SINTECHCOM: Science, Technology and Communication Journal
Vol. 6, No. 3, June 2026, pp. 373-384, DOI: 10.59190/stc.v6i3.393

Nutri-score classification of snack products using word embedding and random forest

Onky Wanda Darmawan*, Junadhi, Lusiana Efrizoni, Nurjayadi
Department of Informatics Engineering, Universitas Sains dan Teknologi Indonesia, Pekanbaru 28299, Indonesia

ABSTRACT

The increasing consumption of packaged snack products has raised concerns regarding their nutritional quality and potential health impacts. Although nutritional information is commonly provided on food packaging, many consumers experience difficulties in interpreting ingredient descriptions and nutritional labels, making it challenging to identify whether a product is healthy or unhealthy. Therefore, an automated classification system is needed to assist consumers in understanding nutritional information more effectively. This study proposes a text-based classification framework for categorizing snack products into healthy and unhealthy classes using Natural Language Processing (NLP), word embedding techniques, and the Random Forest algorithm. The dataset was obtained from the Open Food Facts database and filtered to include snack products only. After preprocessing and class balancing, a total of 465 samples were used for model development and evaluation. The preprocessing stage consisted of case folding, tokenization, stopword removal, and stemming. Three word embedding techniques, namely Word2Vec, GloVe, and FastText, were employed to transform textual ingredient descriptions into numerical feature representations. Subsequently, Random Forest was utilized as the classification algorithm, and its performance was evaluated using Accuracy, Balanced Accuracy, Precision, Recall, F1-score, and Macro F1-score. The experimental results show that GloVe achieved the best performance among the evaluated embedding methods, obtaining an accuracy of 86.02%, balanced accuracy of 84.72%, precision of 85.98%, recall of 86.02%, F1-score of 85.91%, and macro F1-score of 85.19%. The findings indicate that GloVe provides a more effective semantic representation of food-related textual information compared to Word2Vec and FastText. Overall, the proposed framework demonstrates the potential of NLP-based approaches for automated nutritional assessment and healthy food classification.

ARTICLE INFO

Article history:

Received Jun 24, 2026
Revised Jun 25, 2026
Accepted Jun 26, 2026

Keywords:

Healthy Food
Word2Vec
GloVe
FastText
Random Forest

* **Corresponding Author**
E-mail address: anjadarmawan07@gmail.com

1. INTRODUCTION

The global consumption of packaged food products has increased substantially over the past decade, driven by changing lifestyles, urbanization, and the growing demand for convenient food options. Among various categories of packaged foods, snack products have become one of the most widely consumed food items due to their accessibility, affordability, and variety of flavors [1]. Despite their popularity, excessive consumption of snack products has been associated with numerous health concerns, including obesity, diabetes, cardiovascular diseases, and other diet-related chronic conditions. These health risks are often linked to high levels of sugar, saturated fat, sodium, and artificial additives commonly found in processed snack products [2]. To promote healthier food choices, food manufacturers are required to provide nutritional information on product packaging.

Nutritional labels typically contain information regarding ingredients, energy values, carbohydrates, proteins, fats, sugars, and sodium content. Such information is intended to help consumers evaluate the nutritional quality of food products before purchase [3]. However, previous studies have shown that many consumers experience difficulties in interpreting nutritional labels due to their complexity, technical terminology, and the large amount of information presented. As a result, purchasing decisions are often influenced by factors such as brand, taste, and price rather than nutritional quality [4]. The increasing availability of food product databases has created new opportunities for the development of intelligent systems capable of automatically analyzing nutritional information. One notable example is the Open Food Facts database, an open-source repository containing extensive information on food products from various countries, including ingredient lists, nutritional values, product categories, and nutritional quality indicators [5, 6].

The availability of such large-scale food datasets enables the application of Artificial Intelligence (AI) and Machine Learning (ML) techniques to support nutritional analysis and food classification tasks. Among various AI technologies, Natural Language Processing (NLP) has emerged as an effective approach for extracting meaningful information from textual data [7, 8]. NLP techniques have been widely applied in document classification, sentiment analysis, healthcare informatics, and recommendation systems. In the context of food informatics, NLP provides the capability to analyze textual information contained in ingredient lists and food labels, allowing automatic identification of nutritional characteristics and health-related patterns [9, 10].

Consequently, NLP offers significant potential for developing intelligent systems that can assist consumers in understanding food products more effectively. A fundamental challenge in text classification is converting textual information into numerical representations that can be processed by machine learning algorithms [11, 12]. Traditional feature extraction approaches, such as Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF), primarily rely on word frequency statistics and often fail to capture semantic relationships between words [13].

To address this limitation, word embedding techniques have been introduced to represent words as dense numerical vectors that preserve contextual and semantic information. By capturing relationships among words within a corpus, word embedding provides richer feature representations and has demonstrated superior performance in various text classification applications [14, 15]. Once textual data have been transformed into vector representations, machine learning algorithms can be employed to perform classification tasks. Random Forest is one of the most widely used classification algorithms due to its robustness, high predictive performance, and resistance to overfitting. As an ensemble learning method, Random Forest constructs multiple decision trees and combines their predictions to produce a more accurate and stable classification result [16-18].

Furthermore, Random Forest has been shown to perform effectively on high-dimensional datasets, making it suitable for text-based classification problems involving large feature spaces generated by word embedding techniques. Several previous studies have investigated machine learning approaches for food-related applications, including food recommendation systems, dietary assessment, food image recognition, and nutritional prediction [19, 20]. However, studies focusing on the automatic classification of snack products into healthy and unhealthy categories based on textual nutritional information remain limited. In addition, many existing studies rely on conventional feature extraction techniques that may not adequately capture semantic relationships embedded in ingredient descriptions and nutritional labels. This limitation highlights the need for more effective text representation approaches capable of improving classification performance in food-related applications. To address these challenges, this study proposes a text-based classification framework for categorizing snack products into healthy and unhealthy classes using word embedding and Random Forest [21, 22].

The proposed framework utilizes textual information obtained from the Open Food Facts dataset, including ingredient descriptions and nutritional label information. The methodology consists of text preprocessing, word embedding-based feature representation, Random Forest classification, and model evaluation using accuracy, precision, recall, and F1-score metrics.

The primary contribution of this study is the development of an automated classification approach that leverages semantic text representation to identify the health category of snack products. By integrating word embedding and Random Forest, the proposed framework aims to improve the effectiveness of nutritional information analysis and provide a practical solution for supporting

healthier food selection. The findings of this study are expected to contribute to the growing field of food informatics and demonstrate the potential of NLP-based approaches for nutritional assessment and food classification.

2. RESEARCH METHODS

This study proposes a machine learning-based framework for classifying snack products into healthy and unhealthy categories using text=ual nutritional information. The framework integrates Natural Language Processing (NLP), word embedding, and Random Forest classification to automatically identify the health category of snack products based on ingredient descriptions and nutritional labels obtained from the Open Food Facts dataset. The research process consists of eight main stages: (1) problem identification, (2) literature review, (3) data collection, (4) text preprocessing, (5) text representation using word embedding, (6) dataset splitting, (7) Random Forest classification, and (8) model evaluation and analysis. The overall research framework is illustrated in Figure 1.

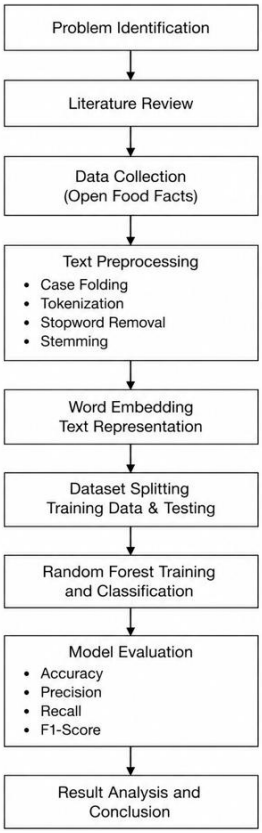

Figure 1. Workflow of the proposed classification method.

2.1. Problem Identification

The first stage of the study involves identifying the research problem. Despite the widespread availability of nutritional information on food packaging, many consumers still face difficulties in interpreting nutritional labels and ingredient lists. As a result, consumers often make purchasing decisions without adequately considering the nutritional quality of food products. This issue highlights the need for an automated classification system capable of categorizing snack products into healthy and unhealthy groups based on textual nutritional information. Therefore, this research investigates the application of NLP and machine learning techniques to support automatic nutritional assessment.

2.2. Literature Review

A comprehensive literature review was conducted to establish the theoretical foundation of the study. Relevant scientific articles, conference papers, books, and technical reports were examined to

understand current developments in food informatics, Natural Language Processing, text classification, word embedding, and machine learning algorithms. The literature review also aimed to identify research gaps in previous studies related to food product classification and nutritional analysis. The findings from the literature review were used to determine the research methodology and experimental design employed in this study.

2.4. Data Collection

The dataset used in this study was obtained from the Open Food Facts database, a publicly accessible repository containing nutritional information and ingredient descriptions of food products from various countries. The dataset includes product names, ingredient lists, nutritional values, and product categories. To ensure consistency with the research objective, only products belonging to the snack category were selected. After the initial filtering process, 1,655 snack product records were obtained from the Open Food Facts database. However, not all records could be utilized for model development. A data quality screening process was conducted to remove products with incomplete nutritional information, missing ingredient descriptions, duplicated entries, and records that could not be assigned to either healthy or unhealthy categories based on the predefined nutritional criteria. After the screening process, 1,346 valid records remained and were used for subsequent analysis, consisting of 186 healthy products and 1,160 unhealthy products.

To mitigate the impact of class imbalance on the classification model, an undersampling technique was applied to the majority class. After the balancing process, the final dataset consisted of 279 unhealthy products and 186 healthy products, resulting in a total of 465 samples used for model development and evaluation. Table 1 summarizes the dataset characteristics used in this study.

Table 1. Dataset characteristics description.

Description	Value
Source	Open food facts
Product category	Snack products
Initial records after filtering	1,655
Number of attributes	163
Records removed during data screening	309
Valid records for labeling	1,346
Numerical features used	6
Healthy class (before balancing)	186
Unhealthy class (before balancing)	1,160
Healthy class (after balancing)	186
Unhealthy class (after balancing)	279
Final dataset size	465

As shown in Table 1, the initial filtering process yielded 1,655 snack product records from the Open Food Facts database. To ensure data quality and consistency, a data screening procedure was performed to remove records with incomplete nutritional information, missing ingredient descriptions, duplicate entries, and products that could not be assigned to either healthy or unhealthy categories based on the predefined labeling criteria. After the screening process, 1,346 valid records remained for analysis, consisting of 186 healthy products and 1,160 unhealthy products.

Table 2. Class distribution before and after balancing.

Class	Before balancing	After balancing
Healthy	186	186
Unhealthy	1,160	279
Total	1,346	465

To construct a reliable classification model, the class distribution of the valid dataset was examined. The dataset exhibited a substantial class imbalance between healthy and unhealthy snack products, with the unhealthy category containing considerably more samples than the healthy

category. Such imbalance may bias the learning process toward the majority class and negatively affect classification performance. Therefore, a balancing strategy based on random undersampling was applied to reduce the dominance of the majority class while preserving sufficient information for model training. The class distributions before and after balancing process are summarized in Table 2.

As shown in Table 2, the original dataset contained considerably more unhealthy products than healthy products. To mitigate the impact of class imbalance, a random undersampling technique was applied to the majority class. Rather than creating a perfectly balanced dataset, a moderate undersampling strategy was adopted to preserve a larger proportion of majority-class information while substantially reducing class imbalance. Consequently, the number of unhealthy products was reduced from 1,160 to 279, resulting in a final dataset of 465 samples. This balancing process is essential to improve the model's ability to learn representative patterns from both classes, reduce classification bias, and minimize information loss that may occur under aggressive undersampling. Distribution of healthy and unhealthy snack products in final dataset is further illustrated in Figure 2.

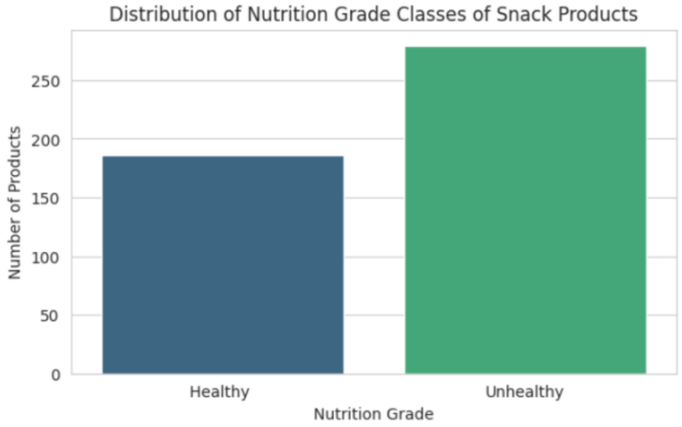

Figure 2. Class distribution of healthy and unhealthy snack products in the final dataset.

Figure 2 shows that the final dataset consists of 186 healthy products and 279 unhealthy products. Although the dataset remains slightly imbalanced, the difference between classes has been substantially reduced compared to the original distribution. This balanced distribution provides a more suitable dataset for training and evaluating the proposed Random Forest classification model.

2.4. Text Preprocessing

Text preprocessing was performed to clean, normalize, and transform raw textual data into a structured format suitable for machine learning analysis. The ingredient descriptions obtained from the Open Food Facts dataset contain various inconsistencies, including uppercase letters, special characters, punctuation marks, numbers, and stopwords that may negatively affect the classification performance. Therefore, preprocessing is necessary to improve data quality and reduce noise before feature extraction using word embedding. In this study, the preprocessing stage consists of four main steps: case folding, tokenization, stopword removal, and stemming. An example of the preprocessing process is presented in Table 3.

Table 3. Example of text preprocessing process.

Process	Output
Original text	Maíz*, aceite de oliva virgen extra* 15%, sal. (*de cultivo ecológico).
Case folding	maíz*, aceite de oliva virgen extra* 15%, sal. (*de cultivo ecológico).
Cleaning	ma z aceite de oliva virgen extra sal de cultivo ecol gico
Tokenization	['ma', 'z', 'aceite', 'de', 'oliva', 'virgen', 'extra', 'sal', 'de', 'cultivo', 'ecol', 'gico']
Stopword removal	['aceite', 'oliva', 'virgen', 'extra', 'sal', 'cultivo', 'ecol', 'gico']
Stemming	['aceit', 'oliva', 'virgen', 'extra', 'sal', 'cultivo', 'ecol', 'gico']

2.4.1. Case Folding

Case folding is the process of converting all characters in the text into lowercase letters to ensure consistency in word representation. This step eliminates differences between uppercase and lowercase forms of the same word, preventing them from being treated as distinct tokens during analysis. For example, the word "Sugar" and "sugar" would be transformed into the same representation after case folding. In addition to converting text to lowercase, a cleaning process was performed to remove punctuation marks, special characters, numbers, and unnecessary symbols that do not contribute meaningful information to the classification task.

2.4.2. Tokenization

Tokenization is the process of splitting a text sequence into individual tokens or words. This step transforms a sentence into smaller textual units that can be processed independently by machine learning algorithms. After tokenization, the cleaned ingredient description is represented as a list of individual terms. This representation facilitates subsequent preprocessing operations and feature extraction.

2.4.3. Stopword Removal

Stopword removal aims to eliminate common words that occur frequently but contribute little semantic information to the classification process. Examples of stopwords include articles, conjunctions, prepositions, and other function words that do not significantly influence the nutritional characteristics of a product. By removing stopwords, the dimensionality of the text data is reduced while preserving the most informative terms related to ingredient composition and nutritional content.

2.4.4. Stemming

Stemming is the process of reducing words to their root or base forms. This technique helps consolidate different morphological variations of a word into a single representation, thereby reducing vocabulary size and improving feature consistency. For instance, the token "aceite" is transformed into "aceit" during the stemming process. As a result, semantically related words can be represented more consistently, enabling the classification model to learn more meaningful patterns from the textual data.

The output generated from the preprocessing stage serves as the input for the word embedding process. By reducing noise and standardizing textual information, preprocessing contributes to the creation of more representative feature vectors, which are essential for improving the performance of the subsequent Random Forest classification model.

2.5. Text Representation Using Word Embedding

After the preprocessing stage, the textual data were transformed into numerical representations using word embedding techniques. This step is essential because machine learning algorithms require numerical inputs rather than raw text. Word embedding enables words to be represented as dense vectors while preserving semantic and contextual relationships among terms appearing in the corpus. Unlike traditional text representation approaches such as Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF), word embedding captures semantic similarities between words by mapping them into a continuous vector space. Consequently, words that frequently appear in similar contexts tend to have similar vector representations.

In this study, three word embedding techniques were employed, namely Word2Vec, GloVe, and FastText. The objective of using multiple embedding methods is to evaluate their effectiveness in representing nutritional and ingredient-related textual information for healthy and unhealthy snack classification. The preprocessed ingredient descriptions were first combined into a textual corpus. Subsequently, each embedding method generated numerical vector representations for the words contained in the corpus. The resulting word vectors were then aggregated to obtain document-level representations that served as input features for the Random Forest classifier.

2.5.1. Word2Vec

Word2Vec is a neural network-based word embedding technique introduced by Mikolov et al. It learns vector representations by analyzing contextual relationships between words within a corpus. Word2Vec employs two architectures, namely Continuous Bag-of-Words (CBOW) and Skip-Gram.

The generated vectors are capable of capturing semantic and syntactic relationships among words. In this study, Word2Vec was utilized to generate dense vector representations of ingredient descriptions and nutritional information. The resulting vectors were subsequently used as features for classification.

2.5.2. GloVe

Global Vectors for Word Representation (GloVe) is a word embedding technique that combines global statistical information and local contextual information from a corpus. Unlike Word2Vec, which focuses primarily on neighboring words, GloVe utilizes a word co-occurrence matrix to learn vector representations. By incorporating global corpus statistics, GloVe can effectively capture relationships among words appearing across the entire dataset. Therefore, GloVe was included in this study to evaluate its capability in representing nutritional-related textual data.

2.5.3. FastText

FastText is an extension of Word2Vec developed by Facebook AI Research. The main advantage of FastText is its ability to represent words using character-level subword information. Instead of treating words as indivisible units, FastText decomposes words into smaller character n-grams and generates embeddings based on these subword representations. This characteristic makes FastText particularly effective in handling rare words, spelling variations, and domain-specific terminology. Since ingredient descriptions often contain uncommon food-related terms, FastText was included to investigate whether subword-based representations improve classification performance. The numerical representation of a word can be expressed as:

$$w_i = [x_1, x_2, x_3, \dots, x_n] \tag{1}$$

where, w_i denotes the vector representation of the i-th word, $x_1, x_2, \dots, x_n$ are numerical values in the embedding space, n represents the embedding dimension.

The similarity between two word vectors can be measured using cosine similarity:

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} \tag{2}$$

where, A and B represent two word vectors, $A \cdot B$ is the dot product between vectors, $\| A \| \| B \|$ denote the vector magnitudes.

The document vectors generated by Word2Vec, GloVe, and FastText were used separately as input features for the Random Forest classifier. The performance of each embedding method was subsequently compared using accuracy, precision, recall, and F1-score to determine the most effective representation technique for healthy and unhealthy snack product classification.

2.6. Data Splitting and Class Balancing

Following the word embedding process, the resulting document vectors were divided into training and testing datasets to evaluate the generalization capability of the proposed classification model. Prior to dataset splitting, class imbalance was addressed using a random undersampling strategy applied to the majority class, resulting in a final dataset of 465 samples. The balanced dataset consisted of 186 healthy products and 279 unhealthy products. The dataset was subsequently divided into training and testing sets using an 80:20 ratio. Based on this configuration, 372 samples were allocated to the training set and 93 samples were reserved for the testing set. The training dataset was used to develop the Random Forest classification model, while the testing dataset was used exclusively for performance evaluation on previously unseen data. Since the balancing process had already been completed before dataset partitioning, no additional oversampling or resampling techniques were applied during model training. Table 4 presents the dataset splitting configuration used in this study.

As shown in Table 4, the dataset was divided into 372 training samples and 93 testing samples using an 80:20 train-test split ratio. The training set was utilized to develop the Random Forest classification model, while the testing set was reserved for evaluating the model's performance on

previously unseen data. This partitioning strategy provides sufficient data for model learning while ensuring reliable and unbiased performance assessment.

Table 4. Dataset splitting configuration.

Dataset	Percentage	Number of samples
Training set	80%	372
Testing set	20%	93
Total	100%	465

2.7. Random Forest Classification

Random Forest was employed as the classification algorithm in this study. Random Forest is an ensemble learning method that constructs multiple decision trees using bootstrap sampling and random feature selection. Each tree independently generates a prediction, and the final class label is determined using a majority voting mechanism. The prediction function of Random Forest can be expressed as:

$$\hat{Y} = \text{mode}\{h_1(x), h_2(x), \dots, h_n(x)\} \tag{3}$$

where, $\hat{Y}$ represents the final predicted class, $h_i(x)$ denotes the prediction of the i-th decision tree, n is the number of decision trees.

The algorithm was selected due to its ability to handle high-dimensional feature spaces, reduce overfitting, and achieve robust classification performance.

2.8. Model Evaluation

The performance of the proposed model was evaluated using a confusion matrix and four commonly used classification metrics: (1) Accuracy, (2) Precision, (3) Recall, and (4) F1-score.

Accuracy measures the overall correctness of predictions:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

Precision evaluates the proportion of correctly predicted positive samples:

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

Recall measures the ability of the model to identify positive samples:

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

F1-score evaluates the balance between precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{7}$$

These metrics provide a comprehensive assessment of classification performance and enable comparison with future studies.

2.9. Result Analysis

The final stage involves analyzing the experimental results obtained from the Random Forest model. The evaluation metrics are interpreted to determine the effectiveness of word embedding in representing nutritional text data and the capability of Random Forest in classifying healthy and unhealthy snack products. The analysis also discusses the strengths and limitations of the proposed approach and provides recommendations for future research in food informatics and NLP-based nutritional assessment.

3. RESULTS AND DISCUSSIONS

This section presents the experimental results obtained from the classification of healthy and unhealthy snack products using Random Forest combined with three word embedding techniques, namely Word2Vec, GloVe, and FastText. The performance of each model was evaluated using Accuracy, Balanced Accuracy, Precision, Recall, F1-score, and Macro F1-score metrics. Furthermore, confusion matrix analysis was conducted to assess the classification capability of each embedding method.

3.1. Classification Performance Comparison

The performance comparison of Random Forest using different word embedding techniques is presented in Table 5.

Table 5. Performance comparison of word embedding methods embedding method.

Embedding method	Accuracy	Balanced accuracy	Precision	Recall	F1-score	Macro F1-score
Word2Vec	0.8172	0.8024	0.8160	0.8172	0.8157	0.8064
GloVe	0.8602	0.8472	0.8598	0.8602	0.8591	0.8519
FastText	0.7957	0.7845	0.7949	0.7957	0.7952	0.7858

The results demonstrate that GloVe achieved the best classification performance among the evaluated embedding methods. Specifically, GloVe obtained an accuracy of 86.02%, outperforming Word2Vec by 4.30 percentage points and FastText by 6.45 percentage points. Similar trends can be observed for Precision, Recall, F1-score, and Macro F1-score, indicating that GloVe consistently provided superior feature representations for the classification task.

As illustrated in Figure 4, GloVe consistently achieved the highest scores across all evaluation metrics. The performance improvement suggests that the global word co-occurrence information utilized by GloVe is more effective in capturing semantic relationships among ingredient descriptions and nutritional terms than the local context modeling of Word2Vec and the subword representation employed by FastText.

3.2. F1-Score Analysis

Since the dataset originally exhibited class imbalance, the F1-score was used as an additional evaluation metric to assess the balance between precision and recall.

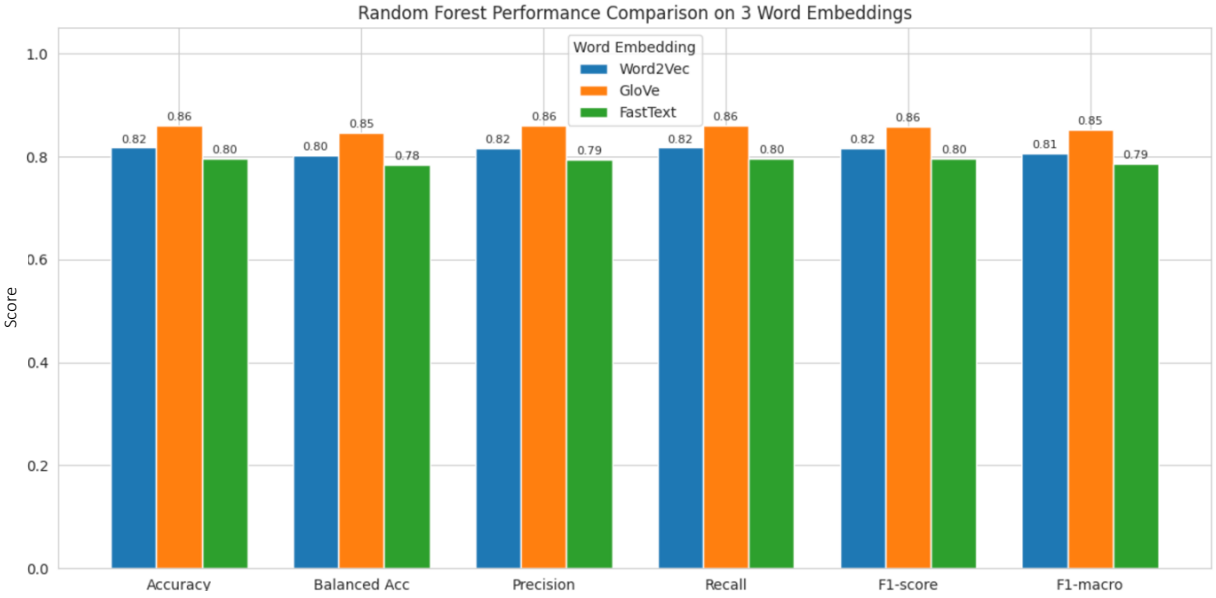

Figure 4. Performance comparison of random forest using different word embedding methods.

Figure 5 shows that GloVe achieved the highest weighted F1-score of 0.859, followed by Word2Vec with 0.816 and FastText with 0.795. The superior F1-score obtained by GloVe indicates that the model was able to maintain a better balance between correctly identifying healthy and unhealthy snack products while minimizing classification errors. The results further confirm that GloVe provides a more discriminative semantic representation of food ingredient descriptions, leading to improved classification performance.

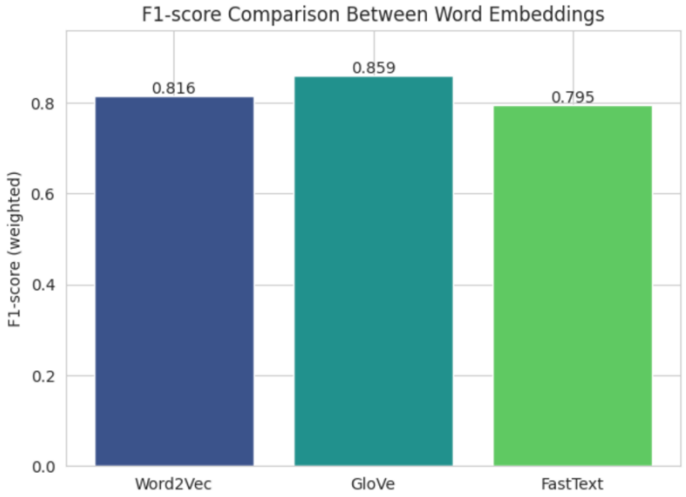

Figure 5. F1-score comparison between word embedding methods.

3.3. Confusion Matrix Analysis

To further evaluate the classification capability of each embedding method, confusion matrices were generated for Word2Vec, GloVe, and FastText.

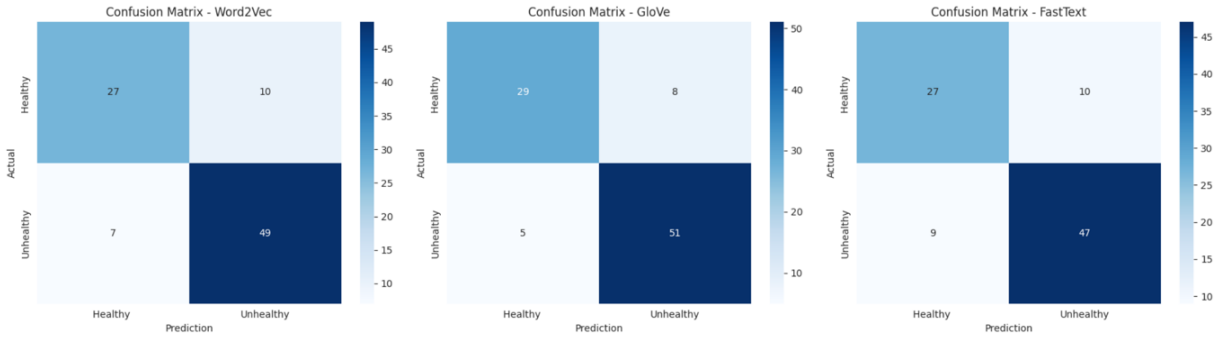

Figure 6. Confusion matrices of the evaluated models.

For the Word2Vec-based model, 27 healthy products and 49 unhealthy products were correctly classified, while 17 samples were misclassified. Although the model demonstrated good performance in identifying unhealthy products, several healthy products were incorrectly classified as unhealthy. The GloVe-based model achieved the best classification results, correctly identifying 29 healthy products and 51 unhealthy products. Only 13 samples were misclassified, resulting in the highest overall accuracy and F1-score among all evaluated methods. The confusion matrix indicates that GloVe was able to distinguish the two classes more effectively than the other embedding techniques. Meanwhile, the FastText-based model correctly classified 27 healthy products and 47 unhealthy products. However, it produced more classification errors than GloVe, particularly in distinguishing healthy products from unhealthy ones. This result contributed to the lower accuracy and F1-score achieved by FastText. Overall, the confusion matrix analysis demonstrates that GloVe provides the most reliable representation for distinguishing healthy and unhealthy snack products based on textual nutritional information.

3.4. Discussion

The experimental results indicate that the choice of word embedding significantly influences the performance of snack product classification. Among the evaluated methods, GloVe consistently achieved the best performance across all evaluation metrics. This finding suggests that global co-occurrence statistics are particularly effective for representing ingredient descriptions and nutritional information contained in food labels. Unlike Word2Vec, which primarily learns local contextual relationships, GloVe incorporates global corpus information, enabling it to capture broader semantic associations among food-related terms. Such capability appears beneficial for identifying nutritional patterns that distinguish healthy and unhealthy snack products. Although FastText is designed to capture subword information and is generally effective for handling rare words, its performance was lower in this study. One possible explanation is that ingredient descriptions in the Open Food Facts dataset contain relatively consistent vocabulary, reducing the advantage provided by character-level representations. Based on the obtained results, the combination of GloVe and Random Forest can be considered the most effective approach for healthy and unhealthy snack product classification. The model achieved an accuracy of 86.02% and an F1-score of 85.91%, demonstrating its potential for supporting automated nutritional assessment systems and assisting consumers in making healthier food choices.

4. CONCLUSION

This study proposed a machine learning framework for classifying snack products into healthy and unhealthy categories using textual nutritional information from the Open Food Facts dataset. The proposed framework integrates Natural Language Processing (NLP), word embedding techniques, and the Random Forest classification algorithm to automatically analyze ingredient descriptions and nutritional information contained in snack product labels. The experimental results demonstrate that the selection of word embedding techniques significantly affects classification performance. Among the evaluated methods, GloVe achieved the best performance, obtaining an accuracy of 86.02%, balanced accuracy of 84.72%, precision of 85.98%, recall of 86.02%, F1-score of 85.91%, and macro F1-score of 85.19%. In comparison, Word2Vec achieved an accuracy of 81.72%, while FastText obtained 79.57%. These findings indicate that GloVe provides a more effective semantic representation of food-related textual data than Word2Vec and FastText for the healthy and unhealthy snack classification task. The confusion matrix analysis further confirmed the superiority of the GloVe-based model, which produced the highest number of correctly classified samples and the lowest classification error among all evaluated approaches. The ability of GloVe to capture global word co-occurrence information appears to contribute significantly to its superior performance in representing ingredient descriptions and nutritional terminology. Overall, the results demonstrate that the combination of GloVe and Random Forest is an effective approach for classifying healthy and unhealthy snack products based on textual nutritional information. The proposed framework has the potential to support automated nutritional assessment systems and assist consumers in making more informed food choices. Future work may explore the integration of deep learning architectures, contextual embeddings such as BERT, and multimodal information combining textual and numerical nutritional features to further improve classification performance.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to Universitas Sains dan Teknologi Indonesia (USTI) for providing academic support and research facilities that contributed to the completion of this study. The authors also acknowledge the Open Food Facts community for making the food product dataset publicly available, which served as the primary data source for this research. Finally, the authors would like to thank all individuals who provided valuable support, guidance, and encouragement throughout the research process.

REFERENCES

[1] Fitriani, R. J. (2025). The Trend of Ultra-Processed Food Consumption and Its Impact on Obesity Risk in Indonesia. *Journal Nutrizone*, **2**(3), 50–61.

[2] Zou, L., Fu, X., Huang, J., Liu, W., Zhou, L., Zhou, Y., Lin, Y., & Liu, L. (2025). Health assessment of snacks and desserts in Guizhou Province: Analysis of fatty acids and sugar content. *PLoS One*, **20**(6), e0321857.

[3] Shangguan, S., Afshin, A., Shulkin, M., Ma, W., Marsden, D., Smith, J., Saheb-Kashaf, M., Shi, P., Micha, R., Imamura, F. & Mozaffarian, D. (2019). A meta-analysis of food labeling effects on consumer diet behaviors and industry practices. *Am. J. Prev. Med.*, **56**(2), 300–314.

[4] Wang, X. (2024). The impact of food nutrition labels on consumer behavior: A cross-national survey and quantitative analysis. *Int. J. Public Health Res.*, **1**(2), 18–27.

[5] Gbashi, S. & Njobeh, P. B. (2024). Enhancing Food Integrity through Artificial Intelligence and Machine Learning: A Comprehensive Review. *Applied Sciences*, **14**(8).

[6] Yang, H., Jiao, W., Zouyi, L., Diao, H. & Xia, S. (2025). Artificial intelligence in the food industry: innovations and applications. *Discover Artificial Intelligence*, **5**(1), 60.

[7] Alkalbani, N., Shahin, L., Benzeghiba, H., Obaid, R. S., Osaili, T. M., Cheikh Ismail, L., Al qasssimi, G., Rauf, M., Abdulrahim, K., Almashgouni, A., Ashuweihi, F. & AL-Fuqaha, D. (2026). Artificial intelligence in functional food innovation: Bioactive enhancement and formulation optimization: A quasi-systematic review. *Food Chemistry: X*, **34**, 103628.

[8] Arslan, S. (2025). Artificial intelligence in food safety and nutrition practices: opportunities and risks. *Academia Nutrition and Dietetics*, **2**(3).

[9] Jerfy, A., Selden, O., & Balkrishnan, R. (2024). The growing impact of natural language processing in healthcare and public health. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, **61**, 00469580241290095.

[10] Özen, N., Papadopoulou, F., Mutlu, O., Öztürk, B., Mu, W., Bulk, L. van den, Velden, B. van der & Hürriyetoğlu, A. (2026). Natural Language Processing In Food Safety Research: A Systematic Review, *Research Square*, 9305529.

[11] Hu, G., Ahmed, M. & L'Abbé, M. R. (2023). Natural language processing and machine learning approaches for food categorization and nutrition quality prediction compared with traditional methods. *The American Journal of Clinical Nutrition*, **117**(3), 553–563.

[12] Xiong, S., Tian, W., Si, H., Zhang, G. & Shi, L. (2024). A Survey of the Applications of Text Mining for the Food Domain. *Algorithms*, **17**(5).

[13] Elhosary, E. & Moselhi, O. (2025). Evaluating Natural Language Processing Algorithms for Improved Hazard and Operability Analysis. *Geodata and AI*, **4**, 100026.

[14] Abimbola, J. O., Kuaban, G. S. & Ajayi, S. A. (2026). Open-Source Embedding Models: A Comprehensive Survey of Techniques, Benchmarks, and Applications. *IEEE Access*, **14**, 41284.

[15] Li, Y. & Yang, T. (2017). Word embedding for understanding natural language: a survey. Guide to Big Data Applications, 83–104.

[16] Khan, I. A., Birkhofer, H., Kunz, D., Lukas, D. & Ploshikhin, V. (2023). A random forest classifier for anomaly detection in laser-powder bed fusion using optical monitoring. *Materials*, **16**(19).

[17] Khan, A. A., Chaudhari, O. & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications*, **244**, 122778.

[18] Salman, H. A., Kalakech, A. & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, 2024, 69–79.

[19] Adugna, T., Xu, W. & Fan, J. (2022). Comparison of Random Forest and Support Vector Machine Classifiers for Regional Land Cover Mapping Using Coarse Resolution FY-3C Images. *Remote Sensing*, **14**(3).

[20] Jalal, N., Mehmood, A., Choi, G. S. & Ashraf, I. (2022). A novel improved random forest for text classification using feature ranking and optimal number of trees. *Journal of King Saud University - Computer and Information Sciences*, **34**(6, Part A), 2733–2742.

[21] Huang, X., Li, Z., Li, Z., Shi, J., Zhang, N., Qin, Z., Du, L., Shen, T. & Zhang, R. (2025). Application of Image Computing in Non-Destructive Detection of Chinese Cuisine. *Foods*, **14**.

[22] Tahtouh, T., Salman, H., Eissa, N., Nassar, N. Al, Maghaydah, S., Alhalabi, M., Yaghi, M., Gad, A., Abdallah, D., Elberry, S., Alhosani, A., Alshehhi, S., Alkhedher, M., Ramadan, M. & Ghazal, M. (2025). Technological enhancements in personalized dietary management for chronic conditions. *Biomedical Engineering Advances*, **10**, 100181.