Data Sheet: Fabric Stain Classification Dataset (v1.0)

1. Dataset Overview
This dataset was created for the purpose of training and evaluating an image classification model that identifies different types of stains on fabric or textile surfaces.

- **Dataset name**: Fabric Stain Classification Dataset
- **Version**: v1.0
- **Task**: Multi-class image classification
- **Data type**: RGB images of fabric/textile surfaces

Classes
The dataset contains the following nine classes:

1. coffee
2. wine
3. tomato_sauce
4. ink
5. chocolate
6. blood
7. juice
8. dirt_mud
9. clean

---

2. Data Sources and Collection
Images were collected from a combination of publicly available web sources and manually curated
examples for academic use.

- **Collection method**: Web search and manual selection
- **Capture conditions**:
  - Varying lighting conditions
  - Different fabric textures and colors
  - Multiple stain shapes and sizes

No personal or identifiable information is present in the dataset.

---

3. Licensing and Usage
- The dataset is intended **strictly for academic and educational purposes**.
- Original image licenses vary depending on the source.

- Where explicit license information was unavailable, images were used under the assumption of
  non-commercial, academic fair use.
- Redistribution or commercial use is **not recommended** without further license verification.

---

4. Labeling Process
Each image was manually assigned a single class label based on visual inspection.
Label definitions were guided by stain color, texture, thickness, and spread pattern.

Detailed label definitions are provided in `label_definitions.txt`.

Labeling Guidelines
- Only one dominant stain type is labeled per image.
- Images with no visible stain are labeled as `clean`.
- Ambiguous or low-quality images were excluded from the final dataset.

---

5. Dataset Split
The dataset was split into training and validation sets prior to model training.

- **Training set**: 80%
- **Validation set**: 20%

To prevent data leakage:
- Duplicate or near-duplicate images were removed across splits.
- Images from the same original source were assigned to only one split when possible.

---

6. Preprocessing and Augmentation
Before training, images underwent the following preprocessing steps:

- Resizing to a fixed input resolution
- Pixel normalization
- Basic data augmentation during training:
  - Random horizontal flip
  - Small rotations
  - Minor color jitter

No augmentation was applied during validation.

---

7. Known Limitations and Biases
- Some stain categories have visual similarities (e.g., coffee vs. chocolate, juice vs. wine).
- Background fabric color may influence model predictions.
- Very small, faint, or partially cleaned stains are more difficult to classify accurately.
- Real-world stains with mixed substances are not well represented.

---

8. Intended Use
- Academic coursework and demonstrations
- Prototype computer vision applications
- Model evaluation and benchmarking

---

9. Not Intended Use
- Medical or forensic analysis (e.g., blood identification)
- Safety-critical or legal decision-making
- Commercial laundry or industrial deployment without further validation

---

10. Ethical Considerations
This dataset does not contain personal data or human subjects.
All images were collected and used solely for educational and research purposes.

---

11. Contact and Maintenance
This dataset was created and maintained as part of a university project.
Future versions may expand class balance and improve coverage of real-world stain conditions.