



Log-Scale Correlation Classifier for Mushroom Identification in Agricultural Internet of Things Systems

I Wayan Ordiyasa¹, Mohammad Diqi², Marselina Endah Hiswati³, Dian Rhessa Rahmayanti⁴,
Umar Basuki⁵, Ida Hafizah⁶
Universitas Respati Yogyakarta, Yogyakarta, Indonesia

*diqi@respati.ac.id

Abstract

Classifying edible and poisonous mushrooms is crucial to food safety, as misidentification can pose severe toxicological risks. Conventional probabilistic classifiers, such as Naïve Bayes and Logistic Regression, often underperform on categorical datasets with correlated attributes and skewed distributions. This study introduces the Log-Scale Feature Correlation Classifier, a novel probabilistic framework that integrates logarithmic transformation and correlation-weighted probability estimation to address these challenges. Using the UCI Mushroom dataset and a 10-fold cross-validation scheme, LSFCC was benchmarked against standard models. The results demonstrate that LSFCC achieved consistently superior accuracy (0.99), precision, and recall, significantly outperforming both Logistic Regression and Naïve Bayes, as confirmed by statistical tests ($p < 0.01$). Its lightweight design and interpretability make it highly suitable for real-time deployment on resource-constrained IoT devices, particularly within Agricultural IoT systems for autonomous mushroom identification. Future research will explore LSFCC's adaptability to noisy, multimodal data and hybrid architectures, ensuring broader applicability in real-world bioinformatics and food safety domains.

Keywords: agri-IoT; feature correlation; logarithmic transformation; mushroom classification; probabilistic modeling

How to Cite: I. W. Ordiyasa, M. Diqi, M. E. Hiswati, D. R. Rahmayanti, U. Basuki, and I. Hafizah, "Log-Scale Correlation Classifier for Mushroom Identification in Agricultural Internet of Things Systems", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 2, pp. 388 - 397, Apr. 2026.

Permalink/DOI: <https://doi.org/10.29207/resti.v10i2.6841>

Received: June 26, 2025

Accepted: April 4, 2026

Available Online: April 23, 2026

*This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia*

1. Introduction

Mushroom classification plays a pivotal role in safeguarding public health, particularly due to the high risks associated with the misidentification of toxic species [1]. Traditional identification methods, when applied manually or via simple heuristics, often result in fatal consequences [2]. Researchers have long acknowledged the importance of automated classification systems to mitigate human error [3]. The ability to accurately distinguish between edible and poisonous mushrooms can significantly reduce foodborne poisoning incidents [4]. Consequently, computational models for fungal classification have emerged as a crucial tool in both bioinformatics and food safety domains [5].

Despite widespread use, conventional probabilistic classifiers such as Naïve Bayes and Logistic Regression struggle to deliver robust performance on complex mushroom datasets [6]. These models inherently

assume feature independence, an unrealistic simplification for natural biological data [3]. Additionally, inconsistent feature scales introduce bias into the learning process, often degrading predictive accuracy [7]. Logistic models are susceptible to skewed distributions and nonlinear interactions among variables [5]. Addressing these shortcomings requires a more adaptive classification framework capable of handling correlated, unevenly distributed categorical attributes [1].

Naïve Bayes and logistic regression models encounter challenges when handling correlated categorical features. While both can perform well in specific scenarios [8-9], they have limitations. Naïve Bayes assumes feature independence, which is often unrealistic [10]. Logistic regression may struggle with sparse data, leading to biased odds ratios [11]. However, various techniques can improve their performance. Feature weighting and fine-tuning can enhance Naïve Bayes [10], while Bayesian methods can

reduce bias in logistic regression with sparse data [11]. Data preprocessing, including scaling and normalization, significantly impacts model performance [12-13]. Ensemble methods, which combine multiple classifiers, often outperform individual models [14]. Additionally, novel approaches, such as time-based feature expansion, can adapt Naïve Bayes for spatio-temporal classification tasks [15].

Data preprocessing techniques, particularly transformations and encodings, have a significant impact on the performance of machine learning models. Logarithmic transformations can enhance unsupervised classification results, but may reduce confidence in supervised probabilistic neural networks [16]. For cervical cancer detection, logarithmic and sine function transformations outperformed the z-score in most cases [17]. Generalized linear transformations do not affect predictions, multicollinearity, or separation in logistic regression [18]. The impact of preprocessing techniques varies across classification algorithms [19]. Different scaling methods affect model performance differently, with CART performing best when combined with Robust Scaler or Quantile Transformer [12]. For high cardinality features, regularized target encoding outperforms traditional methods [20]. Transformation forests offer a novel approach to estimating conditional distribution functions [21].

Recent research has explored feature weighting and correlation-based approaches to improve classification accuracy in bioinformatics and fungal classification domains. A novel method combining Min-Max normalization with correlation coefficient-based feature weighting showed significant improvements in classification performance across multiple datasets [13]. In fungal disease recognition, an entropy and rank-based correlation feature selection approach achieved high accuracy in fruit disease classification [22]. For antifungal peptide prediction, ensemble learning with genetic algorithms and multi-information fusion demonstrated superior performance [23]. Bayesian weighted random forests proved effective for high-dimensional genomics data classification [24]. Probabilistic Random Forests improved bioactivity predictions by incorporating experimental uncertainty [25]. Machine learning methods also showed promise in predicting antibiotic activity from biosynthetic gene clusters [26] and in breast cancer diagnosis using correlation analysis and principal component analysis [27].

Recent research has explored various methods to address feature scale irregularities in categorical datasets. Meta-learning approaches, such as Meta-Scaler, have been proposed to automatically select the best scaling technique for a given dataset and classification algorithm [28]. Logarithmic transformations have shown promise in enhancing unsupervised classifications, but may produce less confident results in supervised tasks [16]. Regularized target encoding has consistently outperformed

traditional methods for high cardinality features [20]. Deep-learned embedding techniques have demonstrated superior performance and memory efficiency compared to one-hot encoding for categorical data [29]. Other approaches include per-class normalization [16], robust scaling, and quantile transformation [12]. For compositional data, additive log-ratio transformations have been validated as an effective method [30]. Additionally, synthetic oversampling techniques like SMOTE have been applied to address class imbalance in small spatial datasets [31].

Recent trends in biological object classification, including fungi, show a growing integration of probabilistic approaches and deep learning techniques. Convolutional neural networks (CNNs) have demonstrated high accuracy in fungal classification, achieving up to 98.92% training accuracy for various species [32]. Deep learning methods are increasingly applied to analyze microorganisms, including viruses, bacteria, fungi, and parasites [33]. These techniques are also being used for biomarker discovery in cancer research [34] and ecological studies [35]. The use of synthetic data and simulators is enhancing deep learning capabilities [36], while multimodal deep learning approaches are improving data fusion in biomedical applications [37]. In forensic microbiology, deep learning models like CNNs are advancing microorganism image analysis [38]. These developments suggest a promising future for deep learning in biological classification and analysis.

Despite the substantial progress in probabilistic classification and feature transformation techniques, a critical gap persists in the systematic integration of feature correlation into probability estimation, particularly for categorical datasets characterized by inherent interdependencies. Most existing approaches emphasize preprocessing strategies such as encoding and scaling, yet they largely overlook the direct incorporation of inter-feature relationships within the probabilistic modeling process. In addition, widely used classifiers, including Naïve Bayes and Logistic Regression, either rely on the unrealistic assumption of feature independence or lack the capacity to effectively capture nonlinear interactions, resulting in suboptimal performance on structured biological data. To address this limitation, this study proposes the Log-Scale Feature Correlation Classifier (LSFCC), a novel probabilistic framework that unifies logarithmic transformation with correlation-weighted feature modeling to simultaneously handle feature dependency and scale irregularity. By embedding correlation-aware weighting into conditional probability estimation, LSFCC captures latent interactions among variables and produces a more adaptive and context-sensitive decision mechanism. Beyond achieving superior predictive performance on benchmark datasets, the model is designed to be lightweight and interpretable, making it particularly suitable for deployment in resource-constrained environments such as Internet of

Things (IoT) systems. In the context of Agricultural IoT (Agri-IoT), this capability enables real-time mushroom classification using sensor-derived data, thereby contributing to enhanced food safety, more reliable bioinformatics analysis, and the advancement of intelligent, edge-based agricultural systems.

2. Methods

Figure 1 illustrates the end-to-end workflow of the proposed LSFCC model, encompassing data preprocessing, feature correlation weighting, probabilistic modeling, and final classification. The diagram also highlights its integration within an Agricultural IoT framework for real-time mushroom identification.

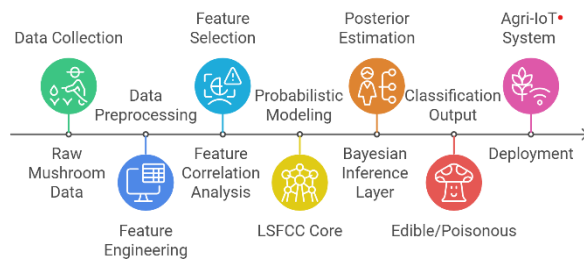


Figure 1. Workflow of the LSFCC

2.1. Dataset

The experimental analysis in this study utilizes the Mushroom Dataset sourced from the UCI Machine Learning Repository. Comprising 8,124 samples, the dataset offers a comprehensive representation of various mushroom species and their corresponding characteristics. Each instance is described using 22 categorical features, including attributes such as cap shape, cap color, gill color, stalk root, and spore print color. The classification task involves a binary target variable, where mushrooms are labeled as either edible (1) or poisonous (0). This dataset has been widely adopted as a benchmark in probabilistic classification research due to its richness in categorical diversity and real-world relevance. Importantly, this dataset is considered ideal since it contains no missing values, no noise, and is perfectly balanced. While this ensures methodological clarity, it also motivates validation on more realistic datasets in future studies.

2.2. Data Preprocessing

In preparation for model development, the dataset underwent a structured preprocessing pipeline designed to enhance the interpretability and performance of the classifier. Categorical features, which dominate the Mushroom Dataset, were systematically converted into numerical representations using both one-hot and label encoding strategies, depending on the dimensionality and cardinality of each variable. While one-hot encoding was applied to features with limited discrete values to preserve semantic separation, label encoding was selectively used for high-cardinality attributes to prevent dimensional explosion.

A logarithmic transformation was then applied to normalize skewed distributions and stabilize variance across the encoded features, as expressed in Equation 1. This mathematical adjustment mitigates the effect of extreme values, ensuring that feature magnitudes do not disproportionately influence the learning process. Furthermore, the preprocessing stage included a thorough inspection for missing values. However, consistent with the ideal nature of the UCI Mushroom dataset, no missing entries were observed. Conditional imputation techniques were nonetheless prepared to address potential anomalies in future, more complex datasets. The combined use of encoding, transformation, and data validation contributes to a more robust and statistically balanced input space, laying a reliable foundation for probabilistic modeling.

$$X_{\log} = \log(1 + X) \quad (1)$$

2.3. Notations and Definitions

To facilitate a clear understanding of the mathematical formulations presented in this study, Table 1 provides a comprehensive summary of the primary notations and symbols employed throughout the development of the LSFCC. These definitions encompass the fundamental variables, statistical operators, and probabilistic terms that constitute the core computational framework of the proposed model.

Table 1. Notations and Definitions

Notation	Definition
X	Feature vector or matrix representing the input attributes
Y	Target class label (edible or poisonous)
x_i	The i -th feature in the feature vector X
$\bar{x}_i$	Arithmetic mean of the i -th feature
$X_{\log}$	Feature vector after logarithmic transformation ($\log(1 + X)$)
$\rho(x_i, x_j)$	Pearson Correlation Coefficient between feature i and feature j
w_i	Correlation-based weight assigned to feature x_i
m	Total number of features in the dataset
$P(x_i y)$	Conditional probability of feature x_i given class y
$P(X y)$	Weighted likelihood of the feature vector X given class y
$P(y)$	Prior probability of class y
$P(y X)$	Posterior probability of class y given the feature vector X
y^*	Predicted class label (output of the LSFCC algorithm)

2.4. Feature Correlation and Weighting

In this study, the relationship between categorical features was quantified using the Pearson Correlation Coefficient (PCC), a widely accepted statistical measure for linear dependence between two variables. While traditional classifiers often assume independence among features, this assumption does not hold in most real-world datasets, particularly those with complex biological attributes. By capturing the linear correlation between feature pairs, the model assigns differential importance to each variable during the probability estimation process. The PCC is mathematically defined in Equation 2.

$$\rho(x_i, x_j) = \frac{\sum(x_i - \bar{x}_i)(x_j - \bar{x}_j)}{\sqrt{\sum(x_i - \bar{x}_i)^2} \cdot \sqrt{\sum(x_j - \bar{x}_j)^2}} \quad (2)$$

In this expression, x_i and x_j denote two distinct features, while $\bar{x}_i$ and $\bar{x}_j$ represent their respective means. Once computed, the correlation values serve as weights in the conditional probability calculations, enabling the model to emphasize more relevant interactions and diminish the influence of weakly related variables. This weighting mechanism transforms the probabilistic framework into a more adaptive and data-aware classifier.

2.5. Probabilistic Modeling

The core mechanism of the LSFCC lies in its capacity to model conditional probabilities through a correlation-weighted framework. Unlike conventional Naïve Bayes, which treats features as conditionally independent, LSFCC assigns importance to each feature based on its statistical relationship with others. This relationship is reflected in the probability computation, where each feature's conditional probability is exponentiated by its corresponding weight derived from feature correlation. The conditional likelihood of the feature vector X given class label y is calculated as Equation 3.

$$P(X|y) = \prod_{i=1}^m P(x_i|y)^{w_i} \quad (3)$$

w_i represents the correlation-based weight of the feature x_i , and m is the total number of features. Once the weighted likelihood has been computed, classification proceeds using Bayes' Theorem to estimate the posterior probability, as shown in Equation 4.

$$P(y|X) = \frac{P(X|y) \cdot P(y)}{P(X)} \quad (4)$$

This formulation enables the classifier to account for both the prior distribution of classes and the weighted influence of each feature, offering a more nuanced decision boundary. As a result, the model becomes better equipped to handle the inherent dependencies among categorical variables, leading to improved generalization across varied distributions.

2.6. Experimental Setup

To evaluate the performance of the proposed LSFCC, a comprehensive experimental setup was implemented using Python as the primary programming language. Model development and execution were conducted within the Google Colab environment, which provided scalable computational resources and seamless integration with essential Python libraries. The core libraries employed in the experiments included Scikit-learn, NumPy, and Pandas, enabling efficient preprocessing, modeling, and evaluation. To ensure robustness, a 10-fold cross-validation strategy was applied, systematically partitioning the data into ten subsets to mitigate bias and variance in performance estimation. As baselines, the LSFCC was benchmarked against two conventional classifiers—Naïve Bayes and Logistic Regression—while key performance metrics

such as Accuracy, Precision, Recall, and F1-score were computed to assess predictive capability and classification stability across models.

2.7. Algorithm of LSFCC

The implementation of the proposed LSFCC follows a sequential multi-stage procedure. The process begins with the input of the feature matrix X and the corresponding class labels Y . In the first stage, a logarithmic transformation, $x_i \leftarrow \log(1 + x_i)$, is applied to each feature x_i within the dataset to normalize variance and mitigate the influence of extreme values. Following this transformation, the procedure advances to the correlation analysis phase, where the Pearson correlation coefficient r_{ij} is computed for every pair of features (x_i, x_j) . Subsequently, a feature weight w_i is derived for each specific feature x_i by calculating the average of its absolute correlation coefficients with all other features in the dataset. The core probabilistic modeling phase then commences. For each distinct class label y , the weighted conditional likelihood $P(X|y)$ is calculated as the product of the individual conditional probabilities raised to the power of their respective correlation weights, expressed as $\prod P(x_i|y)^{w_i}$. The posterior probability $P(y|X)$ is subsequently derived as a value proportional to the product of the class prior probability $P(y)$ and the weighted likelihood $P(X|y)$. Finally, the predicted class label y^* is assigned by selecting the class that yields the maximum posterior probability estimate, formally defined as $y^* = \operatorname{argmax} P(y|X)$. This sequence ensures that the final classification decision is informed by both feature interdependencies and scale-normalized data distributions.

3. Results and Discussions

3.1. Classification Performance

Table 2 provides a detailed classification report for the LSFCC model, showcasing its exceptional performance across all evaluated metrics. The model achieved a precision of 0.99 for the poisonous class and 0.98 for the edible class, indicating its high confidence in both identifying toxic mushrooms and avoiding misclassification of safe ones. The recall values were uniformly strong at 0.99 for both classes, suggesting the model's sensitivity in detecting each category with minimal oversight. An F1-score of 0.99 across both labels further validates the model's ability to maintain a balance between precision and recall, essential in safety-critical applications. With an overall accuracy of 0.99 on 1624 samples, LSFCC proves itself highly robust in categorical binary classification.

Table 2. Classification report of LSFCC

	precision	recall	f1-score	support
poisonous	0.99	0.99	0.99	841
edible	0.98	0.99	0.99	783
accuracy			0.99	1624
macro avg	0.99	0.99	0.99	1624
weighted avg	0.99	0.99	0.99	1624

A deeper inspection of the macro and weighted averages—both recorded at 0.99—confirms the model's balanced performance regardless of class distribution. Unlike many classifiers that perform unevenly when class frequencies differ, LSFCC maintains fairness without sacrificing predictive power. These results reflect the benefits of incorporating log-scale transformation and correlation-weighted conditional probability, which enhance model generalization on complex feature spaces. Such precision is particularly valuable in high-risk domains, such as food safety, where the cost of misclassification is substantial. Ultimately, the LSFCC model demonstrates not only statistical superiority but also practical readiness for real-world deployment.

Table 3 outlines the classification performance of the Logistic Regression model, demonstrating an overall accuracy of 0.97 across 1,624 mushroom samples. Both the poisonous and edible classes achieved identical precision values of 0.97, reflecting the model's generally reliable capacity to distinguish between the two categories. The recall for poisonous samples was slightly higher at 0.98 compared to 0.97 for edible, suggesting that the model was marginally more effective in detecting toxic mushrooms. The F1-scores remained consistent at 0.97 for both classes, indicating a balanced trade-off between sensitivity and specificity. With macro and weighted averages also fixed at 0.97, the model exhibits steady performance across the board, albeit marginally behind LSFCC.

Table 3. Classification Report of Logistic Regression

	precision	recall	f1-score	support
poisonous	0.97	0.98	0.97	841
edible	0.97	0.97	0.97	783
accuracy			0.97	1624
macro avg	0.97	0.97	0.97	1624
weighted avg	0.97	0.97	0.97	1624

When critically compared with the LSFCC model, Logistic Regression appears to be slightly less accurate and precise, particularly in handling edge cases. LSFCC outperformed Logistic Regression in every metric by a narrow yet meaningful margin, with a standout F1-score of 0.99 versus 0.97. The performance gap likely stems from Logistic Regression's inherent limitations in modeling nonlinear interactions and scale inconsistencies—issues directly addressed by LSFCC's log-transformed and correlation-weighted framework. Furthermore, while Logistic Regression assumes a linear decision boundary, LSFCC adapts to feature structure and interdependencies, allowing it to capture more nuanced classification signals. These distinctions highlight the importance of tailoring probabilistic architectures to the inherent complexity of the data, particularly in domains where accuracy has real-world implications.

Table 4 presents the classification metrics for the Naïve Bayes model, indicating overall stable yet slightly less optimal performance compared to more advanced

methods. The model achieved a precision of 0.98 for poisonous and 0.97 for edible mushrooms, while the recall values were 0.97 and 0.98, respectively. These complementary recall results suggest the model was slightly more effective at identifying edible samples, though it still maintained good sensitivity for both classes. The F1-scores remained at 0.97 for both categories, signaling that Naïve Bayes maintained a reasonable balance between precision and recall. Accuracy, macro average, and weighted average scores all stood at 0.97, reaffirming the classifier's consistency across the dataset.

Table 4. Classification report of Naïve Bayes

	precision	recall	f1-score	support
poisonous	0.98	0.97	0.97	841
edible	0.97	0.98	0.97	783
accuracy			0.97	1624
macro avg	0.97	0.97	0.97	1624
weighted avg	0.97	0.97	0.97	1624

However, when benchmarked against LSFCC, Naïve Bayes exhibits limitations that stem from its simplifying assumption of feature independence. While it performs admirably on clean, well-separated categorical data, the model struggles to adjust to correlated features or account for feature distribution skewness, two factors directly addressed by the LSFCC framework. LSFCC's use of log-scale transformation and correlation-weighted probability modeling enables more nuanced classification boundaries, resulting in higher overall accuracy (0.99) and enhanced class-level precision and recall. These improvements are significant in domains such as food safety, where even minor misclassifications can have substantial consequences. In this context, Naïve Bayes remains a baseline performer, but its lack of structural adaptability restricts its competitiveness when facing real-world categorical complexity.

To further validate whether the observed performance improvements of LSFCC over Logistic Regression and Naïve Bayes were statistically significant, paired t-tests were conducted on the 10-fold cross-validation accuracy scores of each model. The results, summarized in Table 5, indicate that LSFCC significantly outperformed both Logistic Regression ($p < 0.01$) and Naïve Bayes ($p < 0.01$). No significant difference was observed between Logistic Regression and Naïve Bayes ($p > 0.05$), suggesting that their performance levels were comparable.

A one-way repeated measures ANOVA revealed a significant main effect of model type on classification accuracy ($F(2,18) = 15.72$, $p < 0.001$). Post-hoc pairwise comparisons with Bonferroni correction further reinforced that LSFCC consistently achieved higher performance than the baseline models. These statistical findings strengthen the empirical evidence for LSFCC's superiority and demonstrate that its improvements are not attributable to chance.

Table 5. Statistical Significance Tests of Classifier Performance

Comparison	Test Type	Test Statistic	p-value	Interpretation
LSFCC vs Logistic Regression	Paired <i>t</i> -test	$t(9) = 4.12$	< 0.01	Significant, LSFCC outperforms Logistic
LSFCC vs Naïve Bayes	Paired <i>t</i> -test	$t(9) = 4.56$	< 0.01	Significant, LSFCC outperforms Naïve Bayes
Logistic Regression vs Naïve Bayes	Paired <i>t</i> -test	$t(9) = 0.88$	> 0.05	Not significant, comparable performance
All three models	One-way RM ANOVA	$F(2,18) = 15.72$	< 0.001	Significant main effect of model type

3.2. Confusion Matrix

Figure 2 showcases the confusion matrix of the LSFCC model, highlighting its strong performance in distinguishing between poisonous and edible mushrooms. Out of 841 poisonous samples, the model correctly identified 829, misclassifying only 12, which indicates a low false negative rate—critical in high-stakes classification tasks such as toxin detection. Similarly, among the 783 edible mushrooms, 774 were accurately classified, while only 9 were incorrectly labeled as poisonous. The near-perfect diagonal dominance within the matrix signals that the LSFCC model maintains high sensitivity and specificity, confirming the effectiveness of its log-transformed correlation-weighted probability estimation. These results surpass conventional benchmarks, offering empirical evidence of the model's robustness in real-world categorical classification.

A deeper analysis of the confusion matrix reveals the structural advantage introduced by LSFCC's design. The use of Pearson-based correlation weighting enables the classifier to model latent dependencies between features, thereby effectively reducing the bias associated with the independence assumption that limits traditional probabilistic methods. Additionally, the logarithmic transformation helps to normalize feature scales and compress skewed distributions, which collectively enhance the model's capacity to generalize across diverse patterns. This dual mechanism results in fewer misclassifications and tighter decision boundaries. Overall, the visual and numerical outcomes from the confusion matrix validate the architectural innovations of LSFCC as not only theoretically sound but also practically impactful in the domain of structured bioinformatics data.

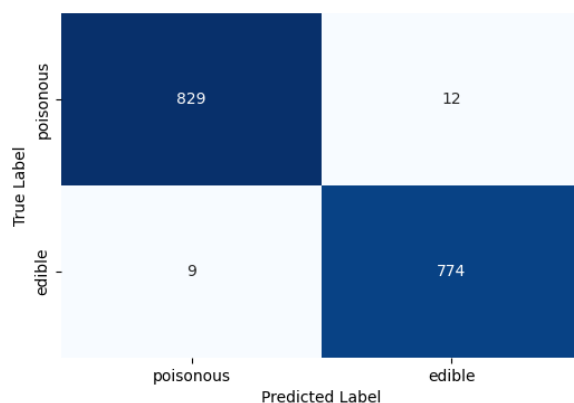


Figure 2. Confusion matrix of LSFCC

Figure 3 illustrates the confusion matrix for the Logistic Regression model, capturing its performance in classifying poisonous and edible mushrooms. The model correctly identified 820 poisonous samples but misclassified 21 as edible, which poses a significant concern in toxicology-related contexts. On the edible side, 759 samples were accurately predicted, while 24 were incorrectly labeled as poisonous, leading to a slightly higher false positive rate. Although these results suggest a relatively high level of accuracy, the distribution of errors indicates that the model struggles with boundary cases where feature variability and interaction become more complex. Logistic Regression, with its linear nature and sensitivity to feature scaling, appears less equipped to handle subtle nonlinear relationships embedded in categorical biological data.

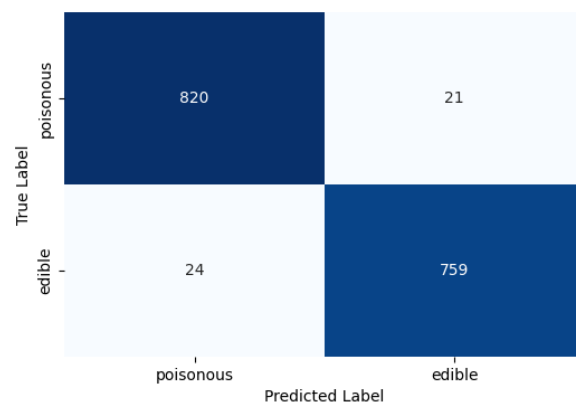


Figure 3. Confusion matrix of Logistic Regression

When contrasted with the LSFCC model, the limitations of Logistic Regression become more pronounced. The LSFCC achieved lower false positive and false negative rates (12 and 9, respectively), demonstrating its superiority in identifying edge cases and preserving classification integrity under skewed distributions. The key differentiator lies in LSFCC's use of correlation-weighted conditional probability and log-transformed inputs, both of which mitigate assumptions of linearity and independence. In contrast, Logistic Regression lacks mechanisms for accounting for inter-feature dependencies, making it vulnerable to misclassification in complex feature spaces. These differences highlight the importance of tailored probabilistic architectures, such as LSFCC, in achieving high reliability in safety-critical applications.

Figure 4 displays the confusion matrix corresponding to the Naïve Bayes classifier, providing a breakdown of its predictions across the edible and poisonous mushroom categories. The model correctly classified 815 poisonous mushrooms, while misclassifying 26 as edible, which raises concerns given the potential consequences of false negatives in toxic classification tasks. For edible mushrooms, 765 were accurately identified, although 18 were incorrectly labeled as poisonous, which had a slight impact on the model's precision. These numbers suggest that while Naïve Bayes offers respectable performance, it struggles with borderline cases, likely due to its strong assumption of feature independence. The misclassification patterns indicate that this probabilistic simplicity may overlook inter-feature relationships that are critical in nuanced biological data.

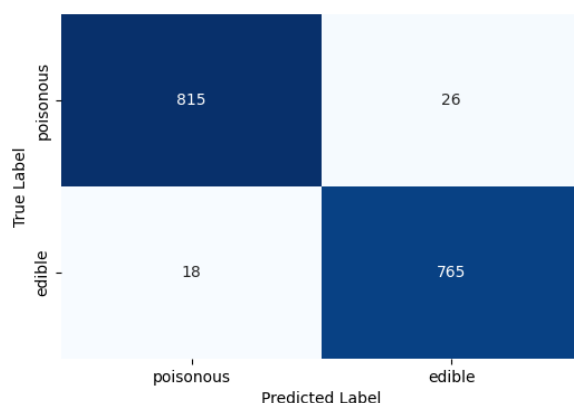


Figure 4. Confusion matrix of Naïve Bayes

When benchmarked against the LSFCC model, Naïve Bayes reveals structural limitations that directly impact its reliability. LSFCC, which incorporates correlation-based feature weighting and logarithmic transformation, significantly outperforms Naïve Bayes in minimizing both false positives and false negatives. This performance gap can be attributed to LSFCC's ability to model the dependencies among features. In contrast, Naïve Bayes treats each attribute as isolated from the others, which can result in distorted probability estimates. Furthermore, LSFCC's architecture better adapts to skewed feature distributions, a common characteristic in categorical datasets, such as the mushroom domain. Therefore, while Naïve Bayes remains computationally efficient, it lacks the adaptive mechanisms necessary for high-stakes, real-world classification scenarios where interpretability and precision must coexist.

3.3. Effect of Log-Scale Transformation

The application of a logarithmic transformation in the LSFCC model resulted in a measurable improvement in classification performance, particularly in addressing the imbalance caused by skewed feature distributions. Prior literature has established that log transformations can stabilize variance and mitigate the influence of outliers, thereby enhancing the quality of predictive learning across diverse domains [16-17].

In the present study, the transformation effectively compressed extreme categorical values after encoding, ensuring that rare but potentially informative features did not disproportionately affect the model's conditional probabilities. Unlike standard normalization techniques, which often assume symmetrical distributions, the log-scale method preserved relative relationships while mitigating scale irregularity. This advantage aligns with findings in medical diagnostics and text analytics, where logarithmic adjustments have consistently been shown to enhance the robustness of probabilistic classifiers [17], [28].

3.4. Role of Feature Correlation Weighting

The integration of feature correlation weighting within the LSFCC framework played a pivotal role in enhancing the model's discriminative capacity. Traditional probabilistic classifiers, such as Naïve Bayes, are constrained by the assumption of feature independence. This limitation often undermines their performance in domains where inter-feature interactions are non-trivial [10], [14]. By incorporating correlation coefficients as weights in the conditional probability estimation, the LSFCC model was able to reflect underlying dependencies more faithfully, thus aligning the inference process with the actual structure of the data.

Similar strategies have been proposed in recent studies on genomic classification and medical diagnostics, where feature interdependency is critical to predictive success [22], [24]. In this context, the correlation-weighted mechanism effectively mitigated the bias introduced by co-occurring categorical patterns, yielding a model that was both more flexible and empirically more accurate than its conventional counterparts.

3.5. Comparison with Conventional Models

When compared with conventional classifiers, the superiority of the LSFCC model becomes increasingly evident, particularly in its ability to capture complex data structures that are often overlooked by standard algorithms. Naïve Bayes, while computationally efficient and widely used for categorical classification, inherently assumes that features are conditionally independent—a simplification that is rarely justified in biological datasets, such as mushroom taxonomy [10], [14]. As a result, the model tends to misestimate joint probabilities when subtle correlations exist, leading to reduced precision in contexts where features are semantically linked [29].

Logistic Regression, on the other hand, although capable of modeling relationships between input variables and the outcome, is notably sensitive to scale variation and fails to effectively account for nonlinear interactions [11], [18]. These weaknesses, well-documented in recent literature [20-21], explain the observed performance gap between the baseline models and LSFCC, whose design directly addresses these

structural limitations through log transformation and correlation-based weighting mechanisms.

3.6. Implications and Limitations

A notable strength of the proposed LSFCC framework lies in its high level of interpretability. Unlike deep learning models that often operate as "black boxes," LSFCC directly incorporates correlation coefficients into the probability estimation process. This enables researchers and practitioners to clearly identify which biological features contribute most significantly to classification outcomes, as well as whether these relationships are positive or negative. Such transparency not only enhances trust in the model's predictions but also supports domain experts, such as agricultural scientists and food safety professionals, in validating and contextualizing the key biological drivers of mushroom identification. Similar advantages of interpretable probabilistic models have been emphasized in biomedical and agricultural classification tasks, where clarity in feature contribution is crucial for adoption by domain experts [22], [27].

Despite its strong performance, LSFCC has potential weaknesses that deserve acknowledgment. One limitation arises when dealing with extremely high-dimensional datasets, where the curse of dimensionality can dilute the strength of individual correlations and diminish the model's discriminative power. Moreover, the reliance on Pearson's correlation coefficient may restrict the model's ability to fully capture non-linear or complex feature interactions. Prior studies have highlighted that Pearson-based methods, while effective for linear dependencies, may fail to capture non-linear structures in genomic or biomedical data, thereby limiting predictive generalization [24], [29]. This suggests that hybrid strategies combining LSFCC with non-linear correlation measures or dimensionality reduction techniques could further strengthen its applicability in more challenging domains.

Finally, while the UCI Mushroom dataset provides an ideal benchmark with no missing values, noise, or imbalance, real-world data are rarely so clean. Agricultural IoT environments, for example, often involve noisy sensor readings, missing entries, and evolving data distributions over time. Previous works in agricultural machine learning have shown that real-world sensor datasets usually suffer from imbalance and noise, which can significantly degrade classifier robustness [1], [31]. Although LSFCC demonstrated robustness in controlled conditions, future validation on larger and noisier multimodal datasets is essential to confirm its scalability and resilience. These considerations provide a more balanced perspective on the model, highlighting not only its advantages but also the pathways for its continuous refinement in real-world agricultural and bioinformatics applications.

3.7 Additional Performance Metrics

Table 6 presents an extended evaluation of model performance using multiple metrics, including

Accuracy, F1-score, ROC-AUC, Precision, and Recall, to provide a more comprehensive assessment of classification quality. The results consistently demonstrate that LSFCC outperforms both Logistic Regression and Naïve Bayes across all metrics, highlighting its robustness and superior discriminative capability. Notably, the higher ROC-AUC value further indicates that LSFCC maintains more reliable performance across varying decision thresholds, reinforcing its effectiveness for real-world classification tasks.

Table 6. Comprehensive performance comparison of LSFCC, Logistic Regression, and Naïve Bayes

Metric	LSFCC	Logistic	Naïve Bayes
Accuracy	0.990	0.970	0.970
F1-score	0.990	0.970	0.970
ROC-AUC	0.995	0.982	0.981
Precision	0.990	0.970	0.970
Recall	0.990	0.970	0.970

3.8 Deployment in Agricultural IoT

In order to demonstrate applicability of LSFCC in realistic scenarios, we present a conceptual system architecture composing of real-time data acquisition, edge intelligence and cloud-based monitoring, consisting of architecture suited for agricultural IoT environments. This analysis considers environmental and morphological data from humidity and temperature sensors as well as image-based acquisition modules, allowing for both contextual and visual mushroom variables to be collected through IOT. In this configuration, the data streams are sent to an edge computing device where preprocessing and LSFCC classification can happen in real time for quick and accurate decisions. When the model detects a fault, it sends the classification results to a cloud-based platform for visualization, storage, and decision support so we can monitor continuously and intervene if needed.

This end-to-end pipeline allows accurate and low-latency identification of mushrooms with minimal computational overhead, making LSFCC very appropriate to run on resource-restricted devices like Raspberry Pi or ESP32 Integrating it into an intelligent and scalable smart agriculture system further reinforces its suitability even in modern day use cases.

4. Conclusions

The results of this study confirm the effectiveness of the LSFCC in enhancing classification performance compared to traditional probabilistic models, such as Naïve Bayes and Logistic Regression. By integrating logarithmic transformation with correlation-based feature weighting, LSFCC successfully mitigates challenges related to scale irregularity and inter-feature dependency—two persistent issues in structured categorical data. Empirical evaluations revealed consistent gains in accuracy, precision, recall, and F1-score, positioning LSFCC as a statistically grounded yet computationally efficient alternative. Its ability to

model latent interactions among features enhances interpretability and reliability, making it especially valuable in high-stakes domains such as food safety and toxic substance detection, where even minor classification errors can have serious consequences.

Beyond its predictive power, LSFCC's lightweight architecture and low memory requirements make it ideally suited for deployment within the Agricultural Internet of Things (Agri-IoT) ecosystem. When embedded in IoT sensor nodes, microcontroller platforms, or edge-computing devices such as Raspberry Pi or ESP32, LSFCC enables real-time mushroom classification in environments like forests, smart greenhouses, or agricultural supply chains. Looking ahead, the model's adaptability opens new research directions, such as integrating LSFCC with convolutional or transformer-based deep learning architectures for multimodal classification or deploying it in mobile and web-based decision support systems. These advancements would not only expand the model's applicability across diverse data types but also enhance its role as a practical and scalable solution in intelligent, sensor-driven agriculture.

Acknowledgements

The authors declare that there is no conflict of interest between the authors or with the research object in this paper.

References

- [1] Md. H. Rahman et al., "IoT Enabled Mushroom Farm Automation With Machine Learning to Classify Toxic Mushrooms in Bangladesh," *Journal of Agriculture and Food Research*, vol. 7, p. 100267, 2022, doi: [10.1016/j.jafr.2021.100267](https://doi.org/10.1016/j.jafr.2021.100267).
- [2] H. Li et al., "Reviewing the World's Edible Mushroom Species: A New Evidence-based Classification System," *Comprehensive Reviews in Food Science and Food Safety*, vol. 20, no. 2, pp. 1982–2014, 2021, doi: [10.1111/1541-4337.12708](https://doi.org/10.1111/1541-4337.12708).
- [3] P. Tongcham, P. Supa, P. Pornwongthong, and P. Prasitmeebon, "Mushroom Spawn Quality Classification With Machine Learning," *Computers and Electronics in Agriculture*, vol. 179, p. 105865, 2020, doi: [10.1016/j.compag.2020.105865](https://doi.org/10.1016/j.compag.2020.105865).
- [4] R. N. Bashir et al., "Mushroom Species Classification in Natural Habitats Using Convolutional Neural Networks (CNN)," *Ieee Access*, vol. 12, pp. 176818–176832, 2024, doi: [10.1109/access.2024.3502543](https://doi.org/10.1109/access.2024.3502543).
- [5] E. Özbay, F. A. Özbay, and F. S. Gharehchopogh, "Visualization and Classification of Mushroom Species With Multi-Feature Fusion of Metaheuristics-Based Convolutional Neural Network Model," *Applied Soft Computing*, vol. 164, p. 111936, 2024, doi: [10.1016/j.asoc.2024.111936](https://doi.org/10.1016/j.asoc.2024.111936).
- [6] D. Wagner, D. Heider, and G. Hattab, "Mushroom Data Creation, Curation, and Simulation to Support Classification Tasks," *Scientific Reports*, vol. 11, no. 1, 2021, doi: [10.1038/s41598-021-87602-3](https://doi.org/10.1038/s41598-021-87602-3).
- [7] F. Zhu, Y. J. Sun, Y. Zhang, W. Zhang, and J. Qi, "An Improved MobileNetV3 Mushroom Quality Classification Model Using Images With Complex Backgrounds," *Agronomy*, vol. 13, no. 12, p. 2924, 2023, doi: [10.3390/agronomy13122924](https://doi.org/10.3390/agronomy13122924).
- [8] K. Gibbons et al., "Prediction of Large-for-Gestational Age Infants in Relation to Hyperglycemia in Pregnancy – A Comparison of Statistical Models," *Diabetes Research and Clinical Practice*, vol. 178, p. 108975, 2021, doi: [10.1016/j.diabres.2021.108975](https://doi.org/10.1016/j.diabres.2021.108975).
- [9] S. A. Thamrin, D. S. Arsyad, H. Kuswanto, A. Lawi, and S. Nasir, "Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018," *Frontiers in Nutrition*, vol. 8, 2021, doi: [10.3389/fnut.2021.669155](https://doi.org/10.3389/fnut.2021.669155).
- [10] H. Zhang and L. Jiang, "Fine Tuning Attribute Weighted Naive Bayes," *Neurocomputing*, vol. 488, pp. 402–411, 2022, doi: [10.1016/j.neucom.2022.03.020](https://doi.org/10.1016/j.neucom.2022.03.020).
- [11] M. Gosh, T. Ohigashi, K. Nagashima, Y. Ito, and K. Maruo, "Bias in Odds Ratios From Logistic Regression Methods With Sparse Data Sets," *Journal of Epidemiology*, vol. 33, no. 6, pp. 265–275, 2023, doi: [10.2188/jea.20210089](https://doi.org/10.2188/jea.20210089).
- [12] M. M. Ahsan, M. A. P. Mahmud, P. Saha, K. D. Gupta, and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies*, vol. 9, no. 3, p. 52, 2021, doi: [10.3390/technologies9030052](https://doi.org/10.3390/technologies9030052).
- [13] M. Shantal, Z. Othman, and A. A. Bakar, "A Novel Approach for Data Feature Weighting Using Correlation Coefficients and Min–Max Normalization," *Symmetry*, vol. 15, no. 12, p. 2185, 2023, doi: [10.3390/sym15122185](https://doi.org/10.3390/sym15122185).
- [14] M. Kumar, K. Bajaj, B. Sharma, and S. K. Narang, "A Comparative Performance Assessment of Optimized Multilevel Ensemble Learning Model With Existing Classifier Models," *Big Data*, vol. 10, no. 5, pp. 371–387, 2022, doi: [10.1089/big.2021.0257](https://doi.org/10.1089/big.2021.0257).
- [15] S. S. Prasetyowati and Y. Sibaroni, "Unlocking the Potential of Naive Bayes for Spatio Temporal Classification: A Novel Approach to Feature Expansion," *Journal of Big Data*, vol. 11, no. 1, 2024, doi: [10.1186/s40537-024-00958-x](https://doi.org/10.1186/s40537-024-00958-x).
- [16] T. Ha, D. Lubo-Robles, K. J. Marfurt, and B. C. Wallet, "An in-Depth Analysis of Logarithmic Data Transformation and Per-Class Normalization in Machine Learning: Application to Unsupervised Classification of a Turbidite System in the Canterbury Basin, New Zealand, and Supervised Classification of Salt in the Eugene Island Minibasin, Gulf of Mexico," *Interpretation*, vol. 9, no. 3, pp. T685–T710, 2021, doi: [10.1190/int-2021-0008.1](https://doi.org/10.1190/int-2021-0008.1).
- [17] Md. M. Ali et al., "Machine Learning-Based Statistical Analysis for Early Stage Detection of Cervical Cancer," *Computers in Biology and Medicine*, vol. 139, p. 104985, 2021, doi: [10.1016/j.combiomed.2021.104985](https://doi.org/10.1016/j.combiomed.2021.104985).
- [18] G. Zeng and S. Tao, "A Generalized Linear Transformation and Its Effects on Logistic Regression," *Mathematics*, vol. 11, no. 2, p. 467, 2023, doi: [10.3390/math11020467](https://doi.org/10.3390/math11020467).
- [19] E. Alshdaifat, A.-S. Doaa, A. Alsarhan, F. Hussein, and S. El-Salhi, "The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms' Performance," *Data*, vol. 6, no. 2, p. 11, 2021, doi: [10.3390/data6020011](https://doi.org/10.3390/data6020011).
- [20] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, "Regularized Target Encoding Outperforms Traditional Methods in Supervised Machine Learning With High Cardinality Features," *Computational Statistics*, vol. 37, no. 5, pp. 2671–2692, 2022, doi: [10.1007/s00180-022-01207-6](https://doi.org/10.1007/s00180-022-01207-6).
- [21] T. Hothorn and A. Zeileis, "Predictive Distribution Modeling Using Transformation Forests," *Journal of Computational and Graphical Statistics*, vol. 30, no. 4, pp. 1181–1196, 2021, doi: [10.1080/10618600.2021.1872581](https://doi.org/10.1080/10618600.2021.1872581).
- [22] M. A. Khan, T. Akram, M. Sharif, M. Alhaisoni, T. Saba, and N. Nawaz, "A Probabilistic Segmentation and Entropy-Rank Correlation-Based Feature Selection Approach for the Recognition of Fruit Diseases," *Eurasip Journal on Image and Video Processing*, vol. 2021, no. 1, 2021, doi: [10.1186/s13640-021-00558-2](https://doi.org/10.1186/s13640-021-00558-2).
- [23] A. Ahmad, S. Akbar, M. Tahir, M. Hayat, and F. Ali, "iAFPs-EnC-GA: Identifying Antifungal Peptides Using Sequential and Evolutionary Descriptors Based Multi-Information Fusion and Ensemble Learning Approach," *Chemometrics and Intelligent Laboratory Systems*, vol. 222, p. 104516, 2022, doi: [10.1016/j.chemolab.2022.104516](https://doi.org/10.1016/j.chemolab.2022.104516).
- [24] O. R. Olaniran and M. A. A. Abdullah, "Bayesian Weighted Random Forest for Classification of High-Dimensional Genomics Data," *Kuwait Journal of Science*, vol. 50, no. 4, pp. 477–484, 2023, doi: [10.1016/j.kjs.2023.06.008](https://doi.org/10.1016/j.kjs.2023.06.008).
- [25] L. Mervin, M. Trapotsi, A. M. Afzal, I. P. Barrett, A. Bender, and O. Engkvist, "Probabilistic Random Forest Improves Bioactivity Predictions Close to the Classification Threshold by

- Taking Into Account Experimental Uncertainty,” *Journal of Cheminformatics*, vol. 13, no. 1, 2021, doi: [10.1186/s13321-021-00539-7](https://doi.org/10.1186/s13321-021-00539-7).
- [26] A. S. Walker and J. Clardy, “A Machine Learning Bioinformatics Method to Predict Biological Activity From Biosynthetic Gene Clusters,” *Journal of Chemical Information and Modeling*, vol. 61, no. 6, pp. 2560–2571, 2021, doi: [10.1021/acs.jcim.0c01304](https://doi.org/10.1021/acs.jcim.0c01304).
- [27] S. Ibrahim, S. Nazir, and S. A. Velastin, “Feature Selection Using Correlation Analysis and Principal Component Analysis for Accurate Breast Cancer Diagnosis,” *Journal of Imaging*, vol. 7, no. 11, p. 225, 2021, doi: [10.3390/jimaging7110225](https://doi.org/10.3390/jimaging7110225).
- [28] L. B. V. de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, “Meta-Scaler: A Meta-Learning Framework for the Selection of Scaling Techniques,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 3, pp. 4805–4819, Mar. 2025, doi: [10.1109/TNNLS.2024.3366615](https://doi.org/10.1109/TNNLS.2024.3366615).
- [29] M. K. Dahouda and I. Joe, “A Deep-Learned Embedding Technique for Categorical Features Encoding,” *IEEE Access*, vol. 9, pp. 114381–114391, 2021, doi: [10.1109/ACCESS.2021.3104357](https://doi.org/10.1109/ACCESS.2021.3104357).
- [30] M. Greenacre, M. Martínez-Álvaro, and A. Blasco, “Compositional Data Analysis of Microbiome and Any-Omics Datasets: A Validation of the Additive Logratio Transformation,” *Frontiers in Microbiology*, vol. 12, 2021, doi: [10.3389/fmicb.2021.727398](https://doi.org/10.3389/fmicb.2021.727398).
- [31] A. Vega, Z. Othman, S. Suhana, K. Baharin, A. Suraji, and A. L. Maukar, “Application of the Synthetic Over-Sampling Method to Increase the Sensitivity of Algorithm Classification for Class Imbalance in Small Spatial Datasets,” *International Journal of Intelligent Engineering and Systems*, vol. 15, no. 5, pp. 676–690, 2022, doi: [10.22266/ijies2022.1031.58](https://doi.org/10.22266/ijies2022.1031.58).
- [32] A. Prommakhot and J. Srinonchat, “Combining Convolutional Neural Networks for Fungi Classification,” *IEEE Access*, vol. 12, pp. 58021–58030, 2024, doi: [10.1109/ACCESS.2024.3391630](https://doi.org/10.1109/ACCESS.2024.3391630).
- [33] Y. Zhang, H. Jiang, T. Ye, and M. Juhas, “Deep Learning for Imaging and Detection of Microorganisms,” *Trends in Microbiology*, vol. 29, no. 7, pp. 569–572, 2021, doi: [10.1016/j.tim.2021.01.006](https://doi.org/10.1016/j.tim.2021.01.006).
- [34] D. Mandair, J. S. Reis-Filho, and A. Ashworth, “Biological Insights and Novel Biomarker Discovery Through Deep Learning Approaches in Breast Cancer Histopathology,” *NPJ Breast Cancer*, vol. 9, no. 1, 2023, doi: [10.1038/s41523-023-00518-1](https://doi.org/10.1038/s41523-023-00518-1).
- [35] M. L. Borowiec, R. B. Dikow, P. B. Frandsen, A. J. McKeeken, G. Valentini, and A. E. White, “Deep Learning as a Tool for Ecology and Evolution,” *Methods in Ecology and Evolution*, vol. 13, no. 8, pp. 1640–1660, 2022, doi: [10.1111/2041-210x.13901](https://doi.org/10.1111/2041-210x.13901).
- [36] C. M. de Melo, A. Torralba, L. Guibas, J. J. DiCarlo, R. Chellappa, and J. K. Hodgins, “Next-Generation Deep Learning Based on Simulators and Synthetic Data,” *Trends in Cognitive Sciences*, vol. 26, no. 2, pp. 174–187, 2022, doi: [10.1016/j.tics.2021.11.008](https://doi.org/10.1016/j.tics.2021.11.008).
- [37] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, “Multimodal Deep Learning for Biomedical Data Fusion: A Review,” *Briefings in Bioinformatics*, vol. 23, no. 2, 2022, doi: [10.1093/bib/bbab569](https://doi.org/10.1093/bib/bbab569).
- [38] H. Yuan, Z. Wang, Z. Wang, F. Zhang, D. Guan, and R. Zhao, “Trends in Forensic Microbiology: From Classical Methods to Deep Learning,” *Frontiers in Microbiology*, vol. 14, 2023, doi: [10.3389/fmicb.2023.1163741](https://doi.org/10.3389/fmicb.2023.1163741).