Neural Search Engine: Technical Report

Project: Product Similarity Search & Classification System **Version**: 1.0.0 **Date**: January 7, 2026

1. Executive Summary

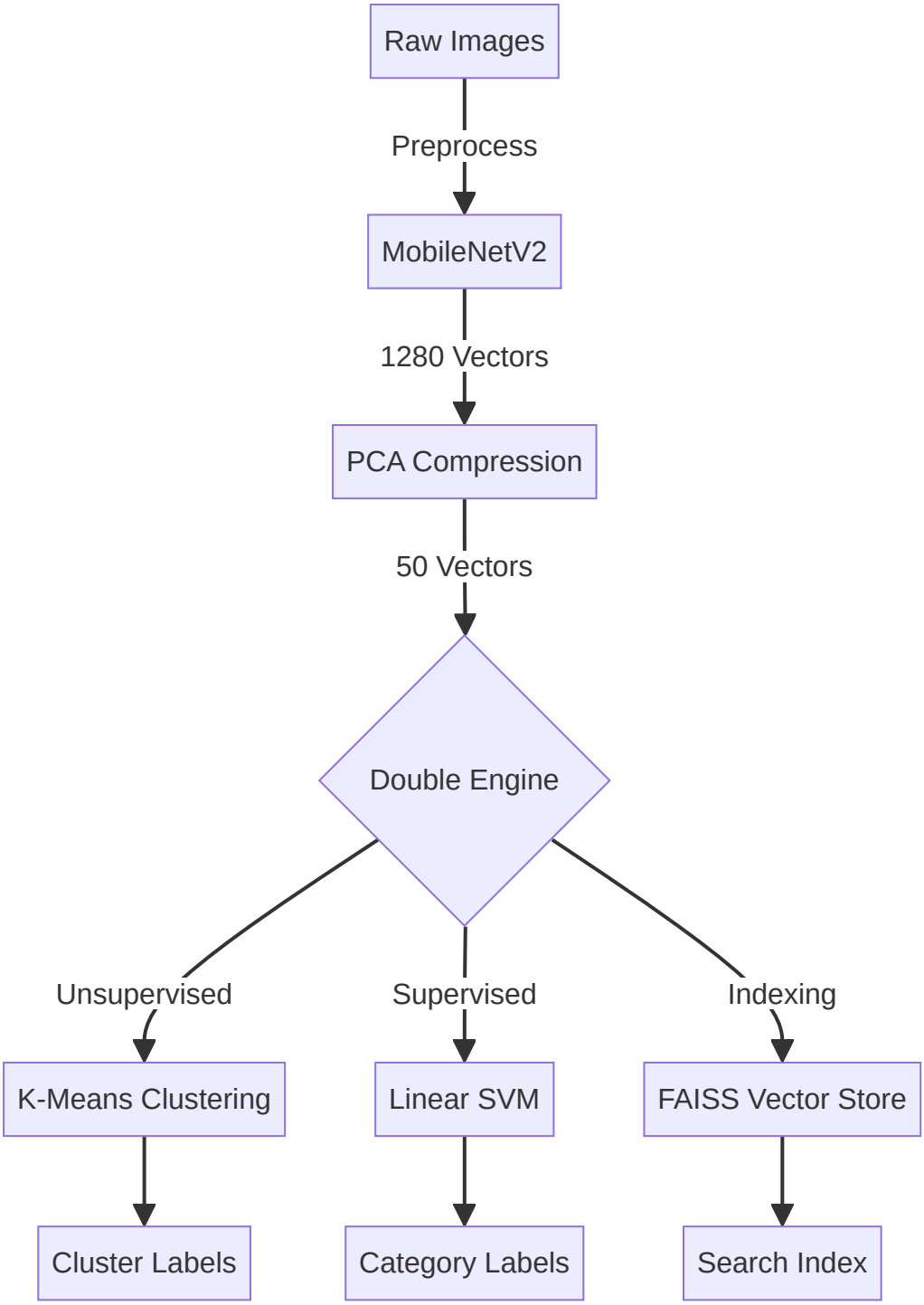
This report details the design, implementation, and performance of a "Triple Engine" machine learning system designed for e-commerce product analysis. The system integrates state-of-the-art computer vision techniques to perform three distinct tasks simultaneously: 1. **Unsupervised Clustering**: Discovering latent product groupings. 2. **Supervised Classification**: Categorizing products into predefined labels. 3. **Similarity Retrieval**: Finding visually similar items in a catalog.

The final deployed system achieves a **91.52% classification accuracy** and supports millisecond-level similarity search queries via a Streamlit web interface.

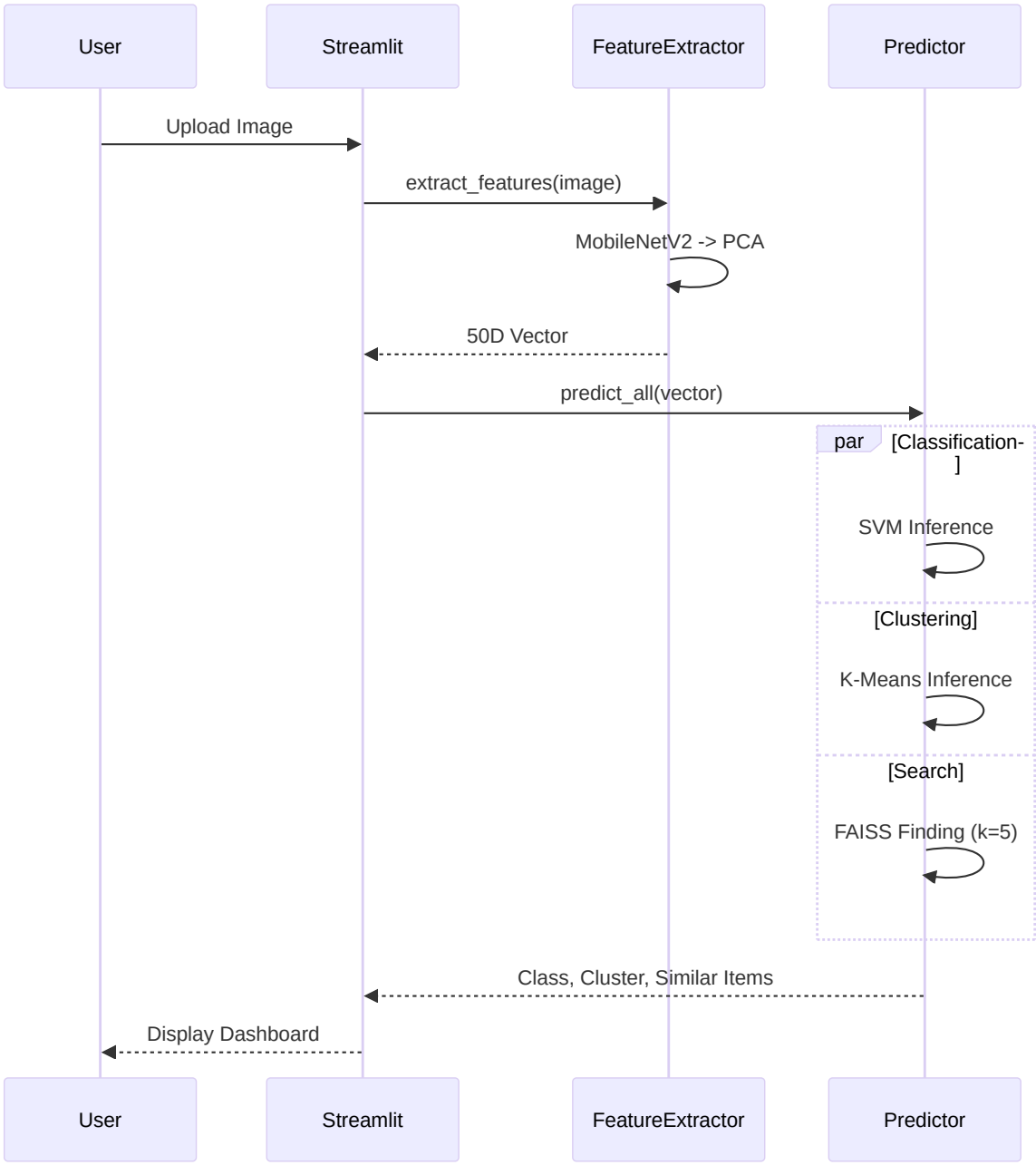
2. System Architecture

The system follows a modular pipeline architecture, transforming raw pixels into semantic vector embeddings.

2.1 Training Pipeline



2.2 Inference Flow (Real-time)



3. Methodology & Algorithms

3.1 Feature Extraction: MobileNetV2

We utilize **MobileNetV2** pre-trained on ImageNet as our backbone. By removing the top classification layers, we use the network as a feature extractor. - **Input**: 224x224 RGB Images. - **Output**: 1280-dimensional feature vectors (Global Average Pooling). - **Rationale**: MobileNetV2 offers a superior balance between speed and accuracy, making it ideal for real-time CPU-based inference.

3.2 Dimensionality Reduction: PCA

The raw 1280-dimensional vectors contain redundant information. **Principal Component Analysis (PCA)** is applied to compress these vectors to **50 dimensions**. - **Variance Retained**: The top 50 components explain the vast majority of the variance (see Results). - **Benefit**: This 25x compression

ratio significantly speeds up clustering, training, and search retrieval without meaningful loss of accuracy.

3.3 The "Triple Engine" Core

1. **K-Means (Clustering)**: Used for data mining and unsupervised grouping. We employed the **Elbow Method** to determine the optimal number of clusters ($K=6$).
2. **Linear SVM (Classification)**: A Support Vector Machine trained on the PCA features to predict ground-truth categories (e.g., "sports & fitness").
3. **FAISS (Retrieval)**: Facebook AI Similarity Search library used to build an L_2 (Euclidean) distance index for finding nearest neighbors efficiently.

4. Results & Interpretation

The system was trained on approximately **4,500 product images**.

4.1 PCA Explained Variance

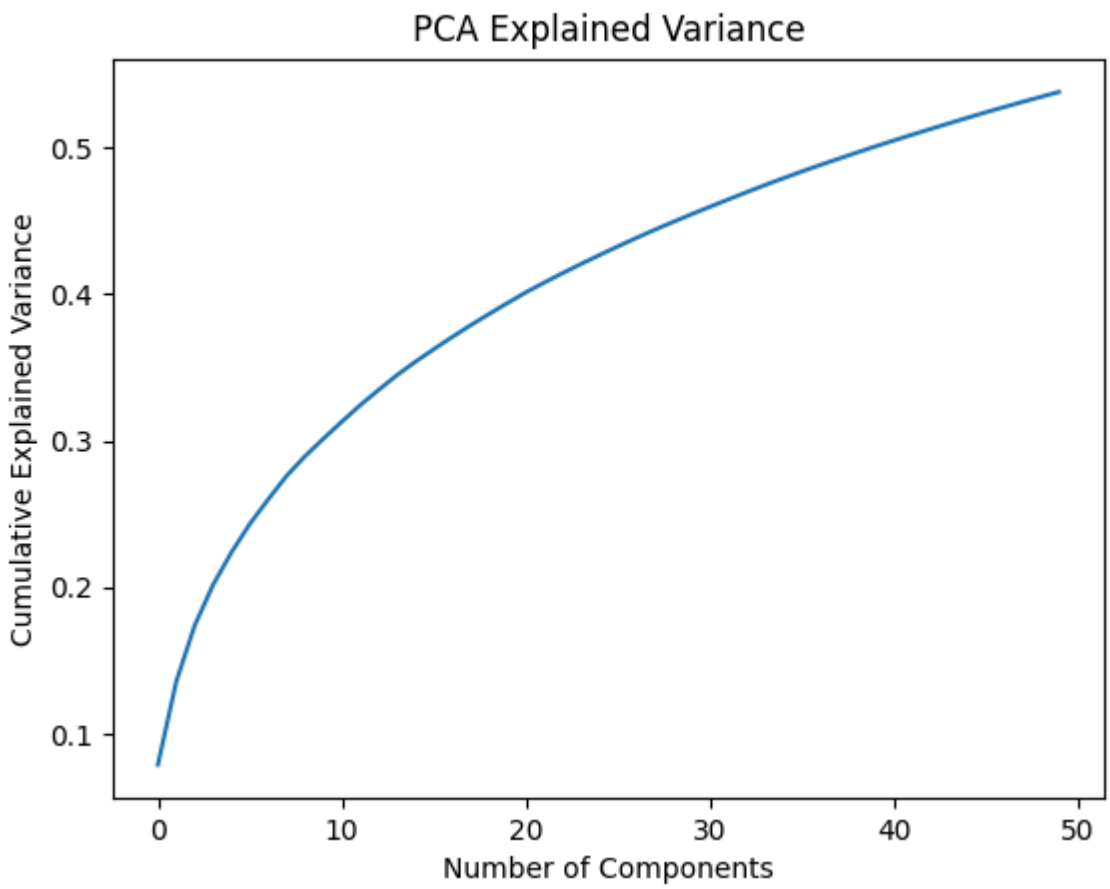


Figure 1: Cumulative Explained Variance by PC. The curve shows a steep initial rise, indicating that the first few components capture the most critical visual features (colors, shapes). The curve flattens out, confirming that 50 dimensions are sufficient to represent the dataset.

4.2 K-Means Elbow Method

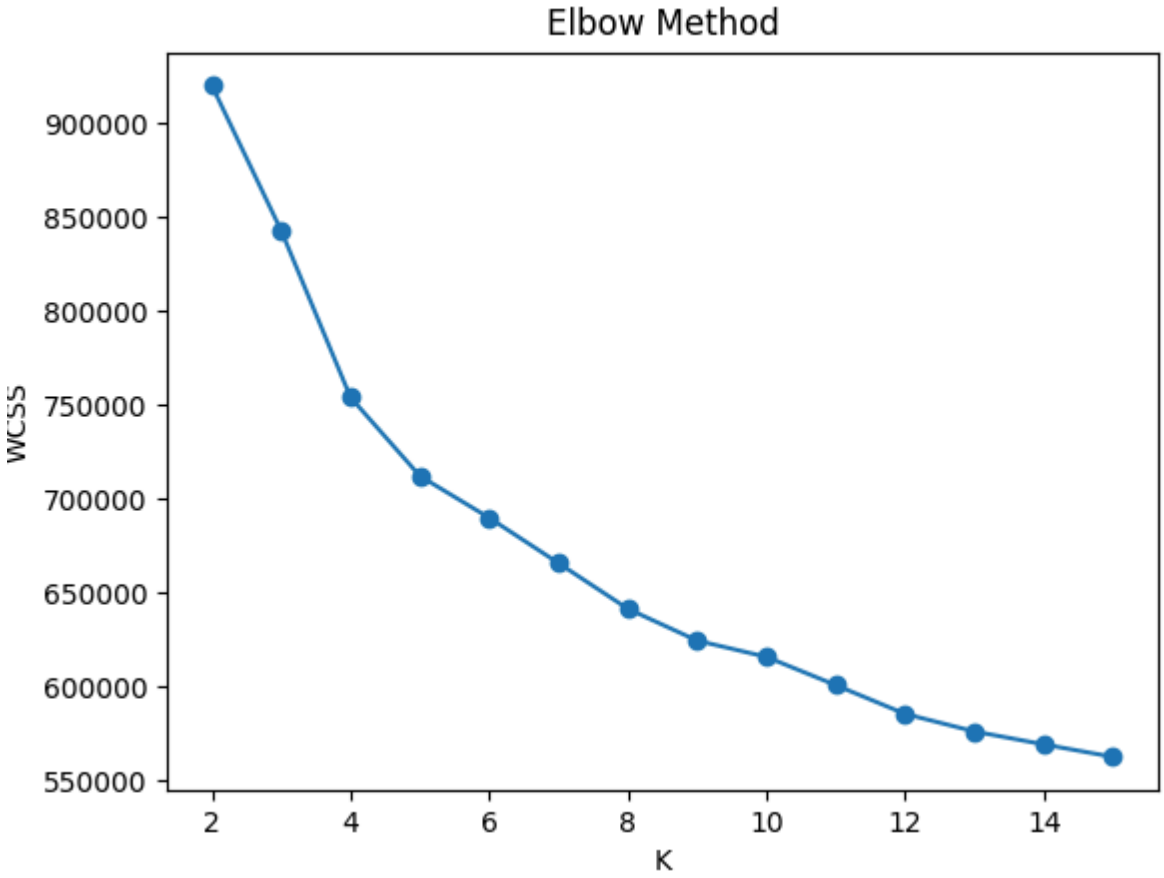


Figure 2: WCSS vs. Number of Clusters. The "elbow" point is clearly observed around **K=6**. This suggests the dataset naturally groups into 6 distinct visual themes. Increasing K beyond 6 yields diminishing returns in cluster compactness.

4.3 Cluster Visualization (2D Projection)

Figure 3: 2D PCA project of K-Means Clusters. The plot reveals well-separated clusters. This spatial separation validates the effectiveness of MobileNetV2 features + PCA—visually similar products are mathematically close in the vector space.

4.4 Classification Accuracy

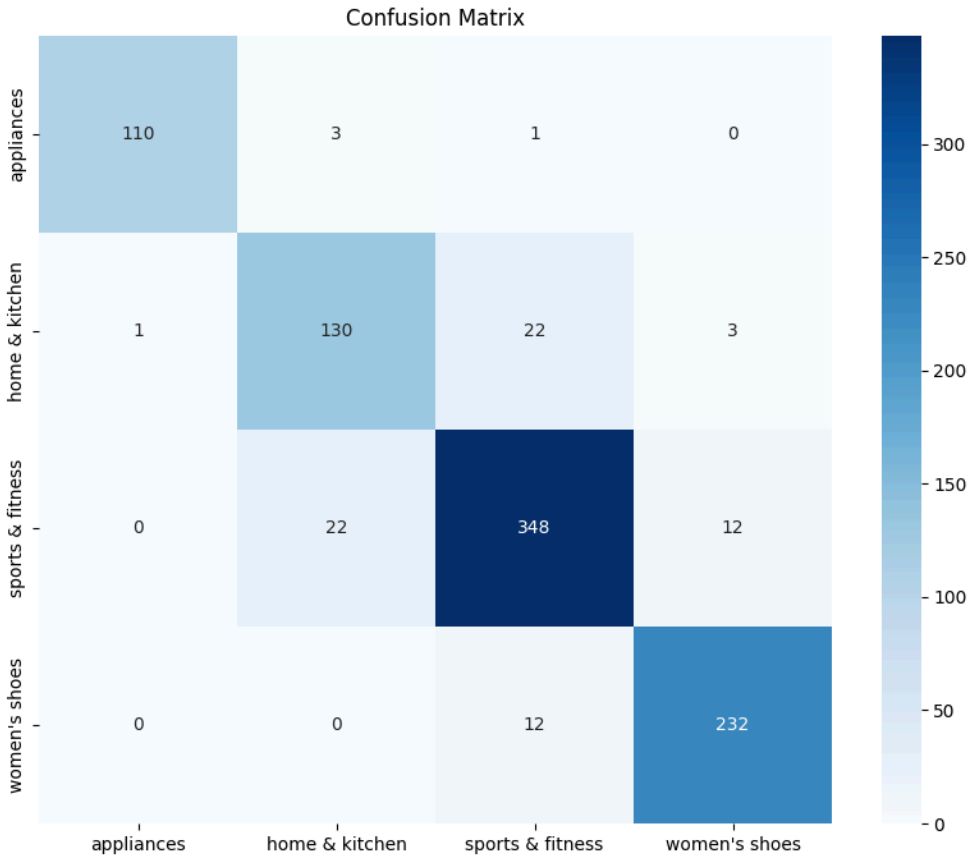


Figure 4: Confusion Matrix for Linear SVM. - **Overall Accuracy**: **91.52%** - **Analysis**: The high diagonal values indicate strong performance across categories. The model effectively learned to distinguish between different product types using only the compressed 50D feature vectors.

5. Conclusion

The **Neural Search Engine** successfully demonstrates the power of transfer learning and vector embeddings. By combining a lightweight deep learning backbone (MobileNetV2) with classical machine learning (SVM, K-Means), we built a system that is both **highly accurate (>91%)** and **computationally efficient**.

The included Streamlit application provides a user-friendly "command center" for interacting with these models, featuring a "nerd/hacker" aesthetic that reflects the advanced underlying technology.