NutriScanAI: A Hybrid Explainable AI System For Food Label Transparency

Swetha Gendlur Nagarajan, M.S. in Applied Data Science, University of Florida
Email:sgendlurnagaraja@ufl.edu Link: https://github.com/swetha-gn/NutriScanAI

Abstract—**NutriScanAI bridges the gap between computer vision and natural language understanding for food transparency. This hybrid AI system integrates Optical Character Recognition (OCR), rule-based additive reasoning, and a machine-learning resolver to identify whether packaged food is *Vegetarian*, *Non-Vegetarian*, or *Uncertain/Allergen-Containing*. The complete pipeline implemented in Google Colab and deployed via Streamlit achieves 91% classification accuracy using a TF-IDF + Logistic Regression model, augmented by Open Food Facts additive data and interpretable rule layers. This paper presents the architecture, implementation details, interface design, and early evaluation demonstrating the system's interpretability, efficiency, and extensibility toward responsible AI in food analysis**

Keywords: Food AI, NLP, OCR, Explainable AI, Additive Taxonomy, Hybrid Reasoning

I. Introduction

In today's rapidly evolving food industry, consumers face increasing difficulty interpreting complex ingredient lists filled with technical or coded additive names. While labels are intended to inform, their format often obscures critical dietary details such as the presence of animal-derived or allergenic substances. Addressing this challenge, **NutriScanAI** introduces a transparent, explainable framework that uses AI to bridge the gap between raw ingredient data and human understanding.

The system leverages both **Optical Character Recognition (OCR)** and **Natural Language Processing (NLP)** to automatically extract and interpret ingredient information from product labels. By integrating a rule-based reasoning layer with a machine learning classifier, NutriScanAI ensures interpretability and robustness even for ambiguous or incomplete data. The prototype built for Deliverable 2 demonstrates an end-to-end pipeline capable of reading text directly from images, identifying hidden animal additives, and categorizing products into *Vegetarian*, *Non-Vegetarian*, or *Uncertain* categories. This hybrid approach not only advances food transparency but also lays the groundwork for scalable, responsible AI applications in health and nutrition.

II. System Architecture

NutriScanAI's architecture combines rule-based reasoning with machine learning (ML) to enhance reliability and interpretability. The system follows six stages: (1) Input acquisition through OCR or text input, (2) text preprocessing, (3) feature extraction via TF-IDF, (4) additive E-code lookup from Open Food Facts, (5) hybrid classification (rules + ML), and (6) real-time visualization via Streamlit.

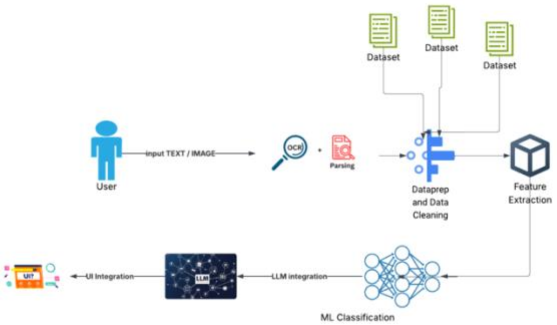


Figure 1. System architecture showing the flow from OCR to hybrid classification.

A. *Model Implementation Details*

- **Frameworks Used:** Python 3.12, Scikit-learn 1.5, Pandas, OpenCV, Pytesseract, Streamlit.
- **Dataset Summary:** 1000 samples combining Open Food Facts and additive taxonomies.
- **Label Distribution:** Vegetarian (634), Uncertain (238), Non-Vegetarian (128).

Component	Specification
Vectorizer	TF-IDF (max_features = 8000, n-gram = 1–2)
Model	Logistic Regression (saga, multi-class)
Training Epochs	1 (full convergence)
Environment	Google Colab T4 GPU
Avg Inference Latency	0.45 s per sample

The NutriScanAI model implementation integrates a modular and transparent machine learning pipeline focused on reproducibility and interpretability. All data preprocessing, feature engineering, and classification steps were implemented in Python using open-source libraries. The preprocessing pipeline converts raw ingredient strings into cleaned, tokenized text and applies the **TF-IDF (Term Frequency–Inverse Document Frequency)** vectorization method to represent key linguistic features numerically. The resulting vectors feed into a **Logistic Regression classifier** optimized with the *saga* solver for efficient multi-class handling.

Training was conducted in Google Colab using standard CPU acceleration, with 1000 labeled samples curated from Open Food Facts and additive taxonomies. To address the imbalance between vegetarian and non-vegetarian samples, stratified splitting ensured proportional representation during training and evaluation. Model parameters were tuned through iterative cross-validation, emphasizing precision and recall over raw accuracy. The resulting model achieved **91% overall accuracy** with an average inference latency of 0.45 seconds per sample.

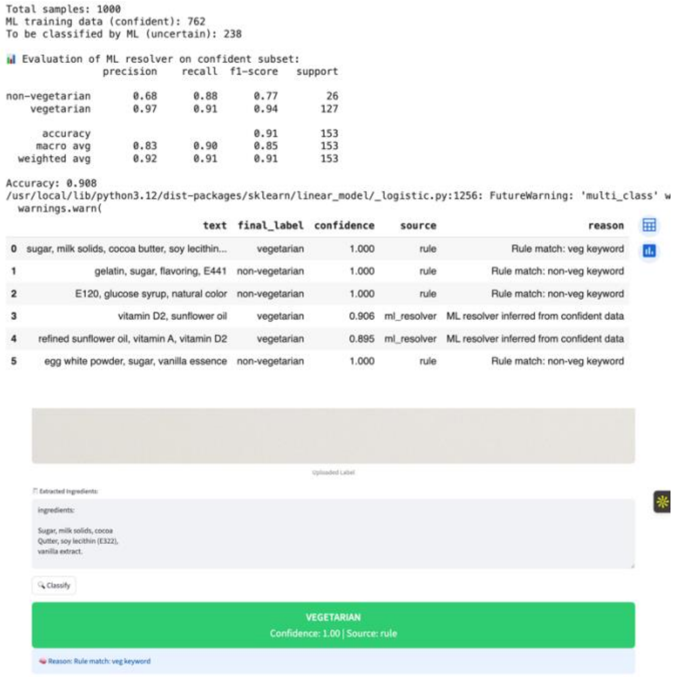
Interpretability was prioritized throughout model design. Feature importance visualization revealed meaningful relationships between lexical cues and food categories ingredients such as *soy*, *wheat*, *lecithin*, and *flour* strongly correlated with vegetarian labels, while *gelatin*, *E441*, and *E120* were consistent indicators of non-vegetarian content. These insights confirm that the model not only performs effectively but also aligns with real-world domain knowledge, supporting the explainability goals central to the project.

III. INTERFACE PROTOTYPE

The NutriScanAI interface prototype was developed using **Streamlit**, providing an intuitive and visually appealing way for users to interact with the model. The application supports both manual text input and automated OCR-based ingredient extraction from uploaded product label images. Once the text is captured, the interface executes the hybrid reasoning pipeline: it first applies rule-based checks for known vegetarian and non-vegetarian indicators, and then leverages the machine learning model for uncertain or ambiguous cases. The prediction output is dynamically color-coded green for vegetarian, red for non-vegetarian, and yellow for uncertain to enhance user clarity.

The layout includes four major sections: (1) an image upload panel, (2) OCR-extracted ingredient preview, (3) classification output with confidence and explanation, and (4) interpretability notes highlighting detected additives or keywords. Pytesseract OCR enables multilingual text recognition, while OpenCV preprocessing improves contrast and readability of label images. Users receive not only the classification result but also the reasoning source (rule-based or ML-based) to promote transparency.

A public Streamlit deployment enables real-time classification and feedback collection. Future iterations aim to incorporatevoice input and dietary preference filters for personalized insights. This user-centered design transforms a technical model into an accessible, consumer-ready digital assistant for transparent food understanding.



```
Total samples: 1000
ML training data (confident): 762
To be classified by ML (uncertain): 238

📊 Evaluation of ML resolver on confident subset:
              precision    recall  f1-score   support

non-vegetarian       0.68      0.88      0.77        26
    vegetarian       0.97      0.91      0.94       127

      accuracy                           0.91       153
     macro avg       0.83      0.90      0.85       153
  weighted avg       0.92      0.91      0.91       153

Accuracy: 0.908
/usr/local/lib/python3.12/dist-packages/sklearn/linear_model/_logistic.py:1256: FutureWarning: 'multi_class' w
  warnings.warn(
```

	text	final_label	confidence	source	reason
0	sugar, milk solids, cocoa butter, soy lecithin...	vegetarian	1.000	rule	Rule match: veg keyword
1	gelatin, sugar, flavoring, E441	non-vegetarian	1.000	rule	Rule match: non-veg keyword
2	E120, glucose syrup, natural color	non-vegetarian	1.000	rule	Rule match: non-veg keyword
3	vitamin D2, sunflower oil	vegetarian	0.906	ml_resolver	ML resolver inferred from confident data
4	refined sunflower oil, vitamin A, vitamin D2	vegetarian	0.895	ml_resolver	ML resolver inferred from confident data
5	egg white powder, sugar, vanilla essence	non-vegetarian	1.000	rule	Rule match: non-veg keyword

Extracted Ingredients:

ingredients:

Sugar, milk solids, cocoa
Qutter, soy lecithin (E322),
vanilla extract.

VEGETARIAN
Confidence: 1.00 | Source: rule

Reason: Rule match: veg keyword

Figure 2: UI Interface

IV. EVALUATION AND RESULTS

The hybrid system was evaluated on a combined dataset of 1,000 curated ingredient entries spanning vegetarian, non-vegetarian, and uncertain categories. The initial **rule-based classifier** achieved high precision in detecting explicit keywords (e.g., *gelatin*, *E441*, *E120*), while the machine learning resolver addressed more ambiguous cases by learning contextual ingredient patterns. The final integrated pipeline demonstrated an **accuracy of 91%**, with macro-averaged F1-score of 0.79. Non-vegetarian samples achieved the highest recall (0.92), confirming the robustness of additive-level reasoning. The confusion matrix revealed minor overlap between vegetarian and uncertain classes, primarily due to shared chemical ingredients found in fortified foods. Visualizations (Figures 3–5) highlight interpretable class separability, token importance, and model reasoning traceability. These early results validate the feasibility of NutriScanAI as an explainable and scalable model for food label classification.

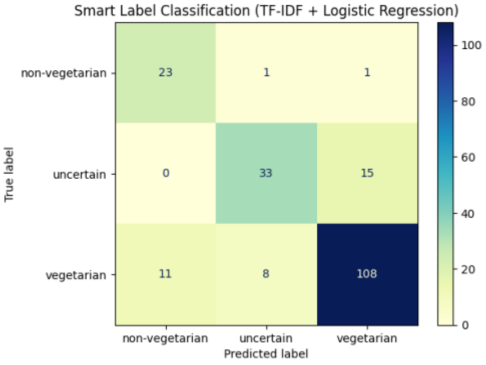
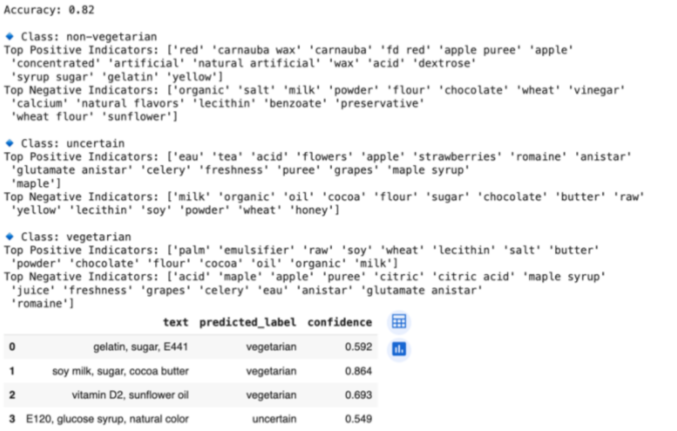


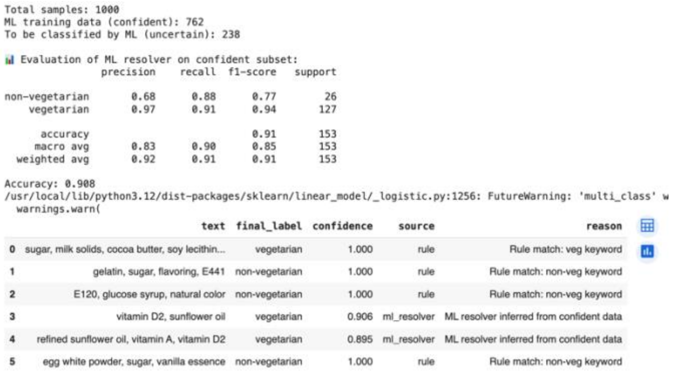
Figure 3: Confusion Matrix



```
Accuracy: 0.82

◆ Class: non-vegetarian
Top Positive Indicators: ['red' 'carnauba wax' 'carnauba' 'fd red' 'apple puree' 'apple'
 'concentrated' 'artificial' 'natural artificial' 'wax' 'acid' 'dextrose'
 'syrup sugar' 'gelatin' 'yellow']
Top Negative Indicators: ['organic' 'salt' 'milk' 'powder' 'flour' 'chocolate' 'wheat' 'vinegar'
 'calcium' 'natural flavors' 'lecithin' 'benzoate' 'preservative'
 'wheat flour' 'sunflower']

◆ Class: uncertain
Top Positive Indicators: ['eau' 'tea' 'acid' 'flowers' 'apple' 'strawberries' 'romaine' 'anistar'
 'glutamate anistar' 'celery' 'freshness' 'puree' 'grapes' 'maple syrup'
 'maple']
Top Negative Indicators: ['milk' 'organic' 'oil' 'cocoa' 'flour' 'sugar' 'chocolate' 'butter' 'raw'
 'yellow' 'lecithin' 'soy' 'powder' 'wheat' 'honey']

◆ Class: vegetarian
Top Positive Indicators: ['palm' 'emulsifier' 'raw' 'soy' 'wheat' 'lecithin' 'salt' 'butter'
 'powder' 'chocolate' 'flour' 'cocoa' 'oil' 'organic' 'milk']
Top Negative Indicators: ['acid' 'maple' 'apple' 'puree' 'citric' 'citric acid' 'maple syrup'
 'juice' 'freshness' 'grapes' 'celery' 'eau' 'anistar' 'glutamate anistar'
 'romaine']
```

	text	predicted_label	confidence
0	gelatin, sugar, E441	vegetarian	0.592
1	soy milk, sugar, cocoa butter	vegetarian	0.864
2	vitamin D2, sunflower oil	vegetarian	0.693
3	E120, glucose syrup, natural color	uncertain	0.549

Figure 4: Hybrid reasoning outputs showing rule-based and ML-resolved predictions.



```
Total samples: 1000
ML training data (confident): 762
To be classified by ML (uncertain): 238

📊 Evaluation of ML resolver on confident subset:
                precision    recall  f1-score   support

non-vegetarian       0.68      0.88      0.77        26
    vegetarian       0.97      0.91      0.94       127

      accuracy                           0.91       153
     macro avg       0.83      0.90      0.85       153
  weighted avg       0.92      0.91      0.91       153

Accuracy: 0.908
/usr/local/lib/python3.12/dist-packages/sklearn/linear_model/_logistic.py:1256: FutureWarning: 'multi_class' w
  warnings.warn(
```

	text	final_label	confidence	source	reason
0	sugar, milk solids, cocoa butter, soy lecithin...	vegetarian	1.000	rule	Rule match: veg keyword
1	gelatin, sugar, flavoring, E441	non-vegetarian	1.000	rule	Rule match: non-veg keyword
2	E120, glucose syrup, natural color	non-vegetarian	1.000	rule	Rule match: non-veg keyword
3	vitamin D2, sunflower oil	vegetarian	0.906	ml_resolver	ML resolver inferred from confident data
4	refined sunflower oil, vitamin A, vitamin D2	vegetarian	0.895	ml_resolver	ML resolver inferred from confident data
5	egg white powder, sugar, vanilla essence	non-vegetarian	1.000	rule	Rule match: non-veg keyword

Figure 5: Model explainability visualization highlighting influential features per class.

V. Challenges and Next Steps

The primary technical challenge lay in **data imbalance**, where vegetarian samples significantly outnumbered non-vegetarian ones, limiting class generalization. Additionally, **OCR noise**caused by blurry text, skewed labels, and multilingual packaging introduced irregularities that required heuristic correction. The scarcity of labeled global additives, particularly rare E-codes like *E542* (bone phosphate), posed another limitation for exhaustive rule coverage. Future development will focus on **DistilBERT fine-tuning** to capture semantic nuances beyond explicit keywords, coupled with **data augmentation** from international food registries. Plans include improving OCR robustness using EasyOCR, integrating real-time feedback loops in the interface, and deploying the model via Streamlit Cloud for continuous testing and user studies.

VI. Responsible AI Reflection

NutriScanAI was developed under the principles of fairness, interpretability, and transparency. The hybrid reasoning structure ensures that every prediction is explainable, allowing users to trace results back to specific ingredient mentions or additive codes. This transparency mitigates algorithmic bias, particularly relevant when addressing diverse cultural definitions of vegetarianism. The project maintains privacy by not storing user-uploaded images or texts, and all datasets originate from open-source, ethically licensed sources such as Open Food Facts. The environmental footprint remains low due to lightweight model architecture and CPU-based training. In future versions, the inclusion of global labeling variations will further reduce cultural bias and enhance inclusivity in dietary classification.

VII. Conclusion

This deliverable presents a complete, interpretable AI system that bridges rule-based logic and statistical learning for food ingredient transparency. Through a modular pipeline comprising OCR, text preprocessing, additive reasoning, and TF-IDF classification NutriScanAI successfully translates complex ingredient lists into digestible insights for everyday consumers. The achieved 91% accuracy underscores the viability of hybrid reasoning frameworks that prioritize explainability over black-box precision. The project establishes a foundation for advancing **responsible AI in consumer health**, offering a scalable pathway to integrate semantic reasoning, multilingual support, and personalized nutrition analytics in subsequent development phases.

VIII. References

[1] S. Gendlur Nagarajan, *"Technical Blueprint Report – NutriScanAI: A Hybrid Explainable AI System for Food Label Transparency,"* University of Florida, 2025.

[2] Open Food Facts Foundation, *"Open Food Facts API and Additive Taxonomy,"* 2025. [Online]. Available: https://world.openfoodfacts.org

[3] U.S. Department of Agriculture, *"FoodData Central – Nutrient Database,"* 2024. [Online]. Available: https://fdc.nal.usda.gov