1 Extended Summary

This work presents a pioneering AI-driven framework designed to detect artificial food colorants in processed foods within Bangladeshi markets, with a focus on assessing their potential health risks. The project addresses a critical public health issue by leveraging multimodal machine learning, combining advanced image processing and natural language processing to classify products as "High Risk" or "Low Risk" based on the presence of harmful colorants. The primary aim is to develop a robust, scalable system that enhances food safety monitoring, particularly in regions with limited regulatory oversight. The framework integrates a modified VGG-16 architecture, a custom Convolutional Neural Network (CNN), BERT/RoBERTa transformer models, and the GATE algorithm for text extraction, supported by a comprehensive dataset and an interactive dashboard for real-time risk visualization.

The project's objectives are twofold: first, to curate a comprehensive dataset of processed foods, including chips, drinks, biscuits, and chocolates, tailored for colorant identification; and second, to develop a multimodal model that achieves high accuracy in distinguishing safe and hazardous products by analyzing both visual and textual data. The dataset captures the diversity of processed foods in Bangladesh, encompassing a wide range of products and their packaging information. This dataset serves as a valuable resource for future research and regulatory efforts, providing a foundation for analyzing food safety across different product categories. The multimodal model integrates visual and textual analysis to deliver reliable classifications, achieving an impressive accuracy of 94.5%, as demonstrated through rigorous performance metrics such as precision, recall, and F1-score.

The proposed architecture is meticulously designed, with each component optimized for performance and efficiency. The pre-processing stage is critical for preparing input images, ensuring they meet quality thresholds for analysis. Images are evaluated using a quality metric $Q = \frac{\text{Sharpness}(I)+\text{Contrast}(I)}{\text{NoiseLevel}(I)}$, where preprocessing is triggered if $Q > \theta$. Resizing standardizes images to 128 × 128 pixels using bilinear interpolation, defined as $I_{\text{resized}}(x, y) = \sum_{i,j} I_{\text{original}}(i, j) \cdot w(x - i, y - j)$, with scaling factors $s_x = \frac{128}{\text{width}}$ and $s_y = \frac{128}{\text{height}}$. Noise reduction employs a Gaussian filter with $\sigma = 1.5$, formulated as $I_{\text{filtered}}(x, y) = \frac{1}{2\pi\sigma^2} \sum_{m,n} I_{\text{resized}}(x - m, y - n) \cdot e^{-\frac{(m^2+n^2)}{2\sigma^2}}$. Lossy compression via discrete cosine transform (DCT) minimizes data size, and augmentation techniques, such as rotation ($I_{\text{rotated}}(x', y') = I_{\text{compressed}}(x \cos\theta - y \sin\theta, x \sin\theta + y \cos\theta)$) and flipping, enhance model robustness by increasing data variability.

The modified VGG-16 architecture comprises 13 convolutional layers, 5 max-pooling layers, and 3 dense layers, with convolution defined as $O_l = \text{ReLU}(W_l * I_{l-1} + b_l)$ and max-pooling as $P_l = \max_{m,n \in 2\times 2} O_l(m, n)$. This structure extracts spatial features critical for identifying visual patterns associated with food packaging. The custom CNN further refines feature extraction through additional convolutional layers, batch normalization ($\hat{F}_l = \frac{F_l - \mu_l}{\sqrt{\sigma_l^2 + \epsilon}}$), and fully connected layers with softmax activation ($O_{\text{CNN}} = \text{Softmax}(W_{\text{FC}} \cdot F_{\text{pooled},L} + b_{\text{FC}})$). The CNN detects low-level features like edges and textures, reducing computational load and preventing overfitting through max-pooling.

Text extraction is facilitated by the GATE algorithm, which processes keyframes to identify and extract text from product packaging. The algorithm employs non-maxima suppression with a threshold $\tau = 0.3$ to generate bounding boxes, followed by OCR tailored for Bangla script. The process is formalized as a sequence of steps, computing scores $S(x, y) = \sigma(W_s \cdot B_{\text{blob},t}(x, y) + b_s)$ and geometries $G(x, y) = W_g \cdot B_{\text{blob},t}(x, y) + b_g$. This enables efficient analysis of ingredient lists, enhancing the model's ability to identify colorant-related risks.

The BERT/RoBERTa integration processes textual data using multi-head attention and feedforward networks. Input embeddings combine token, segment, and positional components ($E = E_{\text{token}} + E_{\text{segment}} + E_{\text{pos}}$), with positional encoding defined as $E_{\text{pos}}(i) = \sin\left(\frac{i}{10000^{2i/d}}\right)$ or cosine variants. Multi-head attention is computed as $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$, enabling the model to capture complex linguistic patterns. The final classification uses softmax activation ($P = \text{Softmax}(W_{\text{class}} \cdot H_L + b_{\text{class}})$), integrating visual and textual features for robust risk assessment.

Performance evaluation is comprehensive, utilizing precision ($\text{Precision} = \frac{TP}{TP+FP}$), recall ($\text{Recall} = \frac{TP}{TP+FN}$), and F1-score ($F1 = 2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$). The multimodal model outperforms visual (90.5%) and textual (91.0%) models, achieving 94.5% accuracy with 25.0 GFLOPs, indicating high efficiency. Confusion matrices ($C_{ij} = \sum_{n=1}^{N} 1(\hat{y}_n = j, y_n = i)$) and ROC curves ($\text{TPR} = \frac{TP}{TP+FN}, \text{FPR} = \frac{FP}{FP+TN}$) provide detailed insights into classification performance. Visualizations, including bar charts for model comparisons, scatter plots for efficiency-accuracy trade-offs, and pie charts for risk distribution (64% high-risk, 36% low-risk), enhance interpretability. Category-wise analysis highlights variations across drinks, chips, biscuits, and chocolates, informing targeted safety measures.

The impact of augmentation is analyzed, demonstrating reduced loss variance and improved accuracy over 50 epochs. Augmentation mitigates overfitting, as evidenced by stable training and validation loss curves, with accuracy improvements visualized in dedicated figures. The interactive dashboard is a key contribution, offering a user-friendly interface for uploading product images and receiving real-time risk classifications. This tool empowers consumers to make informed choices and supports regulators in monitoring food safety, with visualizations showcasing risk assessments for individual products.

The study's contributions are multifaceted. The curated dataset provides a robust foundation for future research, capturing the diversity of processed foods in Bangladesh. The multimodal model advances automated food safety assessment, integrating state-of-the-art image and text analysis techniques. The GATE-based text extraction system offers a scalable solution for processing large datasets, while the dashboard enhances accessibility and practical implementation. By addressing a pressing public health issue, this research establishes a pioneering framework for food safety, with potential applications in other developing countries facing similar challenges.