Iranian Food Science and Technology Research Journal
Homepage: https://ifstrj.um.ac.ir

Research Article
Vol. 21, No. 3, Aug.-Sep. 2025, p. 271-286

Implementation of Several Data Mining Strategies on Electronic Nose Data for Identifying Gluten in Cheese

M. Nasiri-Galeh[1*], M. Ghasemi-Varnamkhasti[2]

1- Department of Information Technology Management, Faculty of Management and Economics, Tarbiat Modares University, Tehran, Iran
(*- Corresponding Author Email: m_nasiri@modares.ac.ir)
2- Department of Biosystems Mechanical Engineering, Faculty of Agriculture, Shahrekord University, Shahrekord, Iran

Received: 27.08.2024
Revised: 28.12.2024
Accepted: 29.12.2024
Available Online: 17.06.2025

How to cite this article:
Nasiri-Galeh, M., & Ghasemi-Varnamkhasti, M. (2025). Implementation of several data mining strategies on electronic nose data for Identifying gluten in cheese. *Iranian Food Science and Technology Research Journal*, *21*(3), 271-286. https://doi.org/10.22067/ifstrj.2024.89599.1360

Abstract

Electronic nose is an electronic device for smell detection. The data obtained from this device are stored in the form of numbers in different columns, which are related to the data of two types of cheese namely gluten-free cheese and cheese with gluten. It is not enough to make decisions and judge the data unless discovering the relationships and patterns between the data obtained to determine the relation of new data recorded by the device to the type of cheese, for this purpose, data mining and machine learning methods have been used in this research. Data mining includes various algorithms such as classification, clustering, and obtaining association rules. To get a better result from the data, a data mining process was performed on 105 different permutations of the models, and 13 models with the highest accuracy in understanding the relationships between the data were chosen. In this research, with data mining methods, cheese with gluten and gluten-free cheese data were classified into separate categories, and a model was created to predict the type of new input data in terms of the nature of cheese (gluten-free and with gluten). With analyzing 105 Permutations, Finally, the best suitable model to be used for data classification using the Random Forest algorithm and MinMaxScaler for scaling was selected with a prediction accuracy of 99.8% for both test and training datasets.

Keywords: Data classification, Data mining, Decision tree, Electronic nose, Machine learning

Introduction

Celiac disease is one of the most common diseases related to nutrition (Gh. Shekari, 2024). People suffered by this disease are allergic to gluten in food, so it is necessary to design a method to detect gluten in food with high accuracy, and one of these methods is the application of electronic nose. Smell collection technology was first created in 1982 with the invention of an array of sensors (Persaud & Dodd, 1982).

In recent years, electronic nose (E-nose) technology has been widely utilized in the food industry to assess quality, authenticity, and safety of products (Wilson, 2009). These devices, by mimicking the human olfactory system, can detect and differentiate volatile compounds present in food products. E-nose consists of an array of chemical sensors that respond to volatile organic compounds, generating unique olfactory patterns. These patterns are analyzed using data mining techniques and multivariate analysis methods, such as Principal Component Analysis (PCA),

https://doi.org/10.22067/ifstrj.2024.89599.1360

Linear Discriminant Analysis (LDA), and Artificial Neural Networks (ANN), enabling precise differentiation and identification of various samples (Zhang, 2023).

The applications of E-nose technology in the food industry are vast, including the evaluation of meat quality, detection of spoilage, determination of shelf life, and identification of food fraud (Zhao, 2024). Recently, there has been increasing attention on employing this technology to evaluate the chemical and sensory properties of dairy products, particularly cheese. Studies have demonstrated that E-nose systems can identify different types of cheese based on their olfactory patterns and even predict the intensity of aroma with high accuracy (Fernandez, 2023).

On the other hand, the increasing prevalence of gluten-related diseases, particularly celiac disease, has brought significant attention to the production and monitoring of gluten-free food products. Celiac disease is a chronic autoimmune disorder triggered by the consumption of gluten, a protein found in wheat, barley, and rye, which damages the small intestine's villi. In individuals with celiac disease, even trace amounts of gluten can cause severe complications, including malabsorption of nutrients, weight loss, gastrointestinal problems, and long-term risks such as osteoporosis and intestinal cancers (Thompson, 2023).

Given the global rise in the number of celiac patients and the growing demand for accurate and rapid detection of gluten in food products, advanced technologies like the electronic nose, combined with data mining methods, offer a promising non-destructive solution for quality control and the identification of potential gluten contamination, especially in dairy products. Moreover, utilizing advanced data analysis techniques, such as machine learning algorithms and sophisticated modeling, can significantly enhance the accuracy and efficiency of E-nose systems for gluten detection (Yu, 2024).

One of the key innovations of this study is that the employed methods can be integrated into the design and development of next-generation electronic nose devices. Incorporating advanced algorithms and precise modeling will improve the performance of these systems, increase detection accuracy, and reduce potential errors. Such improvements will contribute to the development of smarter and more efficient devices that can better meet the demands of the food industry.

This paper investigates gluten detection in cheese samples using data generated by an electronic nose and several data mining methods. The study is structured to include the following sections: an Introduction, providing the background and motivation for the research; Materials and Methods, detailing the data collection and preprocessing steps; Algorithms Used, describing the machine learning techniques applied; Modeling, outlining the process of building and validating the models; and Results and Discussion, presenting the findings and their implications.

Justification for the Study

The use of electronic noses (E-noses) in food quality control has become increasingly significant due to their ability to provide rapid and non-invasive detection of various food components. In the context of gluten detection in dairy products such as cheese, this technology offers a promising alternative to conventional methods. Traditional gluten detection methods, such as ELISA (Enzyme-Linked Immunosorbent Assay) and PCR (Polymerase Chain Reaction), although accurate, are often labor-intensive, time-consuming, and require specialized laboratory conditions. Given the growing concerns about gluten contamination in foods and the rising prevalence of gluten-related disorders, there is a critical need for more efficient, accurate, and user-friendly detection methods. This study aims to apply advanced data mining strategies to E-nose data, providing a novel approach for the reliable identification of gluten in cheese.

Research Gaps

1. **Insufficient Application of E-noses in Gluten Detection**: While E-noses have been widely applied in various fields of food analysis, their specific use for gluten detection in

dairy products remains underexplored. Existing studies have primarily focused on other food matrices, leaving a significant gap in the research related to dairy products, particularly cheese (Bhattacharya, 2008).

2. **Limited Integration of Data Mining Techniques**: Previous research utilizing E-noses has often relied on basic statistical analysis rather than employing advanced data mining techniques. This has limited the potential of E-noses in detecting complex food contaminants like gluten. There is a need for studies that explore a broader range of data mining algorithms to improve the accuracy and robustness of gluten detection (Wilson, 2013)
3. **Challenges in Ensuring Cross-Product Applicability**: Most existing studies on gluten detection are focused on specific food products or use limited datasets, which hampers the generalizability of the findings. There is a lack of comprehensive studies that validate the effectiveness of these techniques across different dairy products, making it difficult to apply these findings in a broader context (Karoui, 2011). This study aims to address this gap by using diverse data sets and evaluating multiple data mining techniques to develop a more generalizable model.

Materials and Methods

In this study, all data mining processes were performed by manual coding in the Python programming language (Python Software Foundation), which are explained in the following algorithms used in this research. The tested samples are gluten-free cheese and gluten-containing cheese, prepared from reputable stores. Data collection was done by the electronic nose device made by the authors, which will be briefly explained below.

Data Collection Electronic Nose Device

In this study, both gluten-free and gluten-containing cheese products were examined. For the gluten-free samples, commercially available products specifically designed for individuals with celiac disease were utilized. These products were verified to be gluten-free through reference methods prior to their inclusion in the analysis, ensuring their suitability for the study.

For the gluten-containing samples, regular cheese products available in the market, known to contain gluten, were used. These products were selected to represent typical gluten-containing cheeses and provided a reliable basis for comparison in the analysis.

The data collection process was completely carried out by the electronic nose device made by the authors, which is briefly explained below. Fig. 1 shows a schematic diagram of an electronic nose.

Data Processing and Preprocessing

Data obtained from the electronic nose was initially in raw form and stored in JSON format. To prepare the data for analysis, visualization, and modeling, it was first necessary to convert it into formats compatible with data modeling and visualization tools. This conversion ensured that the data could be effectively interpreted and utilized for subsequent processes.

Following this, a comprehensive data preprocessing phase was conducted to enhance data quality and ensure accuracy. This process included:

1. Data Cleaning: Outlier values were identified and removed, and missing or invalid data points were appropriately handled to prevent them from compromising the results.
2. Normalization: To ensure consistency across features, scaling methods such as MinMaxScaler were applied to normalize the data. This step allowed all features to have comparable ranges, which is crucial for the optimal performance of many machine learning algorithms.

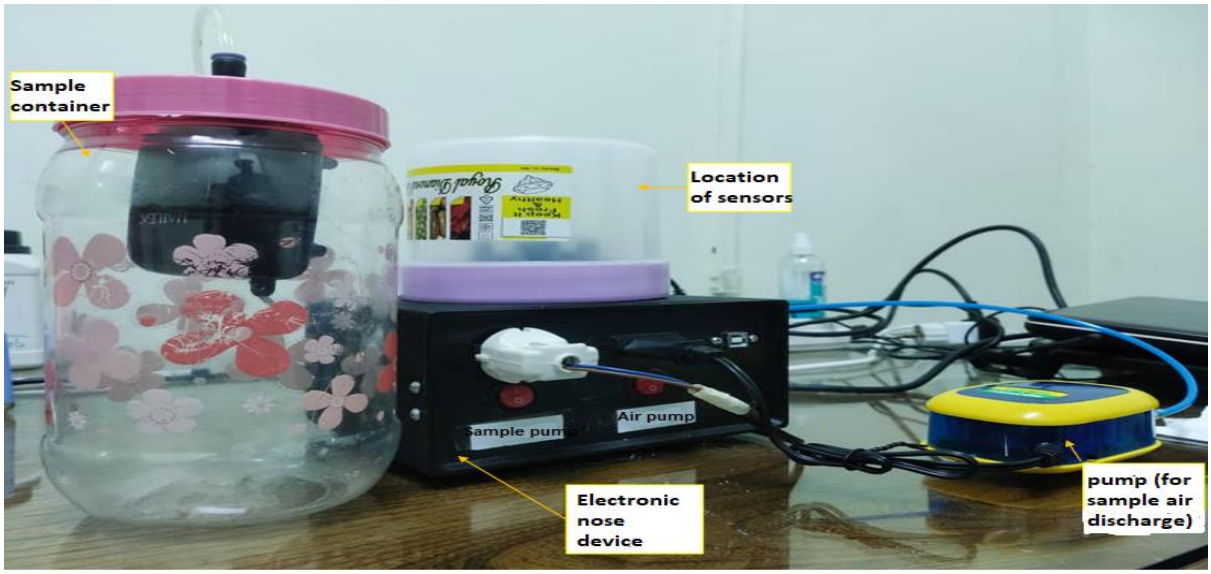

Fig. 1. The electronic nose device used in the research

The sample to be tested is placed in the sample container, then by the sample air discharge pump, the sample air, which is accompanied by the smell emitted from the sample, is directed to the location of sensors, and the sensors then store the desired data.

In addition, specific preprocessing strategies were tailored to meet the requirements of different machine learning models. For certain models, dimensionality reduction techniques were applied to simplify the dataset and enhance performance without sacrificing accuracy. This was done when it was determined that reducing the number of features would lead to better outcomes. Conversely, for models where dimensionality reduction was unnecessary or unsuitable, alternative methods of data preparation were employed.

After completing these preprocessing steps, the data was fully processed, cleaned, normalized, and formatted. The resulting output, as shown in Fig. 2, served as the foundation for the machine learning processes described in subsequent sections.

Furthermore, due to the high accuracy achieved with the applied machine learning methods, additional techniques, such as the area under the curve (AUC) or integration-based methods, were deemed unnecessary. The robustness and precision of the employed algorithms ensured reliable results, eliminating the need for supplementary approaches and streamlining the analysis.

Fig. 2 shows an example of a graph obtained from a sampling device.

Used Algorithms

AdaBoost

Boosting is an approach to machine learning based on the idea of creating a strong rule from by combining of relatively weak rules (Schapire, 2013) which was introduced in 1990 by Freund and Schapire (Schapire, 197-227). Booting algorithms for classification and regression issues provide a solution for easier comparison of algorithms (Hu, 2008) The AdaBoost algorithm was the first Boosting algorithm developed as an application algorithm and used in a variety of applications such as classification issues (Freund, 1997).

The AdaBoost algorithm is a boosting classification method designed to enhance weak classifiers and transform them into strong ones. It typically begins with a basic classification algorithm, which is used as a template to train an initial classifier on the training data. The algorithm then adjusts the sample weights based on the performance of this classifier. These re-weighted samples are subsequently used to train the next-level learner. An iterative design of this process was

conducted and the algorithm assigns optimized weights to each learner was achieved, ultimately forming the final, robust classification model (Wang, 2019).

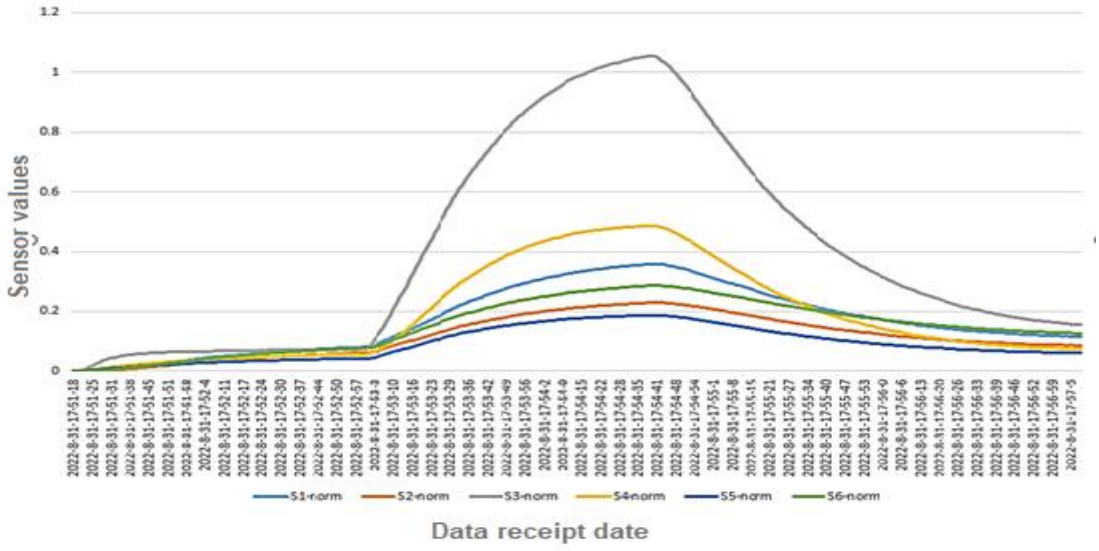

Fig. 2. A diagram drawn with the data collected by the 6 sensors of the device
This chart displays data collected by six sensors, with each sensor's data represented in a different color. Data were plotted based on date and categorized by the type of sensor.

Decision Tree

Decision Tree classification is one of the most well-known machine-learning techniques presented by Quinlan (Quinlan, 1993). The Decision Tree in classification is one of the multi-stage decision-making methods. The general method in the Decision Tree is that a complex decision is divided into a set of simple decisions and finally, by solving a set of these simple decisions, the desired output for the main complex decision is reached.

After the Decision Tree is created, it can be used to classify the test data that have the same characteristics as the training data (Stein, 2005). The general method of the Decision Tree can be displayed in Fig. 3. As shown in this figure, the complex decision, which is a set of pixels on the left side, is transformed into simple decisions on the right side at each stage, so that by solving those simple decisions, the complex decision is finally solved.

Decision Tree models are suitable for data mining due to their acceptable accuracy and low computational cost (Du, 2002). Most Decision Tree classifiers (such as C4.5) perform the classification in two steps: tree construction and tree pruning.

In tree construction, the decision tree model is built by recursive partitioning. Tree pruning is used to improve the generalization of a decision in a Decision Tree, as well as to prune the leaves and branches responsible for classifying single or very small vector data (Du, 2002).

Random Forest

The Random Forest was invented in the early 2000s by L. Breiman (Breiman, 2001). Random Forest is a collection of trees that are taught independently (Breiman, 2001). For the final prediction, the Random Forest combines the predictions of all trees with the average, which is a generalization property (Criminisi, 2011) .

By random sampling, a subset of educational data is used to learn a separate tree (Ren, 2015).

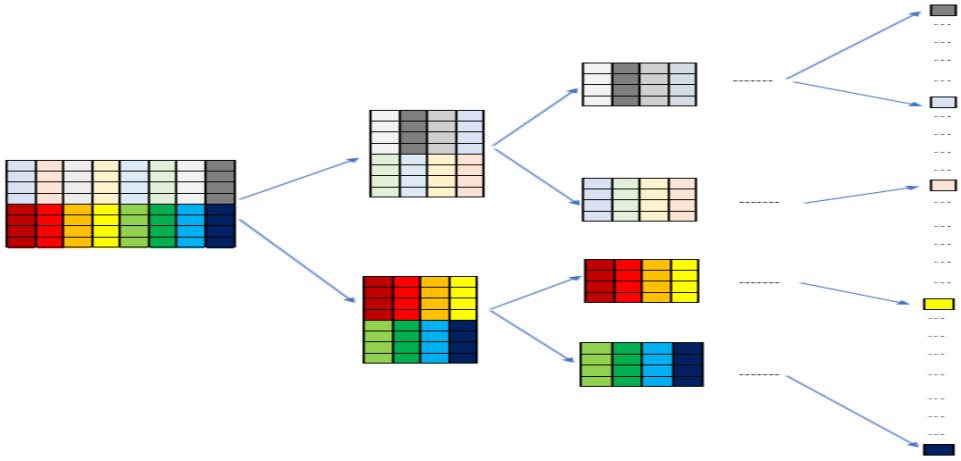

Fig. 3. A general image of the division of decision making in the Decision Tree

Deep trees are the main source of the power of a Random Forest; although reduction randomness can also improve the robustness of an individual tree, it is preferred to increase the depth of trees because a special randomization procedure is required to ensure complementarity with other trees (Ren, 2015).

Among machine learning algorithms, the Random Forest works very well in terms of the accuracy of forecasting and interpretation of the model (Qi, 2012). The Random Forest algorithm for classification and regression is based on the accumulation of a large number of decision trees.

Random Forest focuses on three features (Breiman, 2001):

1. Provides accurate predictions for many applications.
2. It can measure the importance of each variable in the model made by teaching the model and ranking the variables according to their ability to predict the response.
3. The pairwise closeness between samples can be measured by the training model.

The Random Forest method can be used for a set of forecasting issues and receives quantitative parameters as input. It can also deal with really large systems (Biau, 2016).

Support Vector Machine (SVM)

Support vector machines (SVM) were presented by Vladimir Vapnik (Vapnik, 1998) in the field of statistical learning theory and minimizing structural risk. It works successfully on various classification and forecasting problems.

In machine learning, the Support Vector Machine (SVM) method is a supervised learning approach used for both classification and regression tasks. An SVM classifier works by distinguishing between different data sets, often through the creation of a non-linear decision boundary.

The algorithm takes a labeled training dataset as input, where each data point belongs to a specific category. During the training process, the SVM constructs a model that can determine the category of new examples, enabling accurate classification of unseen data.

Modeling

Data classification was utilized in this research because the dataset comprises two distinct types of data, and the objective is to define separate regions that can classify new data accurately. The Python programming language was used for modeling, along with various machine learning algorithms. Specifically, the sklearn library was employed for implementing the algorithms and defining the search space, while the matplotlib library was used to generate graphical outputs. Additionally, the NumPy and pandas libraries facilitated numerical operations on the data. Various algorithms were applied for

classification, scaling, and dimensionality reduction, resulting in 105 permutations of different algorithm combinations to identify the optimal output.

Results and Discussion

Explanation of the Data Entry into the Analysis

One way to ensure the accuracy of the output data and verify the results during testing is to repeat the test. In this data collection, both gluten-containing and gluten-free cheeses were tested, and the data was recorded. For each type of cheese, the test was repeated 7 times, resulting in a total of 14 tests. The outputs from these tests were stored in separate files across 10 columns for each test, which included data on time, temperature, humidity, device voltage, and readings from 6 gas sensors.

For the machine learning analysis, there are two types of data: training data and testing data. To improve the algorithm's accuracy and ensure precise results, one of the 7 experiments for each type of cheese was used as the test data, while all 14 experiments were used as the training data.

The Results of Machine Learning

Almost all well-known models and algorithms were used interchangeably for three tasks: classification, scaling, and dimensionality reduction. To select the optimal model, it is crucial that the accuracy percentages of both the test and training data are high and similar. A significant discrepancy between the accuracy of the test and training data indicates that the model is unsuitable for prediction and may be overfitting, that means it is not a good candidate for final predictions. To determine the best model, 105 permutations were tested, of which 13 permutations with the highest prediction accuracy are described below, along with their corresponding outputs.

Explanation of the Table

The output table generated from the implementation of machine learning models contains several key performance metrics that provide insights into the efficacy of the models. Below is an explanation of the metrics included in the table:

- **Precision**: This metric indicates the accuracy of the model in predicting correct outputs. It represents the probability that a positive prediction by the model is actually correct. A higher precision value reflects fewer false positives, emphasizing the reliability of the predictions.
- **Recall**: Also known as sensitivity or coverage, this metric shows the percentage of actual positive instances in the data that were correctly identified by the model. It measures the extent to which the model has successfully captured relevant data points during the modeling process.
- **F1-Score**: This metric is the harmonic mean of Precision and Recall. It provides a single performance score that balances both metrics, particularly useful when there is an uneven distribution of classes or when both false positives and false negatives need to be considered equally.

These metrics collectively offer a comprehensive evaluation of the model's performance, ensuring that both accuracy and coverage are accounted for. The detailed results from the table guide the selection of the most effective model for further analysis or application.

1-<Classifier: AdaBoost> <Scaler: MinMaxScaler>

This model uses the AdaBoost algorithm to classify data and MinMaxScaler to scale.

Table 1- Prediction accuracy of data classification model with AdaBoost and MinMaxScaler scaling

Type of cheese	Precision	Recall	F1-score
Gluten	0.996	0.993	0.995
Gluten-free	0.993	0.996	0.995

The prediction accuracy of the model is 99.6% for gluten-containing cheese and 99.3% for gluten-free cheese.

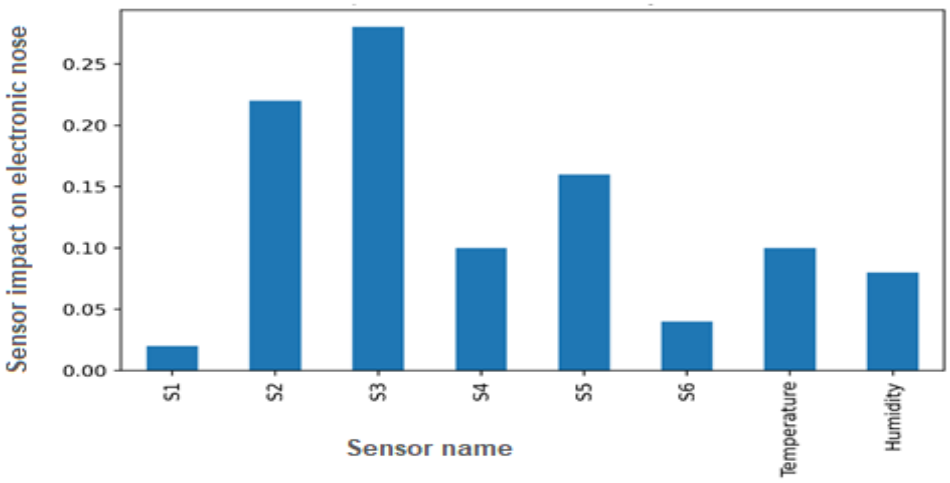

Fig. 4. The effect of each of the electronic nose sensors on the data classification model with AdaBoost and MinMaxScaler scaling

In model 1, the S3 and S2 sensors have the most impact, respectively.

2-<Classifier: AdaBoost> <Scaler:StandardScaler>

In this model, the AdaBoost algorithm is used for data classification and StandardScaler is used for scaling.

In model 2, the S3 and S2 sensors have the most impact, respectively.

3-<Classifier: AdaBoost>

In this model, the AdaBoost algorithm is used to classify data.

In model 3, sensors S3 and S2 have had the greatest impact, respectively.

4-<Classifier: DecisionTree-entropy > <Scaler: StandardScaler>

In this model, DecisionTree-entropy algorithm is used for data classification and StandardScaler is used for scaling.

Table 2- Prediction accuracy of data classification model with AdaBoost and StandardScaler scaling

Type of cheese	Precision	Recall	F1-score
Gluten	0.986	0.991	0.988
Gluten-free	0.991	0.986	0.988

The prediction accuracy of the model is 98.6% for gluten-containing cheese and 99.1% for gluten-free cheese.

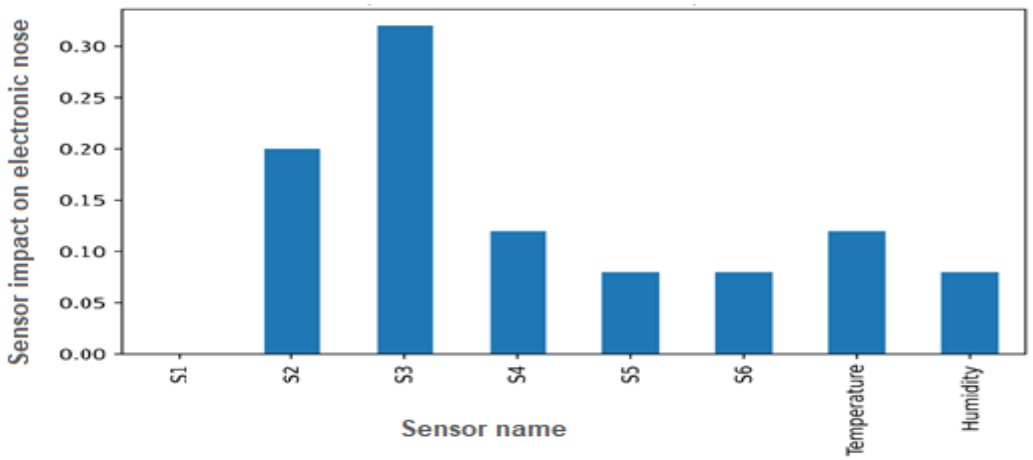

Fig. 5. The effect of each of the electronic nose sensors on the data classification model with AdaBoost and StandardScaler scaling

Table 3- Accuracy of data classification model with AdaBoost

Type of cheese	Precision	Recall	F1-score
Gluten	0.996	0.987	0.992
Gluten-free	0.988	0.997	0.992

The prediction accuracy of the model is 99.6 % for gluten-containing cheese and 98.8 % for gluten-free cheese.

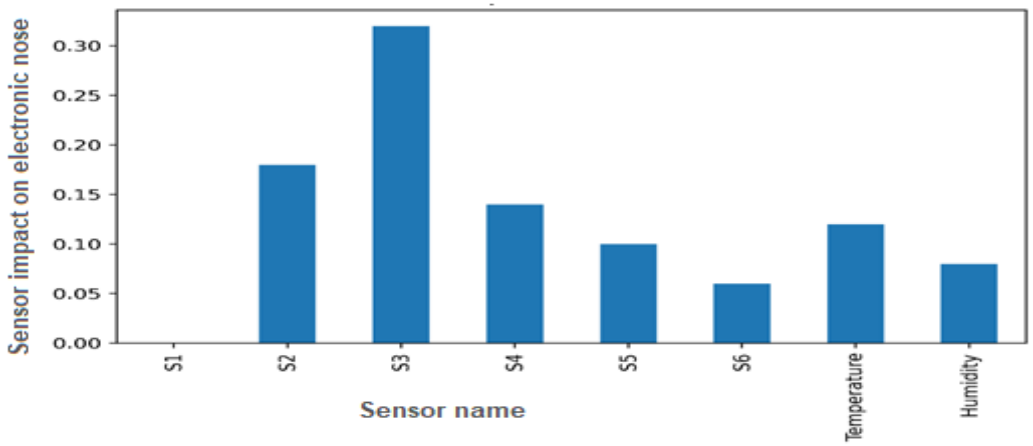

Fig. 6. The effect of each of the electronic nose sensors on the data classification model with AdaBoost

Table 4- Prediction accuracy of classification model with DecisionTree-entropy and scaling with StandardScaler

Type of cheese	Precision	Recall	F1-score
Gluten	0.995	0.995	0.995
Gluten-free	0.995	0.995	0.995

The prediction accuracy of the model is 99.5% for gluten-containing cheese and 99.5% for gluten-free cheese.

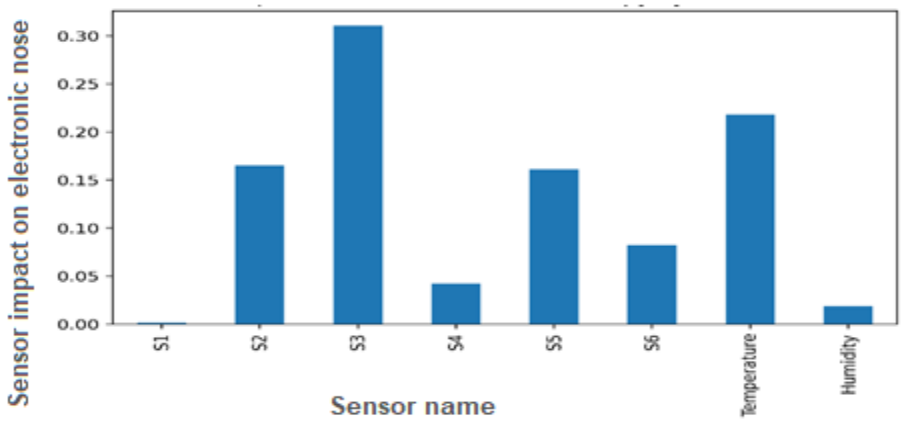

Fig. 7. The effect of each of the electronic nose sensors in the classification model with DecisionTree-entropy and scaling with StandardScaler

In model 4, S3 and Temperature sensors have had the greatest impact, respectively.

5-<Classifier: DecisionTree-entropy >

In this model, DecisionTree-entropy algorithm is used for data classification.

Table 5- Prediction accuracy of classification model with DecisionTree-entropy

Type of cheese	Precision	Recall	F1-score
Gluten	0.997	0.995	0.996
Gluten-free	0.995	0.996	0.995

The prediction accuracy of the model is 99.7% for gluten-containing cheese and 99.5% for gluten-free cheese.

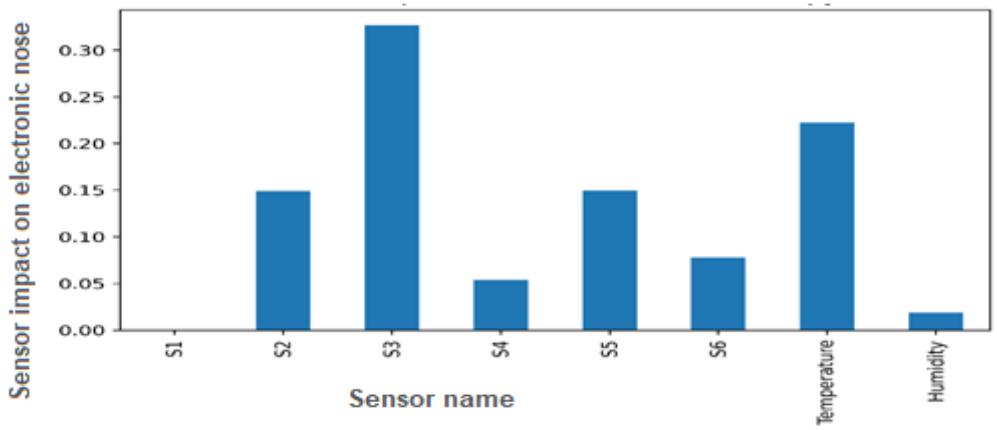

Fig. 8. The effect of each of the electronic nose sensors in the classification model with DecisionTree-entropy

In Model 5, the S3 and Temperature sensors had the most impact, respectively.

6-<Classifier: DecisionTree-gini> <Scaler: MinMaxScaler>

In this model, DecisionTree-gini algorithm is used for data classification and MinMaxScaler is used for scaling.

In Model 6, S3 and Temperature sensors have had the greatest impact, respectively.

7-<Classifier: DecisionTree-gini> <Scaler: StandardScaler>

In this model, DecisionTree-gini algorithm is used for data classification and StandardScaler is used for scaling.

The prediction accuracy of the model is 99.3% for gluten-containing cheese and 99.1% for gluten-free cheese.

8-<Classifier: DecisionTree-gini>

In this model, DecisionTree-gini algorithm is used for data classification.

Table 6- Prediction accuracy of classification model with DecisionTree-gini and scaling with MinMaxScaler

Type of cheese	Precision	Recall	F1-score
Gluten	0.996	0.996	0.996
Gluten-free	0.996	0.996	0.996

The prediction accuracy of the model is 99.6% for gluten-containing cheese and 99.6% for gluten-free cheese.

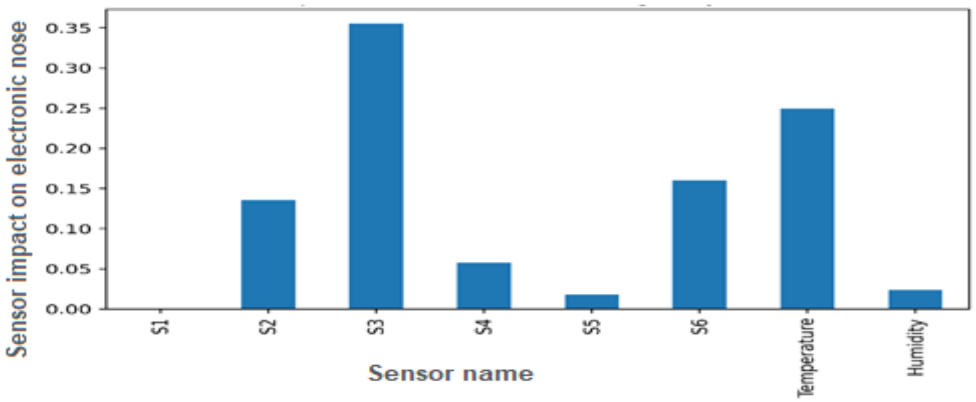

Fig. 9. The effect of each of the electronic nose sensors in the classification model with DecisionTree-gini and scaling with MinMaxScaler

Table 7- Prediction accuracy of classification model with DecisionTree-gini and scaling with StandardScaler

Type of cheese	Precision	Recall	F1-score
Gluten	0.993	0.991	0.992
Gluten-free	0.991	0.993	0.992

Table 8- Prediction accuracy of classification model with DecisionTree-gini

Type of cheese	Precision	Recall	F1-score
Gluten	0.995	0.993	0.994
Gluten-free	0.993	0.995	0.994

The prediction accuracy of the model is 99.5% for gluten-containing cheese and 99.3% for gluten-free cheese.

9-<Classifier: RandomForest> <Scaler: MinMaxScaler> <DimReducer: KernelPCA-poly>

In this model, RandomForest algorithm is used for data classification, MinMaxScaler is used for scaling, and KernelPCA-poly is used for dimensionality reduction.

Table 9- Prediction accuracy of classification model with RandomForest, scaling with MinMaxScaler and dimensionality reduction with KernelPCA-poly

Type of cheese	Precision	Recall	F1-score
Gluten	0.985	0.969	0.977
Gluten-free	0.971	0.986	0.978

The prediction accuracy of the model is 98.5% for gluten-containing cheese and 97.1% for gluten-free cheese.

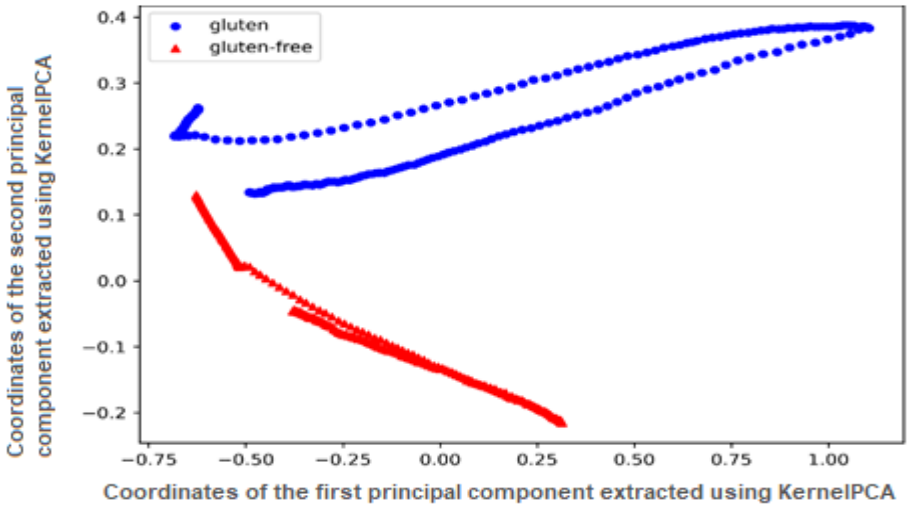

Fig. 10. Classification of data and drawing of the separating area in the classification model with RandomForest, MinMaxScaler scaling and dimensionality reduction with KernelPCA-poly

Gluten-containing and gluten-free cheeses are separated by delineating the area and with different colors with classification by RandomForest model, MinMaxScaler scaling, and KernelPCA-poly dimensionality reduction. In the classification method, as shown in Fig. 10, with the entry of new data, its area and, as a result, the type of data division that belongs to cheese with gluten or gluten-free is determined.

10-<Classifier: RandomForest> <Scaler: MinMaxScaler>

In this model, RandomForest algorithm is used for data classification and MinMaxScaler is used for scaling.

Table 10- Prediction accuracy of classification model with RandomForest and scaling with MinMaxScaler

Type of cheese	Precision	Recall	F1-score
Gluten	0.998	0.998	0.998
Gluten-free	0.998	0.998	0.998

The prediction accuracy of the model is 99.8% for gluten-containing cheese and 99.8% for gluten-free cheese.

11-<Classifier: RandomForest> <Scaler: StandardScaler>

In this model, RandomForest algorithm is used for data classification and StandardScaler is used for scaling.

Table 11- Prediction accuracy of classification model with RandomForest and scaling with StandardScaler

Type of cheese	Precision	Recall	F1-score
Gluten	0.998	0.997	0.997
Gluten-free	0.998	0.998	0.997

The prediction accuracy of the model is 99.8% for gluten-containing cheese and 99.6% for gluten-free cheese.

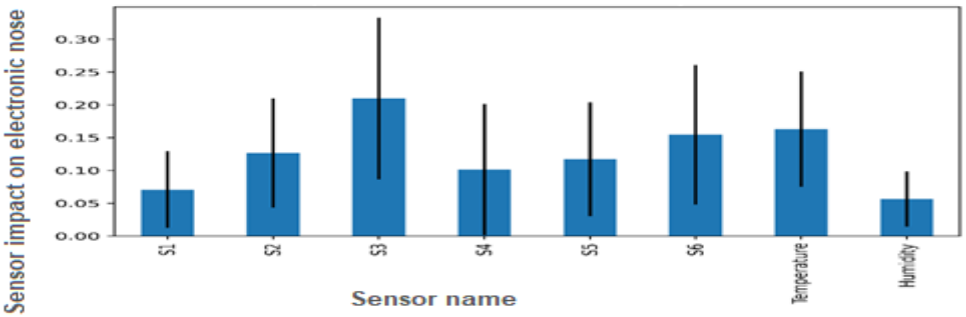

Fig. 11. The effect of each of the electronic nose sensors in the classification model with RandomForest and scaling with StandardScaler

In Model 11, S3 and Temperature sensors have had the most impact, respectively. The black line in each bar shows the influence interval of each sensor on the model, which is finally calculated as the average of this interval and is drawn as a bar graph.

12-<Classifier: RandomForest>

In this model, RandomForest algorithm is used for data classification.

Table 12- Prediction accuracy of classification model with RandomForest

Type of cheese	Precision	Recall	F1-score
Gluten	1.0	0.996	0.998
Gluten-free	0.997	1.0	0.998

The prediction accuracy of the model is 100% for gluten-containing material and 99.7% for gluten-free material.

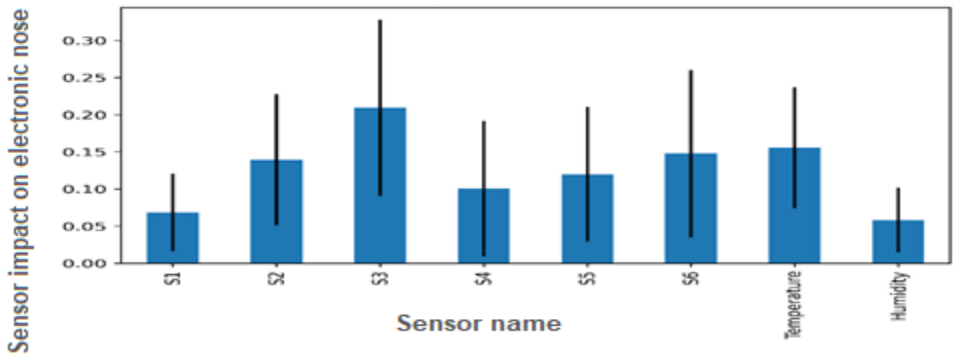

Fig. 12. The effect of each of the electronic nose sensors in the classification model with RandomForest

In Model 12, S3 and Temperature sensors have had the most impact, respectively. The black line in each bar shows the influence interval of each sensor on the model, which is

finally calculated as the average of this interval and is drawn as a bar graph.

13-<Classifier: SVC > <Scaler: MinMaxScaler>

In this model, SVC algorithm is used for data classification and MinMaxScaler is used for scaling.

Table 13- Prediction accuracy of classification model with SVC and scaling with MinMaxScaler

Type of cheese	Precision	Recall	F1-score
Gluten	0.993	0.968	0.98
Gluten-free	0.969	0.993	0.981

The prediction accuracy of the model is 99.3% for gluten-containing cheese and 96.9% for gluten-free cheese.

By observing the set of conditions, one of the models with 99.8% prediction accuracy for two datasets, was selected as the best model. In this model, the Random Forest algorithm was used for data classification, and MinMaxScaler was applied for scaling.

Conclusion

This study explored the application of electronic nose technology combined with advanced data mining and machine learning techniques for the detection of gluten in cheese samples. The primary focus was on distinguishing between gluten-free and gluten-containing cheeses using data collected from electronic nose sensors. The raw data, stored in structured tagged tables, was preprocessed to ensure quality and accuracy. Given the large volume of data and the necessity of repeated experiments for reliability, traditional methods of analysis were deemed inefficient due to their high time and cost requirements, as well as their limited accuracy.

In contrast, this research demonstrated that modern data mining and machine learning techniques provide a more effective solution. These methods not only reduce time and cost but also enhance the accuracy and reliability of gluten detection. By analyzing and modeling the sensor data using these innovative approaches, it was possible to accurately classify the cheese samples into gluten-free and gluten-containing categories.

The findings of this research highlight the potential of leveraging advanced computational techniques to improve food safety and quality assessment processes. Specifically, this study offers a cost-effective and efficient method for identifying gluten in food products, which could significantly benefit individuals with celiac disease or gluten sensitivity. The ability to distinguish gluten content in cheese with high accuracy using electronic nose technology is a step forward in addressing the dietary needs of this population.

In conclusion, the methodology developed in this research can be adapted for broader applications in food safety, potentially contributing to the development of new, efficient devices for gluten detection. These advancements could pave the way for more precise and accessible gluten-testing technologies, thereby improving the quality of life for people who need to adhere to a strict gluten-free diet.

Author Contributions

Mohammad Nasiri-Galeh: was responsible for data curation, formal analysis, investigation, methodology, resources, visualization, and writing – original draft. **Mehdi Ghasemi-Varnamkhasti:** contributed to conceptualization, supervision, and writing – review and editing.

Funding Source

The sampling device used in this study was developed under the Shahid Ahmadi Roshan project with financial support from the Iranian Elite Foundation. The device was designed and constructed by a dedicated team. In this study, samples were collected using this device, and data analysis was performed on the collected data. All costs related to sampling and data analysis were covered by the authors.

Acknowledgments

The authors would like to express their sincere appreciation to Mohammad Hossein Shams and Dariush Valipour, members of the electronic nose development team, for their valuable assistance in the construction of the device.

References

1. Criminisi, J.S. (2011). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision, 7(2-3),* 81-227. https://doi.org/10.1561/0600000035
2. Bhattacharya, N.T. (2008). Preemptive identification of optimum fermentation time for black tea using electronic nose. Sensors and Actuators B: *Chemical, 131*(1), 110-116. https://doi.org/10.1016/j.snb.2007.12.032
3. Biau, G. (2016). A random forest guided tour. *Test 25.2*, 197-227. https://doi.org/10.1007/s11749-016-0481-7
4. Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.
5. Du, W. (2002). Building decision tree classifier on private data.
6. Fernandez, L.Z. (2023). Applications of electronic noses in cheese quality assessment. *Journal of Food Science and Technology, 60*(3), 1234-1245.
7. Freund, Y.S. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119-139. https://doi.org/10.1006/jcss.1997.1504
8. Gh. Shekari, E.M. (2024). Evaluation of the quantitative and qualitative characteristics of gluten-free chicken nuggets containing quinoa flour and hydroxypropyl methyl cellulose (HPMC). (HPMC). *Iranian Food Science & Technology Research Journal/Majallah-i Pizhūhishhā-yi ʿUlūm va Sanāyiʿ-i Ghaẕāyī-i Īrān, 20*(1), 47-62.
9. Hu, W.H. (2008). Adaboost-based algorithm for network intrusion detection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 38*(2), 577-583. https://doi.org/10.1109/TSMCB.2007.914695
10. Karoui, R. (2011). Fluorescence spectroscopy measurement for quality assessment of food systems—a review. *Food and Bioprocess Technology, 4*, 364-386. https://doi.org/10.1007/s11947-010-0370-0
11. Persaud, K., & Dodd, G. (1982). Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. *Nature, 299*, 352-355. https://doi.org/10.1038/299352a0
12. Qi, Y. (2012). Random forest for bioinformatics. *Ensemble machine learning. Springer, Boston, MA*, 307-323. https://doi.org/10.1007/978-1-4419-9326-7_11
13. Quinlan, J.R. (1993). *C4.5, Programs for Machine Learning. Morgan Kaufmann San Mateo Ca.*
14. Ren, S.C. (2015). Global refinement of random forest. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
15. Schapire, R.E. (197-227). The strength of weak learnability. *Machine Learning*, *5*(2), 1990. https://doi.org/10.1007/BF00116037
16. Schapire, R.E. (2013). Explaining adaboost. In: Empirical inference. *Springer, Berlin, Heidelberg*, p. 37-52.
17. Stein, G.C. (2005). Decision tree classifier for network intrusion detection with GA-based feature selection. Proceedings of the 43rd annual Southeast regional conference-Volume 2. https://doi.org/10.1145/1167253.1167288
18. Thompson, T.S. (2023). Advances in gluten detection methods for celiac disease management. *Nutrients*, *15*(2), 789.
19. Vapnik, V. (1998). *Statistical Learning Theory.* John Wiley and Sons.

20. Wang, F.L. (2019). Feature learning viewpoint of AdaBoost and a new algorithm. *IEEE Access, 7*, 149890-149899. https://doi.org/10.1109/ACCESS.2019.2947359
21. Wilson, A.D. (2009). Applications and advances in electronic-nose technologies. *Sensors, 9*(7), 5099-5148. https://doi.org/10.3390/s90705099
22. Wilson, A.D. (2013). Diverse applications of electronic-nose technologies in agriculture and forestry. *Sensors, 13(2)*, 2295-2348. https://doi.org/10.3390/s130202295
23. Yu, H.L. (2024). Rapid detection of gluten contamination in food products using advanced sensor technologies. *Food Chemistry, 420*, 136042.
24. Zhang, Y.C. (2023). Data mining approaches in electronic nose technology for food quality control. *Trends in Food Science & Technology, 135*, 245-258.
25. Zhao, X.L. (2024). Emerging sensor-based technologies for food safety and quality monitoring. *Sensors and Actuators B: Chemical, 389*, 134934.

نشریه پژوهشهای علوم و صنایع غذایی ایران
https://ifstrj.um.ac.ir

مقاله پژوهشی
جلد ۲۱، شماره ۳، مرداد- شهریور ۱٤۰٤، ص. ۲۸٦-۲۷۱

پیاده‌سازی چندین استراتژی داده‌کاوی روی داده‌های بینی الکترونیکی برای شناسایی گلوتن در پنیر

محمد نصیری گله[1*]- مهدی قاسمی ورنامخواستی[2]

تاریخ دریافت: ۱۴۰۳/۰۶/۰۶
تاریخ پذیرش: ۱۴۰۳/۱۰/۰۹

چکیده

بینی الکترونیکی یک دستگاه الکترونیکی برای تشخیص بو است. داده‌های به‌دست‌آمده از این دستگاه به‌صورت عددی و در ستون‌های مختلف ذخیره می‌شوند که مربوط به داده‌های دو نوع پنیر بدون گلوتن و پنیر حاوی گلوتن هستند. این داده‌ها به‌تنهایی برای تصمیم‌گیری و قضاوت کافی نیستند و لازم است روابط و الگوهای میان آن‌ها کشف شود تا مشخص شود داده‌های جدید ثبت‌شده توسط دستگاه به کدام دسته از پنیرهای دارای گلوتن یا بدون گلوتن تعلق دارند. به همین منظور، در این تحقیق از روش‌های داده‌کاوی و یادگیری ماشین استفاده شده است. داده‌کاوی شامل الگوریتم‌های متنوعی مانند طبقه‌بندی، خوشه‌بندی و استخراج قوانین وابستگی است. برای دستیابی به نتایج بهتر، فرآیند داده‌کاوی بر روی ۱۰۵ ترکیب مختلف از مدل‌ها انجام شد و ۱۳ مدلی که بالاترین دقت را در درک روابط میان داده‌ها داشتند، در تحقیق ذکر شده‌اند. در این پژوهش، با استفاده از روش‌های داده‌کاوی، داده‌های مربوط به پنیرهای دارای گلوتن و بدون گلوتن در دسته‌های جداگانه طبقه‌بندی شدند و مدلی جهت پیش‌بینی نوع داده‌های جدید از نظر ماهیت پنیر (دارای گلوتن یا بدون گلوتن) ایجاد شد. پس از تحلیل ۱۰۵ ترکیب مختلف، در نهایت مدلی که از الگوریتم Random Forest برای طبقه‌بندی و از MinMaxScalerبرای مقیاس‌بندی داده‌ها استفاده می‌کرد، به‌عنوان بهترین مدل با دقت پیش‌بینی ۹۹.۸ درصد برای هر دو مجموعه داده‌های آموزش و آزمون انتخاب شد.

واژه‌های کلیدی: بینی الکترونیکی، داده کاوی، درخت تصمیم، طبقه‌بندی داده‌ها، یادگیری ماشین

۱- گروه مدیریت فناوری اطلاعات، دانشکده مدیریت و اقتصاد، دانشگاه تربیت مدرس، تهران، ایران
(*- نویسنده مسئول: Email: m_nasiri@modares.ac.ir)
۲- گروه مهندسی مکانیک بیوسیستم ، دانشکده کشاورزی، دانشگاه شهرکرد، شهرکرد، ایران

https://doi.org/10.22067/ifstrj.2024.89599.1360