CHAPTER I
INTRODUCTION

1.1 Background of the Study

Soybean (Glycine max) occupies a strategic position within Indonesia's agricultural and food processing sectors [1]. As a key plant-based protein source, it underpins the production of widely consumed foods such as tempeh, tofu, and soy milk, and is also used as a raw material for livestock feed [2]. This protein content makes soybean an important contributor to public nutrition, especially for people who limit their intake of animal-derived protein [3]. Given this role, checking seed quality before the production stage becomes an essential step to guarantee that the raw material used meets the required standards [4].

Quality assessment of soybean seeds typically relies on visual cues such as color, shape, and overall physical condition [5]. From these cues, seeds are commonly grouped into categories such as intact, broken, skin-damaged, spotted, and immature [6]. The quality of soybean seeds itself varies due to a combination of factors that occur throughout the growth cycle and after harvest [7]. ranging from pest infestation and mechanical damage during harvesting to soil condition, humidity levels, and temperature fluctuations in the surrounding environment [8].

At present, soybean seed grading is still largely carried out manually through visual inspection performed by human workers [9]. This conventional approach tends to be subjective, prone to inconsistency, and time-consuming, particularly when applied to large volumes of data [10]. At the same time, the visual traits found on soybean seeds actually carry valuable information that could support a more systematic classification process [11]. This gap points to the need for a classification method that is not only accurate but also efficient and reliable enough to deal with the wide range of visual variation found in soybean seeds variation that stems from biological, mechanical, and environmental influences alike [8].

The growth of Artificial Intelligence, particularly in the area of deep learning, opens up the possibility of replacing manual grading with automated systems that are more objective and consistent [12]. Among the techniques most frequently applied for this purpose is the Convolutional Neural Network (CNN), which is well known for its ability to extract visual features from image data efficiently [13]. A number of earlier studies have shown that CNN-based methods can be applied successfully to classify

the quality of soybean seeds. One such study, conducted by Reza Al Husna and Tubagus Maulana Kusuma in 2025 [14]. applied the EfficientNet-B0 architecture to a dataset containing 5,513 soybean seed images. Using the Adam optimizer with a learning rate of 0.001 and a batch size of 32, the model achieved a classification accuracy of 95%. This finding suggests that EfficientNet-B0 is capable of delivering strong results when applied to soybean seed quality classification tasks.

A related study by Ning Zhang et al. (2023) [15]. examined a different set of CNN architectures AlexNet, VGG19, ResNet101, and DenseNet121 while also addressing data imbalance through augmentation along with loss function adjustments based on focal loss and label smoothing. The use of focal loss was found to meaningfully boost testing accuracy, with DenseNet121 reaching an accuracy of 98.48%.

Beyond architecture selection, pairing CNNs with more advanced data augmentation strategies has also been shown to strengthen model performance and generalization within agricultural applications. Rohit Maheshwari, Amit Sharma, Seema Kumari Nagar, and Jagdeesh Prasad Meena (2025) [16]. for instance, used EfficientNet-B0 as a pretrained feature extractor together with Mixup and Cutmix augmentation to classify diseases on plant leaves. Their proposed model reached a validation accuracy of 95.11%, reinforcing the idea that combining EfficientNet with sophisticated augmentation methods can improve both performance and generalization in agricultural image classification.

Hyperparameter tuning has likewise been shown to improve CNN model outcomes. Sri Rahayu and Sayyid Faruk Romdoni (2025) [17]. explored this angle by applying Bayesian Optimization to a pretrained CNN built for predicting mango leaf disease. Working with the MangoLeafBD dataset 4,000 images spanning 8 classes, processed through resizing, normalization, and partitioning into training, validation, and testing sets—they compared several pretrained CNNs, including DenseNet121, ResNet50V2, MobileNetV3 Small, MobileNetV3 Large, and InceptionV3. MobileNetV3 Large produced the strongest results, reaching a baseline accuracy of 0.995 that climbed to a perfect 1.0 once Bayesian optimization was applied, yielding predictions that were both more accurate and more efficient for mango leaf disease classification.

Taken together, these findings suggest that there is still meaningful room to improve how soybean seed quality is classified. Building on that opportunity, this study introduces an image-based classification approach centered on the EfficientNet-

B0 architecture. EfficientNet-B0 was chosen for its lightweight yet efficient design, which tends to perform well even on datasets of small to medium size. That said, the performance of any deep learning model remains heavily dependent on hyperparameter choices and how varied the available training data is.

To work around these constraints, this study pairs Bayesian Optimization used to systematically search for the most suitable hyperparameter configuration, since hyperparameter effectiveness tends to be data-dependent with the Mixup technique, an interpolation-based augmentation method. Mixup is applied on-the-fly during training, functioning as a regularization mechanism that increases variation in the training data, helps curb overfitting, and improves the model's ability to generalize. By bringing together the EfficientNet-B0 architecture, Bayesian-optimized hyperparameters, and Mixup augmentation, this research aims to build a soybean seed quality classification model that performs more accurately, stably, and consistently than previous approaches. The resulting best-performing model will then be deployed in a web-based system capable of automatically classifying soybean seed quality from uploaded images. Based on this overall direction, the study is titled "Optimization of Soybean Quality Classification Using EfficientNet-B0 with Mixup and Bayesian Hyperparameter Optimization."

1.2 Formulation of the Problem

Drawing from the background outlined above, this study addresses the following research questions:

How can the Mixup technique and Bayesian Optimization-based hyperparameter tuning be implemented on the EfficientNet-B0 architecture for classifying soybean seed images?

How does the classification performance for soybean seed quality differ across the various testing scenarios, as measured through the confusion matrix, recall, precision, and F1-score?

How can the best-performing model identified through testing be deployed into a website-based system for automatic soybean seed quality detection?

1.3 Research Objectives

In line with the problem formulation above, this study sets out to:

1. Analyze how applying the Mixup technique and Bayesian Optimization-based hyperparameter tuning affects the performance of EfficientNet-B0.
2. Assess model performance using the confusion matrix, recall, precision, and F1-score.
3. Deploy the best-performing model into a website to enable fast and efficient soybean seed quality detection.

1.4 Research Significances

Based on the objectives above, this research is expected to offer the following benefits:

1. Providing insight for the agricultural and food industry sectors into how deep learning can support automated, objective soybean seed quality classification.
2. Offering a clearer picture of how Bayesian Optimization-based tuning of EfficientNet-B0 can enhance soybean seed quality classification performance.
3. Clarifying the extent to which the Mixup data augmentation technique influences the performance of the EfficientNet-B0 model.
4. Serving as a reference point for future research exploring deep learning-based approaches to soybean seed quality classification.

1.5 Scope and Limitation

This study is bounded by the following scope and limitations:

1. The dataset combines secondary data sourced from Kaggle, titled "Soybean SeedsClassificationDataset"(https://www.kaggle.com/datasets/aryashah2k/soybean-seedsclassification-dataset), with primary data consisting of 15 images captured directly by the researcher to test the model's implementation within the web application.
2. The dataset is organized into five soybean seed quality classes intact, spotted, immature, broken, and skin-damaged comprising a total of 5,513 secondary images alongside 15 primary images used specifically for testing the web application.
3. The research object is restricted to imported soybean seeds under standard conditions, and therefore does not reflect the characteristics of locally grown soybean seeds.

4. Hyperparameter optimization is confined to Bayesian Optimization using a Gaussian Process (GP) surrogate model paired with an Expected Improvement (EI) acquisition function.
5. Model training is carried out using the TensorFlow framework with Keras as the high-level API, within the Google Colaboratory (Google Colab) computing environment.
6. The model is deployed as a simple web service built with FastAPI, functioning as a soybean seed quality classification system, with its dataset usage limited strictly to the data used in this research.