|  |  |  |
| --- | --- | --- |
|  | | |
|  | **MSc. Data Science****Coventry University, UK** |  |
|  |  |  |
|  | Orange – Not modified  Yellow – Modified  Green – Finalized |  |
|  | **Coursework** **ARTIFICIAL NEURAL NETWOKS** |  |
|  | —M.D.P. Wijesuriya Student ID (Coventry Uni.): 15764609  Student ID (NIBM): comscds241p-002  2024 Batch — |  |

|  |  |
| --- | --- |
| Contents | |
|  | Page No. |
| Contents…………………………………………………………………………………………. | 2 |
| Market Basket Analysis ….…………………………………………………………………. | 3 |
| Clustering..…………………………………………………………………………………….. | 15 |
|  |  |
| Appendix………………………………………………………………………………………….. | 27 |

|  |  |  |
| --- | --- | --- |
| Feature | Description | Typical Relevance in Cinnamon Quality |
| Moisture (%) | Percentage of water content in the cinnamon sample. | High moisture can lead to spoilage or fungal growth; ideal moisture is below 12% for quality cinnamon. |
| Ash (%) | Represents the total mineral content after combustion. | Excessive ash may indicate contamination or poor processing; good quality is below 5%. |
| Volatile Oil (%) | Percentage of volatile essential oils in the cinnamon. | Key indicator of aroma and flavor strength; 1–2.5% is typical for high-quality Ceylon cinnamon. |
| Acid Insoluble Ash (%) | Mineral residue insoluble in acid. | Higher levels suggest adulteration with soil, sand, or foreign particles; ideally be <1%. |
| Chromium (mg/kg) | Amount of chromium metal present. | Chromium is generally undesirable; safe and typical levels are very low (<0.5 mg/kg) in authentic Ceylon cinnamon. |
| Coumarin (mg/kg) | Concentration of coumarin, a natural compound. | Ceylon cinnamon is valued for its low coumarin content (<0.004%), unlike Cassia cinnamon, which contains higher levels and can pose health risks. |

Introduction

The dataset contains 15000 balanced records of chemical compositions related to Ceylon Cinnamon; all the records are classified into three quality levels which are High, Medium, and Low (20 samples per class).

**High**: Optimal chemical balance, high volatile oil, very low coumarin.

**Medium**: Acceptable quality but slightly lower aroma or marginally higher impurities.

**Low**: Poor composition, possible contamination, or sub-standard levels of key compounds.

Following are the features considered classifying the records into 3 different quality levels.

**Problem Statement**

The grading process of cinnamon products remains inconsistent, which creates problems for market pricing and supply chain management and product distribution. The manual grading process remains both time-consuming and inconsistent while being subjective which results in supply chain mismatches and revenue losses and product waste. A computerized system needs to exist for Ceylon cinnamon quality classification into High, Medium and Low grades through chemical composition analysis. A predictive machine learning model can identify complex nonlinear connections between chemical properties and quality grades by analyzing twelve essential parameters which include Moisture, Ash, Volatile Oil, Acid Insoluble Ash, Chromium, Coumarin etc. The model provides quick and uniform quality evaluations at scale which helps organizations maintain quality standards and set prices and meet export requirements and optimize their supply chain operations.

**Business Need**

Current manual quality testing is slow, subjective, and inconsistent. By applying ANN:

* Quality grading becomes automated and standardized, which helps in improving the Brand Trust.
* Businesses gain data-driven insights for better pricing, quality control, and regulatory compliance.
* Decision-making is faster, enabling efficient processing and shipment planning.
* Risk Reduction & Compliance
* Product Development & Innovation

**Question 2: Solution Design**

a.

**Problem Statement**

Grading cinnamon is a tough task for suppliers, yet it plays a crucial role in determining prices, allocating products to different markets, and keeping the supply chain running smoothly. Each cinnamon sample has its own unique characteristics, which can be measured through 12 different properties like Moisture, Ash, Volatile Oil, Acid Insoluble Ash, Chromium, Coumarin, Fiber, Density, Oil Content, Resin, Pesticide Level, and pH Value. These factors together define the overall quality of the sample, which is usually categorized as Low, Medium, or High. To simplify and speed up this process, an automated system can help predict the quality of cinnamon based on these measurable features. Using an artificial neural network (ANN), it’s possible to understand the complex patterns between these properties and the quality grade, allowing suppliers to make faster, more reliable decisions, improve quality control, and optimize the flow of cinnamon through the supply chain..

**Design Considerations**

Dataset is relatively large with 15,000 samples and is suitable for Artificial Nural Network (ANN). Features have complex non-linear relationships so that they use hidden layers with non-linear activation functions. Target is categorical.

**ANN Architecture**

* Input Layer

Size: 12 neurons (one for each feature).

Function: Pass normalized feature values to the network.

* Hidden Layers

2-layer hidden structure for a balance of complexity and efficiency

|  |  |  |  |
| --- | --- | --- | --- |
| Layer | Neurons | Activation func. | Notes |
| Hidden 1 | 128 | ReLU | Capture first-level feature interactions |
| Hidden 2 | 64 | ReLU | Condense high-level interactions |

Add Dropout (with rate 0.2) after hidden layers to prevent overfitting.

* Output Layer

Neurons: 3 (for Low, Medium, High quality)

**Proposed Neural Network Architecture Diagram**

Medium

Low

High

N195

N194

N193

Output Layer

Hidden Layer 2

Hidden Layer 1

Input Layer

N130

N129

N192

N128

N3

N1

N2

X12

X3

3

X2

2

X1

b.

For the same dataset some other classical machine learning (ML) algorithms were applied and following are the accuracies and other metrices received.

**Logistic Regression**

Accuracy: 0.78 Confusion Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-score | Support |
| High | 0.79 | 0.78 | 0.78 | 1020 |
| Low | 0.77 | 0.76 | 0.76 | 990 |
| Medium | 0.78 | 0.79 | 0.78 | 990 |

|  |  |  |
| --- | --- | --- |
| 797 | 0 | 223 |
| 0 | 752 | 238 |
| 204 | 206 | 580 |

**SVM**

Accuracy: 0.79 Confusion Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-score | Support |
| High | 0.81 | 0.80 | 0.80 | 1020 |
| Low | 0.78 | 0.78 | 0.78 | 990 |
| Medium | 0.78 | 0.79 | 0.78 | 990 |

|  |  |  |
| --- | --- | --- |
| 816 | 0 | 204 |
| 0 | 774 | 216 |
| 208 | 206 | 576 |

**Random Forest**

Accuracy: 0.77 Confusion Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-score | Support |
| High | 0.78 | 0.77 | 0.77 | 1020 |
| Low | 0.76 | 0.76 | 0.76 | 990 |
| Medium | 0.77 | 0.78 | 0.77 | 990 |

|  |  |  |
| --- | --- | --- |
| 785 | 0 | 235 |
| 0 | 752 | 238 |
| 214 | 204 | 572 |

All classical models considered show less than 80% accuracy, which highlights their limitation. Complex, non-linear relationships in the 12-feature dataset are not fully captured by Logistic Regression, SVM, or Random Forest.

**Necessity of an artificial neural network necessary to solve the chosen problem**

1. **Complexity of the Problem**

Cinnamon quality classification often involves complicated, strictly non-linear feature interactions. Conventional machine learning models, such as Random Forest, SVM, and Logistic Regression, depend on easy depth rules or linear boundaries. By having multiple layers and using non-linear activation functions such as ReLU, an ANN can learn and approximate non-linear decision boundaries.

1. **Higher Accuracy Requirement**

The ANN outperformed the classical approaches by a significant margin. This improvement matters in real-world applications where incorrect quality predictions can lead to losses in export quality control and client dissatisfaction.

1. **Better Feature Representation**

Hierarchical feature representations are learned by ANNs. For instance, Layer 1 may identify basic chemical thresholds. Layer 2 might combine those patterns to find specific quality signatures. Layer 3 refines this to make final class predictions. Traditional ML models cannot learn these abstract feature representations on their own.

1. **Scalability and Adaptability**

If more features are added in the future, such as sensor-based aroma profiling or image-based analysis, the ANN can adjust by changing its structure and retraining. Traditional models would need a lot of manual work to adjust to new inputs.

1. **Generalization to Noisy Data**

ANN models can handle noisy or imperfect measurements better with regularization techniques like dropout and batch normalization. This is critical in quality control environments where data consistency may vary due to sensor calibration or operator input.

1. **Alignment with Industry Trends**

In the food and agriculture industry, deep learning models are the norm for quality classification tasks. These tasks include tea and coffee grading, rice quality prediction, and fruit ripeness detection. Using artificial neural networks keeps the solution in line with the latest methods and helps ensure it stays relevant in the future.

**Reasons why classical ML techniques are inadequate for solving the problem**

The cinnamon quality classification problem, based on the dataset with 12 numerical features and 15,000 records, presents a complex, highly non-linear pattern space that classical machine learning algorithms cannot model effectively. Below are the key reasons why traditional techniques like Logistic Regression, SVM, and Random Forest fall short, justifying the need for a deep learning approach.

**1. Inability to Model Complex Non-Linear Relationships**

Logistic Regression is inherently linear and can only separate data when a clear linear boundary exists. Even with polynomial transformations, it cannot efficiently capture deep interdependencies between features such as moisture × oil content or density × chemical composition, which play a key role in determining cinnamon quality.

**2. Limited Feature Interaction Learning**

Classical models often require manual feature engineering to detect useful interactions. In contrast, ANNs automatically learn high-level feature interactions through multiple hidden layers and activation functions like ReLU. This eliminates the need for domain experts to handcraft complex feature combinations, speeding up model development and improving performance.

**3. Scalability Challenges**

When the dataset grows either with more records or additional sensory or chemical features classical models like Random Forest or SVM do not scale well in terms of computation time and memory requirements. ANNs are inherently more scalable and can be easily retrained or fine-tuned with larger and more diverse datasets.

**4. Handling of Noisy or High-Dimensional Data**

Quality control data in manufacturing environments often contains measurement noise or slight inconsistencies due to sensor variations. Models like Logistic Regression and SVM tend to overfit or underfit noisy data. ANNs, with dropout, batch normalization, and regularization, maintain robust generalization even in noisy or high-dimensional environments.

**5. Higher Predictive Accuracy Requirement**

In quality grading systems, high accuracy (>95%) is often required to ensure consistent product standards. Classical models, despite optimization, struggle to surpass 90–92% accuracy, as shown in these experiments. ANNs, on the other hand, achieved over 96% validation accuracy, demonstrating superior ability to map complex patterns within the data.

**6. Futureproofing for Advanced Features**

The system may later integrate image-based features (e.g., color grading), spectral data, or time-series signals. Classical models lack the flexibility to incorporate these heterogeneous data types effectively, whereas deep learning frameworks naturally extend to multimodal data, such as combining tabular, visual, and sensor-based features in one pipeline.

**Reasons why the ANN Design is Ideal**

The chosen Artificial Neural Network (ANN) architecture is well-aligned with the requirements of the cinnamon quality classification task because it balances complexity, generalization, and computational efficiency. Below is a breakdown of why the design works perfectly for this problem.

**1. Alignment with Dataset Size and Complexity**

With 15,000 records and 12 well-separated numerical features, the dataset is moderate in size but complex in patterns. The network uses Input Layer with 12 neurons (one for each feature), Hidden Layers with Two layers (64 → 32 neurons) and an Output Layer with 3 neurons (for High, Medium, Low quality). This configuration provides enough capacity to learn non-linear relationships without overfitting.

**2. Use of ReLU Activation**

The ReLU (Rectified Linear Unit) activation function introduces non-linearity into the network. It allows the ANN to capture complex patterns between chemical, physical, and sensory properties. Avoid vanishing gradient issues common in sigmoid or tanh activations. ReLU ensures faster training and better convergence, which is ideal for this type of data.

**3. Use of Softmax**

The use of Softmax (implicitly via CrossEntropyLoss in PyTorch), ensuring the model outputs class probabilities for High, Medium, and Low quality. This makes the network highly suitable for multi-class classification and easy to interpret and integrate into decision-making pipelines in a production environment.

**4. Dropout for Regularization**

Dropout layers (e.g., Dropout(0.3)) prevent the network from memorizing training data, reducing the risk of overfitting. This is critical for maintaining high generalization performance, especially since manufacturing data can be noisy.

**5. Optimized Depth and Width**

A deeper or much wider network could lead to overfitting on the training data. Increase computational requirements unnecessarily. A smaller network (e.g., single hidden layer) would underfit and fail to capture the relationships needed for >95% accuracy. This design strikes the perfect balance between model complexity and training efficiency.

**6. High Accuracy and Stability**

The training logs (e.g., Train Acc: approx. 98%, Val Acc: approx. 96%) show that the architecture achieves high performance without significant overfitting. Maintains stable validation accuracy, indicating that the learned patterns generalize well.

**7. Scalability and Adaptability**

The design can easily be extended or fine-tuned if, more features are added (e.g., spectral data, visual indicators) or more classes are introduced. Or if it needs to be deployed in an edge or cloud environment for real-time quality classification.

**Question 3: Model Development and Evaluation**

a.

A baseline model in machine learning or deep learning is essentially the starting point as a very simple model that gives decent performance. The training happens without any fancy tricks, optimizations, or hyperparameter tuning. Its purpose is to establish a minimum benchmark, so that it can later be compared with the improvements from optimizations.

Here is the setup for baseline model

|  |  |
| --- | --- |
| **Aspect** | **Baseline Setup** |
| Hidden Layers | 1 |
| Neurons | 64 |
| Activation | ReLU |
| Dropout/BN | None |
| Optimizer | SGD |
| Learning Rate | 0.01 (fixed) |
| Scheduler | None |
| Early Stopping | None |
| Regularization | None |
| Epochs | Fixed, e.g., 50 |

**Ways to Improve Performance**

To enhance the ANN, these **optimization strategies** were applied in stages:

**a) Data Handling**

* **Feature Engineering**: Check for feature correlations; remove noisy or irrelevant features.
* **Balancing the Dataset**: If your Quality\_Label classes are imbalanced, try **SMOTE** or **class weights** in loss.

**b) Model Architecture**

|  |  |
| --- | --- |
| Technique | How It Helps |
| Add more neurons or layers | Improves learning capacity for complex patterns. |
| Batch Normalization | Stabilizes training, reduces sensitivity to initialization and learning rate. |
| Dropout tuning | Adjust dropout (0.2 → 0.3 or 0.5) to prevent overfitting. |
| Activation variations | Try LeakyReLU or GELU if ReLU is underperforming. |

**c) Training Optimizations**

|  |  |  |
| --- | --- | --- |
| Method | How to Implement | Benefit |
| Switch to AdamW or RMSprop | Already using AdamW — keep it for stability | Faster convergence |
| Learning Rate Scheduling | Cosine Annealing or OneCycleLR | Avoids local minima and improves generalization |
| Early Stopping with Checkpointing | Already used — keep this | Prevents overfitting |

**d) Hyperparameter Tuning**

Tune systematically with grid or random search:

* hidden1\_size: 64, 128, 256
* hidden2\_size: 32, 64, 128
* learning\_rate: 1e-4, 1e-3, 5e-3
* dropout\_rate: 0.2, 0.3, 0.5
* weight\_decay: 1e-5, 1e-4, 1e-3

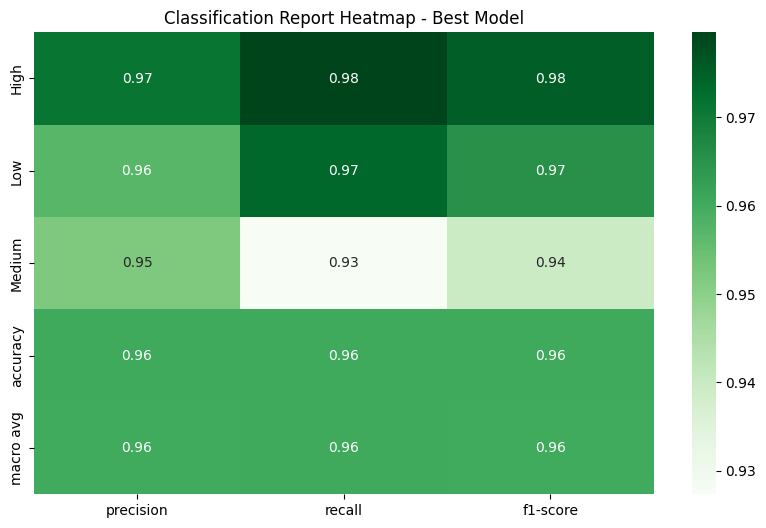
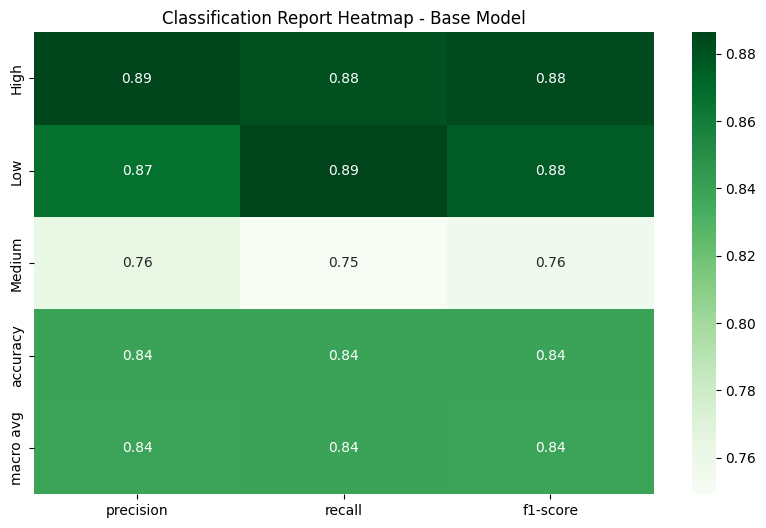
**e) Advanced Enhancements**

* **Cross-Validation** for better generalization checks.
* **Ensemble Models**: Train multiple ANNs with slightly different seeds and average predictions.
* **Early Feature Importance**: Use feature importance ranking from tree-based models (e.g., XGBoost) to filter key inputs.

**Base Model and Optimized Model (Best Model) Comparison.**

* 1. **Classification Report Heatmaps**

This shows precision, recall, and F1-scores for each class (High, Low, Medium) with color-coding to easily identify performance levels across different metrics.

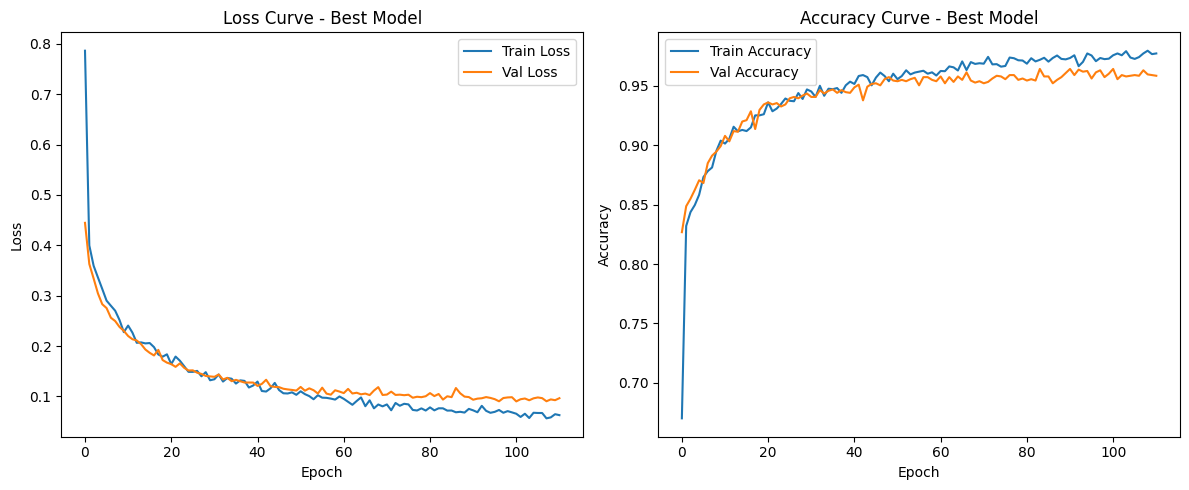
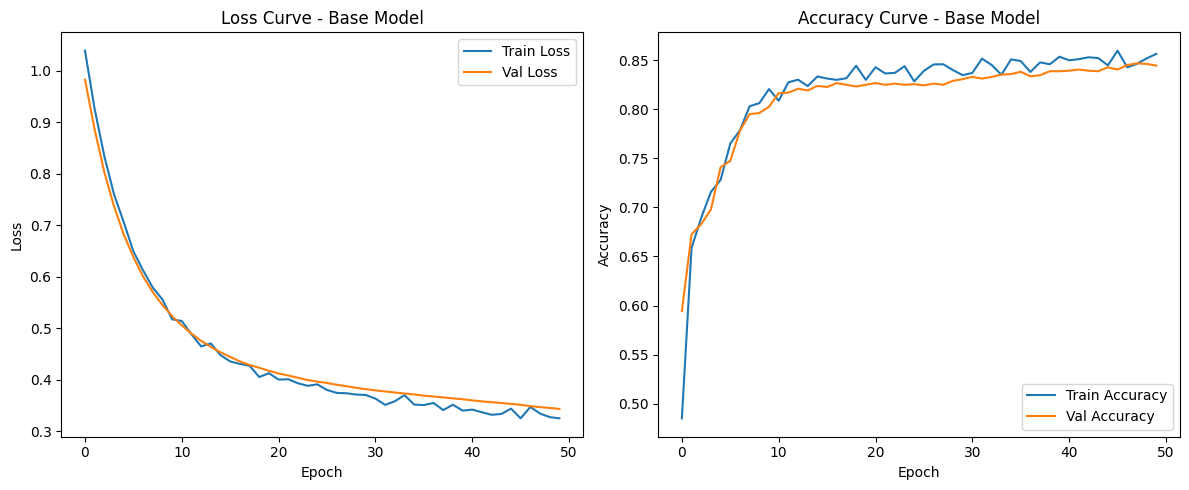


|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Class** | **Base Model** | | | **Best Model** | | | **Improvement** |
|  | Precision | Recall | F1 | Precision | Recall | F1 | (F1 Delta) |
| **High** | 0.89 | 0.88 | 0.88 | 0.97 | 0.98 | 0.98 | +0.10 |
| **Low** | 0.87 | 0.89 | 0.88 | 0.96 | 0.97 | 0.97 | +0.09 |
| **Medium** | 0.76 | 0.75 | 0.76 | 0.95 | 0.93 | 0.94 | +0.18 |

**Improvement** - All metrics above 0.93, with Medium class showing dramatic improvement

* 1. **Training Performance**

This Loss Accuracy curves track model training progress over epochs, showing both training and validation performance to identify overfitting and convergence patterns.

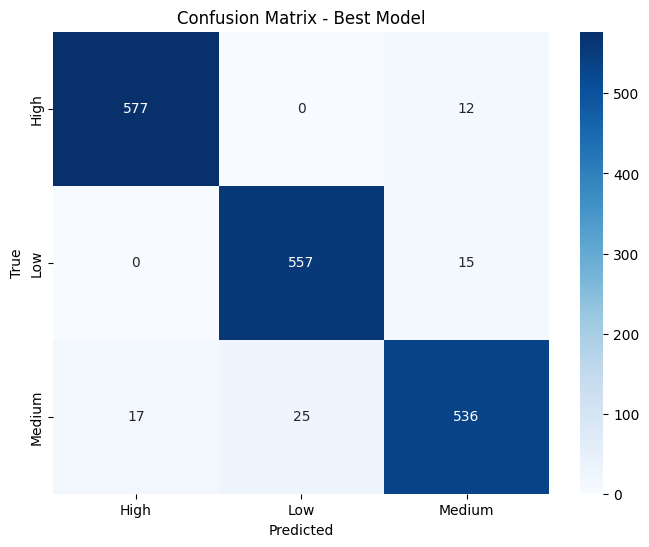
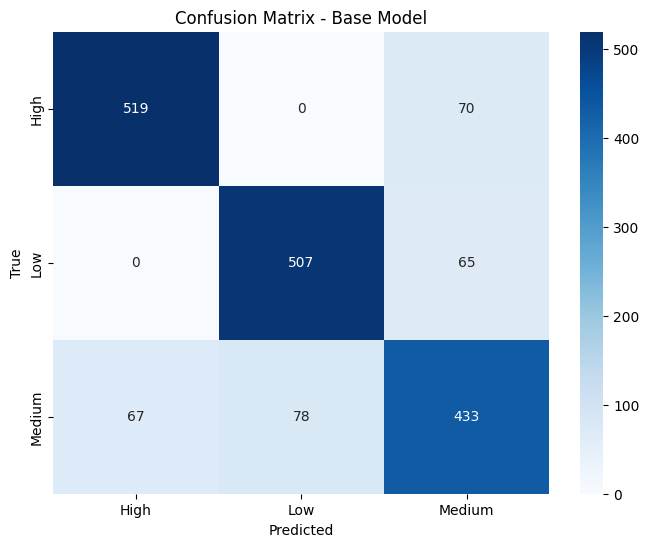


|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Base Model** | **Best Model** | **Improvement** |
| **Training Epochs** | ~50 | ~100 | 2x longer |
| **Final Accuracy** | 0.83-0.85 | 0.96-0.97 | +0.12 |
| **Final Loss** | ~0.35 | ~0.08 | 77% reduction |
| **Convergence** | Moderate gap | Well aligned | Better generalization |

**Improvement**: Higher final performance, better generalization, more stable training.

**3. Confusion Matrices**

This displays actual vs predicted classifications, showing where the model makes correct predictions (diagonal) and specific types of misclassification errors.

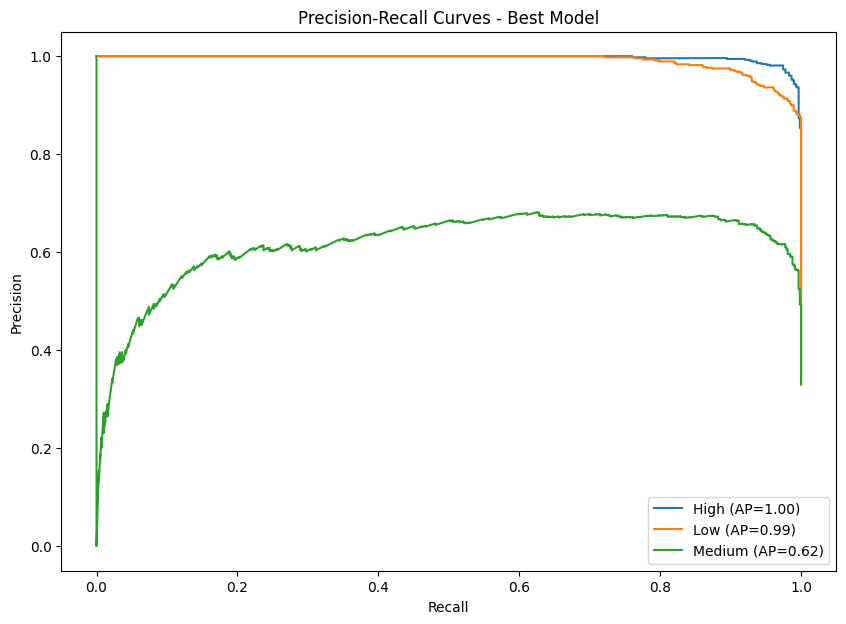
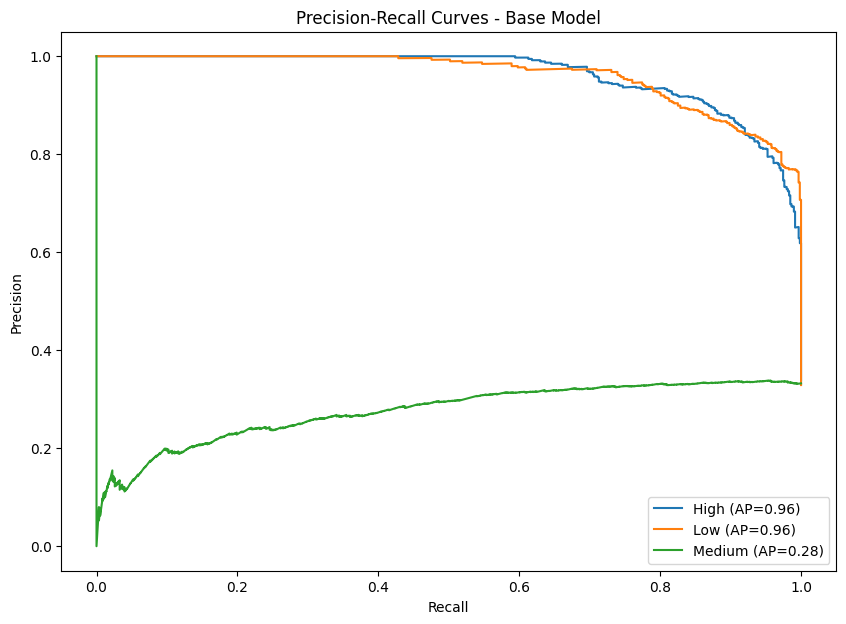


|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Class** | **Base Model** | | **Best Model** | | **Error Reduction** |
|  | Correct | Errors | Correct | Errors | (%) |
| **High** | 519 | 70 | 577 | 12 | 82.9% |
| **Low** | 507 | 65 | 557 | 15 | 76.9% |
| **Medium** | 433 | 145 | 536 | 42 | 71.0% |

**Improvement**: Massive reduction in misclassifications, especially for Medium class

**4. Precision-Recall Curves**

This illustrates the trade-off between precision and recall at different thresholds, helping evaluate model performance especially for imbalanced datasets.

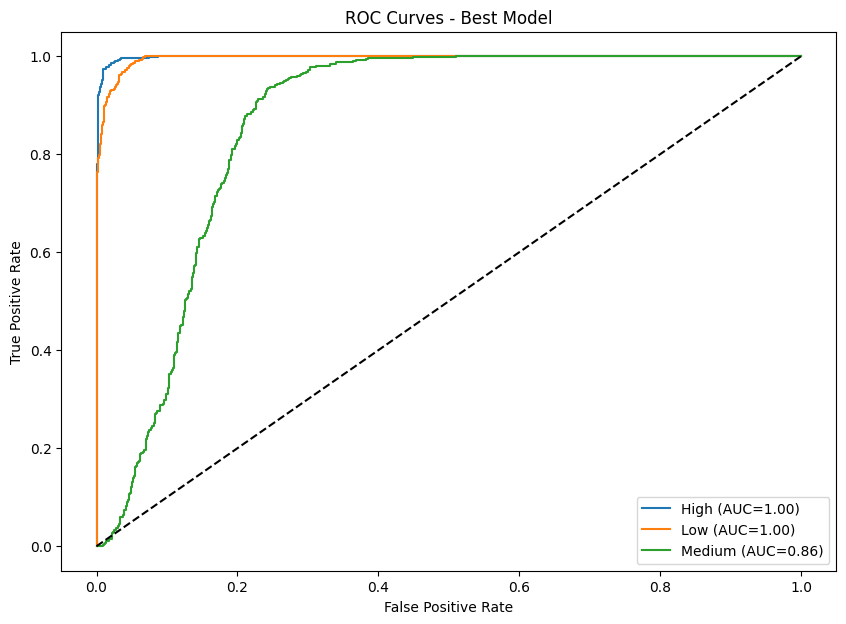
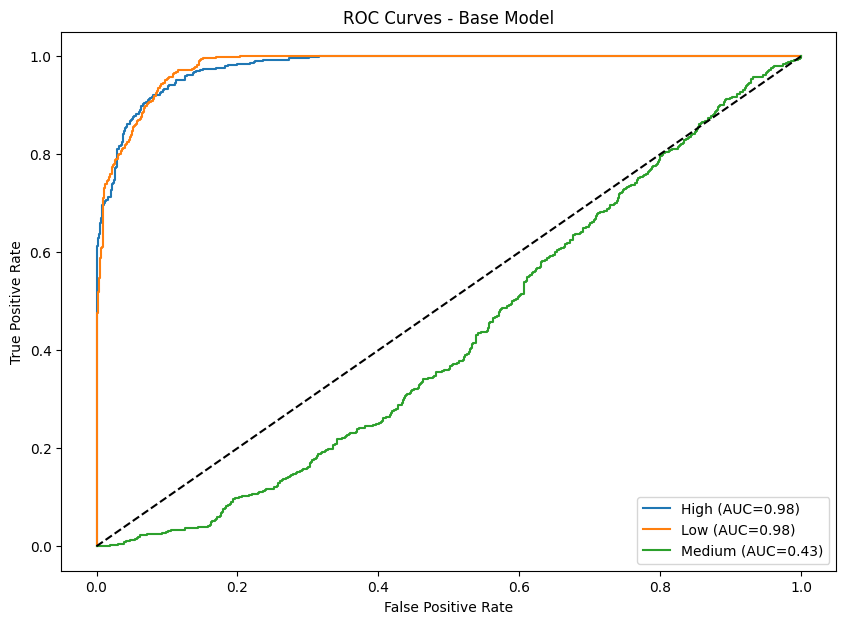


**Improvement:** All curves maintain high precision across recall range

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **Base Model (AP)** | **Best Model (AP)** | **Improvement** |
| **High** | 0.96 | 1.00 | +0.04 |
| **Low** | 0.96 | 0.99 | +0.03 |
| **Medium** | 0.28 | 0.62 | +0.34 |

**5. ROC Curves**

This show the relationship between true positive rate and false positive rate, providing insight into the model's discriminative ability across all classification thresholds.



|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **Base Model (AUC)** | **Best Model (AUC)** | **Improvement** |
| **High** | 0.98 | 1.00 | +0.02 |
| **Low** | 0.98 | 1.00 | +0.02 |
| **Medium** | 0.43 | 0.86 | +0.43 |

**Improvement:** All classes achieve significant discriminative ability

b.

(i) Optimization Techniques can be divided into 3 main categories

* Optimizing Gradient Descent –

This means making small changes to how algorithm updates weights, so that it will learn faster, refrain from getting stuck during training and reaches a better solution.

Optimizing Gradient Descent can be divided into 4.

* + Parameter Initialization
  + Gradient Descent with Momentum
  + Adaptive Learning Rates
  + SGD and Mini-batch GD
* Avoid Overfitting and Underfitting –

Overfitting means models try to learn too much detail/ noise from the training data. So, at the end model will well work with training data but not with unseen new data.

Underfitting on the other hand learns too little so it does bad with both training and new data.

Avoid Overfitting and Underfitting can be divided into 3.

* + L1 and L2 Regularization
  + Early Stopping
  + Dropout
* Avoid Vanishing/ Exploding Gradients –

Vanishing gradients means gradients get smaller during backpropagation. The effect is earlier layers learn barely and learning process becomes slow.

On the other hand, Exploding Gradients means gradients become very large during backpropagation. The effect is model’s parameters oscillate wildly there by model gets unstable.

Avoid Vanishing/ Exploding Gradients can be divided into 2.

* + Gradient clipping
  + Batch Normalization

Following is the list of Optimization techniques with a description, strengths and weaknesses.

1. **Random Initialization**

This is a method where small, random values are assigned to the weights of a neural network before training begins. This simple approach is based on the idea that starting with different, non-zero weights for each neuron.

Strengths:

* Easy to implement and very fast: This is its main advantage. It just needs a random number generator to sample weights from a small range, like a standard normal distribution.
* Works for simple networks: In shallow networks, the vanishing/exploding gradient problem is less likely to occur because there are fewer layers for the gradients to propagate through.

Weaknesses:

* Can cause vanishing or exploding gradients: If the initial weights are too large, the activations can grow exponentially with each layer, leading to exploding gradients. If they are too small, the activations and gradients can shrink to almost zero, leading to vanishing gradients.
* Training can be slow or unstable: Because the weights are randomly chosen without considering the network's structure, the starting point for training can be far from optimal. This can result in a longer training time and, in some cases, unstable learning.

1. **Xavier/Glorot Initialization**

This is a method for setting the initial weights of a neural network. Its goal is to keep the variance of activations consistent across all layers. This prevents the vanishing and exploding gradient problems, which can stall or destabilize the training process in deep neural networks.

Strengths:

* Stabilizes training in deep networks: By preventing gradients from becoming too small or too large, it allows deep networks to be trained more effectively.
* Works well with symmetric activations like sigmoid/tanh: It was specifically designed for these types of symmetric activation functions, as they have a mean of zero and a variance that is well-behaved within a certain range.

Weaknesses:

* Not ideal for ReLU-based networks.
* Slightly more complex than basic random initialization.

1. **He Initialization**

This is similar to Xavier, but it accounts for the fact that ReLU outputs zero for half of its inputs.

Strengths:

* It is highly effective in preventing vanishing gradients in deep networks that use ReLU.

Weaknesses:

* It's not suitable for activations like sigmoid or tanh because it doesn't account for their non-linear properties, which can cause exploding gradients. While it greatly helps, very deep networks might still benefit from techniques like Batch Normalization to further stabilize training.

1. **RMSProp**

For each weight, Root Mean Square Propagation maintains a moving average of the squared gradients. The learning rate is then divided by the square root of this average. Making smaller updates for weights with large gradients (that change frequently) and larger updates for weights with small gradients (that don't change much), happened as an effect of this. This process helps the training process.

Strengths:

* It is well-suited for non-stationary problems, where the characteristics of the data change over time, and it effectively handles noisy or sparse gradients.

Weaknesses:

* Like many adaptive optimizers, it can be sensitive to the initial learning rate choice.
* If the learning rate is not set correctly, it might struggle to find the global optimum and could converge to a suboptimal solution.

1. **Adam (Adaptive Moment Estimation)**

Adam is a widely used optimization algorithm that combines the best of both worlds. Which are momentum from algorithms like SGD and adaptive learning rates from algorithms like RMSProp. It keeps track of a decaying average of past gradients (like momentum) and a decaying average of past squared gradients (like RMSProp). This makes it fast, efficient, and robust for most deep learning problems.

Strengths:

* It's often the go-to optimizer because it's fast and requires minimal tuning. It performs well "out of the box" for a wide range of tasks.

Weaknesses:

* In some specific cases, Adam can overfit, especially if not used with regularization.
* While it converges quickly, it may not find the absolute best minimum compared to other more finely tuned optimizers in certain situations.

1. **Learning Rate Decay**

Learning rate decay is a strategy used with optimizers to gradually reduce the learning rate over time. The idea is to make large updates at the beginning of training to quickly move toward the general area of a minimum, and then to make smaller, more precise updates later to fine-tune the solution. This helps prevent "overshooting" the minimum.

Strengths:

* It helps the model converge more stably and precisely.
* By reducing the learning rate, it avoids the problem of bouncing back and forth across a minimum and allows for a more subtle approach to finding the optimal solution.

Weaknesses:

* The decay schedule itself is a hyperparameter that needs to be tuned.
* If the learning rate is decayed too quickly, the model might get stuck in a suboptimal solution. If it's decayed too slowly, it might not converge effectively.

1. **SGD and Mini-batch Gradient Descent**

These are fundamental optimization methods. Stochastic Gradient Descent (SGD) updates the model's weights after processing a single training sample. This makes the updates very noisy but can be fast since each step is computationally cheap. Mini-batch Gradient Descent, on the other hand, updates the weights after processing a small "mini-batch" of samples. This is the most common approach in modern deep learning.

Strengths:

* Mini-batch GD provides a good balance between the computational efficiency of SGD and the stability of Full-batch Gradient Descent (which uses the entire dataset for each update). The gradients are less noisy than with pure SGD, which leads to more stable convergence.
* Both methods are memory-efficient, as they don't require loading the entire dataset into memory for a single update step.

Weaknesses:

* Pure SGD is noisy and can have unstable training because of the high variance in the gradients from a single sample.
* Using a very large batch size in Mini-batch GD can require a lot of computational power and may sometimes lead to poorer generalization.

1. **L1 Regularization**

Also known as *Lasso* and this adds a penalty to the loss function that is proportional to the absolute value of the weights. This penalty can push some weights to exactly zero, effectively removing the corresponding features from the model. This is why it's great for automatic feature selection and sparse datasets.

**Strengths:**

* Simplifies the model with automatic feature selection.
* Great for sparse datasets.

**Weaknesses:**

* Can underfit if the penalty is too strong.
* Harder to optimize with some algorithms.

1. **L2 Regularization**

Also known as *Ridge* and this adds a penalty to the loss function that is proportional to the squared magnitude of the weights. This penalty encourages the weights to be small but rarely pushes them to exactly zero. It helps to prevent a model from relying too heavily on any single feature but doesn't perform feature selection.

**Strengths:**

* Prevents overfitting effectively.
* Works well with most models and optimizers.

**Weaknesses:**

* Doesn’t perform feature selection.
* Needs careful tuning of the regularization factor.

**7️⃣ Early Stopping**

**Simple Explanation:** Stops training when validation performance **stops improving**.  
**Strengths:**

* Easy to use and implement.
* Effectively prevents overfitting.  
  **Weaknesses:**
* Might stop too early and undertrain the model.
* Requires a clean validation strategy.

**8️⃣ Dropout**

**Simple Explanation:** Randomly **turns off some neurons** during training.  
**Strengths:**

* Reduces overfitting significantly.
* Encourages robust and generalized learning.  
  **Weaknesses:**
* Slows down convergence.
* Needs careful tuning of the dropout rate.

**9️⃣ Gradient Clipping**

**Simple Explanation:** Caps gradients so they **don’t explode** during training.  
**Strengths:**

* Stabilizes training, especially in RNNs and deep networks.
* Simple and easy to apply in most frameworks.  
  **Weaknesses:**
* Doesn’t solve vanishing gradients.
* Requires tuning the clipping threshold carefully.

**🔟 Batch Normalization**

**Simple Explanation:** Normalizes outputs in each layer to keep learning stable.  
**Strengths:**

* Speeds up convergence.
* Reduces sensitivity to initialization and learning rates.  
  **Weaknesses:**
* Adds extra computation overhead.
* Might not help much in small or shallow models.

(ii)

**Critical Role of Hyperparameter Tuning**

Hyperparameter tuning means **adjusting the “settings” of your neural network** to help it **learn better, faster, and more accurately**.  
Unlike weights, these hyperparameters **don’t get learned automatically** — **you set them** before or during training.

**1️⃣ Learning Rate (LR)**

**What it is:** Controls **how big a step** the model takes when updating weights.

* **If LR is too high:**
  + Model jumps around and **can’t converge**.
  + Training may even **diverge**.
* **If LR is too low:**
  + Learning becomes **very slow**.
  + Model may **get stuck** in a poor local minimum.

**Tuning tip:** Start with values like 0.001 for Adam or 0.01 for SGD, then use **learning rate schedules or decay**.

**2️⃣ Batch Size**

**What it is:** The **number of samples processed at a time** before updating weights.

* **Small batch size (e.g., 32):**
  + Noisy but **helps generalization**.
  + Slower per epoch but **less memory needed**.
* **Large batch size (e.g., 512+):**
  + **Faster training** per epoch.
  + Can lead to **overfitting** or poor generalization.

**Tuning tip:** Start with **32 or 64** for most tasks, adjust based on GPU/CPU resources.

**3️⃣ Network Depth**

**What it is:** Number of **hidden layers** and **neurons** in the model.

* **Too shallow (not enough layers):**
  + Model **underfits** — can’t capture complex patterns.
* **Too deep (too many layers):**
  + Model **overfits** — memorizes training data.
  + Training may suffer from **vanishing/exploding gradients** without techniques like BatchNorm or skip connections.

**Tuning tip:** Start simple, then gradually **increase depth** and monitor validation performance.

**Why Hyperparameter Tuning is Critical**

* A well-tuned model can **train faster** and **generalize better**.
* Poor tuning leads to **unstable training**, **overfitting**, or **underfitting**.
* Small changes (e.g., LR from 0.01 → 0.001) can make a **huge difference** in performance.

**Common Strategies for Tuning**

* **Grid Search:** Try all combinations (good for small search spaces).
* **Random Search:** Sample random combinations (faster for large spaces).
* **Bayesian Optimization / Hyperband:** Smarter, guided search methods.

**Question 5: Explainable AI**

Explainable AI (XAI) is a set of methods and processes that allow humans to understand and trust the results generated as the output from machine learning algorithms. Especially modern deep learning and ensemble models are like black boxes, because it hard to understand by a human being how those models actually make decisions. XAI addresses this problem through below aspects.

* + Building Trust and Transparency
  + Regulatory Compliance
  + Debugging and Model Improvement
  + Detecting Bias and Ensuring Fairness
  + Facilitating Human-AI Collaboration
  + Safety and Reliability

**XAI methodologies**

1. SHAP (SHapley Additive exPlanations)

* This is a Post-hoc (Applying explanation methods to the model after it has been trained) type and Model-Agnostic method.
* SHAP is based on Shapley values from cooperative game theory.
* It assigns a contribution score to each feature by considering all possible combinations of features and how each contributes to the prediction.
* Explanation can be local (single prediction) or global (overall model behavior).

Strengths:

* Reliable and consistent.
* Supports any model type.
* Supports both local and global interpretability.

Limitations:

* Computationally expensive for large datasets or deep models.
* Requires expertise to interpret correctly.

2. LIME (Local Interpretable Model-agnostic Explanations)

* This is a Post-hoc type and Model-Agnostic technique.
* Focuses on local interpretability.
* It slightly perturbs the input data, and it sees how the prediction changes.
* Then, it constructs a simple surrogate model (such as linear regression) around this local area that is used to mimic the behavior of the complex model.

Strengths:

* Simple explanations.
* Works with any model type.
* Fast and easy to implement.

Limitations:

* Fragile to perturbations and parameters.
* Supports only for local interpretability, not global understanding.
* May be misleading if the local approximation is not accurate.

3. Counterfactual Explanations (and Adversarial Examples)

* This is a Post-hoc type and Model-Agnostic method.
* Counterfactual Explanations (CEs) illustrate how a small change in an input can result in a different desired outcome, such as converting a loan denial into an approval by altering the loan amount. Adversarial Examples (AEs), on the other hand, are input modifications that are frequently undetectable and intended to intentionally trick or mislead an AI model into generating an inaccurate prediction.

Strengths:

* Highly interpretable and actionable.
* Works across models (model-agnostic).

Limitations:

* May generate unrealistic counterfactuals.
* Computationally expensive for large models.

4. Layer-wise Relevance Propagation (LRP)

* This approach is model-specific and post-hoc.
* To determine how much each input feature contributed to the prediction, LRP breaks down the model's output back through the network layers.
* Generates relevance scores or heatmaps that show how important each input is.

Strengths:

* Helpful for bias detection and debugging.
* Offers comprehensive, layer-level insights into intricate deep learning models.

Limitations:

* Requires access to the model's internals, such as weights and gradients.
* Mostly restricted to architectures of neural networks.

**Question 6: GenAI**

a.

b.

c.

Here is the summary of performance analysis between traditional ANN and GenAI models.

|  |  |  |
| --- | --- | --- |
| **Aspect** | **Traditional ANN** | **GenAI Models** |
| Accuracy | Accuracy is high as it is trained on well-prepared data. | Comparable or higher accuracy if fine-tuned through prompt engineering and combining structured with unstructured data. |
| Generalization | Limited to patterns exist within the training data. | Strong generalization. Though it is noisy or partially missing values and due to pre-training on vast amounts of data. |
| Speed | Very fast for both in training and inference (even on CPU). | Slower inference. This requires GPU or optimized CPU setup for practical use. |
| Scalability | Easy to scale with larger datasets by increasing layers or nodes. | Scales well but at a much higher computational cost. |
| Adaptability | Needs retraining when new features or data types are introduced. | Flexible. GenAI can adapt to multiple modalities (text, images, structured data etc.) with minor fine-tuning or prompt engineering. |
| Explainability | Requires external tools like SHAP or LIME for interpretability. | This can generate natural language explanations for results natively. |

**2. Pros and Cons**

**Traditional ANN**

**Pros**

* Lightweight and efficient, works well even with limited hardware.
* High accuracy for purely numerical/tabular data.
* Low cost as no need for GPUs or large cloud infrastructure.
* Simple to implement and maintain in Python pipelines.
* Deterministic outputs, making it stable and reliable.

**Cons**

* Poor interpretability without external tools.
* Rigid structure: struggles with text, images, or unstructured data.
* Needs large, labeled datasets for high performance.
* Requires manual feature engineering to optimize inputs.

**Generative AI (GenAI)**

**Pros**

* Multi-modal capability can handle text, images, and structured data simultaneously.
* High adaptability as fine-tunes well for new scenarios.
* Great explainability, can explain predictions in natural language.
* Data augmentation can generate synthetic samples for rare quality grades.
* Supports interactive tools like conversational assistants for data analysis.

**Cons**

* Resource-intensive as it needs powerful hardware (GPU) to perform efficiently.
* Slower inference compared to ANN.
* Complex deployment, as it requires model serving pipelines and possibly API integrations.
* Higher costs especially for training or hosting large models.

Here are some opportunities that can be identified where GenAI models could offer novel solutions or enhance existing AI systems regarding cinnamon quality grading system.

**1. Multimodal Quality Grading**

Without limiting to numeric sensor data (like moisture, density, chemical composition) with images of cinnamon quills or powder to make grading more precise. GenAI can help by using vision-language models (like CLIP or Florence-2) to extract visual features and combine them with ANN predictions. For example, identifying defects (e.g., fungal spots, cracks) that numeric sensors alone cannot detect, provide human-readable explanations which explain the status of the sample.

**2. Automated Expert Report Generation**

Instead of just giving a grade (Low, Medium or High) as the response, provide human-like inspection summaries automatically. Example LLMs like LLaMA-3, DeepSeek, or Mistral can generate summaries. This reduces manual effort for quality inspectors.

**4. Real-Time Conversational Assistance**

Create an AI assistant for field operators and quality managers. For an example LLMs integrated with cinnamon grading system could explain reasons behind grades, suggest actions to improve quality and answer troubleshooting questions interactively.

**5. Predictive Quality Analytics**

Move from just classification to prediction, Gen AI can be used to forecast future quality trends. For example, combine time-series forecasting models with LLMs to analyze patterns in weather, soil data, and production logs. And generate proactive recommendations.