Project Proposal

Machine Learning
BE2100 - Section 003
Group 7 - Members:

- Lokman Nasri
- Abdoul Bah
- Brandon Dong
- Akrm Al-Hakimi
- Caitlin Mann
- Ali Slim
- Meenaz Shaikh
- Andriy Babiychuk

Image Classification for Defective Materials

Objective

Develop a supervised machine learning model to automatically classify industrial casting products as *good* or *defective*. The aim is to support quality control processes by detecting defects with high accuracy.

Dataset

We will use the publicly available **Real-Life Industrial Dataset of Casting Products** (Kaggle link).

- Contains images of casting parts labeled as “defective” or “okay.”
- Provides a realistic setting for quality assurance in manufacturing.

**Methodology*

1. **Data Pre-processing**:
   - Clean and normalize images.
   - Split into training, validation, and testing sets.
2. Model Training
   - Use existing image-classification libraries such as scikit-learn, TensorFlow/Keras, or PyTorch.
   - Start with simple, pre-built architectures (e.g., MobileNet, ResNet, or a basic CNN template) rather than designing networks from scratch.

- Focus on selecting an appropriate model, adjusting straightforward parameters (learning rate, batch size), and applying transfer learning if needed.

Keeping the pipeline lightweight so it's easy to reproduce and debug within the semester is important.

1. **Evaluation & Analysis**:
   - Assess accuracy, precision, recall, and F1-score (a rating of how our classification model performs).
   - Construct a confusion matrix to visualize performance(A simple table that compares our model's predictions with the actual labels).
   - Use hypothesis testing (e.g., comparing baseline vs. model accuracy) to validate significance.
2. **Reporting**
   - Summarize findings, key metrics, and insights for a final report and presentation.

Expected Outcome
A robust image-classification system that can detect defective materials with measurable accuracy, along with a clear statistical analysis of model performance.

Scope & Feasibility

- Data is readily available and labeled, reducing collection overhead.
- The workflow (data cleaning → model training → evaluation) is straightforward and entirely possible within the given timeline this semester.
- Since we will use established libraries and pre-built algorithms, the heavy mathematical work (e.g., deriving formulas or building networks from scratch) will be handled by existing tools. Thus, most of the cognitive load will be on providing results for the final project.