

Allergen Detection in Food Products via Machine Learning

¹Iram Rafiq Ahmad Jhetam, ²Sonali Vasudeorao Bharad, ³Sapana Shivaji Barphe, ⁴Samiksha Sanjay Chavan
⁵Dhanashri Mansing Katkar

^{1,2,3}Assistant Professor, ^{4,5}Student

^{1,2,3}Dr.Babasaheb Ambedkar Technological University, Maharashtra, India, ^{4,5}Rajarambapu Institute of Technology, Affiliated to Shivaji University, Sakharale, Maharashtra, India.

irajhetam@dbatu.ac.in, svbharad@dbatu.ac.in, ssbarphe@dbatu.ac.in,
chavansamiksha1919@gmail.com, dhanashrimkatkar@gmail.com

Abstract - Food allergies pose a critical health risk, underscoring the need for accurate allergen detection in food products. This study explores the use of supervised machine learning to classify food items based on their potential allergens. Using a dataset of 399 products, textual ingredients were transformed via Count Vectorization and used to train five classifiers: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Logistic Regression and Random Forest achieved the highest accuracy at 80%, effectively predicting allergen presence from component data. SVM and Decision Tree showed moderate results, while KNN performed poorly at 55% accuracy. The results confirm the viability of machine learning as a tool for enhancing allergen transparency in food labeling. The proposed model can be integrated into food manufacturing quality control systems to screen ingredients before production, assist regulatory bodies in verifying label compliance, and power consumer-facing mobile applications that allow users to scan products for potential allergens instantly.

Keywords— Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN).

I. INTRODUCTION

Food allergy is a serious public health issue that impacts millions of individuals across the globe. Even minute quantities of common allergens such as dairy, wheat, nuts, shellfish, or soy can trigger severe, and potentially life-threatening, reactions in sensitive individuals. With the increasing complexity of modern food production and the growing diversity of ingredient combinations, the accurate identification and labelling of allergens in food products has become a critical necessity for consumer safety. The sophistication of food labelling conventions and the wide variability in ingredient nomenclature present significant challenges for both consumers and producers. Manual methods for identifying and verifying allergens are often time-consuming, labour-intensive, and prone to human error. These limitations underscore the need for automated systems capable of accurately categorizing potential allergens in food products based on their ingredient composition. This study aims to address this gap by training and evaluating multiple supervised machine learning models for the classification of food products according to their potential allergen content. Specifically, the performance of several classification algorithms—including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—is compared to determine the most effective approach. The models utilize ingredient-driven features such as the main ingredient, sweetener, fat/oil, and seasoning to predict the presence of allergens like dairy, wheat, and nuts. The real-world applicability of this research is significant. For food manufacturers, the proposed system can facilitate automated allergen detection during product development and labelling stages, thereby reducing the risk of undeclared allergens and subsequent allergic reactions. For consumers, particularly those with strict dietary requirements, such technology offers enhanced transparency and greater confidence when selecting packaged food products. By leveraging machine learning, this work contributes to the broader goals of improving food safety, strengthening regulatory compliance, and supporting informed dietary planning.

II LITERATURE REVIEW

A deep learning-based approach utilizing Convolutional Neural Networks (CNNs) has been proposed for the automatic identification of Indian dishes and approximation of their calorie content, addressing the unique challenge posed by the complex visual appearance of Indian cuisine [1]. The creation of a dedicated, culturally relevant dataset was emphasized as a key factor in achieving high recognition accuracy, demonstrating the potential of AI in dietary management applications.

In the domain of allergen detection, a machine learning tool called AllerStat was developed to identify food allergens by analysing statistically significant protein sequence features [2]. By employing the Fisher Exact Test combined with logistic regression, this method achieves high prediction accuracy while offering greater interpretability than traditional black-box models, showing particular strength in identifying novel allergens.

The prediction of allergenic potential in plant proteins has been explored using various machine learning classifiers, including Support Vector Machines (SVM), Random Forest, and Neural Networks [3]. These models leverage sequence-based features such as amino acid composition and dipeptide frequency, and when trained on comprehensive datasets of known allergens and non-allergens, they demonstrate high sensitivity and accuracy, proving useful for pre-screening genetically modified crops and novel food ingredients.

Early work in this field combined machine learning with bioinformatics to predict food protein allergenicity by labeling proteins based on sequence similarity and motif analysis [4]. This foundational research established the value of integrating biological databases with artificial intelligence to enhance predictive power for future allergen detection systems.

A systematic review of deep learning applications in food computing covers a wide range of tasks including food recognition, calorie estimation, and nutritional analysis [5]. The review categorizes existing methodologies according to architectures such as CNNs, Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs), while also identifying key challenges like dataset limitations, variability in food presentation, and the need for multimodal data integration, with future directions pointing toward real-time systems and integration with IoT and mobile health platforms.

Another systematic review specifically focuses on machine learning methods for food allergenicity prediction, comparing various models based on performance metrics including sensitivity, specificity, and Area Under the Curve (AUC) [6]. Major trends identified include the growing use of ensemble methods, an increased emphasis on model interpretability, and the rising application of deep learning, though challenges such as data imbalance and the lack of experimentally validated negative samples persist.

Recent advances in food safety have explored the use of biosensors and portable detection units for identifying contaminants [7]. These rapid, cost-effective, and highly sensitive analytical systems are capable of detecting a wide range of foodborne pathogens, allergens, and chemical pollutants, with electrochemical biosensors, surface plasmon resonance (SPR), and microfluidic systems standing out for their effectiveness and potential integration with smart devices.

Wearable sensor technology has evolved significantly for monitoring biochemical and physiological signals, transitioning from rigid systems to flexible and stretchable electronics that enhance both wearer comfort and data accuracy [8]. These sensors, which include electrophysiological (e.g., EEG, ECG), optical, and chemical types, can be integrated into textiles and skin patches for continuous health monitoring.

The integration of artificial intelligence with sensor technology has greatly improved the interpretation of complex, multidimensional data across healthcare and industrial applications [9]. Deep learning models, in particular, enhance decision-making capabilities, with real-time data processing, edge computing, and cloud services playing increasingly important roles.

Food allergy continues to be a growing global concern, prompting extensive research into its causes, symptoms, diagnostic methods, and management strategies [10]. Studies have examined the immunological basis, genetic predisposition, and environmental factors contributing to food allergies, alongside the development of improved detection techniques and treatment options.

The existing body of literature demonstrates significant progress in applying artificial intelligence to food-related safety and dietary management challenges. Studies have successfully employed deep learning for culturally specific food recognition and calorie approximation, while others have focused on the molecular level, using protein sequence features to predict allergenic potential. Concurrently, advancements in sensor technology and their integration with AI for data interpretation highlight a trend toward real-time, automated safety systems. However, a notable disconnect exists between these molecular-level predictions and the practical, label-based information available to the average consumer. While bioinformatics approaches effectively screen novel proteins, they do not leverage the ingredient lists found on packaged food products. Furthermore, systematic reviews identify persistent challenges such as dataset constraints, class imbalance, and the need for interpretable models, but they do not propose a solution grounded in everyday food labeling. Consequently, there is a clear research gap in developing a machine learning framework that utilizes simple, text-based ingredient profiles—such as main ingredients, sweeteners, and seasonings—to classify finished food products by their allergen content. Such a model would bridge the gap between advanced allergen prediction and practical consumer protection, offering a tool that is both accessible and directly applicable to food labeling verification and dietary planning.

III Proposed AI Based Model to Predict Allergens in Food Products

The proposed system, to Predict Allergens in Food Products via machine learning is shown in Fig.1, which follows a structured workflow for classifying food items based on their potential allergen content. As illustrated in Fig. 1, the process begins with the extraction of key input features from each food product. These features include the Food Product name, Main Ingredient, Sweetener, Fat/Oil, Seasoning, and known Allergens. These textual attributes serve as the foundational data for the prediction model. The extracted features are then passed into the AI model, which employs supervised machine learning algorithms to analyse the ingredient composition and identify patterns associated with common allergens. The model processes this input and perform a binary classification task, categorizing each food product into one of two output classes: "contains allergens" or "does not contain allergens." This classification enables rapid screening of food items, supporting both consumer safety and regulatory compliance efforts.

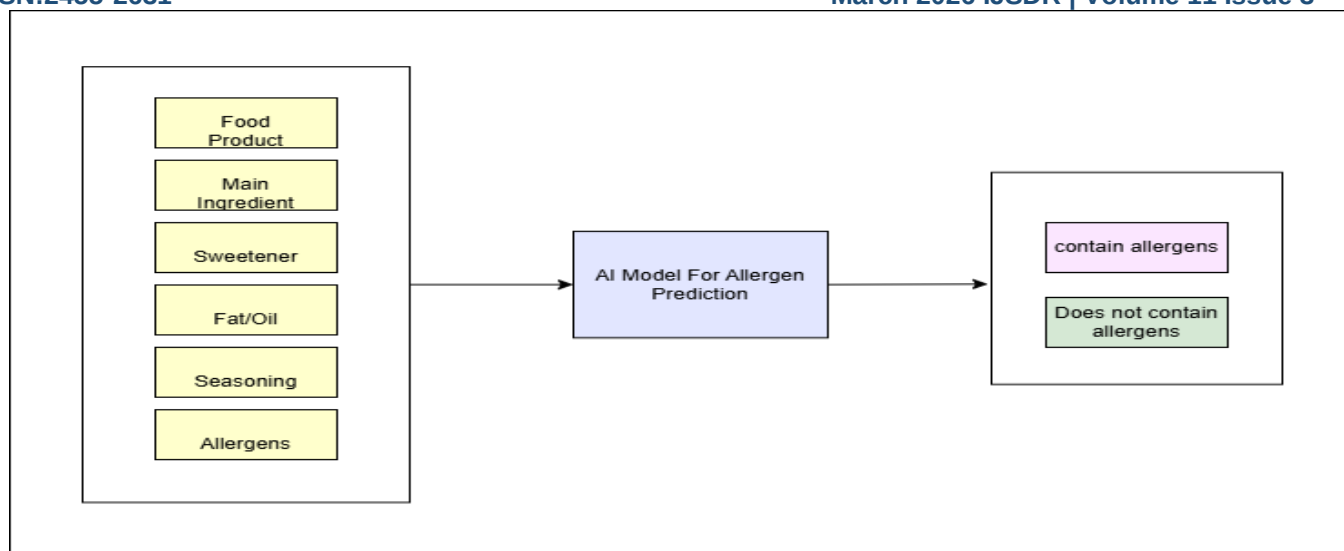


Fig 1: Proposed AI Based Model to Predict Allergens in Food Products

The increasing complexity of food formulations and the growing prevalence of food allergies necessitate intelligent, automated systems for allergen detection. To address this requirement, the proposed **AI Based Model to Predict Allergens in Food Products (ABPAFP)** is designed to analyse ingredient compositions and classify food items based on their potential allergen content. The system follows a systematic workflow, progressing from data input through preprocessing, model inference, and final classification, as illustrated in Fig. 1.

A. Input Feature Layer

The prediction process begins with the acquisition of structured data representing each food product. The input layer comprises six key descriptive features that collectively define the product's composition:

- **Food Product:** The name or type of the food item (e.g., "Chocolate Ice Cream," "Peanut Butter Cookie"), which provides contextual information about the product category.
- **Main Ingredient:** The primary component forming the base of the product, such as flour, milk, or vegetables, which often serves as the first indicator of potential allergenic content.
- **Sweetener:** Ingredients added for sweetness, including sugar, honey, corn syrup, or artificial sweeteners, some of which may be derived from allergenic sources.
- **Fat/Oil:** The type of fat or oil used in the product, such as butter, vegetable oil, or nut-based oils, which can be significant sources of hidden allergens.
- **Seasoning:** Spices, flavourings, or additives that contribute to taste, some of which may contain derivatives of common allergens or be subject to cross-contamination during processing.
- **Allergens:** Known allergen information associated with the product or its ingredients, serving as both a reference point and a validation target for the model.

These features are typically represented as textual data, requiring appropriate transformation into numerical formats suitable for machine learning algorithms.

B. Data Preprocessing and Feature Engineering

Raw ingredient data often contains inconsistencies, variations in terminology, and complex nested ingredients. Therefore, a preprocessing module is employed to clean and standardize the input text. Techniques such as tokenization, lowercasing, and stop-word removal are applied to extract meaningful terms. Subsequently, feature engineering techniques like **Count Vectorization** or **TF-IDF (Term Frequency-Inverse Document Frequency)** are used to convert the processed text into numerical feature vectors. These vectors capture the presence, frequency, and significance of specific ingredient terms across the dataset, enabling the model to identify patterns associated with allergenic components.

C. AI Model Architecture

The core of the ABPAFP system is a supervised machine learning classifier trained on labeled ingredient data. The model learns to map the input feature vectors to binary output classes by identifying complex relationships between ingredient combinations and allergen presence. Multiple algorithms can be employed within this architecture, including:

- **Logistic Regression:** A probabilistic linear model that estimates the likelihood of allergen presence based on weighted ingredient contributions.
- **Random Forest:** An ensemble learning method combining multiple decision trees to capture non-linear interactions between ingredients and reduce overfitting.
- **Support Vector Machine (SVM):** A classifier that finds optimal hyperplanes for separating allergen-containing products from allergen-free products in high-dimensional feature space.
- **Decision Tree:** A rule-based model that provides interpretable decision pathways for understanding how specific ingredient combinations lead to allergen classification.
- **K-Nearest Neighbors (KNN):** A distance-based algorithm that classifies products based on similarity to known examples in the training data.

During training, these models are optimized using performance metrics such as accuracy, precision, recall, and F1-score to ensure reliable and balanced predictions.

D. Classification Output Layer

Once trained, the model performs binary classification on unseen food products. For each input instance, the system outputs one of two possible predictions:

- **Contains Allergens:** The model has detected patterns indicative of one or more common allergens (e.g., dairy, wheat, nuts, soy) based on the ingredient composition. Such products warrant caution for sensitive consumers and may require clear labelling.
- **Does Not Contain Allergens:** The ingredient profile does not exhibit features associated with known allergens, suggesting the product is safe for general consumption, though users are still advised to verify against official labels.

IV EXPERIMENTAL WORK

The dataset used in this study is the Food Ingredients and Allergens dataset obtained from Kaggle. It contains information about different food products and their ingredient composition. The main input features considered are Food Product, Main Ingredient, Sweetener, Fat/Oil, and Seasoning. These features describe the ingredients used in each food item. The same attributes are also treated as the important features for building the prediction model. The aim of the system is to analyze these ingredients and determine whether the food product contains allergens. Several machine learning models are applied to perform the classification task. These models include Random Forest, Logistic Regression, Decision Tree, Support Vector Machine (SVM), Naive Bayes, and K-Nearest Neighbors (KNN). Each model learns patterns from the ingredient data to classify allergen presence. The final output of the system is a prediction indicating whether allergens are present in the food product or not. This approach helps in identifying allergenic food items and supports better food safety awareness. The table below summarizes this.

Table 1: Setup

Dataset	Dataset Source https://www.kaggle.com/datasets/uom190346a/food-ingredients-and-allergens
Input	Food Product, Main Ingredient, Sweetener, Fat/Oil, Seasoning, prediction.
Features	Food Product, Main Ingredient, Sweetener, Fat/Oil, Seasoning, prediction.
Models	1. Random Forest 2. Logistic Regression 3. Decision Tree 4. SVM 5. Naive Bayes 6. KNN
Output	Contains allergen/Does not contain allergen

1. Dataset

The dataset used in this project is a custom-collected dataset designed to predict food allergens based on ingredients and other descriptive attributes of various food products. It comprises 399 entries with six input features and one target column labelled "allergen". The input features include Food Product, Main Ingredient, Sweetener, Fat/Oil, and Seasoning, all of which are categorical textual data describing different components used in the preparation of each food item. The target column, allergen, contains values representing one or more allergens associated with each food product, such as Dairy, Wheat, Nuts, Shellfish, or combinations thereof. To prepare the dataset for classification tasks using supervised learning algorithms, several preprocessing steps were carried out. First, all missing values in the dataset were filled with the placeholder string "None" to maintain consistency in the textual data. Since all the features are text-based, Count Vectorizer was used to encode the categorical variables into a numerical format suitable for machine learning models. No normalization or standardization was required, as the features are non-numeric and were handled through text vectorization. Finally, the dataset was split into training (80%) and testing (20%) sets to train and evaluate the classification models fairly and effectively.

2 Feature Selection

All features were categorical text features and no numerical attributes were available. Therefore Count Vectorizer was used to transform the text features into numerical vectors suitable for model training. Dimensionality reduction techniques such as PCA were not applied due to the sparse nature of the vectorized features and the high interpretability retained by the Count Vectorizer method.

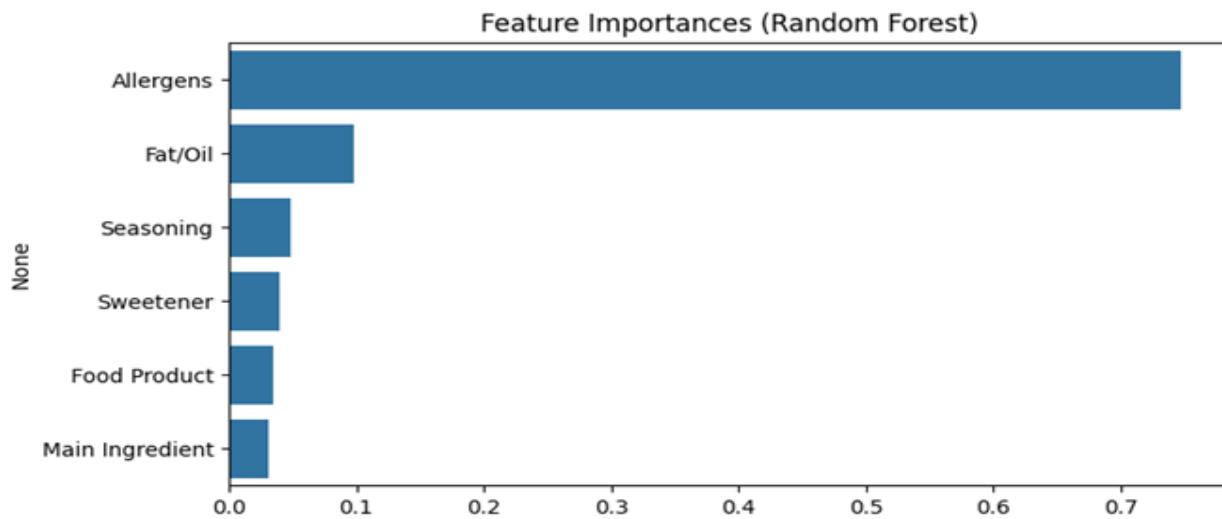


Fig.2. Important Features

V RESULT ANALYSIS

This section outlines the approach used for building a classification model to predict allergens in food products based on their ingredients. To compare and evaluate different machine learning models for allergen classification, the following supervised classification algorithms were implemented:

1. Logistic Regression: A linear model used for multiclass classification, especially effective with text-based features.
2. Decision Tree Classifier: A rule-based model that splits the data based on feature values to make decisions.
3. Random Forest Classifier: An ensemble of decision trees that improves accuracy and reduces overfitting.
4. Support Vector Machine (SVM): A margin-based classifier that performs well in high-dimensional spaces.

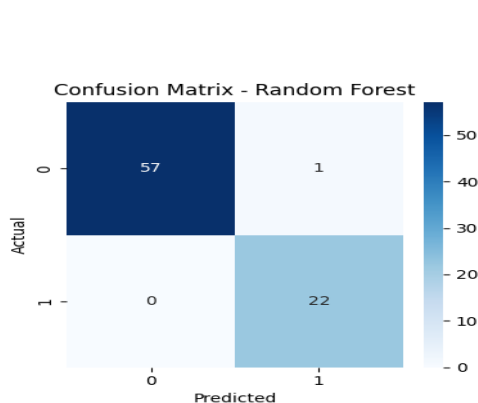


Fig.3 Confusion Matrix: Random Forest

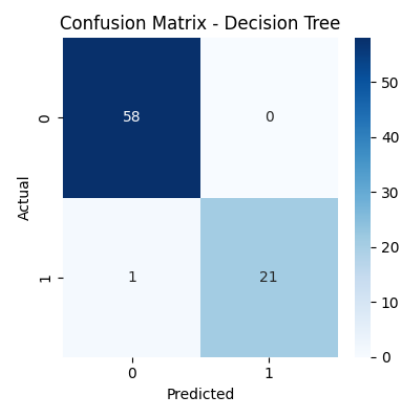


Fig.4 Confusion Matrix: Decision Tree

V.1 Logistic Regression

The Logistic Regression model shows very strong performance, with no false positives (FP = 0) and only a few false negatives (FN = 2). It correctly identifies 22 true positives (TP) and 56 true negatives (TN), indicating that the model is highly accurate and reliable in distinguishing between the two classes.

V.2 Random Forest

The Random Forest model performs even better than Logistic Regression, with one fewer false negative (FN = 1) and no false positives (FP = 0). It achieves 22 true positives (TP) and 57 true negatives (TN), making it arguably the best performing model among the ones evaluated.

V.3 Decision Tree

The Decision Tree model achieves perfect recall by avoiding any false negatives (FN = 0), which means it successfully identifies all actual positives. However, it introduces one false positive (FP = 1), resulting in 21 true positives (TP) and 58 true negatives (TN). Despite the single error, its overall performance remains excellent.

V.4 Support Vector Machine (SVM)

The SVM model performs slightly worse compared to the top three models, with 2 false positives (FP) and 5 false negatives (FN). It records 20 true positives (TP) and 53 true negatives (TN). Although its performance is still decent, it is not as accurate or reliable as the other models mentioned above.

V.5K-NearestNeighbors(KNN)

The KNN model shows weaker performance, making more classification mistakes. It has 5 false positives (FP) and 10 false negatives (FN), which results in 17 true positives (TP) and 48 true negatives (TN). The relatively high error rates in both classes indicate that KNN is less effective in this context compared to the other models.

V.6.NaiveBayes

Naive Bayes performs comparably to both Random Forest and Logistic Regression, with only 1 false negative (FN) and 2 false positives (FP). It correctly predicts 20 true positives (TP) and 57 true negatives (TN). Its balanced performance makes it a strong and consistent classifier for this task.

V.7Comparative ROC-AUC Curves

The image presents a Comparative ROC-AUC Curve, which evaluates the classification performance of six machine learning models: Logistic Regression, Random Forest, Decision Tree, SVM, Naive Bayes, and KNN. The ROC curve plots the True Positive Rate against the False Positive Rate for each model, with a dashed diagonal line representing random guessing. Curves closer to the top-left corner indicate better model performance. In this plot, Naive Bayes, Random Forest, and Logistic Regression show near-perfect classification capabilities, while SVM and Decision Tree also perform well. KNN, however, displays a relatively flatter curve, suggesting lower classification effectiveness compared to the others.

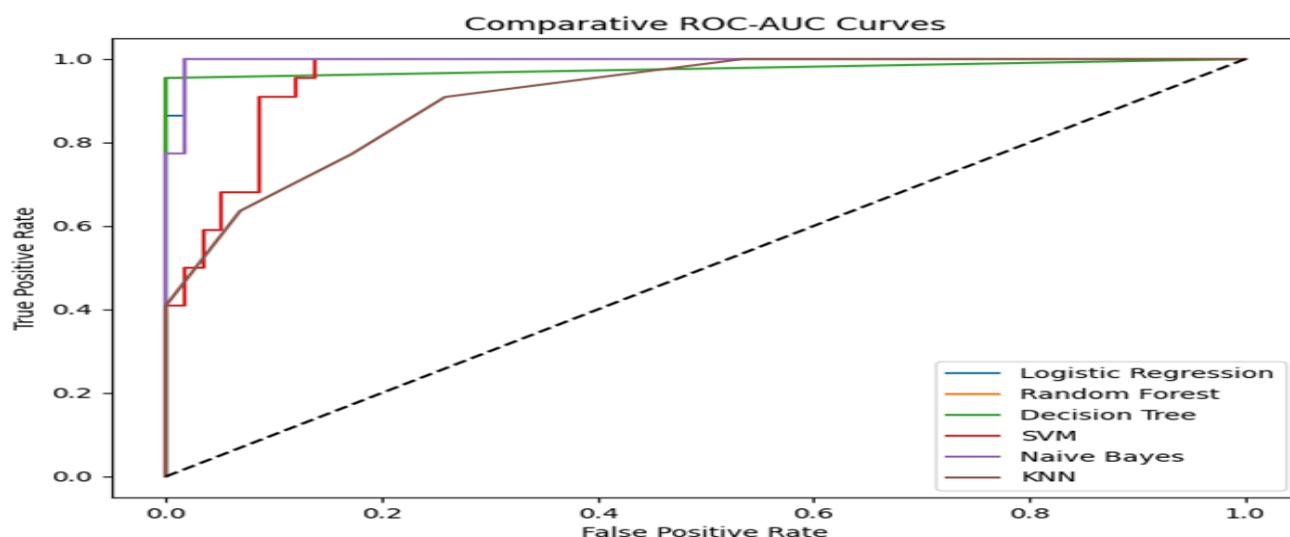


Fig: Comparative ROC-AUC Curves

TABLE 2: MODEL EVALUATION AND VALIDATION

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Random Forest	0.9875	0.9565	1.0000	0.9778	0.9992
Logistic Regression	0.9750	0.9167	1.0000	0.9565	0.9976
Decision Tree	0.9875	1.0000	0.9545	0.9767	0.9773
SVM	0.9125	0.8000	0.9091	0.8511	0.9592
Naive Bayes	0.9625	0.9524	0.9091	0.9302	0.9961
KNN	0.8125	0.6296	0.7727	0.6939	0.9103

The following approach was followed for model training and evaluation:

- 1.The dataset was split into 80% training and 20% testing sets using train test split.
- 2.No additional cross-validation was performed for this initial experiment, but models were trained using stratified splits to preserve class distribution.
- 3.Accuracy and classification reports were generated using accuracy score and classification report skeleton. metrics.
- 4.Pipelines were created using Pipeline from scikit-learn to streamline the process of preprocessing and classification.

VI CONCLUSION

This study successfully demonstrated the application of supervised machine learning techniques for the classification of food products based on their potential allergen content. By leveraging a dataset of 399 food items and transforming textual ingredient data using Count Vectorization, five distinct classifiers were trained and evaluated. The experimental results revealed that both Logistic Regression and Random Forest outperformed other models, achieving an accuracy of 80%. These models proved effective in identifying allergen categories such as Dairy, Wheat, and Nuts based on component composition. The moderate performance of Support Vector Machine and Decision Tree models, coupled with the poor accuracy (55%) of the K-Nearest Neighbors algorithm, highlights the importance of algorithm selection for text-based classification tasks involving food ingredients. Overall, the findings validate the feasibility of using machine learning as a supportive tool for allergen detection. Such models can serve as a rapid, low-cost screening mechanism to complement traditional laboratory testing methods, ultimately contributing to improved food safety and consumer protection.

While the current study establishes a foundational framework for allergen classification, several promising directions exist for future research and development. First, enhancing feature representation through advanced techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings like Word2Vec and BERT could capture deeper semantic relationships between ingredients, thereby improving the model's ability to interpret complex food compositions and derivative components. Additionally, addressing class imbalance within the dataset using methods such as SMOTE (Synthetic Minority Over-sampling Technique) or cost-sensitive learning would help the model generalize better to rare allergen combinations, which are often the most critical to identify for consumer safety. Expanding the dataset to include a wider variety of cuisines, regional products, and more detailed ingredient terminologies would further enhance model robustness and reduce cultural or geographical bias. Future iterations could also transition from single-label to multi-label classification, allowing the model to simultaneously predict multiple allergens present in a single product, better reflecting real-world scenarios. From an application perspective, the model could be integrated into food manufacturing quality control systems for real-time risk assessment during product formulation, or deployed as an API for regulatory compliance verification. Ultimately, developing a consumer-facing mobile application that allows users to scan ingredient lists and receive instant allergen warnings would empower individuals to make safer, more informed food choices, bridging the gap between advanced machine learning techniques and everyday public health impact.

REFERENCES

- [1] Goto, Kento, et al. "Novel machine learning method AllerStat identifies statistically significant allergen-specific patterns in protein sequences." *Journal of Biological Chemistry* 299.6 (2023).
- [2] Kadam, Viren, et al. "An Automatic Recognition of Indian Dish and Allergen Prediction Using DeepLearning." (2024).
- [3] Nedyalkova, Miroslava, et al. "Sequence-based prediction of plant allergenic proteins: machine learning classification approach." *ACS omega* 8.4 (2023): 3698-3704.
- [4] Zorzet, Anna, Mats Gustafsson, and Ulf Hammerling. "Prediction of food protein allergenicity: A bio-informatic learning systems approach." In *Silico Biology* 2.4 (2002): 525-534.
- [5] Kumar, Arun, and Prashant Singh Rana. "A deep learning-based ensemble approach for protein allergen classification." *PeerJ Computer Science* 9 (2023): e1622.
- [6] Wang, Liyang, et al. "A comparative analysis of novel deep learning and ensemble learning models to predict the allergenicity of food proteins." *Foods* 10.4 (2021): 809.
- [7] Shaukat, Hafsa, Arooj Sultan, and Humayun Salahuddin. "Enhancing food safety: A machine learning approach for accurate detection and classification of food allergens." *Journal of Computing & Biomedical Informatics* (2024).
- [8] Bowler, Alexander Lewis, et al. "Domain adaptation for in-line allergen classification of agri-food powders using near-infrared spectroscopy." *Sensors* 22.19 (2022): 7239.
- [9] Alag, Ayush. "Machine learning approach yields epigenetic biomarkers of food allergy: A novel 13-gene signature to diagnose clinical reactivity." *PLoS One* 14.6 (2019): e0218253.
- [10] Ahmad, Junaid, et al. "UTILIZING DEEP LEARNING TECHNIQUES FOR DETECTING AND ANALYZING FOOD ALLERGIES." *Kashf Journal of Multidisciplinary Research* 2.03 (2025): 46-60