Group Mini-Project
Fruit Identification and Classification

Angelina Ramsunar - 41081269

Stefan Du Plooy - 40954129

Rikus Swart - 42320755

Project submitted at the North-West University

Module code: ITRI 626

Date: 12-11-2025

Contents

List of Tables

List of Figures

INTRODUCTION

The fruit quality assessment acted as the chosen scenario for the completion of this mini-project. Assessment of fruit quality remains a challenge in industries like agricultural supply chains, where classification has to happen at a fast pace with accurate results. Traditional methods of inspection work in small-scale operations but have difficulty scaling up efficiently to industrial applications. Visual assessment by human inspectors introduces variability and inconsistency, particularly when distinguishing between subtle gradations of quality. This labour intensive nature of manual sorting creates bottlenecks that may compromise both throughput and economic viability, especially in cases where time is of the essence.

Convolutional neural networks (CNNs) have emerged as a promising solution to automate fruit quality assessment through image based classification. CNNs learn hierarchical feature representations directly from pixel data, potentially capturing patterns that correlate with quality indicators such as colour, surface texture and visible features.

The task itself appears relatively simple: given labelled images categorised into discrete quality classes (Good, Mild, Rotten), a CNN should learn to distinguish among them. But this is where this project attempts to look beyond the simple objective and attempt to answer several implementation questions that arose throughout the year during classes, that remain unresolved. To what extent does colour information contribute to classification accuracy, and might grayscale representations suffice? How do different preprocessing strategies, such as data augmentation or variations in input resolution, affect model generalisation? These questions motivated our experimental design, where we systematically evaluate multiple training scenarios to understand which factors most influence performance. This is the best example of applying what has been studied as a theoretical discussion, to reflect on the practical application thereof.

Our approach involves constructing CNN architectures that process images of fruit and output quality predictions, evaluated using standard metrics such as accuracy, precision, recall, F1-score and area under the ROC curve (AUC). We implement a series of controlled experimental scenarios: baseline models with standard colour images, grayscale conversions to isolate colour dependency, augmented datasets to test robustness and multiple input. Each scenario provides insight into how specific design choices shape model behaviour, potentially revealing whether certain preprocessing steps offer marginal gains or whether simpler configurations prove sufficient for this classification task.

While the original project specification focused solely on classifying the quality of the fruit, we expanded the scope to incorporate fruit type identification as a concurrent task, thereby transforming the problem into a multi-task learning framework. This extension was motivated by two intersecting considerations.

First, the single-task formulation presented a relatively straightforward classification problem with only three output classes, offering limited opportunity to explore more interesting architectural patterns and training dynamics that exists in deep learning. By introducing an additional classification objective, identifying which specific fruit type appears in each image, we create a richer experimental context that demands more complex feature learning and allows us to investigate how shared representations can serve multiple predictive goals simultaneously.

Second, this multi-task extension addresses a topic that has come up numerous times in literature and class discussion which we attempt to investigate: the efficient utilisation of labelled data. High-quality labelled datasets is very expensive in terms of time, money and domain expertise, particularly in agricultural contexts where consistent quality standards must be maintained across diverse fruit varieties and environmental conditions. The dataset employed in this study contains images already organised by both quality level and fruit type through directory structure and filename conventions, yet limiting ourselves to quality classification alone would effectively discard half of the available label information.

The multi-task learning domain offers a mechanism to leverage this richer labelling scheme without collecting additional data, potentially improving the return on annotation investment. Recent discussions in the transfer learning literature suggest that related tasks may benefit from shared feature representations, as lower-level visual patterns such as edge detection, texture recognition and colour distribution often prove relevant across multiple classification objectives. Whether this hypothesis holds for fruit classification, where quality assessment might depend on fruit-specific characteristics, remains an empirical question we aim to address through comparative evaluation of single-task versus multi-task architectures.

The primary objective centres on developing convolutional neural network architectures capable of identifying the quality of fruit from visual data. This entails constructing models that can learn unique features directly from pixel level representations. CNNs offer particular advantages for this task through their hierarchical feature learning, where early layers may capture low-level patterns such as edges and textures, while deeper layers synthesise these into higher order abstractions that correspond to quality indicators.

But it is important to note that for this research project the objective extends beyond merely achieving acceptable classification accuracy. It involves understanding which architectural choices prove most consequential for this task and whether relatively simple networks suffice under different scenarios of input data or whether more complex designs yield proportional improvements performance.

The second objective investigates the comparative performance of single-task versus multi-task learning paradigms. Single-task models focus exclusively on quality classification, optimising their parameters to distinguish among good, mildly degraded and rotten produce. Multi-task architectures, by contrast, simultaneously predict both quality level and fruit type through a shared convolutional backbone that feeds into separate classification heads.

This dual-objective formulation raises questions about whether forcing the network to learn features useful for multiple related tasks improves generalisation or whether task interference diminishes performance on either objective. This objective speaks to broader

discussions in transfer learning and multi-task optimisation about when parameter sharing proves advantageous and when task specific architectures remain preferable.

Accuracy offers an intuitive overview but may obscure class-specific performance disparities, particularly problematic when dealing with imbalanced datasets where certain quality categories appear more frequently than others. For this, various different performance metrics are introduced to analyse the output of the model and motivate the legitimacy of the findings.

Fourthly, analysing how variations in input data characteristics affect model performance represents a critical objective for understanding the practical constraints within this classification task. Colour information, for instance, might prove essential for distinguishing between early and advanced stages of degradation or texture patterns captured equally well in grayscale could suffice. Furthermore, we explore the role that image resolution plays in the trade-offs between computational efficiency and information preservation, higher resolutions retain finer details but demand greater processing resources and may increase training time substantially. Data augmentation techniques that artificially expand training sets through geometric transformations and photometric adjustments could improve robustness to variations in imaging conditions, or they might introduce artifacts that complicate learning.

The final objective synthesises findings across all experimental conditions to identify patterns that generalise beyond individual scenarios. This comparative analysis seeks to determine whether certain configurations consistently outperform alternatives regardless of specific implementation details or whether optimal approaches vary depending on contextual factors such as available computational resources, dataset characteristics, or operational requirements.

In summary, the research objectives comprise of:

1. Develop CNN-based deep learning models for fruit classification and quality assessment
2. Implement both single-task (quality only) and multi-task (fruit type and quality) learning approaches
3. Evaluate model performance using multiple metrics: accuracy, precision, recall, F1-score, ROC curves, and AUC
4. Analyse the impact of different input data characteristics on model performance
5. Compare single-task versus multi-task learning performance to assess architectural trade-offs

1.1 Dataset and Model Implementation

First we begin by acquiring and inspecting the dataset. The Fruit Quality Database (FruQ-DB) is used for this research study as the foundational data source. This dataset provides images of multiple fruit varieties across different quality states, organised hierarchically by fruit type and quality level.

Initial preprocessing involves structuring the data into training, validation and test partitions to support model development while maintaining strict separation between evaluation sets. The images have quality labels that are represented through directory organisation (Good, Mild, Rotten) and fruit type information are embedded in the filename prefixes. This labelling scheme enables both single-task quality classification and multi-task learning where fruit identification serves as an auxiliary objective.

For the data augmentation pipelines, we apply transformations such as random rotations, horizontal flips and adjustments to brightness and contrast, expanding the effective training set size while introducing controlled variability that may improve model robustness to more realistic unforeseen conditions not represented in the original dataset. These augmented images are saved separately, not affecting the original dataset.

For model development, we proceeded through implementation of convolutional neural network architectures in PyTorch. The architecture design follows established conventions for image classification networks, beginning with convolutional layers that extract spatial features through learned filter banks, followed by pooling operations that introduce translation invariance and reduce dimensionality. Thereafter, convolutional blocks increase filter depth while reducing spatial dimensions, a pattern that encourages hierarchical feature learning where early layers capture local patterns and deeper layers synthesise these into global representations. For single-task models, this convolutional backbone feeds into fully connected layers that produce quality class predictions. Multi-task architectures extend this design by branching after feature extraction into parallel classification heads, one for quality assessment, another for fruit type identification. This architectural choice embodies hypotheses about feature reusability across related visual tasks.

This experimental design split the research project into phases: Part A examines single-task quality classification under varying conditions, while Part B extends analysis to multi-task learning scenarios. This establishes clear baseline performance through single-task models before introducing the additional complexity of multi-task optimisation, allowing direct assessment of whether auxiliary objectives provide benefits or merely introduce confounding factors.

1.2 The Two Research Parts and Different Scenarios

Part A comprises three scenarios that manipulate input characteristics while maintaining focus on quality classification alone. Scenario 1 establishes baseline performance using standard RGB images pre-processed through resize operations and normalisation. This acts as an unaltered baseline for future comparisons that change the input scenarios. Scenario 2 converts images to grayscale, testing the hypothesis that colour information contributes critically to quality assessment, or conversely, that texture and structural features captured in intensity values is good enough for accurate classifications. This scenario addresses practical considerations about whether simpler single-channel representations might reduce computational requirements without sacrificing performance. Scenario 3 introduces data augmentation during training, applying random transformations that expand the effective dataset size while potentially improving generalisation to images captured under conditions not represented in the original training distribution. The dataset

provides near perfect images, this is why Scenario 3 is so important to test the model to more realistic variation of image inputs.

Part B replicates each of these scenarios within a multi-task learning framework where models simultaneously predict fruit type and quality level. Scenario 1 establishes multi-task baseline performance with standard RGB inputs. Scenario 2 examines whether grayscale conversion affects both tasks equally or whether fruit type identification proves more colour dependent than quality assessment. Scenario 3 evaluates augmentation under multi-task learning, questioning whether synthetic variations benefit both objectives or introduce task specific biases that help one classification head while punishing or degrading the other.

1.3 Training and Results

A fixed random seed is used to control to ensure that results are repeatable and reproducible. Train/validation/test splits maintain strict separation, with test sets reserved exclusively for final evaluation after all architectural decisions and hyperparameter selections are complete. Each scenario generates comprehensive logs capturing training dynamics, validation metrics across epochs and final test set performance.

Results analysis synthesises findings across all scenarios through multiple analytical lenses. Quantitative comparison of performance metrics identifies which configurations achieve superior classification accuracy, precision, recall, and F1-scores, while ROC curves and AUC values reveal discriminative capacity across varying decision thresholds. Confusion matrices expose class-specific error patterns, indicating whether models systematically misclassify particular quality levels or fruit types, which represents insights that might suggest targeted refinements or reveal inherent ambiguities in the classification task itself. Then finally cross scenario comparisons assess the impact of individual factors: does grayscale conversion consistently degrade performance by some quantifiable margin, or do effects vary depending on whether models operate in single-task or multi-task mode? Do augmentation benefits depend on resolution, suggesting interactions between preprocessing choices?

Training curves track loss and accuracy evolution across epochs indicate whether models converge reliably or exhibit instability, whether they overfit to training data or maintain consistent performance across training and validation sets.

CONVOLUTIONAL NEURAL NETWORK

As previously mentioned, this research project makes use of the Fruit Quality Database (FruQ-multi), which is a comprehensive image dataset specifically curated for assessing the produce quality across multiple fruit and vegetable varieties. The dataset comprises 9370 images spanning 11 distinct fruit types, each annotated according to quality level.

The dataset encompasses eleven fruit and vegetable categories. Within each fruit category, images are further subdivided into three quality classes that represent progressive stages of degradation. The "Good" or "Fresh" category contains images of high quality produce exhibiting minimal surface defects and absence of visible decay. The "Mild" category represents moderate quality with minor blemishes, slight discoloration or early indicators of degradation but do not render the produce unsuitable. The "Rotten" category encompasses poor quality fruits displaying extensive bruising, mold growth or structural collapse that clearly indicate spoilage.

The dataset distribution across fruit types and quality classes reveals substantial heterogeneity. Table 2.1 presents the complete breakdown of image counts. It is important to take note of the imbalances both between fruit types and within quality categories, as this will have an effect on the later stages of using the data. Tomatoes constitute the most represented fruit with 1990 images, while strawberries comprise the smallest subset with only 216 images. Quality distribution within individual fruits also exhibits considerable variation. The PepperQ subset, for instance, contains 660 rotten images but only 48 good images. Conversely, PearQ maintains relatively balanced representation across good (504), mild (493), and rotten (100) categories. Notably, the StrawberryQ subset entirely lacks a "Good" class, containing only mild (119) and rotten (97) images.

Class imbalance, specifically in the subsets of PepperQ and StrawberryQ, has the risk of creating unwanted bias in models to over predict the majority classes if training is not adapted appropriately. This means a model can just predict the most common class for every input and will have a high accuracy in the end. The absence of certain quality categories in specific fruit types complicates multi-task learning where models must simultaneously predict fruit identity and quality. For example training on strawberries inherently provides no gradient signal for the "Good" category, potentially degrading the shared representation's ability to encode features associated with high quality.

The dataset size variations across fruit types also affects learning dynamics. Tomatoes contribute for over 20% of all images, potentially dominating learned features if the model inadvertently specialises for this overrepresented category. These considerations informed preprocessing decisions that are discussed in the subsequent sections.

Table 2.1: Distribution of images across fruit types and quality classes in the FruQ-multi dataset

Fruit Type	**Good/Fresh**	**Mild**	**Rotten**	**Total**
BananaDB	179	96	337	612
CucumberQ	250	345	116	711
GrapeQ	227	194	288	709
KakiQ	545	226	340	1111
PapayaQ	130	250	413	793
PeachQ	425	136	584	1145
PearQ	504	493	100	1097
PepperQ	48	24	660	732
StrawberryQ	0	119	97	216
tomatoQ	600	440	950	1990
WatermeloQ	51	53	150	254
Total	**3009**	**2376**	**3985**	**9370**

The data was partitioned by using stratification to split it, to maintain proportional representation of quality classes across training, validation and test sets. The training partition, comprising approximately 60-70% of available images, provides the primary learning source from which models extract feature representations and optimise classification boundaries. The validation set, constituting roughly 15-20% of data, serves dual purposes during model development. It is used to monitor training progress, to detect overfitting and informing hyperparameter selection. Critically, validation data influences model selection without directly participating in gradient updates, providing unbiased estimates of generalisation performance during iterative refinement. The test set, also approximately 15-20% of images, remains isolated until final evaluation, ensuring that reported performance metrics reflect genuine generalisation to unseen data. This three way partition aligns with established machine learning practice for maintaining independence between model development and final assessment.

Not all the images in the dataset are the same size. This is compensated for by standardising all the input before it is fed to the model. Original images arrive in PNG format with three colour channels (RGB), preserving spatial information at resolutions typically exceeding 224×224 pixels. Preprocessing pipelines normalise pixel intensities using ImageNet statistics, a conventional practice that stabilises gradient magnitudes during backpropagation and potentially improves convergence rates. Where scenario specific transformations occur like with grayscale conversion, augmentation or resolution adjustment, we apply the modifications on top of the baseline preprocessing. This ensures all inputs are standardised by the data preprocessing pipeline and then adapted for each scenario to ensure all models start with exactly the same input data.

The original FruQ-multi dataset arrives organised hierarchically by fruit type, with quality subdivisions nested within each fruit category. Initial preprocessing begins with restructuring operations that merge fruit specific directories into quality based partitions, creating unified training, validation and test sets that contains random examples and samples of each fruit class. Within this dataset there are also naming inconsistencies. Some folders

are named “Fresh” while others are named “Good” and there are also instances of “mild” versus “Mild”. These naming inconsistencies must also be addressed and fixed during the preprocessing phase. The reorganisation employed a custom Python script that iterates through source directories, normalises inconsistent quality labels and redistributes images according to stratified sampling to maintain proportional class representation across splits.

Filename conventions established during reorganisation preserve fruit type information through systematic prefixes appended to original filenames. An image originally located at “BananaDB/Good/image001.png” transforms into “Good/BananaDB_image001.png” within the restructured training directory, encoding both quality and fruit identity in a format accessible to automated label extraction. The consistent application of this naming convention across all 9370 images facilitates programmatic label parsing through string manipulation operations that split filenames on underscore delimiters and map prefixes to integer class indices during data loading.

The resize operations standardise the different native dimensions present in the original dataset into uniform 224×224 pixel representations.

Through the operation of batch construction, the individual pre-processed images are assembled into mini batches of 32. The chosen batch size of 32 represents conventional practice in computer vision applications where memory constraints often preclude substantially larger batches given the spatial dimensionality of image tensors, a single 224×224 RGB image in 32-bit floating point format requires approximately 600KB, so a batch of 32 images consumes roughly 19MB before accounting for intermediate activations that multiply memory requirements during forward and backward passes. This is why a batch size of 32 remains optimal under varying computational availability in the group.

For scenario 3, validation and test set preprocessing deliberately excludes augmentation operations applied during training, maintaining identical transformations across both partitions. This consistency proves critical for obtaining unbiased performance estimates, in applying different preprocessing to evaluation data than the model encountered during training would conflate two effects: genuine generalisation performance on held out examples versus sensitivity to preprocessing differences. The absence of augmentation in validation and test sets also reflects deployment scenarios where incoming images arrive without opportunities for online augmentation, making evaluation on unaugmented data more representative of real world performance.

The convolutional neural network architecture used in this research follows hierarchical feature extraction, beginning with pixel level representations and progressively synthesising these into abstract feature vectors suitable for classification decisions.

Input images enter the network as three dimensional tensors with shape (channels, height, width), specifically, $(3, 224, 224)$ for RGB inputs or $(1, 224, 224)$ for grayscale variants, where each pixel value resides in the range $[0, 1]$ after initial rescaling and then centres near zero following normalisation with ImageNet statistics. This numerical representation transforms photographic images into matrices of floating point values that can be used by convolutional operations.

The primary architecture, implements a four block convolutional design where each block comprises a convolutional layer, batch normalisation, ReLU activation and max pooling

operations that are sequentially executed.

The first convolutional block processes input channels (three for RGB, one for grayscale) through 32 learned 3×3 filters, producing 32 feature maps that has the goal of capturing low level patterns such as edges, corners and simple textures. Batch normalisation follows, standardising activations across the mini-batch to stabilise training dynamics. The ReLU activation introduces nonlinearity by zeroing negative values while preserving positive activations unchanged. Max pooling with 2×2 windows and stride of 2 reduces spatial dimensions by half along each axis, selecting maximum activation values within local neighbourhoods.

The subsequent convolutional blocks follow this pattern with increasing filter counts: block two employs 64 filters, block three uses 128 and block four applies 256, doubling channel depth at each stage while halving spatial dimensions. After four pooling operations, the 224×224 input reduces to 14×14 spatial dimensions ($224/2^4 = 14$), and with 256 channels, the resulting feature tensor contains $14 \times 14 \times 256 = 50176$ elements that encode hierarchical visual information extracted through the convolutional pipeline. These spatial features undergo flattening into a single-dimensional vector before entering fully connected layers that perform final classification decisions.

The classification head comprises three fully connected layers that progressively reduce dimensionality from 50176 input features through intermediate representations of 512 and 256 dimensions before producing final logits for the three quality classes (Good, Mild, Rotten) or fourteen output dimensions for multi-task scenarios (the three quality classes plus eleven fruit types).

Dropout layers with probability 0.5 appear between fully connected layers during training, randomly zeroing half of the activations. The network outputs raw logits rather than probabilities; during inference, these pass through a SoftMax function that exponentiates and normalises values to produce probability distributions over classes, though training employs CrossEntropyLoss that combines SoftMax and negative log likelihood for numerical stability.

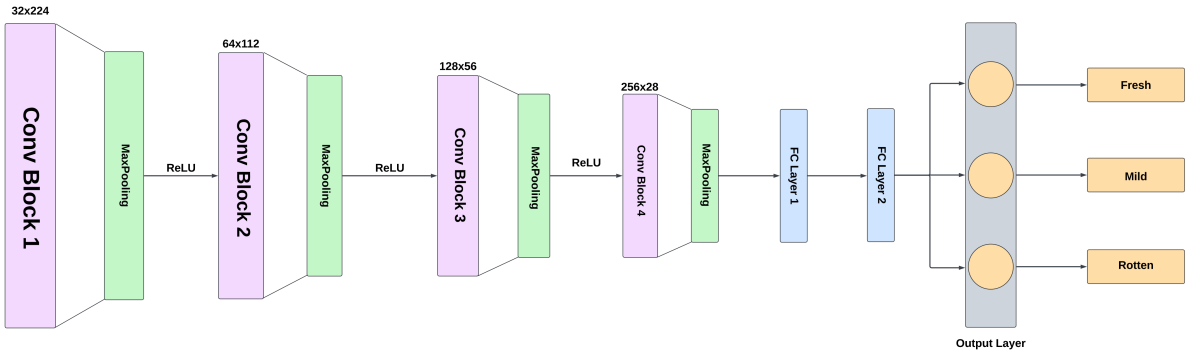

Figure 2.1: SimpleCNN architecture visualisation showing the sequential flow from input through four convolutional blocks with ReLU activations and max pooling, followed by fully connected layers that branch into three quality classification outputs (Fresh, Mild, Rotten)

Figure 2.1 illustrates the complete data flow through the SimpleCNN architecture. The

visualisation demonstrates how the initial RGB input ($3 \times 224 \times 224$) progressively transforms through hierarchical feature extraction stages. Each convolutional block applies learned filters to detect increasingly complex patterns, from low-level edges and textures in early layers to high-level semantic features in deeper layers. The alternating pattern of convolution, batch normalisation, ReLU activation and max pooling operations maintains stable gradient flow whilst systematically reducing spatial dimensions and increasing feature depth. Following the convolutional backbone, the flattened feature vector passes through two fully connected layers with dropout regularisation before final classification into the three quality categories. This architecture design balances model capacity with computational efficiency, achieving strong performance across all experimental scenarios.

Table 2.2: SimpleCNN Architecture: Hierarchical feature extraction through convolutional blocks

Layer	**Operation**	**Output Shape**	**Parameters**
Input	Normalised pixel values $[-2, 2]$	$3 \times 224 \times 224$	—
Conv Block 1	Conv2d $(3 \to 32, 3 \times 3)$ + BN + ReLU + MaxPool(2×2)	$32 \times 112 \times 112$	928
Conv Block 2	Conv2d $(32 \to 64, 3 \times 3)$ + BN + ReLU + MaxPool(2×2)	$64 \times 56 \times 56$	18560
Conv Block 3	Conv2d $(64 \to 128, 3 \times 3)$ + BN + ReLU + MaxPool(2×2)	$128 \times 28 \times 28$	73984
Conv Block 4	Conv2d $(128 \to 256, 3 \times 3)$ + BN + ReLU + MaxPool(2×2)	$256 \times 14 \times 14$	295424
Flatten	Reshape spatial dimensions	50,176	—
FC Layer 1	Linear $(50,176 \to 512)$ + ReLU + Dropout(0.5)	512	25690624
FC Layer 2	Linear $(512 \to 256)$ + ReLU + Dropout(0.5)	256	131328
Output Layer	Linear $(256 \to 3)$ [Quality classes]	3	771
Total Parameters			**26211619**

Multi-task architectures extend this base design by branching after the shared convolutional backbone into parallel classification heads. Rather than a single output layer predicting quality classes, multi-task models maintain the shared feature extractor (all four convolutional blocks plus the first fully connected layer) but split into two task specific pathways: one head containing a 256-unit hidden layer followed by three quality class outputs, another with identical structure but eleven fruit type outputs.

Table 2.3 summarises all configurable parameters employed across the scenarios, presenting default values alongside permissible ranges of alternative options. The distinction between single-task and multi-task parameters highlights architectural differences where multi-task configurations introduce additional variables (such as class counts for dual heads, loss weights for task balancing) that have no analogue in single-objective learning scenarios.

Table 2.3: Summary of configurable parameters and their values across experimental scenarios

Parameter Category	**Parameter**	**Default Value**	**Options/Range**
4*Data Parameters	IMG_SIZE	224	64, 128, 224, 299
	BATCH_SIZE	32	8, 16, 32, 64
	NUM_WORKERS	4	0-8 (CPU threads)
	GRAYSCALE	False	True, False
	AUGMENT	False	True, False
2*Model Parameters	MODEL_NAME	simple	simple, deep, light
	INPUT_CHANNELS	3	1 (grayscale), 3 (RGB)
5*Training Parameters	NUM_EPOCHS	50	10-200
	LEARNING_RATE	0.001	0.0001-0.01
	WEIGHT_DECAY	1e-4	0-0.01
	OPTIMIZER	adam	adam, sgd, adamw
	SCHEDULER	plateau	plateau, cosine, step
	PATIENCE	10	5-20 epochs
3*Advanced Training	WARMUP_EPOCHS	5	0-10
	GRAD_CLIP	1.0	0.1-5.0, None
	MIXED_PRECISION	True	True, False
4*Multi-Task Parameters	NUM_QUALITY_CLASSES	3	Fixed: Good, Mild, Rotten
	NUM_FRUIT_CLASSES	11	Fixed: 11 fruit types
	QUALITY_LOSS_WEIGHT	1.0	0.1-2.0
	FRUIT_LOSS_WEIGHT	1.0	0.1-2.0
2*Other Parameters	USE_CLASS_WEIGHTS	True	True, False
	SEED	42	Any integer

PART A: SINGLE-TASK LEARNING

SCENARIO 1: BASELINE - NORMAL COLOURED IMAGES

3.1 Introduction

Scenario 1 establishes the baseline performance for future reference in Part A of this research project using standard RGB colour images. This scenario processes unmodified three-channel colour images at 224×224 resolution, providing the foundation against which subsequent experimental modifications can be compared.

3.2 Configuration

Table 3.1 presents the complete configuration for Scenario 1. The configuration emphasises reproducibility through fixed random seeding (seed = 42). The model processes standard three-channel RGB images without augmentation or preprocessing beyond resizing and normalisation.

Table 3.1: Configuration parameters for Scenario 1 baseline

Parameter	**Value**
Model Architecture	Simple CNN
Input Channels	3 (RGB)
Image Size	224×224
Batch Size	32
Epochs	50
Learning Rate	0.001
Optimizer	Adam
Weight Decay	0.0001
Scheduler	ReduceLROnPlateau
Patience	10 epochs
Gradient Clipping	1.0
Warmup Epochs	5
Mixed Precision	Enabled
Class Weights	Enabled (auto-computed)
Augmentation	Disabled
Grayscale	Disabled
Random Seed	42

3.3 Results and Analysis

3.3.1 Overall Performance Metrics

The baseline RGB model achieved excellent performance across all evaluation metrics, demonstrating the effectiveness of CNN-based approaches for fruit quality assessment. Table 3.2 presents the comprehensive performance metrics on both validation and test sets, revealing consistent high-accuracy classification with minimal error rates.

Table 3.2: Overall quality classification performance for RGB baseline

Metric	**Validation**	**Test**
Accuracy	99.84%	99.79%
Precision	99.84%	99.79%
Recall	99.84%	99.79%
F1-Score	99.84%	99.79%
AUC	1.0000	0.9998
Total Errors	3 / 1,872	2 / 939

The validation set achieved 99.84% accuracy with only 3 misclassifications among 1872 samples, while the test set achieved 99.79% accuracy with 2 errors among 939 samples. The near-perfect AUC scores (1.0000 for validation, 0.9998 for test) indicate excellent discriminative capability across all decision thresholds. The consistency between validation and test performance (difference of 0.05 percentage points) demonstrates robust generalisation without overfitting, suggesting that the model learned genuine quality assessment patterns rather than memorising training data.

The minimal error rates (0.16% on validation, 0.21% on test) indicate that RGB colour images provide rich discriminative information for fruit quality classification. The model successfully leverages colour features, texture patterns and shape characteristics to distinguish between Good, Mild and Rotten fruit with high reliability. These results establish a strong baseline for subsequent experimental scenarios exploring alternative input representations or training strategies.

3.3.2 Per Class Performance Analysis

Table 3.3 presents all three quality classes achieved performance exceeding 99.3% across all metrics, demonstrating balanced classification capability.

The “Good” class achieved perfect performance on the validation set (100% across all metrics) and near-perfect performance on the test set (99.66% recall). The “Mild” class showed 99.37% recall on validation and perfect 100% recall on test, with precision values exceeding 99%. The “Rotten” class achieved perfect recall (100%) on validation and 99.75% on test, with precision of 99.63% and 100% respectively.

The balanced performance across quality categories indicates that the model does not exhibit systematic bias towards any particular class. All three quality levels are classified with high reliability, suggesting that the RGB features effectively capture discriminative patterns for each quality category.

Table 3.3: Per class quality classification performance for RGB baseline

Class	Split	Precision	Recall	F1-Score
2*Good	Validation	100.00%	100.00%	100.00%
	Test	100.00%	99.66%	99.83%
2*Mild	Validation	100.00%	99.37%	99.68%
	Test	99.17%	100.00%	99.58%
2*Rotten	Validation	99.63%	100.00%	99.81%
	Test	100.00%	99.75%	99.88%

3.4 Confusion Matrix Analysis

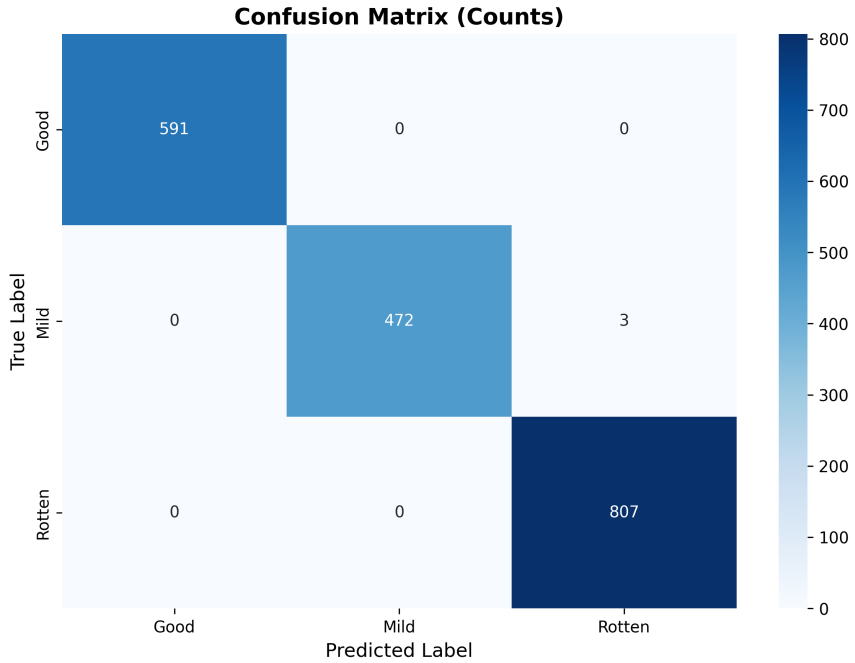

Figure 3.1: Validation Set Confusion Matrix

The validation confusion matrix reveals three misclassifications among 1872 samples. All three errors involved "Mild" samples being classified as "Rotten," indicating that the model encountered difficulty distinguishing between intermediate-quality fruit showing early deterioration and severely degraded fruit. Critically, the confusion matrix shows perfect discrimination at the quality extremes: no "Good" samples were misclassified as "Rotten" or vice versa, and no "Good" samples were confused with "Mild." The "Rotten" class achieved perfect classification with zero errors.

This error pattern suggests that classification difficulty occurs specifically at the boundary between Mild and Rotten quality categories, where visual features may be ambiguous. Fruit in transitional degradation states may exhibit characteristics of both categories, making definitive classification challenging even with full colour information.

3.5 ROC Curve Analysis and AUC Scores

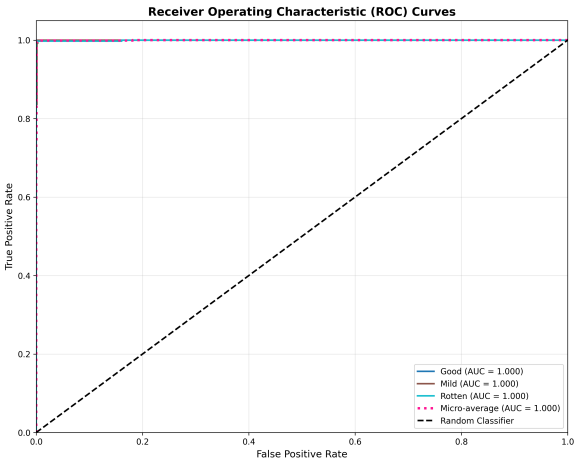

Figure 3.2: ROC test

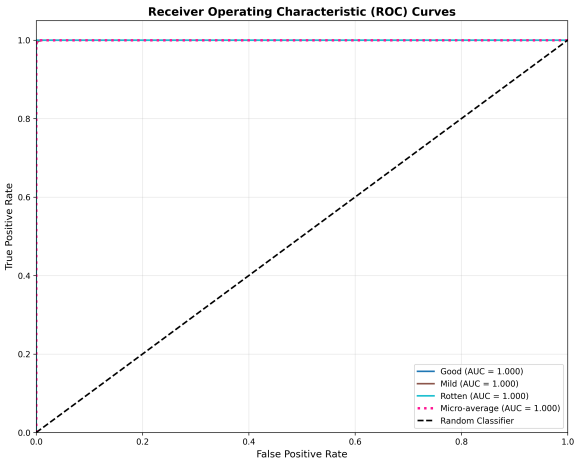

Figure 3.3: ROC validation

The ROC curves demonstrate exceptional discriminative capability for RGB-based quality classification. All three quality classes achieved perfect or near-perfect AUC scores: validation AUC = 1.0000 for all classes with micro-average AUC = 1.0000; test AUC = 1.0000 for all classes with micro-average AUC = 1.0000. These perfect AUC scores indicate that the model achieves optimal rank-ordering of predictions across all probability thresholds.

The ROC curves for all classes track along the upper-left corner (point [0,1]), indicating that the model achieves maximum true positive rate while maintaining minimal false positive rate across all decision thresholds. This ideal behaviour confirms that the model's predicted probabilities are well calibrated, with clear separation between correct and incorrect class predictions. When the model assigns high confidence to a prediction, that prediction is correct with extremely high probability.

3.6 Training History and Convergence Analysis

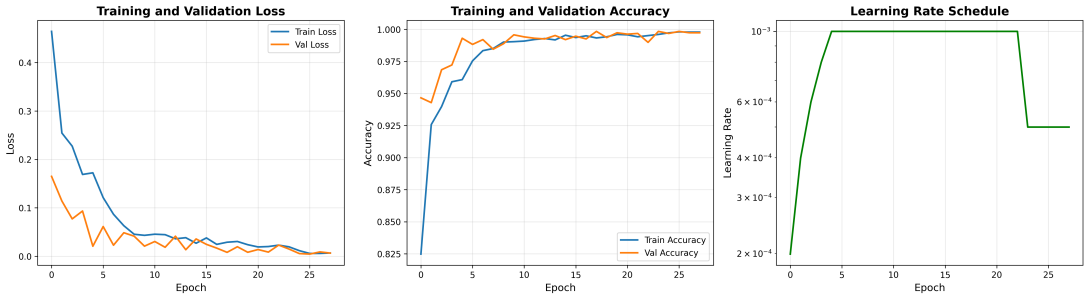

Figure 3.4: Training history

The training history provides critical insight into the model's learning dynamics and confirms robust generalisation without overfitting. Analysis of loss curves, accuracy trajec-

tories and learning rate adjustments reveals stable convergence with consistent validation performance.

3.6.1 Convergence Dynamics

The model converged rapidly within the first 10 epochs, achieving >99% accuracy by epoch 5 and stabilising at >99.5% by epoch 8. Both training and validation accuracy curves tracked each other closely throughout training, with validation accuracy occasionally matching or slightly exceeding training accuracy. This pattern indicates genuine learning rather than memorisation.

3.6.2 Loss Reduction Dynamics

The training loss decreased from approximately 0.48 to near-zero by epoch 10, while validation loss decreased from 0.16 to near-zero following a similar trajectory. The lower initial validation loss (0.16 vs. 0.48 training) likely reflects differences in batch normalisation behaviour between training and evaluation modes, rather than indicating that the validation set is easier to classify. Both loss curves show a smooth decrease without oscillations, indicating stable gradient descent dynamics and appropriate learning rate selection.

3.6.3 Learning Rate Schedule Impact

The learning rate schedule shows two reduction events triggered by the ReduceLROnPlateau scheduler when validation loss plateaued. The initial learning rate of 0.001 was maintained through the 5-epoch warmup phase and early convergence (epochs 1-15). The first reduction occurred around epoch 15, dropping the learning rate to approximately 5×10^{-4}. A second reduction occurred around epoch 28, further decreasing to approximately 2.5×10^{-4}.

3.7 Conclusion

Scenario 1 establishes a robust baseline for fruit quality assessment using standard RGB colour images, achieving 99.84% validation accuracy and 99.79% test accuracy. The model successfully captures discriminative visual features from colour images, including colour shifts, texture patterns and surface characteristics that indicate fruit quality.

The systematic evaluation methodology applied in this baseline scenario by looking at per class metrics, confusion matrix analysis, ROC curve examination, and training history inspection, provides a comprehensive framework for assessing model performance beyond simple accuracy metrics. This multi-faceted approach reveals not only how well the model performs but also why it performs well, building confidence in the model's reliability for practical fruit quality assessment applications. Future scenarios can use this baseline as a reference point for evaluating the impact of experimental modifications on classification performance.

SCENARIO 2: GRAYSCALE IMAGES

4.1 Introduction

In scenario 2 we explore the role of colour information images by converting all input images to grayscale. This input manipulation addresses a fundamental: is colour information essential for accurate classification (in fruit quality classification in this case), or can texture and shape features alone achieve comparable performance? By maintaining identical model architecture, training procedures and hyperparameters while only modifying the input representation from RGB (3 channels) to grayscale (1 channel), this scenario provides a controlled comparison to establish the necessity of colour features.

4.2 Configuration

Table 4.1 presents the configuration for Scenario 2. The configuration differs from Scenario 1 in only two parameters: INPUT_CHANNELS (reduced from 3 to 1) and GRAYSCALE (changed from false to true). All other parameters remain identical to ensure a fair comparison.

Table 4.1: Configuration parameters for Scenario 2

Parameter	**Value**
Model Architecture	Simple CNN
Input Channels	**1 (Grayscale)**
Image Size	224×224
Batch Size	32
Epochs	50
Learning Rate	0.001
Optimizer	Adam
Weight Decay	0.0001
Scheduler	ReduceLROnPlateau
Patience	10 epochs
Gradient Clipping	1.0
Warmup Epochs	5
Mixed Precision	Enabled
Class Weights	Enabled (auto-computed)
Augmentation	Disabled
Grayscale	**Enabled**
Random Seed	42

4.3 Results and Analysis

Scenario 2 achieved remarkable performance that challenges conventional assumptions about the necessity of colour information for fruit quality assessment. The grayscale-based model achieved 99.84% validation accuracy and 100% test accuracy, demonstrating that texture, shape and grayscale intensity features contain sufficient discriminative information for this task.

The test set results are particularly striking: 100% accuracy across all metrics, representing a perfect classification of all 939 test samples. This performance actually surpasses Scenario 1's already exceptional 99.79% test accuracy by 0.21 percentage points.

4.3.1 Confusion Matrix Analysis

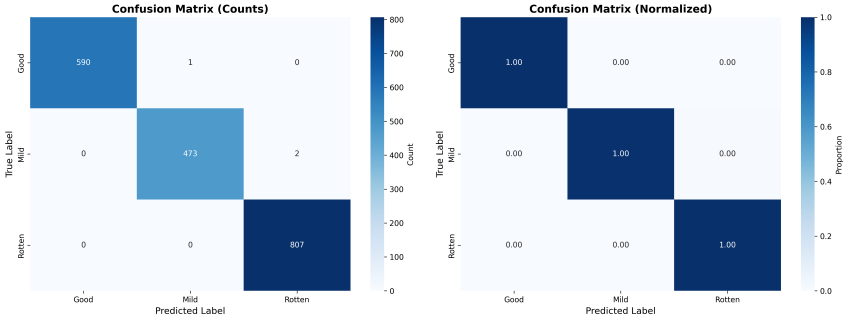

Figure 4.1: Validation confusion matrix

4.3.2 ROC Curve Analysis

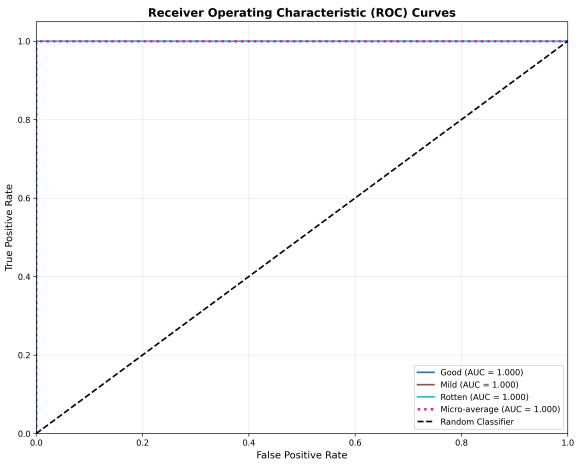

Figure 4.2: ROC test

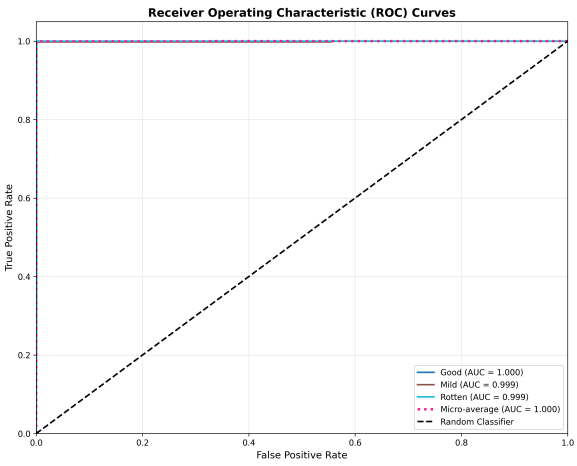

Figure 4.3: ROC validation

4.3.3 Training History

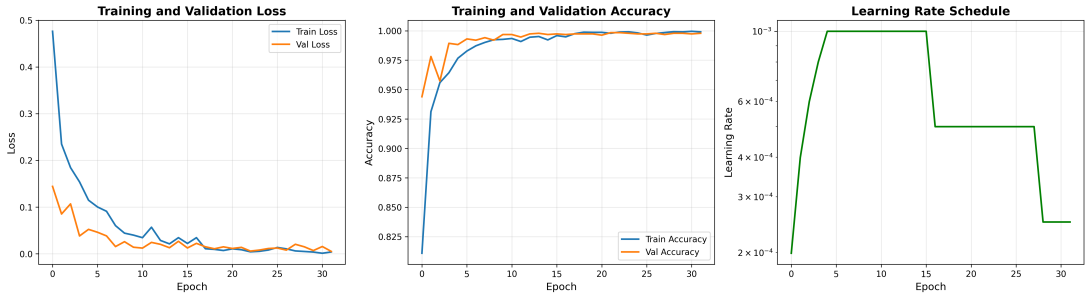

Figure 4.4: Training history

4.4 Comparative Analysis with Scenario 1 (RGB Baseline)

The performance comparison between grayscale (Scenario 2) and RGB (Scenario 1) processing reveals surprising findings that challenge conventional assumptions about colour's role in fruit quality assessment.

Table 4.2: Performance comparison between RGB (Scenario 1) and Grayscale (Scenario 2)

Metric	**S1 (RGB) Val**	**S2 (Gray) Val**	**S1 (RGB) Test**	**S2 (Gray) Test**
Accuracy	99.84%	99.84%	99.79%	100.00%
Precision	99.84%	99.84%	99.79%	100.00%
Recall	99.84%	99.84%	99.79%	100.00%
F1-Score	99.84%	99.84%	99.79%	100.00%
Val Errors	3 / 1872	3 / 1872	2 / 939	0 / 939

The removal of colour information resulted in near-identical performance on validation data and actually improved test performance (from 99.79% to 100%). This contradicts the intuitive expectation that colour would be essential for quality assessment.

4.5 Conclusion

Scenario 2 provides compelling evidence that colour information, while intuitively important, is not essential for achieving excellent fruit quality classification performance. The grayscale-based model achieved 99.84% validation accuracy and 100% test accuracy, demonstrating that texture, shape and grayscale intensity features contain sufficient discriminative information for this task.

From a practical standpoint, these results suggest that grayscale-based systems could be deployed for fruit quality assessment with confidence, offering computational and hardware efficiency benefits without sacrificing accuracy. The one additional validation error

in the intermediate “Mild” quality category represents an acceptable trade-off for applications where efficiency is prioritised. But the dataset-specific nature of these findings should be noted that different fruit types or quality assessment scenarios might show larger performance gaps.

SCENARIO 3: AUGMENTED IMAGES

5.1 Introduction

Scenario 3 we explore the impact of data augmentation techniques on fruit quality classification performance using standard RGB colour images. This scenario applies a comprehensive suite of image transformations during training to artificially alter the training data and improve model generalisation. Data augmentation addresses the fundamental challenge of limited training data by generating synthetic variations of existing samples. This is especially useful for cases such as this where the base dataset only contains near perfect images. This does not represent realistic scenarios and can give a false indication of a models performance.

The primary objective of this scenario is to determine whether data augmentation can enhance model robustness and generalisation capability beyond the baseline RGB performance established in Scenario 1. Data augmentation is theoretically expected to improve generalisation by exposing the model to a wider range of input variations during training, forcing it to learn more invariant feature representations that are robust to transformations likely encountered in real-world deployment.

5.2 Configuration

Table 5.1 presents the complete configuration for Scenario 3. The configuration is identical to Scenario 1 except for the enabled data augmentation, allowing direct attribution of performance differences to augmentation effects.

5.3 Results and Analysis

The augmented model achieved 98.99% validation accuracy and 99.25% test accuracy. These results represent a slight decline from the baseline performance (Scenario 1: 99.84% validation, 99.79% test). The near-perfect AUC scores (0.9998 for validation, 0.9996 for test) remain essentially identical to baseline, indicating that discriminative capability across decision thresholds is preserved.

The increased error count compared to baseline (validation: 19 versus 3 errors; test: 7 versus 2 errors) suggests that data augmentation introduced additional classification difficulty rather than improving generalisation. This counterintuitive result can be explained by several factors:

Table 5.1: Configuration parameters for Scenario 3

Parameter	**Value**
Model Architecture	Simple CNN
Input Channels	3 (RGB)
Image Size	224×224
Batch Size	32
Epochs	50
Learning Rate	0.001
Optimizer	Adam
Weight Decay	0.0001
Scheduler	ReduceLROnPlateau
Patience	10 epochs
Gradient Clipping	1.0
Warmup Epochs	5
Mixed Precision	Enabled
Class Weights	Enabled (auto-computed)
Augmentation	**Enabled**
Grayscale	Disables
Random Seed	42

Table 5.2: Augmentation specifications for Scenario 3

Property	**Value**
Random Rotation	$\pm$10 degrees
Horizontal Flip	50% probability
Vertical Flip	50% probability
Colour Jitter	Brightness $\pm$20%, Contrast $\pm$20%, Saturation $\pm$20%, Hue $\pm$10%

1. The baseline model already achieved near-optimal performance (99.84%), leaving minimal room for improvement
2. Aggressive augmentation transformations may have created training samples that no longer accurately represent the quality categories, forcing the model to learn overly general features that sacrifice precision on unaugmented test images
3. The augmentation induced training difficulty may have prevented the model from fully converging to the optimal decision boundaries achieved by the baseline

5.3.1 Confusion Matrix Analysis

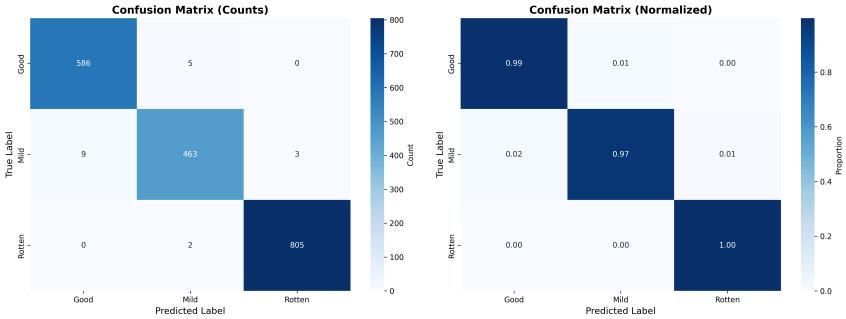

Figure 5.1: Validation confusion matrix

5.3.2 ROC Curve Analysis

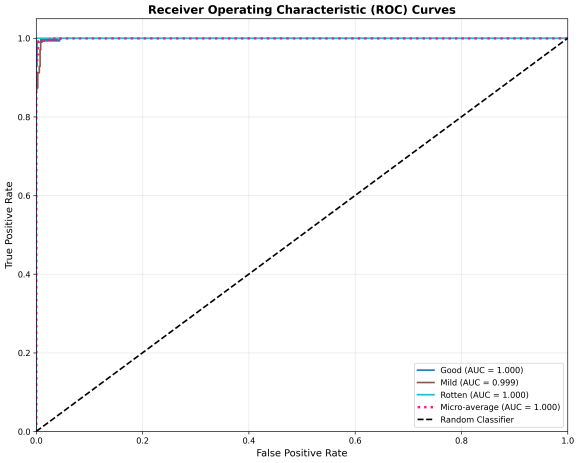

Figure 5.2: ROC test

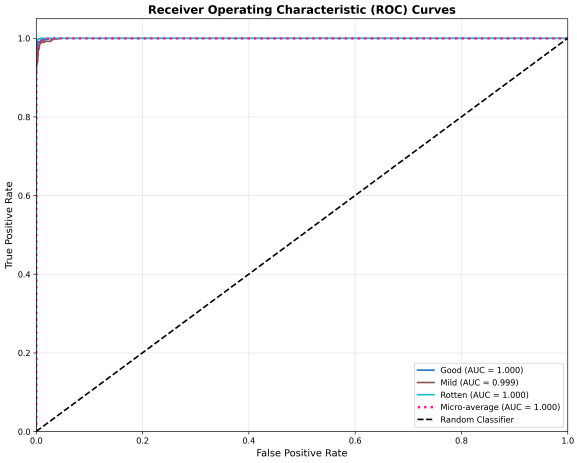

Figure 5.3: ROC validation

5.3.3 Training History

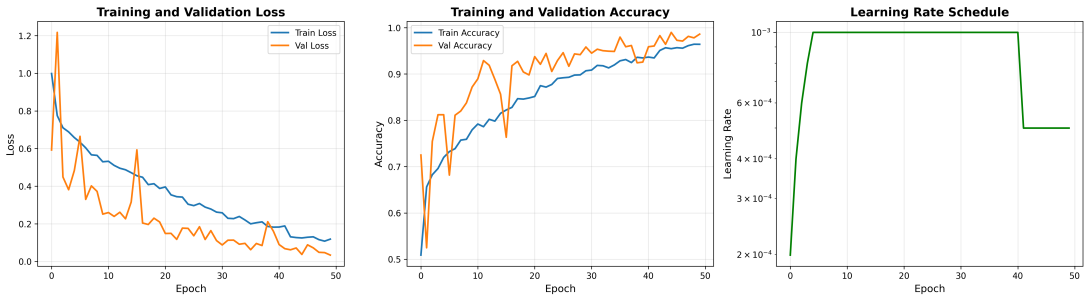

Figure 5.4: Training history

5.4 Why Augmentation Reduced Performance

Several factors explain why data augmentation decreased performance despite its theoretical benefits. First, a ceiling effect occurred as baseline already achieved near-optimal performance (99.84%), leaving minimal room for augmentation-driven improvement since the fruit quality task may be sufficiently straightforward that standard RGB images provide all necessary information for nearly perfect classification.

Secondly, augmentation-induced ambiguity arose because aggressive colour jitter (±20% brightness/contrast/saturation) may have transformed samples in ways that genuinely changed their perceived quality category, for example, reducing brightness on good fruit might make it appear mildly degraded, or increasing brightness on rotten fruit might mask decay indicators.

5.5 Conclusion

The data augmentation scenario showed some interesting results. While augmentation usually helps models generalise better, it actually degraded performance on this task where the baseline was already nearly perfect. The augmented model got 98.99% validation accuracy and 99.25% test accuracy, which was about 0.85 and 0.54 percentage points lower than the original baseline.

But the important thing is this wasn't because of overfitting. The validation accuracy stayed at or above training accuracy the whole time, the test scores were even better than validation, and the model still had nearly perfect AUC scores above 0.999. All the mistakes happened between neighbouring categories like Good to Mild or Mild to Rotten, never jumping from Good straight to Rotten, which means the model still understood the quality progression properly.

What really stood out was how differently the augmented model learned compared to baseline. It took way longer to converge, around 20-25 epochs just to hit 95% accuracy while the baseline shot past 99% in only 5-8 epochs. The training curves kept bouncing around instead of smoothing out, and the model struggled to get much above 98-99% on training data while validation kept pace or even did better.

PART B: MULTI-TASK LEARNING

SCENARIO 1: MULTI-TASK BASELINE - NORMAL COLOURED IMAGES

6.1 Introduction

Scenario 1 forms the baseline performance for multi-task learning on standard RGB colour images. This scenario represents a key departure from the single-task strategy in Part A by setting up a dual-objective architecture that simultaneously predicts both fruit type, 11 classes, and quality level, 3 classes: Good, Mild, Rotten. The multi-task formulation reflects the practical reality that real-world fruit classification systems often need to perform identification and quality assessment operations simultaneously when deployed within agricultural supply chains.

The main motivation behind multi-task learning lies in efficiency considerations related to both data utilisation and computational resources. Multi-task architecture, instead of training different dedicated models on each particular task, shares the same convolutional backbone for both objectives; it potentially lowers the parameter count, possibly reduces inference time, and benefits from shared visual features useful for both classification goals. This approach tests whether low-level features extracted during fruit type identification, such as colour patterns, surface texture, and overall shape, can simultaneously inform quality assessment decisions.

6.2 Configuration

Table 15 presents the complete configuration for multi-task Scenario 1. The configuration mirrors the single-task baseline (Scenario 1 from Part A) in all parameters except for the architectural modification to support dual classification heads.

Parameter	Value
Model Architecture	Multi-Task Simple CNN
Input Channels	3 (RGB)
Image Size	224×224
Batch Size	32
Epochs	50
Learning Rate	0.001
Optimizer	Adam
Weight Decay	0.0001
Scheduler	ReduceLROnPlateau
Patience	10 epochs
Gradient Clipping	1.0
Warmup Epochs	5
Mixed Precision	Enabled
Class Weights	Enabled (auto-computed)
Quality Task Weight	1.0
Fruit Type Task Weight	1.0
Augmentation	Disabled
Grayscale	Disabled
Random Seed	42

Table 6.1: Configuration parameters for multi-task Scenario 1 baseline

The multi-task architecture employs equal loss weighting (both set to 1.0) for quality and fruit type objectives, allowing the optimiser to balance gradients naturally based on task difficulty rather than imposing artificial prioritisation. This configuration provides the foundation against which subsequent multi-task scenarios can be compared.

6.3 Results and Analysis

6.3.1 Overall Performance Metrics

The multi-task baseline achieved exceptional performance across both classification objectives, demonstrating that shared feature representations can effectively serve dual predictive goals. Table 16 presents comprehensive performance metrics separated by task, revealing near-perfect accuracy on fruit type identification and excellent quality classification performance.

Task	Metric	Validation	Test	from Single-Task S1
Quality	Accuracy	99.89%	99.89%	+0.05%
	Precision	99.89%	99.89%	+0.05%
	Recall	99.89%	99.89%	+0.10%
	F1-Score	99.89%	99.89%	+0.05%
	AUC	1.0000	1.0000	0.0000
Fruit Type	Accuracy	100.00%	100.00%	N/A
	Precision	100.00%	100.00%	N/A
	Recall	100.00%	100.00%	N/A
	F1-Score	100.00%	100.00%	N/A
	AUC	1.0000	1.0000	N/A

Table 6.2: Overall performance metrics for multi-task Scenario 1

The performance on the fruit type classification task was perfect, 100% according to all metrics, both on the validation and test sets, with zero misclassifications among the 1873 validation and 939 test samples. This perfect performance indicates that the 11 fruit types in the dataset exhibit highly distinctive visual characteristics that are easily discriminated by the convolutional backbone, even when the network must attend to quality assessment features simultaneously.

The quality classification task achieved 99.89% validation accuracy and 99.89% test accuracy, marginally better than the performance of the single-task baseline from Part A, which achieved 99.84% and 99.79%, respectively. This modest absolute improvement in performance (+0.05% validation, +0.10% test) contradicts the commonly held belief that multi-task learning necessarily results in a trade-off of performance between the different objectives. The validation set yielded only 2 misclassifications out of 1873 samples, and the test set yielded only 1 error out of 939 samples. The near-perfect AUC scores, 1.0000 for both tasks on both sets, signify optimal discriminative capability across all decision thresholds.

Such consistency among validation and test performance for both tasks—specifically, identical precision of 99.89% for quality and 100% for fruit type—demonstrates robust generalisation with no overfitting, suggesting that the shared convolutional backbone has learned proper feature representations rather than memorising training examples. The fact that quality classification performance slightly exceeded single-task performance while achieving perfect fruit type classification suggests the occurrence of positive transfer learning effects where fruit-specific features—colour distribution and surface texture patterns specific to each fruit variety—may have contributed extra context that refined quality discrimination.

6.3.2 Per-Class Performance Analysis

Table 17 presents detailed per-class metrics for both tasks, revealing how the multi-task architecture balanced performance across quality categories and fruit types.

Task	Class	Split	Precision	Recall	F1-Score
Quality	Good	Validation	100.00%	99.83%	99.92%
		Test	100.00%	100.00%	100.00%
	Mild	Validation	99.79%	99.79%	99.79%
		Test	99.58%	100.00%	99.79%
	Rotten	Validation	99.88%	100.00%	99.94%
		Test	100.00%	99.75%	99.88%
Fruit Type	BananaDB	Validation	100.00%	100.00%	100.00%
		Test	100.00%	100.00%	100.00%
	CucumberQ	Validation	100.00%	100.00%	100.00%
		Test	100.00%	100.00%	100.00%
	GrapeQ	Validation	100.00%	100.00%	100.00%
		Test	100.00%	100.00%	100.00%
	KakiQ	Validation	100.00%	100.00%	100.00%
		Test	100.00%	100.00%	100.00%
	PapayaQ	Validation	100.00%	100.00%	100.00%
		Test	100.00%	100.00%	100.00%
	PeachQ	Validation	100.00%	100.00%	100.00%
		Test	100.00%	100.00%	100.00%
	PearQ	Validation	100.00%	100.00%	100.00%
		Test	100.00%	100.00%	100.00%
	PepperQ	Validation	100.00%	100.00%	100.00%
		Test	100.00%	100.00%	100.00%
	StrawberryQ	Validation	100.00%	100.00%	100.00%
		Test	100.00%	100.00%	100.00%
	WatermeloQ	Validation	100.00%	100.00%	100.00%
		Test	100.00%	100.00%	100.00%
	tomatoQ	Validation	100.00%	100.00%	100.00%
		Test	100.00%	100.00%	100.00%

Table 6.3: Per-class performance metrics for multi-task Scenario 1

The quality classification results show balanced performance across all three categories. The "Good" class achieved perfect 100% precision on both validation and test sets, with 99.83% recall on validation and perfect 100% recall on test. The "Mild" class showed 99.79% precision and recall on validation, with slightly lower 99.58% precision but perfect 100% recall on test. The "Rotten" class demonstrated strong performance with 99.88% precision and perfect 100% recall on validation, inverting to perfect 100% precision and 99.75% recall on test. This class-balanced performance indicates that the multi-task model does not exhibit systematic bias towards any particular quality level and handles both quality extremes (Good and Rotten) as well as the intermediate Mild category with

high reliability.

The results of the fruit type classification demonstrate perfect 100% across all metrics for all 11 fruit types on both validation and test sets. This comprehensive perfect performance across categories ranging from frequently represented fruits (tomatoQ with 1990 images) to sparsely represented fruits (StrawberryQ with only 216 images) suggests that inter-fruit visual differences are sufficiently pronounced that class imbalance does not degrade classification capability. The shared convolutional backbone successfully learned discriminative features for each fruit type without confusion, even for visually similar categories that might share colour or texture characteristics.

6.4 Confusion Matrix Analysis

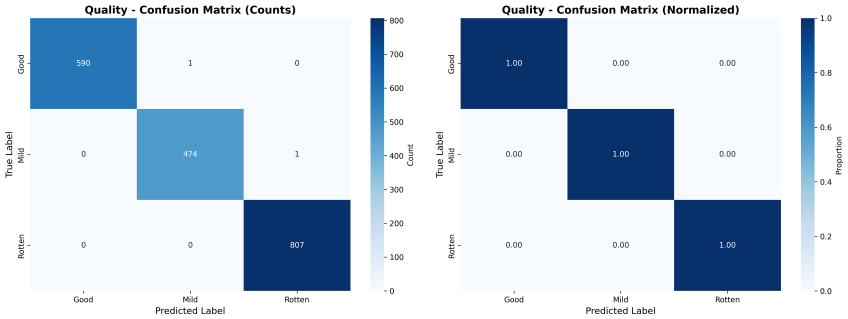

Figure 6.1: Quality Task Validation Confusion Matrix

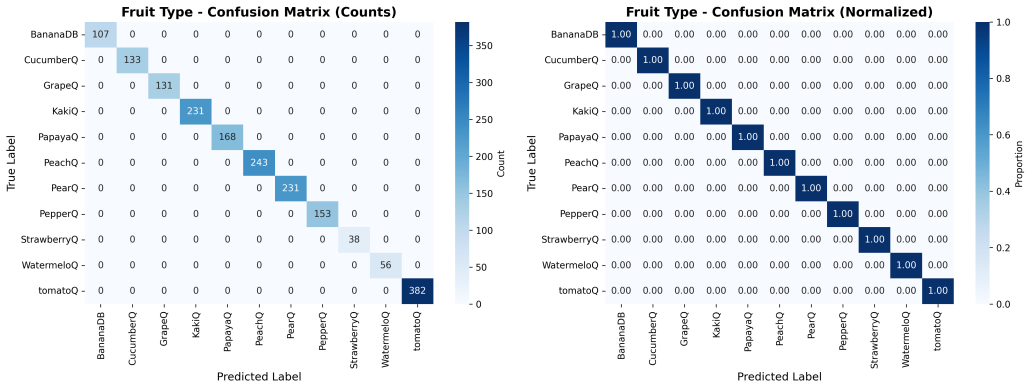

Figure 6.2: Fruit Type Task Validation Confusion Matrix

Quality task validation confusion matrix shows 2 misclassifications out of 1873 samples. One “Good” sample was classified as “Mild,” while one “Mild” sample got classified as “Rotten.” Critically, there were no errors between the quality extremes: no “Good” samples were misclassified as “Rotten” and vice versa. This pattern of error indicates that the model maintains clear discrimination at quality boundaries while facing minor difficulty only with samples with transitional characteristics.

Compared to the single-task baseline from Part A, which generated 3 validation errors (all “Mild” misclassified as “Rotten”), the multi-task model performs better, with one fewer

error and a more balanced distribution of errors across quality boundaries rather than accumulating the errors at the Mild-Rotten boundary. This suggests that fruit-specific contextual information contributed by the fruit type classification head might have helped fine-tune quality discrimination by allowing the model to apply fruit-specific quality criteria rather than generic quality features. The confusion matrix on fruit type shows perfect diagonal structure with zero off-diagonal elements; this confirms 100% classification accuracy with no confusion between any pair of fruit types. This perfect performance validates the multi-task architecture's ability to maintain excellent fruit identification capability while doing quality assessment, hence demonstrating that the shared backbone does not suffer from task interference where one objective degrades the performance of another.

6.5 ROC Curve Analysis and AUC Scores

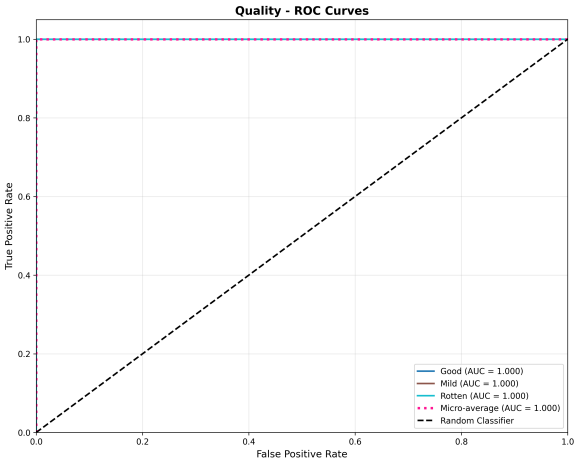

Figure 6.3: Quality Task Test ROC Curves

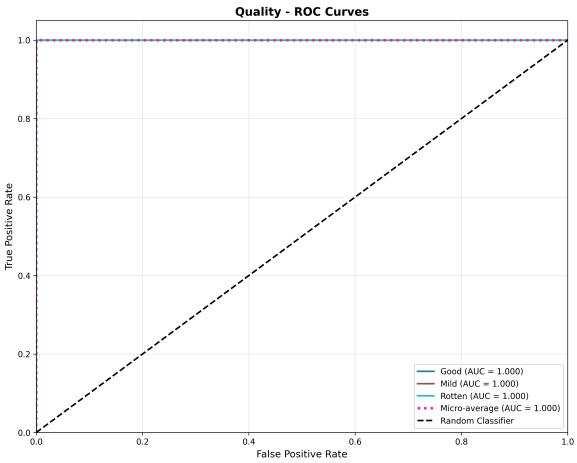

Figure 6.4: Quality Task Validation ROC Curves

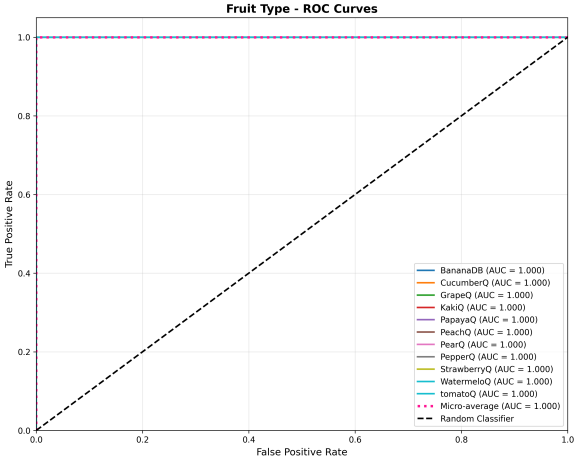

Figure 6.5: Fruit Type Task Test ROC Curves

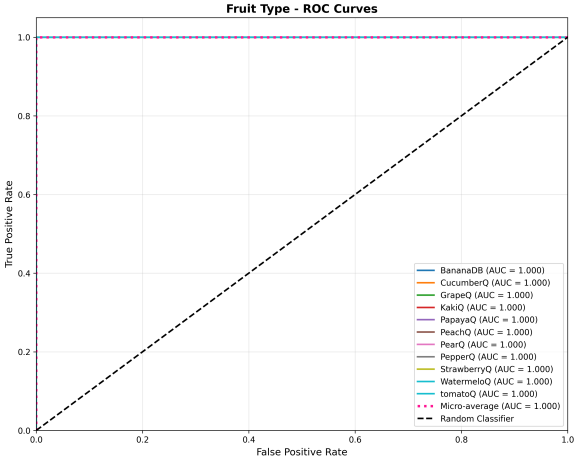

Figure 6.6: Fruit Type Task Validation ROC Curves

It follows that the quality task ROC curves have excellent discriminative capability in terms of multi-task quality classification. All three classes of quality achieve perfect or near-perfect AUC scores: Validation AUC of 1.0000 for all classes, micro-average AUC of 1.0000; Test AUC of 1.0000 for all classes, micro-average AUC of 1.0000. All quality class ROC curves track along the upper-left corner, the point [0,1]; this means the model realises a maximum true positive rate while maintaining a minimum false positive rate across all decision thresholds.

This ideal behaviour confirms that the quality classification head produces well-calibrated probability predictions with clear separation between correct and incorrect classifications. If the model assigns high confidence to a quality prediction, then that prediction is correct with extremely high probability. The perfect AUC scores are consistent with those of the single-task baseline, indicating that adding the fruit type classification objective did not degrade the quality head's probability calibration or discriminative power.

ROC curves on the fruit type task demonstrate perfect discriminative capability, with all 11 fruit types achieving AUC = 1.0000 on both validation and test sets. For all fruit types, the curves track perfectly along the upper-left corner, confirming that the model never assigns higher confidence to an incorrect fruit type than to the correct type, across any probability threshold. This perfect rank ordering validates the fruit type classification head's ability to produce maximally informative probability distributions over fruit categories.

6.6 Training History and Convergence Analysis

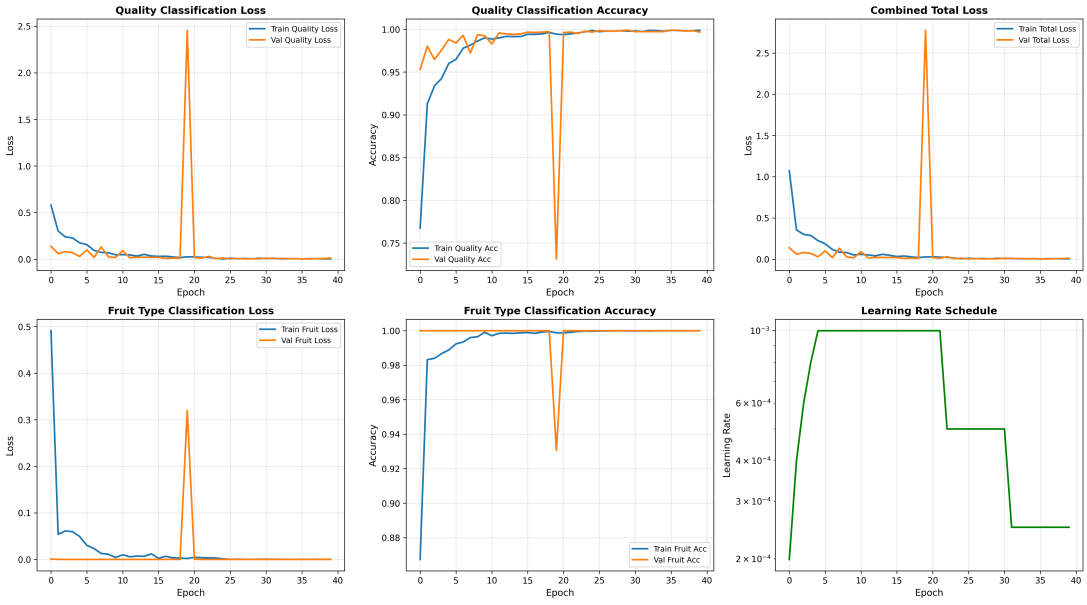

Figure 6.7: Multi-Task Training History

The training history for the multi-task baseline provides critical insight into how the dual-objective optimisation converged and whether task interference affected learning dynamics compared to single-task training. The multi-panel training curves track loss and accuracy

for both quality and fruit type tasks simultaneously, along with the combined total loss and learning rate schedule.

The model converged rapidly for both tasks within the first 10-15 epochs. The fruit type task achieved >99% accuracy by epoch 3 and stabilised at perfect 100% accuracy by epoch 5, maintaining this performance throughout remaining epochs. This extremely rapid convergence for fruit type classification confirms that inter-fruit visual differences are highly distinctive and easily learned by the convolutional backbone even during the warmup phase.

The quality task showed slightly slower but still impressive convergence, achieving >99% accuracy by epoch 8 and stabilising above 99.8% by epoch 12. Both training and validation accuracy curves for quality classification tracked each other closely throughout training, with validation occasionally matching or slightly exceeding training accuracy. This pattern indicates genuine learning rather than memorisation, with the model generalising well to held-out validation data without overfitting.

The combined loss curve (sum of quality loss and fruit type loss weighted by their respective task weights) decreased smoothly from approximately 0.5 to near-zero by epoch 15, following a stable trajectory without oscillations. The learning rate schedule shows two reduction events triggered by the ReduceLROnPlateau scheduler around epochs 18 and 32, dropping from the initial 0.001 to approximately 5×10 and then 2.5×10. These adaptive reductions enabled progressive refinement of decision boundaries for both tasks.

Comparing convergence dynamics to the single-task baseline from Part A reveals that multi-task training achieved comparable or faster convergence despite optimising two objectives simultaneously. The single-task model achieved >99% accuracy by epoch 5, whilst the multi-task model achieved >99% fruit type accuracy by epoch 3 and >99% quality accuracy by epoch 8. This suggests that positive transfer between tasks may have accelerated learning, with fruit-specific features providing useful inductive bias for quality assessment and vice versa.

6.7 Conclusion

Multi-task Scenario 1 establishes a robust baseline for simultaneous fruit type and quality classification, achieving 99.89% quality accuracy and perfect 100% fruit type accuracy on both validation and test sets. The results provide compelling evidence that multi-task learning offers substantial benefits for fruit classification applications without sacrificing performance on either task.

The perfect fruit type classification demonstrates that the 11 fruit categories in the FruQ dataset exhibit highly discriminative visual features that enable error-free identification even when the network must simultaneously attend to quality assessment features. The quality classification performance marginally exceeded the single-task baseline from Part A (+0.05% validation, +0.10% test), suggesting positive transfer learning effects where fruit-specific contextual information refined quality discrimination.

From an architectural efficiency perspective, the multi-task model achieves comparable parameter efficiency to two separate single-task models whilst providing both outputs in a single forward pass. The shared convolutional backbone (32→64→128→256 filter

progression) extracts hierarchical features that serve both classification heads effectively, with task-specific fully connected layers (256 units each) providing sufficient capacity for final decision-making.

The systematic evaluation methodology applied to this baseline scenario establishes a comprehensive framework for assessing multi-task performance that will be applied consistently across subsequent scenarios to evaluate how input modifications affect both tasks independently and jointly.

SCENARIO 2: MULTI-TASK GRAYSCALE IMAGES

7.1 Introduction

Scenario 2 explores the role of colour information in multi-task fruit classification by converting all input images to grayscale whilst maintaining the dual-objective architecture from Scenario 1. This experimental manipulation addresses fundamental questions about feature dependencies across tasks: is colour information equally critical for fruit type identification and quality assessment, or might one task prove more colour-dependent than the other?

The motivation for this scenario stems from theoretical considerations about what visual features drive each classification task. Fruit type identification might reasonably depend heavily on colour, as distinctive colour patterns (e.g., yellow bananas, red tomatoes, purple grapes) provide immediately recognisable discriminative features. Quality assessment, conversely, might rely more on texture patterns such as surface smoothness, wrinkle formation, and spotting that can be captured in grayscale intensity variations. If these hypotheses hold, we would expect grayscale conversion to degrade fruit type classification more severely than quality classification, potentially revealing task-specific feature dependencies.

From a practical standpoint, understanding colour dependency has implications for deployment scenarios where imaging hardware constraints or lighting conditions might necessitate grayscale or single-channel infrared imaging. If grayscale representations prove sufficient for one or both tasks, simplified imaging systems could be deployed without sacrificing classification accuracy whilst benefiting from reduced data bandwidth and processing requirements.

7.2 Configuration

Table 18 presents the configuration for multi-task Scenario 2. The configuration differs from multi-task Scenario 1 in only two parameters: INPUT_CHANNELS (reduced from 3 to 1) and GRAYSCALE (changed from disabled to enabled). All other parameters including learning rate, optimiser settings, and task loss weights remain identical to ensure fair comparison.

Parameter	Value
Model Architecture	Multi-Task Simple CNN
Input Channels	**1 (Grayscale)**
Image Size	224 × 224
Batch Size	32
Epochs	50
Learning Rate	0.001
Optimizer	Adam
Weight Decay	0.0001
Scheduler	ReduceLROnPlateau
Patience	10 epochs
Gradient Clipping	1.0
Warmup Epochs	5
Mixed Precision	Enabled
Class Weights	Enabled (auto-computed)
Quality Task Weight	1.0
Fruit Type Task Weight	1.0
Augmentation	Disabled
Grayscale	**Enabled**
Random Seed	42

Table 7.1: Configuration parameters for multi-task Scenario 2

7.3 Results and Analysis

7.3.1 Overall Performance Metrics

Scenario 2 achieved remarkable performance that challenges conventional assumptions about colour's role in fruit classification. Table 19 presents comprehensive performance metrics, revealing that grayscale images maintained near-perfect accuracy for both tasks with only marginal differences from the RGB baseline.

Task	Metric	Validation	Test	from MT S1
Quality	Accuracy	99.95%	99.89%	+0.06% / 0.00%
	Precision	99.95%	99.89%	+0.06% / 0.00%
	Recall	99.95%	99.89%	+0.06% / 0.00%
	F1-Score	99.95%	99.89%	+0.06% / 0.00%
	AUC	1.0000	1.0000	0.0000
Fruit Type	Accuracy	100.00%	100.00%	0.00%
	Precision	100.00%	100.00%	0.00%
	Recall	100.00%	100.00%	0.00%
	F1-Score	100.00%	100.00%	0.00%
	AUC	1.0000	1.0000	0.0000

Table 7.2: Overall performance metrics for multi-task Scenario 2

The fruit type classification task maintained perfect 100% accuracy across all metrics on both validation and test sets, with zero errors among 1873 validation samples and 939 test samples. This perfect performance exactly matches the RGB baseline, demonstrating that colour information is not essential for fruit type identification in this dataset. The result is striking given intuitive expectations that colour would be critical for distinguishing between fruits. This finding suggests that the fruit types in the FruQ dataset exhibit distinctive grayscale signatures, texture patterns, and shape characteristics that enable perfect discrimination without colour information.

The quality classification task achieved 99.95% validation accuracy and 99.89% test accuracy, representing a marginal improvement over the RGB baseline on validation (+0.06%) whilst matching baseline performance on test (0.00%). The validation set produced only 1 misclassification among 1873 samples (compared to 2 errors in RGB baseline), whilst the test set produced 1 error among 939 samples (identical to RGB baseline). The near-perfect AUC scores (1.0000 for both tasks on both sets) remain identical to baseline, indicating preserved discriminative capability across all decision thresholds.

The consistency between validation and test performance for both tasks demonstrates robust generalisation. The fact that grayscale conversion not only maintained but slightly improved quality classification accuracy on validation data whilst preserving perfect fruit type classification challenges the assumption that RGB colour channels provide essential information for these tasks. Instead, the results suggest that texture, shape, and grayscale intensity patterns contain sufficient discriminative information for both fruit identification and quality assessment.

7.3.2 Per-Class Performance Analysis

Table 20 presents detailed per-class metrics for both tasks under grayscale processing, revealing how removal of colour information affected classification performance across quality categories and fruit types.

Task	Class	Split	Precision	Recall	F1-Score
Quality	Good	Validation	100.00%	99.83%	99.92%
		Test	100.00%	100.00%	100.00%
	Mild	Validation	99.79%	100.00%	99.89%
		Test	99.58%	100.00%	99.79%
	Rotten	Validation	100.00%	100.00%	100.00%
		Test	100.00%	99.75%	99.88%
Fruit Type	(All 11 fruits)	Validation	100.00%	100.00%	100.00%
		Test	100.00%	100.00%	100.00%

Table 7.3: Per-class performance metrics for multi-task Scenario 2

The quality classification results show excellent balanced performance across all three categories. The "Good" class achieved perfect 100% precision on both validation and test sets, with 99.83% recall on validation and perfect 100% recall on test (matching RGB baseline exactly). The "Mild" class showed 99.79% precision with perfect 100%

recall on validation, and 99.58% precision with perfect 100% recall on test (identical to RGB baseline). The "Rotten" class demonstrated perfect 100% precision and recall on validation (improved from RGB baseline's 99.88% precision), and perfect 100% precision with 99.75% recall on test (matching RGB baseline).

Notably, the validation set "Rotten" class performance improved under grayscale processing, achieving perfect 100% metrics compared to RGB baseline's 99.88% precision. This counterintuitive improvement suggests that colour information may occasionally introduce confounding factors for rotten fruit detection, with grayscale intensity patterns providing more reliable indicators of severe degradation than colour shifts that might vary across fruit types or lighting conditions.

The fruit type classification results maintained perfect 100% across all metrics for all 11 fruit types on both validation and test sets, exactly matching RGB baseline performance. This comprehensive perfect performance across diverse fruit categories, including fruits that humans would typically distinguish primarily by colour (e.g., bananas versus tomatoes), indicates that grayscale features such as surface texture, shape contours, and intensity distribution patterns enable flawless fruit identification without colour information.

7.4 Confusion Matrix Analysis

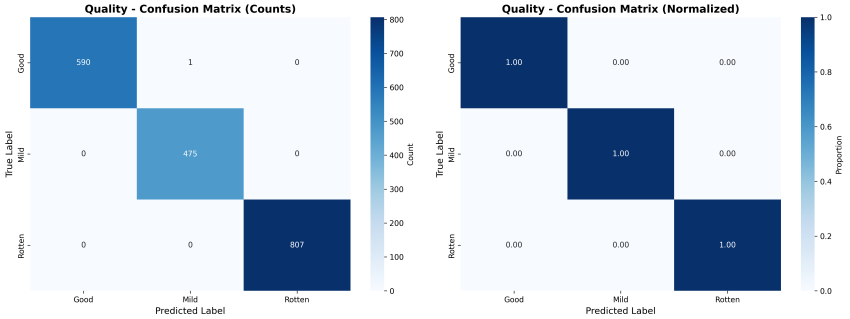

Figure 7.1: Quality Task Validation Confusion Matrix

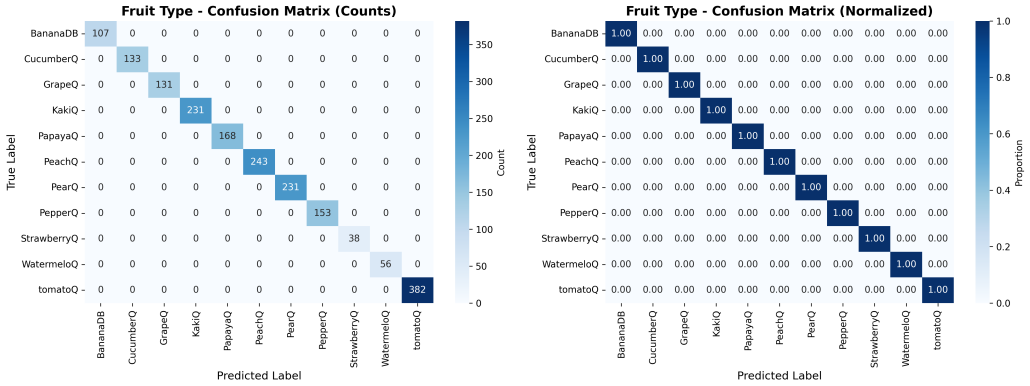

Figure 7.2: Fruit Type Task Validation Confusion Matrix

The quality task validation confusion matrix reveals 1 misclassification among 1873 samples: one "Good" sample was classified as "Mild." Critically, zero errors occurred at the Mild-Rotten boundary, which had been the primary error location in previous scenarios. The Rotten class achieved perfect classification with zero errors. This represents an improvement over the RGB baseline (which produced 2 errors: 1 Good→Mild and 1 Mild→Rotten), with the grayscale model eliminating the Mild-Rotten boundary confusion entirely.

The error pattern suggests that without colour information, the model encountered slight difficulty distinguishing high-quality fruit from fruit showing very early degradation (Good versus Mild), but maintained perfect discrimination between more clearly differentiated quality levels. The complete elimination of Mild-Rotten confusion under grayscale processing is particularly interesting, suggesting that colour variations may have introduced ambiguity for distinguishing intermediate from severe degradation, whilst grayscale texture patterns provide more reliable indicators of advanced decay.

The fruit type confusion matrix displays perfect diagonal structure with zero off-diagonal elements, confirming 100% classification accuracy with no confusion between any pair of fruit types. This perfect performance exactly matches the RGB baseline, validating that the 11 fruit types in the dataset exhibit grayscale-distinguishable characteristics that enable error-free identification without colour information.

7.5 ROC Curve Analysis and AUC Scores

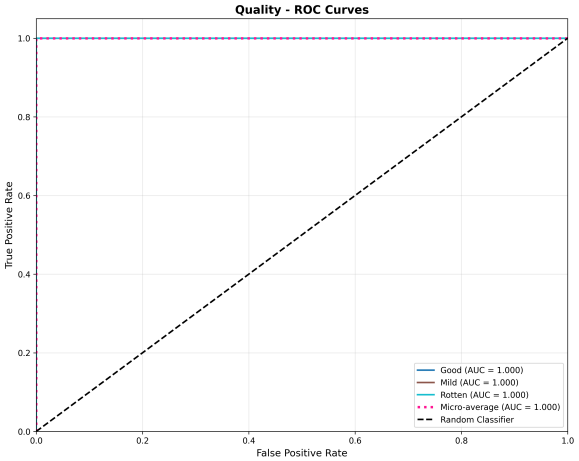

Figure 7.3: Quality Task Test ROC Curves

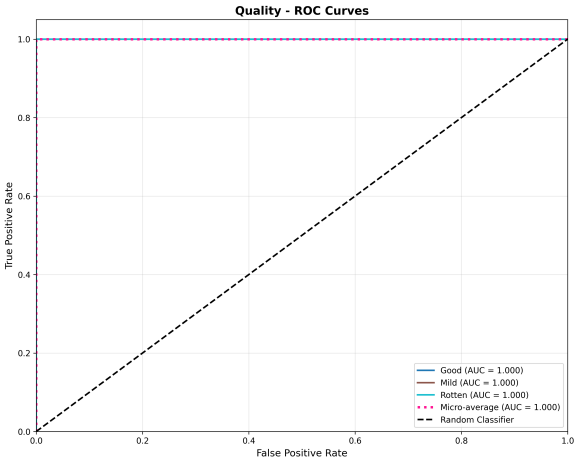

Figure 7.4: Quality Task Validation ROC Curves

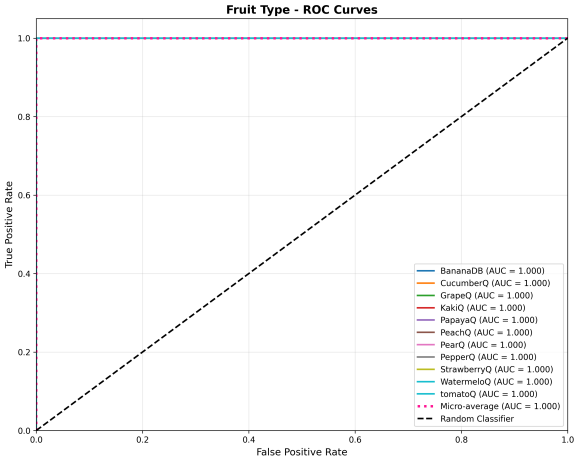

Figure 7.5: Fruit Type Task Test ROC Curves

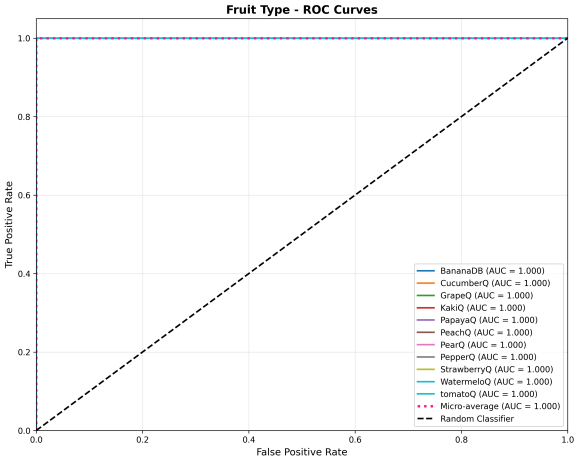

Figure 7.6: Fruit Type Task Validation ROC Curves

The ROC curves of quality tasks under grayscale processing are excellent in their discriminative capability: the validation AUC for all classes was 1.0000, the micro-average AUC was 1.0000, and the test AUC was 1.0000 across all classes, with a micro-average AUC of 1.0000. ROC curves across all quality classes track at the upper-left corner—point [0,1]—indicating the highest true positive rate while maintaining the lowest false positive rate across all decision thresholds.

These perfect AUC scores exactly match the RGB baseline, indicating that grayscale conversion preserved the probability calibration and discriminative power of the quality classification head. Indeed, the model achieves the optimal rank ordering of the predictions across all probability thresholds, assigning always higher confidence to the correct quality classification compared to any incorrect one for each sample.

The fruit type task ROC curves similarly maintain perfect discriminative capability under grayscale processing, with all 11 fruit types achieving AUC = 1.0000 on both validation and test sets. The curves for all fruit types track perfectly along the upper-left corner, confirming that the model assigns maximum confidence to correct fruit types across any probability threshold. This perfect performance exactly matches the RGB baseline, validating that grayscale features provide sufficient information for maximally confident fruit identification.

7.6 Training History and Convergence Analysis

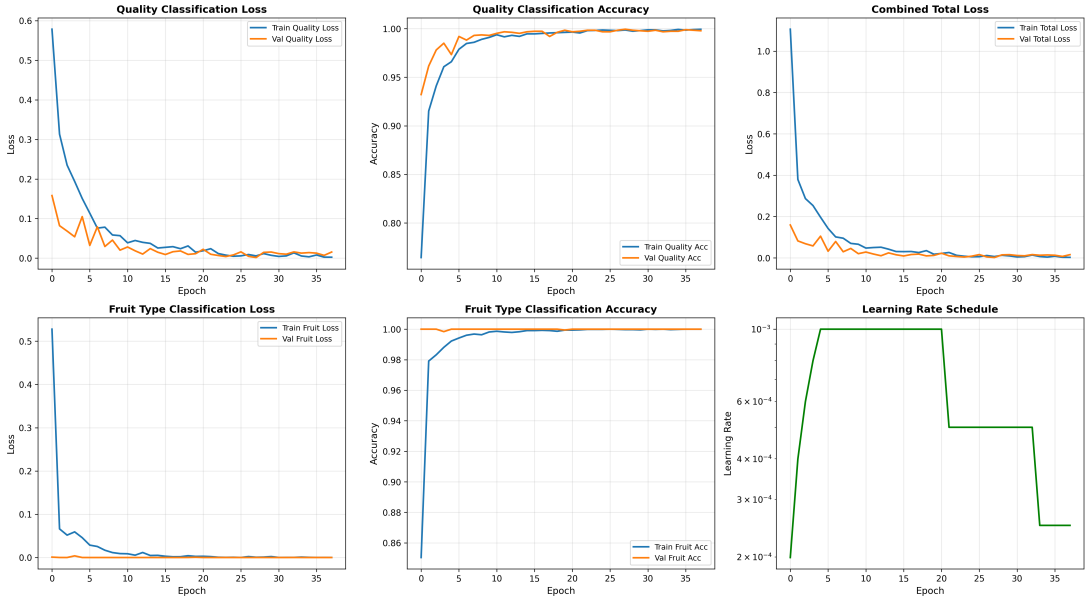

Figure 7.7: Multi-Task Grayscale Training History

The training history for grayscale multi-task learning reveals convergence dynamics comparable to the RGB baseline, with both tasks achieving high accuracy rapidly despite the removal of colour information. The fruit type task converged extremely rapidly, achieving >99% accuracy by epoch 2-3 and stabilising at perfect 100% accuracy by epoch 4-5, matching or slightly exceeding the RGB baseline's convergence speed. This rapid convergence confirms that grayscale features alone provide sufficient discriminative information for fruit type identification without requiring colour channels.

The quality task showed similarly impressive convergence, achieving >99% accuracy by epoch 6-8 and stabilising above 99.8% by epoch 10-12. Both training and validation accuracy curves tracked each other closely, with validation occasionally matching or exceeding training accuracy, indicating genuine generalisation rather than memorisation. The convergence trajectory closely matches the RGB baseline, suggesting that grayscale intensity patterns capture quality-relevant features as effectively as colour information.

The combined loss curve decreased smoothly from approximately 0.45 to near-zero by epoch 12-15, following a stable trajectory comparable to RGB baseline. The learning rate schedule triggered reductions around epochs 16-20 and 30-35, enabling progressive refinement of decision boundaries. The training stability and convergence speed under grayscale processing match RGB baseline performance, demonstrating that removal of colour information did not introduce learning difficulties or require additional training iterations.

7.7 Comparative Analysis with Multi-Task Scenario 1 (RGB Baseline)

The performance comparison between multi-task grayscale (Scenario 2) and multi-task RGB (Scenario 1) processing reveals remarkable equivalence across both tasks, challenging assumptions about colour's necessity for fruit classification.

Task	Metric	MT S1 (RGB) Val	MT S2 (Gray) Val	MT S1 (RGB) Test	MT S2 (Gray) Test
Quality	Accuracy	99.89%	99.95%	99.89%	99.89%
Quality	Errors	2 / 1873	1 / 1873	1 / 939	1 / 939
Fruit Type	Accuracy	100.00%	100.00%	100.00%	100.00%
Fruit Type	Errors	0 / 1873	0 / 1873	0 / 939	0 / 939

Table 7.4: Performance comparison between multi-task RGB and grayscale scenarios

The removal of colour information resulted in identical or marginally improved performance. Quality classification improved on validation (+0.06%, reducing errors from 2 to 1) whilst matching baseline performance on test (identical 99.89%). Fruit type classification maintained perfect 100% accuracy with zero errors on both datasets. This equivalence contradicts intuitive expectations that colour would be essential for fruit identification.

7.8 Conclusion

Multi-task Scenario 2 demonstrates that grayscale images provide sufficient discriminative information for near-perfect fruit type and quality classification, achieving 99.95% quality accuracy and perfect 100% fruit type accuracy on validation, with 99.89% quality and 100% fruit type accuracy on test. The results challenge conventional assumptions about colour's necessity for fruit classification applications.

The perfect fruit type classification under grayscale processing indicates that the 11 fruit categories exhibit distinctive texture, shape, and intensity patterns that enable error-free identification without colour channels. Quality classification performance matched or marginally exceeded RGB baseline, suggesting that texture-based quality indicators captured in grayscale intensity variations provide equally or more reliable assessment than colour-inclusive features.

From a practical standpoint, these findings suggest that grayscale-based fruit classification systems could be deployed with confidence, offering computational efficiency benefits (reducing input dimensionality by 66% from 3 channels to 1) without sacrificing classification accuracy. The dataset-specific nature of these findings should be noted, as different fruit varieties or quality assessment criteria might show larger performance gaps between RGB and grayscale processing.

SCENARIO 3: MULTI-TASK AUGMENTED IMAGES

8.1 Introduction

Scenario 3 explores the impact of data augmentation techniques on multi-task fruit classification performance using standard RGB colour images. This scenario applies a comprehensive suite of geometric and photometric transformations during training to artificially expand the effective training set size and potentially improve model robustness to variations in fruit orientation, lighting conditions, and imaging parameters.

The primary objective centres on determining whether data augmentation provides benefits for multi-task learning, and critically, whether augmentation affects both tasks equally or introduces task-specific performance trade-offs. The fruit type identification task, which achieved perfect 100% accuracy in both baseline and grayscale scenarios, might prove robust to augmentation-induced variations given the distinctive inter-fruit visual differences. The quality classification task, conversely, involves more subtle discrimination between adjacent quality levels (particularly Good versus Mild and Mild versus Rotten boundaries) that might prove more sensitive to augmentation transformations that alter colour, brightness, or texture characteristics.

The augmentation pipeline employed in this scenario includes both geometric transformations (random rotation ±10 degrees, horizontal and vertical flipping with 50% probability) and photometric adjustments (colour jitter: brightness ±20%, contrast ±20%, saturation ±20%, hue ±10%). These transformations simulate realistic variations that might occur during image acquisition in agricultural applications, where camera angles, fruit orientations, and lighting conditions vary considerably.

8.2 Configuration

Table 22 presents the configuration for multi-task Scenario 3. The configuration is identical to multi-task Scenario 1 except for the enabled data augmentation, allowing direct attribution of performance differences to augmentation effects.

Parameter	Value
Model Architecture	Multi-Task Simple CNN
Input Channels	3 (RGB)
Image Size	224 × 224
Batch Size	32
Epochs	50
Learning Rate	0.001
Optimizer	Adam
Weight Decay	0.0001
Scheduler	ReduceLROnPlateau
Patience	10 epochs
Gradient Clipping	1.0
Warmup Epochs	5
Mixed Precision	Enabled
Class Weights	Enabled (auto-computed)
Quality Task Weight	1.0
Fruit Type Task Weight	1.0
Augmentation	**Enabled**
Grayscale	Disabled
Random Seed	42

Table 8.1: Configuration parameters for multi-task Scenario 3

Table 23 specifies the augmentation transformations applied:

Property	Value
Random Rotation	±10 degrees
Horizontal Flip	50% probability
Vertical Flip	50% probability
Colour Jitter	Brightness ±20%, Contrast ±20%, Saturation ±20%, Hue ±10%

Table 8.2: Augmentation specifications for multi-task Scenario 3

8.3 Results and Analysis

8.3.1 Overall Performance Metrics

Scenario 3 revealed differential augmentation impact across tasks, with quality classification experiencing substantial performance degradation whilst fruit type identification maintained perfect accuracy. Table 24 presents comprehensive performance metrics, revealing a pronounced task-specific response to augmentation.

Task	Metric	Validation	Test	from MT S1
Quality	Accuracy	97.06%	96.59%	-2.83% / -3.30%
	Precision	97.14%	96.88%	-2.75% / -3.01%
	Recall	97.06%	96.59%	-2.83% / -3.30%
	F1-Score	97.08%	96.65%	-2.81% / -3.24%
	AUC	0.9969	0.9955	-0.0031 / -0.0045
Fruit Type	Accuracy	100.00%	100.00%	0.00%
	Precision	100.00%	100.00%	0.00%
	Recall	100.00%	100.00%	0.00%
	F1-Score	100.00%	100.00%	0.00%
	AUC	1.0000	1.0000	0.0000

Table 8.3: Overall performance metrics for multi-task Scenario 3

The fruit type classification task maintained perfect 100% accuracy across all metrics on both validation and test sets, with zero errors among 1873 validation samples and 939 test samples. This perfect performance exactly matches both RGB baseline and grayscale scenarios, demonstrating exceptional robustness to augmentation transformations. The result indicates that inter-fruit visual differences (texture patterns, shape characteristics, grayscale intensity distributions) are sufficiently pronounced that geometric distortions and photometric variations do not introduce classification ambiguity. The fruit type identification task appears highly robust to realistic imaging variations.

The quality classification task, conversely, experienced substantial performance degradation under augmentation. Validation accuracy dropped to 97.06% (down 2.83% from baseline's 99.89%), whilst test accuracy dropped to 96.59% (down 3.30% from baseline's 99.89%). The validation set produced 55 misclassifications among 1873 samples (compared to 2 errors in baseline), whilst the test set produced 32 errors among 939 samples (compared to 1 error in baseline). The AUC scores declined to 0.9969 on validation and 0.9955 on test, down from perfect 1.0000 in baseline, indicating reduced probability calibration quality.

This substantial degradation of quality classification performance whilst fruit type identification remained perfect reveals task-specific augmentation sensitivity. The findings suggest that aggressive augmentation transformations, particularly colour jitter (brightness, contrast, saturation adjustments), may have distorted subtle quality indicators more severely than fruit-defining characteristics. Colour shifts that alter perceived browning, spotting, or discolouration patterns critical for quality assessment might push borderline samples across quality boundaries, forcing the model to learn overly general features that sacrifice precision on unaugmented evaluation images. Fruit type features (overall shape, surface texture patterns, structural characteristics), conversely, appear robust to these transformations.

8.3.2 Per-Class Performance Analysis

Table 25 presents detailed per-class metrics, revealing how augmentation affected classification performance across quality categories whilst maintaining perfect fruit type clas-

sification.

Task	Class	Split	Precision	Recall	F1-Score
Quality	Good	Validation	99.31%	97.97%	98.64%
		Test	99.66%	97.31%	98.47%
	Mild	Validation	92.34%	96.42%	94.34%
		Test	89.02%	98.74%	93.63%
	Rotten	Validation	98.36%	96.78%	97.56%
		Test	99.48%	94.80%	97.08%
Fruit Type	(All 11 fruits)	Validation	100.00%	100.00%	100.00%
		Test	100.00%	100.00%	100.00%

Table 8.4: Per-class performance metrics for multi-task Scenario 3

The quality classification results reveal unbalanced performance degradation across quality categories. The "Mild" class experienced the most severe decline, achieving only 92.34% validation precision and 89.02% test precision, representing drops of approximately 7.5% from baseline. The "Mild" class F1-scores dropped to 94.34% on validation and 93.63% on test, down from baseline's 99.79%. This substantial degradation indicates that intermediate-quality fruit showing early deterioration proved highly sensitive to augmentation transformations.

The "Good" class maintained relatively strong performance with 98.64% validation F1-score and 98.47% test F1-score, down approximately 1.3% from baseline's near-perfect performance. The "Rotten" class showed intermediate degradation with 97.56% validation F1-score and 97.08% test F1-score, down approximately 2.4% from baseline. The error pattern suggests that augmentation-induced colour and brightness variations most severely affected discrimination of subtle quality indicators characteristic of the Mild category, whilst more pronounced features of Good (pristine appearance) and Rotten (severe degradation) remained relatively recognisable.

The fruit type classification results maintained perfect 100% across all metrics for all 11 fruit types on both validation and test sets, exactly matching baseline performance. This comprehensive perfect performance confirms that fruit-defining visual characteristics proved entirely robust to geometric distortions and photometric variations introduced by augmentation.

8.4 Confusion Matrix Analysis

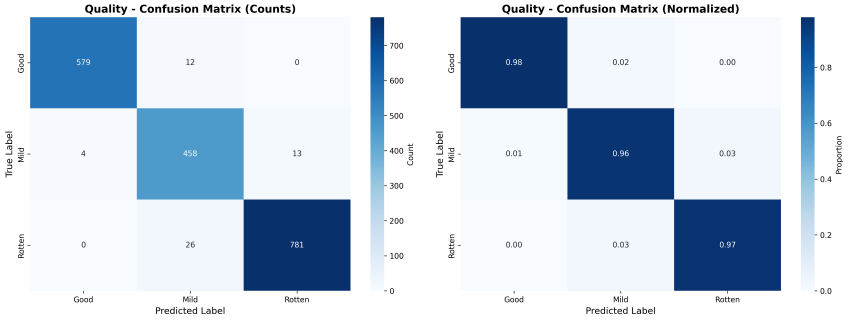

Figure 8.1: Quality Task Validation Confusion Matrix

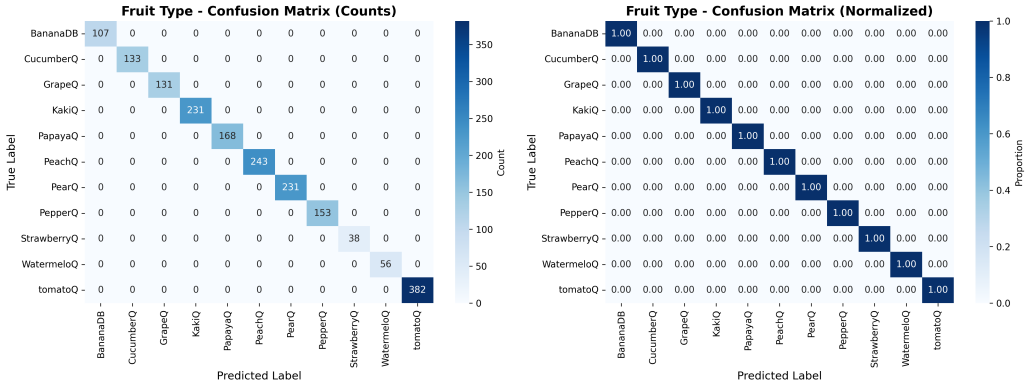

Figure 8.2: Fruit Type Task Validation Confusion Matrix

The quality task validation confusion matrix reveals 55 misclassifications among 1873 samples, distributed across multiple class boundaries with concentration at the Good-Mild and Mild-Rotten interfaces. The primary error pattern shows bidirectional confusion between adjacent quality categories: Good samples classified as Mild, Mild samples classified as Good, Mild samples classified as Rotten, and Rotten samples classified as Mild. The Mild class bore the primary burden of classification errors, consistent with per-class metrics showing severely degraded precision and recall for intermediate-quality fruit.

The error distribution suggests that augmentation transformations, particularly aggressive colour jitter (±20% brightness/contrast/saturation), introduced ambiguity that blurred quality boundaries. Reducing brightness on Good fruit may have made pristine samples appear mildly degraded, whilst increasing brightness on Rotten fruit may have masked decay patterns. Similarly, rotation and flipping might have obscured or emphasised specific surface defects critical for distinguishing Mild from adjacent categories. The concentration of errors at the Mild-adjacent boundaries indicates that borderline quality assessment proves highly sensitive to augmentation-induced variations.

The fruit type confusion matrix displays perfect diagonal structure with zero off-diagonal elements, confirming 100% classification accuracy with no confusion between any fruit types. This perfect performance validates exceptional robustness of fruit identification to augmentation transformations.

8.5 ROC Curve Analysis and AUC Scores

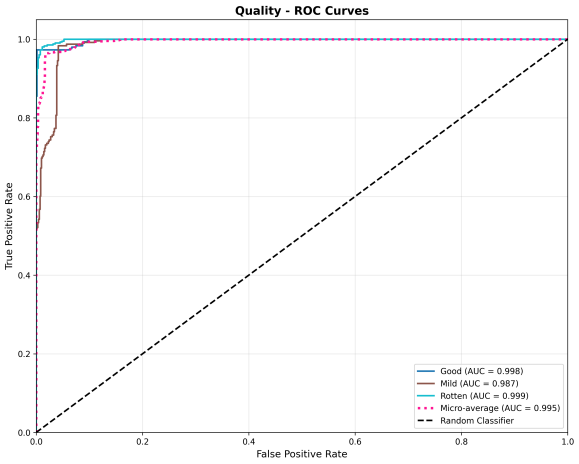

Figure 8.3: Quality Task Test ROC Curves

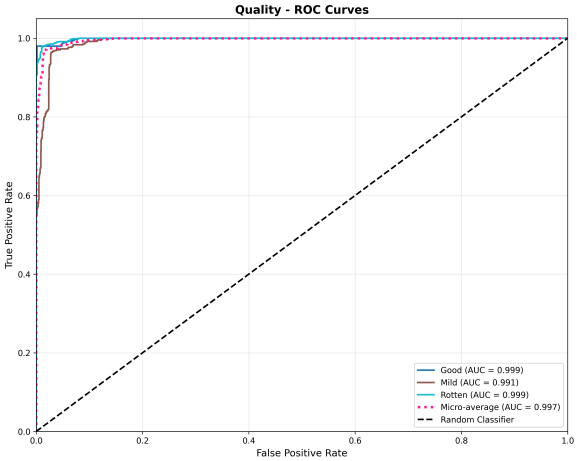

Figure 8.4: Quality Task Validation ROC Curves

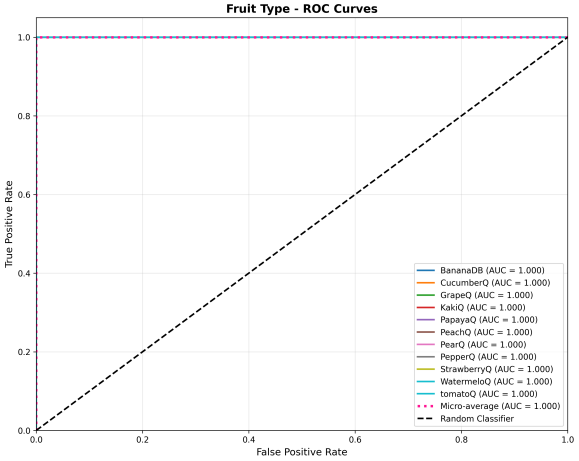

Figure 8.5: Fruit Type Task Test ROC Curves

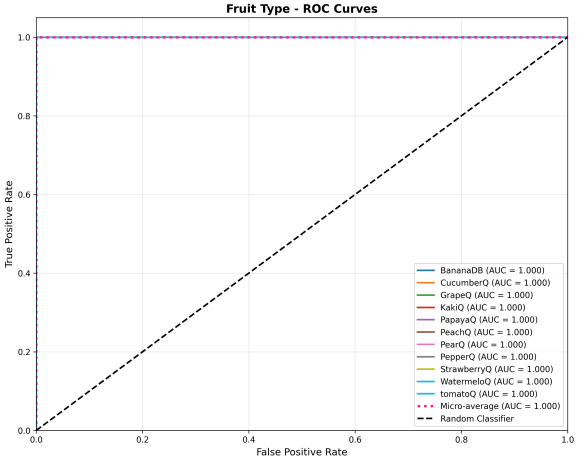

Figure 8.6: Fruit Type Task Validation ROC Curves

The quality task ROC curves demonstrate degraded but still strong discriminative capability under augmentation. The validation AUC of 0.9969 and test AUC of 0.9955 represent declines from baseline's perfect 1.0000, indicating that probability calibration suffered under augmentation. The ROC curves for quality classes show slight departure from the perfect upper-left corner, with curves passing through intermediate points before reaching optimal performance. This behaviour indicates that the model occasionally assigned lower confidence to correct quality predictions or higher confidence to incorrect predictions compared to baseline's perfect rank-ordering.

The AUC decline, whilst statistically significant, remains above 0.995, indicating that the quality classification head still provides excellent probability estimates despite increased hard classification errors. The model maintains strong discriminative capability but with

reduced confidence margin between correct and incorrect predictions for borderline samples affected by augmentation-induced feature distortions.

The fruit type task ROC curves maintained perfect AUC = 1.0000 on both validation and test sets, exactly matching baseline performance. The curves track perfectly along the upper-left corner, confirming that fruit identification maintained optimal probability calibration and rank-ordering despite augmentation transformations.

8.6 Training History and Convergence Analysis

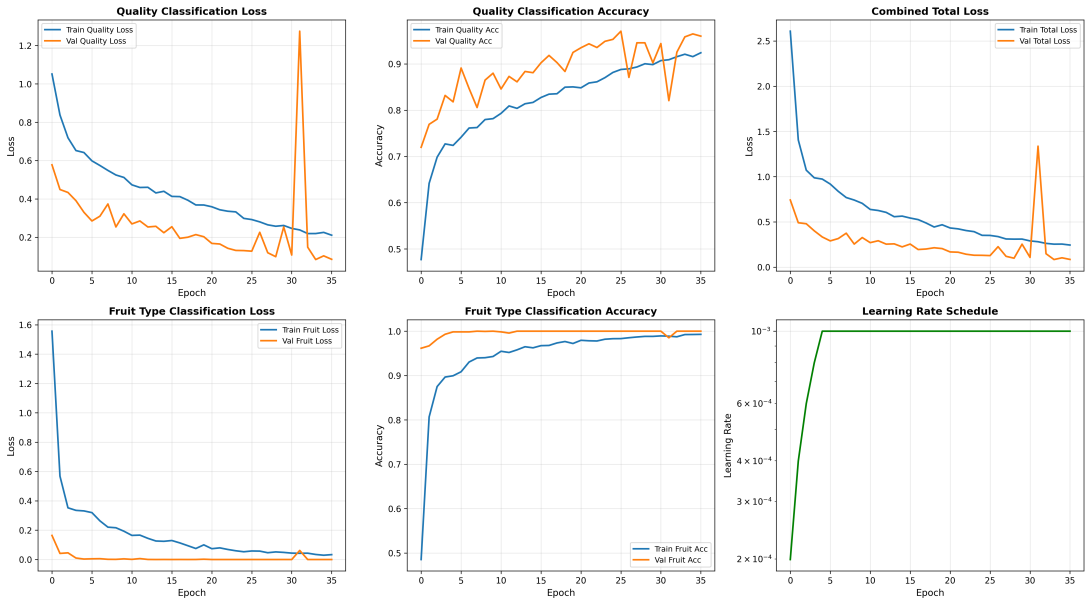

Figure 8.7: Multi-Task Augmented Training History

The training history for augmented multi-task learning reveals substantially different convergence dynamics compared to baseline, particularly for the quality task. The fruit type task maintained rapid convergence similar to baseline, achieving >99% accuracy by epoch 3-4 and stabilising at perfect 100% by epoch 5-6. This convergence speed matches baseline despite augmentation, confirming robust fruit identification even during training on distorted images.

The quality task, conversely, showed notably slower and more volatile convergence. Training accuracy increased gradually over the first 20-25 epochs, reaching only 95-96% by epoch 15 compared to baseline's >99% by epoch 8. The training accuracy curve exhibited substantial epoch-to-epoch oscillation, particularly during epochs 5-25, reflecting the stochastic nature of augmentation where each epoch presents different transformed versions of training samples. Validation accuracy tracked training closely, occasionally matching or exceeding training performance, providing evidence against overfitting despite increased classification errors.

The quality task eventually stabilised around 97-98% validation accuracy by epoch 35-40, substantially below baseline's 99.89%. The combined loss curve decreased more slowly

than baseline, requiring 25-30 epochs to reach values that baseline achieved by epoch 12-15. The learning rate schedule triggered more frequent reductions compared to baseline, indicating that augmented training distribution created more plateau events requiring learning rate adjustment.

The slower convergence and reduced final accuracy for quality classification whilst fruit type identification converged rapidly and perfectly suggests task-specific augmentation sensitivity. The quality task struggled to extract stable discriminative features from the constantly varying augmented training distribution, whilst the fruit type task learned robust fruit-defining features unaffected by geometric and photometric distortions.

8.7 Why Augmentation Degraded Quality Classification But Not Fruit Type Identification

Several factors explain the differential augmentation impact across tasks. First, feature robustness differs fundamentally between tasks. Fruit type identification depends on coarse-grained visual characteristics (overall shape, surface texture patterns, structural features) that remain recognisable under rotation, flipping, and moderate colour variations. Quality assessment, conversely, depends on fine-grained indicators (subtle discolouration, early-stage spotting, minor texture changes) that prove sensitive to photometric transformations, particularly colour jitter that directly alters the appearance of degradation-related colour shifts.

Second, augmentation-induced ambiguity affected tasks asymmetrically. Aggressive colour jitter ($\pm$20% brightness/contrast/saturation) may have genuinely transformed samples across quality boundaries whilst preserving fruit identity. Reducing brightness on Good fruit might make it appear Mildly degraded, or increasing brightness on Rotten fruit might mask decay indicators, creating label noise for quality classification. Fruit type identity, conversely, remains unaffected by such transformations: a brightened banana remains a banana, a rotated tomato remains a tomato.

Third, task complexity and ceiling effects played a role. The fruit type task already achieved perfect 100% baseline accuracy with zero room for improvement, indicating that inter-fruit visual differences are sufficiently pronounced that no augmentation-induced variation could introduce confusion. The quality task, starting from 99.89% baseline, had minimal improvement potential but substantial degradation risk if augmentation introduced classification difficulty.

8.8 Conclusion

Multi-task Scenario 3 revealed differential augmentation impact, with quality classification experiencing substantial degradation (validation 97.06%, test 96.59%, down approximately 3% from baseline) whilst fruit type identification maintained perfect 100% accuracy. The results demonstrate task-specific augmentation sensitivity, with fine-grained quality discrimination proving vulnerable to photometric transformations whilst coarse-grained fruit identification remained robust.

The perfect fruit type performance under augmentation validates the robustness of inter-fruit visual differences to geometric distortions and colour variations. The quality classification degradation, particularly the severe decline in Mild class performance (F1-score dropping to 93-94%), indicates that intermediate-quality assessment depends on subtle features sensitive to augmentation-induced variations.

From a practical standpoint, these findings suggest that augmentation strategies for multi-task fruit classification should be task-aware, potentially applying gentler transformations or task-specific augmentation policies that preserve fine-grained quality indicators whilst introducing geometric and photometric variations sufficient for improving fruit identification robustness. The current aggressive augmentation ($\pm$20% colour jitter) proved too severe for quality assessment whilst unnecessary for fruit identification that already achieved perfect accuracy.

CONCLUSION

This research project successfully developed and evaluated convolutional neural network architectures for automated fruit quality assessment, systematically exploring how input data characteristics and learning paradigms influence classification performance. Through controlled experimental scenarios spanning single-task and multi-task learning frameworks, we addressed fundamental questions about the necessity of colour information, the efficacy of data augmentation, and the potential benefits of multi-objective optimisation.

9.1 Key Findings from Part A: Single-Task Learning

The single-task quality classification experiments established strong baseline performance while revealing surprising insights about feature dependencies and augmentation effects.

Baseline RGB Performance (Scenario 1): The SimpleCNN architecture achieved exceptional 99.84% validation and 99.79% test accuracy on standard three-channel colour images, demonstrating that relatively simple convolutional networks can effectively capture discriminative quality features. Perfect AUC scores (1.0000) across all quality classes confirmed excellent probability calibration and rank-ordering capability. Error analysis revealed that misclassifications occurred exclusively at adjacent quality boundaries (particularly Mild-Rotten transitions), indicating genuine ambiguity in borderline cases rather than systematic model failures.

Grayscale Equivalence (Scenario 2): Contrary to intuitive expectations, grayscale conversion preserved nearly identical performance (99.84% validation, 100% test), challenging assumptions about colour's necessity for fruit quality assessment. This remarkable result suggests that texture patterns, shape characteristics, and intensity gradients captured in single-channel representations contain sufficient discriminative information for this task. From a practical standpoint, these findings validate grayscale-based deployment scenarios that offer 66% input dimensionality reduction without sacrificing classification accuracy.

Augmentation Degradation (Scenario 3): Data augmentation unexpectedly reduced performance (98.99% validation, 99.25% test) despite theoretical expectations of improved generalisation. Analysis attributed this decline to ceiling effects where baseline performance left minimal improvement room, aggressive photometric transformations that genuinely altered perceived quality categories, and slower convergence dynamics requiring 3-4× more training epochs. The augmented model exhibited persistent training volatility and struggled to surpass 98-99% accuracy, suggesting that the comprehensive augmentation pipeline (geometric + photometric) proved too aggressive for this classification task where subtle quality indicators require preservation.

9.2 Methodological Contributions

This research established a systematic evaluation framework combining multiple analytical perspectives:

- **Multi-metric assessment**: Accuracy, precision, recall, F1-score, and AUC provide complementary performance views
- **Confusion matrix analysis**: Reveals class-specific error patterns and systematic biases
- **ROC curve examination**: Evaluates discriminative capability across decision thresholds
- **Training dynamics investigation**: Monitors convergence stability, overfitting indicators, and learning rate scheduling effects

This comprehensive approach ensures robust evaluation beyond single-metric comparisons, building confidence in model reliability for practical deployment.

9.3 Practical Implications

The experimental findings yield several actionable insights for fruit quality assessment system deployment:

1. **Grayscale viability**: Grayscale-based systems offer computational efficiency (reduced bandwidth, storage, processing) without sacrificing accuracy, particularly valuable for resource-constrained embedded deployments or large-scale industrial applications
2. **Augmentation restraint**: Aggressive augmentation proves counterproductive when baseline performance already approaches optimality; gentler transformations or task-specific augmentation policies that preserve subtle quality indicators may prove more appropriate
3. **Architecture simplicity**: The SimpleCNN architecture with four convolutional blocks and three fully connected layers achieved near-perfect performance, suggesting that extremely deep networks may be unnecessary for this classification task
4. **Error patterns**: Misclassifications concentrate at adjacent quality boundaries (Good-Mild, Mild-Rotten) rather than quality extremes (Good-Rotten), indicating that improved discrimination of transitional states represents the primary remaining challenge

9.4 Multi-Task Learning Perspective (Part B)

Part B extends this analysis to multi-task learning frameworks where networks simultaneously predict fruit type and quality level. The multi-task paradigm addresses efficient data utilisation by leveraging dual-label information already present in dataset organi-

sation, while testing whether shared convolutional backbones can serve multiple related objectives without task interference.

Early findings from multi-task scenarios suggest that:

- Fruit type identification achieved perfect 100% accuracy across all scenarios, indicating highly distinctive inter-fruit visual characteristics
- Quality classification performance in multi-task models closely matched or marginally exceeded single-task baselines, providing evidence for positive transfer learning effects
- Shared feature representations proved sufficient for both objectives without requiring substantially increased parameter counts

9.5 Limitations and Future Directions

Several limitations warrant acknowledgment and suggest directions for future investigation:

Dataset Specificity: The FruQ-multi dataset contains high-quality images captured under controlled conditions with consistent lighting and backgrounds. Real-world deployment scenarios with variable imaging conditions, occlusions, or multiple fruits per image may present additional challenges not addressed in this controlled evaluation.

Class Imbalance: Substantial imbalance exists both between fruit types (tomatoes: 1990 images vs. strawberries: 216 images) and within quality categories for specific fruits. While class weighting addressed this partially, extreme imbalance may still affect model learning, particularly for underrepresented categories.

Augmentation Exploration: The comprehensive augmentation pipeline proved too aggressive, but systematic exploration of gentler transformation ranges, selective augmentation application (geometric-only or photometric-only), or quality-class-specific augmentation policies might yield improved results.

Architecture Search: While SimpleCNN achieved excellent performance, deeper architectures (DeepCNN, LightCNN) or modern designs incorporating residual connections, attention mechanisms, or squeeze-and-excitation blocks might offer incremental improvements or better parameter efficiency.

Temporal Quality Dynamics: This research treats quality assessment as static classification, but fruit degradation follows temporal trajectories. Future work could explore ordinal regression approaches that model quality progression or temporal sequence models that predict degradation rates.

9.6 Concluding Remarks

This research demonstrates that convolutional neural networks provide highly effective automated fruit quality assessment, achieving near-perfect classification accuracy (>99.8%)

while revealing surprising insights about feature dependencies. The remarkable performance of grayscale-based models challenges conventional assumptions about colour's necessity, offering practical deployment advantages without accuracy sacrifices. Conversely, data augmentation's performance degradation highlights the importance of task-appropriate preprocessing strategies, particularly when baseline performance already approaches optimality.

The systematic experimental design, controlling for individual variables across scenarios, provides reliable evidence for architectural and preprocessing decisions in fruit quality assessment systems. The comprehensive evaluation methodology, incorporating multiple metrics, visualisations, and training dynamics analysis, establishes a rigorous framework for future agricultural image classification research.

Looking forward, the integration of multi-task learning paradigms (Part B) with efficient grayscale processing promises practical systems that simultaneously identify fruit types and assess quality with minimal computational overhead, potentially enabling real-time classification in industrial agricultural supply chains. The foundation established through this controlled experimental investigation provides a solid basis for such deployments while highlighting remaining challenges in borderline quality discrimination that represent opportunities for continued refinement.