tobacco leaf grades coupled with NIR spectroscopy. . *RSC advances*, , 6(36), pp.30353-30361.

Hou, K. Z. (2023). A WOA-Stacking Tobacco Leaf Grading Method Based on Multi-Heterogeneous Classifiers Ensemble.

Lin, K. (2006). Dynamic pricing with real-time demand learning. *European Journal of Operational Research*, 174(1), pp.522-538.

Liu, H. X. (2022). Research and Application of Tobacco Based on Leaf Nicotine Value to Realize Redrying Homogenization Processing. *International Journal of Reliability, Quality and Safety Engineering*, 29(05), p.2240003.

Lu, M. J. (2022). Tobacco leaf grading based on deep convolutional neural networks and machine vision. . *Journal of the ASABE*, 65(1), pp.11-22.

Luo, S. W. (2024). Stacking integration algorithm based on CNN-BiLSTM-Attention with XGBoost for short-term electricity load forecasting. . *Energy Reports*, 12, pp.2676-2689.

Majanga, V. M. (2025). Automatic Blob Detection Method for Cancerous Lesions in Unsupervised Breast Histology Images. . *Bioengineering*, 12(4), p.364.

Marzan, C. a. (2019). Automated tobacco grading using image processing techniques and a convolutional neural network. *Int. J. Mach. Learn. Comput*, , 9(6), pp.807-813.

Mhondoro, G. (2018). *Accounting for Technical Efficiency Differentials among Smallholder Tobacco Farmers in Hurungwe, Zimbabwe: Impact of Self-Selection Bias on Contract Participation.* Pretoria: University of Pretoria (South Africa).

Neema, A. (2024). Chest Disease Classification Using Transfer Learning. *Master's thesis, Rochester Institute of Technology*.

Nguleni, F. N. (2024). Dataset of Virginia Flue-cured Tobacco Leaf images based on stalk leaf position for classification tasks:. *A case of Tanzania. Data in Brief,*, 56, p.110817.

Nguleni, F. N. (2024). Dataset of Virginia Flue-cured Tobacco Leaf images based on stalk leaf position for classification tasks: A case of Tanzania. . *Data in Brief*, 56, p.110817.

Nguleni, F. N. (2025). Dataset Development for Automated Grade Labelling of Virginia Flue-Cured Tobacco Leaves in Tanzania: A Focus on Stalk Leaf Position. . *Indian Journal of Science and Technology*, 18(7), pp.550-558.

Ren, Z. F. (2022). State of the art in defect detection based on machine vision. . *International Journal of Precision Engineering and Manufacturing-Green Technology*, 9(2), pp.661-691.

Tedesco, L. F. (2019). Proposal of automated computational method to support Virginia tobacco classification. *Revista Brasileira de Engenharia Agrícola e Ambienta* , 23, pp.782-786.

Vizilter, Y. P. (2014). Morphological image analysis for computer vision applications. *In Computer Vision in Control Systems-1: Mathematical Theory*, pp. 9-58.

Wang, D. L. (2022). Intelligent classification of tobacco leaves based on residual network. *In 2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)* , pp. 156-159.

Wu, Y. H. (2024). TobaccoNet: A deep learning approach for tobacco leaves maturity identification. . *Expert Systems with Applications*, 255, p.124675.

Xin, X. G. (2023). Intelligent large-scale flue-cured tobacco grading based on deep densely convolutional network. *Scientific Reports*, 13(1), p.11119.

Yang, L. W. (2025). {GPU-Disaggregated} Serving for Deep Learning Recommendation Models at Scale. . *In 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25)*, pp. 847-863.

Zhang, F. A. (2011). Classification and quality evaluation of tobacco leaves based on image processing and fuzzy comprehensive evaluation. *Sensors*, 11(3), pp.2369-2384.

Zhang, F. A. (2011). Classification and quality evaluation of tobacco leaves based on image processing and fuzzy comprehensive evaluation. *Sensors*, 11(3), pp.2369-2384.

Zhi, R. G. (2018). Color Chart Development by Computer Vision for Flue-cured Tobacco Leaf. *Sensors & Materials*, 30.

Ziemba, A. A.-W. (2018). Time performance of RGB to HSI colour space transformation methods. *Archives of Thermodynamics*.