United States International University

Department of Data Science & Analytics

Project Report

Traffic Density Classification on Kenyan Roads Using Image-Based Machine Learning

Submitted by: Mitchelle Moraa
Course: DSA 2020- Artificial Intelligence
Instructor: Dr. Edward Ombui
Date: July 17, 2025

1. Project Overview

This project explores the application of machine learning techniques to classify traffic density in images captured from Kenyan roads. The core objective is to develop a binary image classification model capable of distinguishing between high-traffic and low-traffic scenes based on visual cues.

This initial phase lays the foundation for a broader traffic monitoring system that can be integrated into urban planning, smart infrastructure, or intelligent transport systems (ITS) in Kenya. It offers a low-cost, scalable approach to road usage monitoring, contributing to data-driven traffic management strategies in developing urban environments.

2. Objectives

The core objectives of this project are as follows:

- To compile a diverse dataset of traffic images representing both low and high congestion scenarios across Kenyan roadways.
- To perform image preprocessing
- To train and evaluate a machine learning classifier to differentiate between traffic levels.
- Test and Analyze
- Interpret predictions using metrics such as ROC AUC and confusion matrix
- To assess the classifier's performance and potential applicability to real-world scenarios.

3. Methodology

Image data was sourced through a combination of automated tools, community-driven contributions, and institutional archives to ensure both diversity and contextual relevance to Kenyan road conditions. The objective was to obtain a balanced and representative dataset for binary traffic density classification.

Sources:

- Automated Web Scraping:
  The `icrawler` Python library was used to programmatically download images via the Bing Image Search API. Keywords such as *"Nairobi traffic jam"*, *"empty Kenyan roads"*, and *"Kenya road congestion"* were employed to retrieve geographically relevant images.
- Crowdsourced Contributions:
  Images were collected from public forums including Reddit, X (formerly Twitter), and Telegram groups dedicated to Kenyan urban issues.
- Media Archives:
  News outlets, online newspapers, and blog posts offered a rich source of contextual traffic images. Particular attention was paid to date stamps and location identifiers to confirm authenticity.
- NGO and Public Sector Platforms:
  Archival material from platforms such as the World Health Organization (WHO) and Kenyan urban planning portals provided access to mobility data in image form.

Dataset Summary:

- Total Images Collected: 229
  - High Traffic: 119 images
  - Low Traffic: 120 images

Each image was manually reviewed and stored in `.jpg` format to maintain consistency.

---

4. Image Preprocessing

In preparation for training a deep learning model on the collected traffic imagery, a comprehensive image preprocessing pipeline was implemented to ensure data quality, consistency, and compatibility with convolutional neural network (CNN) architectures such as VGG16 and ResNet.

Key preprocessing and cleaning steps:

- Image Resizing: All images were resized to 224 × 224 pixels, matching the input dimension requirements of standard pretrained CNN architectures

- Normalization: Pixel values were normalized to a 0–1 scale by dividing by 255. This helps accelerate model convergence and stabilize training.
- Labeling and Directory Structure: Images were organized into binary-labeled directories:
  - `1_high_traffic/` for high-density traffic scenes
  - `0_low_traffic/` for low-density road conditions
- Duplicate Image Removal: Using image hashing techniques, duplicates were identified by comparing hash values. Redundant files were deleted to prevent data leakage and overfitting.
- Quality Filtering:
  - Blurry/Low-Quality Images: Detected using Laplacian variance thresholds and removed.
  - Corrupted Files: Scanned and discarded using file integrity checks and PIL exception handling.
  - Overly Bright or Underexposed Images: Detected through pixel intensity histograms and filtered out to ensure visual clarity.
- Pixel Variation Check: Images with minimal pixel variance, blank or near-monochromatic backgrounds were excluded to enhance dataset informativeness.
- Dataset Integrity Checks: A count of images per class was conducted to ensure class balance, preventing bias during model training.
- Data Structuring for Model Input:
  - All processed images were converted into a single Numpy array for efficient input handling.
  - Batching and shuffling were applied using PyTorch's `DataLoader` to promote randomness during training and reduce variance.
- Color and Channel Standardization:
  - Ensured to remove RGB channel configuration.
  - Increased resolution and contrast where appropriate to enrich feature visibility.

5. Tools & Technologies

The following tools and libraries were used throughout the project:

- Python (icrawler) – Automated image scraping from Bing.
- Manual Downloads – From Kenyan-based sources including:
  - Reddit forums
  - X (formerly Twitter)
  - Local news archives
  - Public sector platforms (e.g., WHO, NGO

- Pillow & OpenCV – Image resizing, filtering, blurriness checks, RGB correction.
- NumPy – Handling arrays and converting image sets into tensors.
- ImageHash – Used to identify and remove duplicate images

6. Significance of the Project

This project holds both practical and academic importance in addressing the pressing issue of traffic congestion in Kenyan urban environments. Traffic buildup in cities such as Nairobi contribute to significant economic losses, air pollution, and daily commuter stress. It proposes a low-cost, scalable alternative for real-time traffic density. It also makes visible the growing field of AI-driven urban analytics.

7. Challenges

Dead Links and Inaccessible Web Pages, a significant number of URLs were either broken or led to pages that no longer existed. This reduced the efficiency of automated data collection and required manual intervention to identify viable sources. *Limited Public Datasets for Kenyan Traffic*. Most existing public datasets are not representative of Kenyan roads and traffic patterns. As a result, additional effort was required to curate a locally relevant dataset through crowdsourcing and regional media. *Unsupported Images*, most images could not be scraped from the web which required collection to be done with manual efforts to get the images.

7. Conclusion

This project will successfully demonstrate the use of machine learning, specifically convolutional neural networks, in classifying traffic density from images of Kenyan roads. By curating a balanced, representative dataset and applying systematic preprocessing and model tuning techniques, a binary image classifier will be trained to distinguish between high-traffic and low-traffic scenes with promising accuracy.

The outcome will highlight the potential for cost-effective, vision-based traffic monitoring system. The project will also contribute academically by documenting a pipeline for image classification grounded in real-world data challenges.

While further improvements are possible; such as deploying the model in a real-time system the foundation laid by this project will prove that localized, data-driven traffic solutions are both feasible and impactful.