



Image-based classification of bolus phases in robotic mastication using hybrid machine learning analysis

Isurie Akarawita^a, Bangxiang Chen^b, Jaspreet Singh Dhupia^a, Martin Stommel^c,
Weiliang Xu^a

^a University of Auckland, New Zealand

^b Xi'an Jiaotong-Liverpool University, China

^c Auckland University of Technology, New Zealand

ARTICLE INFO

Keywords:

Chewing phase classification
Food bolus transformation
Image texture analysis
Machine learning
Mouth process model
Robotic chewing

ABSTRACT

Non-invasive, continuous monitoring of food bolus transformation during robotic mastication remains a significant challenge, as existing methods rely on physical extraction of the bolus or single-modality image features that lack robustness across the continuous and non-linear nature of food breakdown. This paper investigates whether hybrid machine-extracted image features can reliably classify chewing phases during in vitro robotic mastication without physical intervention. A hybrid feature-extraction framework is proposed that integrates deep convolutional features from VGG16 with classical texture descriptors, Grey-Level Co-occurrence Matrix (GLCM), image lightness, and variance, applied to a dataset of 1024 segmented chewed biscuit bolus images captured across two chewing trajectories: vertical motion (T1) and combined vertical and lateral motion (T13). Supervised classifiers, Support Vector Machines, Linear Discriminant Analysis (LDA), and Random Forests were evaluated using trial-level grouped cross-validation. Using a 3-phase schema conceptually aligned with the Hutchings and Lillford Mouth Process Model, LDA achieved accuracies of 75.9% for T1 and 82.3% (T13), with the hybrid feature set outperforming the ablation study. Improved separability with T13 demonstrates that shear effect produces more distinctive bolus image features. The unchewed class was consistently well separated, while structured misclassification patterns occur mainly between adjacent phases highlight the challenging nature of the study. These results establish a non-contact, vision-based measurement framework for classifying bolus evolution as a proof-of-concept foundation for robotic chewing systems.

1. Introduction

Food oral processing is a complex process and food bolus transformation during mastication is critical for efficient digestion, safe swallowing, sensory perception and research in food and therapeutic applications. Robotic chewing systems have emerged as valuable tools for studying food oral processing in controlled environments. Multiple robotic platforms, including the TriChewer (Wang et al., 2025), iBOMS-III (Xu et al., 2024), and a 6-axis bionic chewing robot (Zhou and Yu, 2022), among others, have been designed to replicate human mastication processes. This rapid expansion in the development of biomimetic oral cavities has highlighted the need for robust and reliable techniques to evaluate chewed food bolus properties (Mishellany-Dutour et al., 2011; Chen et al., 2023). Traditional studies, such as extracting the food bolus from the robotic oral cavity, face challenges in standardisation, reproducibility, and limitations on the ability to systematically investigate the complex physical changes that occur during chewing (Ren

et al., 2025; Wang et al., 2022). Therefore, non-invasive systems offer a promising alternative by enabling continuous monitoring of mastication, supported by recent advancements in noncontact airflow-induced deformation and structured light imaging for accurate prediction of food texture parameters such as shear force and firmness (Hu et al., 2023).

Image-based analysis has become a standard approach for characterising food texture and bolus properties (Zhou and Yu, 2022). Classical image processing techniques remain valuable, especially for analysing spatial intensity patterns related to surface roughness, fragmentation, and structural heterogeneity. Grey-Level Co-occurrence Matrix (GLCM) based features have been used for image texture characterisation in various domains, including food analysis (Tournier et al., 2012). It has effectively characterised food texture between beef and plant-based burger pattie boluses, showing consistent trends with increasing in vitro chewing cycles (Akarawita et al., 2024a); however, in that study, bolus

* Corresponding author.

E-mail address: iaka416@aucklanduni.ac.nz (I. Akarawita).

<https://doi.org/10.1016/j.jfoodeng.2026.113324>

Received 8 March 2026; Received in revised form 31 August 2026; Accepted 8 September 2026

Available online 17 September 2026

0260-8774/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Nomenclature

AI	Artificial intelligence
GLCM	Grey-level co-occurrence matrix
LAB	CIELAB colour space
LD1, LD2	First and second linear discriminant axes
RF	Random forest
SD	Standard deviation
TPA	Texture profile analysis
T1	Trajectory1 (only vertical motion)
CV	Cross-validation
HSV	Hue, saturation, value colour space
LDA	Linear discriminant analysis
PCA	Principal component analysis
ROI	Region of interest
SVM	Support vector machine
VGG16	Visual Geometry Group 16-layer convolutional neural network
T13	Trajectory13 (Vertical + lateral motion)

images were acquired by extracting the bolus from the oral cavity after discrete chewing intervals rather than through continuous in-cavity monitoring. Lightness, skewness, and intensity variance measurements have also proven valuable for assessing the saliva content of in vitro chewed food boluses (Akarawita et al., 2024b) as they capture spatial relationships between pixel intensities. However, selecting features that remain robust and interpretable becomes challenging for foods undergoing large structural transitions from solid to paste or bolus during mastication, due to the evolving and non-linear nature of food breakdown. Accurate classification of chewing phases requires an understanding of the progressive nature of food breakdown, where early phases involve large-scale fragmentation and later phases focus on particle size reduction and the integration of saliva. The Mouth Process Model proposed by Hutchings and Lillford (Hutchings and Lillford, 1988) describes food oral processing as a continuous breakdown pathway, governed primarily by a reduction in structural integrity and an increase in lubrication, which progresses over time until a swallowable bolus is formed. As this transformation occurs along a continuum rather than in discrete steps, clear boundaries between chewing phases do not naturally exist.

Recent advancements in computer vision have positioned deep learning as a powerful tool for tasks requiring complex texture and structural interpretation. Convolutional Neural Networks (CNNs), with their ability to learn hierarchical representations, have become central to estimating the mastication phases of human chewed food boluses (Nishi et al., 2019). In a recent study, deep learning-based segmentation techniques such as Mask R-CNN were successfully applied to analyse food bolus formation and particle changes during in vitro chewing, demonstrating the feasibility of real-time, non-invasive monitoring of dynamic texture changes (Ren et al., 2025). Beyond food applications, computer vision frameworks integrating GLCM-based features, statistical descriptors, Hough transform features, and feature-reduction techniques have demonstrated robust performance in challenging environmental conditions, with optimised feature subsets achieving high detection accuracy using Support Vector Machines (SVM) (Li et al., 2017). CNN models pre-trained on large datasets (e.g., ImageNet) such as VGG19, ResNet50, and Inception-based architectures have been adapted via transfer learning for food image recognition and classification tasks, showing effective performance in discriminating food categories and features in diverse food image datasets (Bu et al., 2024). VGG-16 has achieved very high accuracy in classifying rice varieties based on colour and texture features, using a customised

model with additional fully connected layers (Ishaq et al., 2025). Furthermore, studies indicate that combining VGG16 features with SVM often outperforms VGG16 with SoftMax classifiers for texture analysis (Priya et al., 2021). Extending in this line of research, Hassan et al. (2025) proposes a novel hybrid deep learning model that integrates automated deep features with statistical textural descriptors (GLCM) for the classification of 3 Neurological disorders from Magnetic Resonance Images. This hybrid approach has captured global patterns and fine-grained textural details, for accurate and reliable image classification. Collectively, these studies indicate that improved classification performance can be achieved by combining CNN-based image segmentation with classical segmentation and feature-extraction methods, enhancing feature robustness in visually complex biological materials (Qadri et al., 2024) and thereby, motivating hybrid deep learning models for real-time analysis of mastication processes.

Grounded in the Hutchings and Lillford Mouth Process Model (Hutchings and Lillford, 1988), which describes oral processing as a progressive and continuous transformation involving structural breakdown and lubrication, this study adopts the model as a conceptual framework and uses chew-count-based experimental stages to represent bolus transformation during robotic mastication. Specifically, the bolus images are classified into four neutral chewing-stage categories: unchewed state, initial, mid, and final chews, corresponding to 0, 1–5, 6–10, and 11–15 chewing cycles, respectively. The novelty of this study lies in the engineering integration of robotic mastication, cycle-based image acquisition, and machine vision-based classification for monitoring bolus transformation. This paper proposes a feature-extraction framework that integrates deep convolutional features from VGG16 with classical descriptors, including GLCM, lightness, and variance, to classify chew-count-based stages of in vitro mastication across two chewing trajectories. The primary classification scheme comprises three chewing phases and additionally a finer four-class partition based on equal-interval chew-count groupings is examined as an exploratory benchmark to probe the limits of image-based phase discrimination at higher resolution.

2. Methodology

2.1. Vision system

The experiments to generate food bolus were performed using a biomimetic masticating robot with a three-degree-of-freedom (DOF) linkage mechanism with an artificial cavity (Chen et al., 2022). The overall experimental setup and the robotic chewing system are illustrated in Fig. 1. The imaging setup consists of a Raspberry Pi HQ camera. It captures images at a resolution of 2.1 megapixels (1920 × 1080 pixels). Given the camera-to-bolus distance of approximately 10 cm and the 6 mm lens configuration, the effective field of view covering the oral cavity region is approximately 30–35 mm in width, yielding a spatial resolution of approximately 0.016–0.018 mm per pixel. This spatial resolution is appropriate for capturing bolus structural changes, which typically occur at millimetre to sub-millimetre scales in in-vitro oral processing studies (Marconati and Ramaioli, 2020). The LED ring light strip operates under constant-current drive, providing stable illumination without flicker or significant thermal drift across the duration of each experimental trial. The camera and lighting assembly are rigidly fixed relative to the oral cavity, and all image frames are captured at the moment of maximum jaw opening, ensuring consistent geometric and illumination conditions across acquisitions (Akarawita et al., 2025). The vision system acquires high-quality images of the bolus as chewing progresses. This configuration supports the capture of fine-grained textural and structural variations associated with food breakdown.

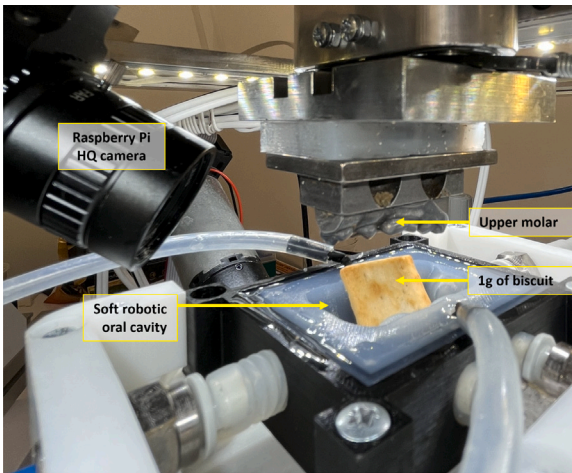


Fig. 1. In-vitro chewing setup.

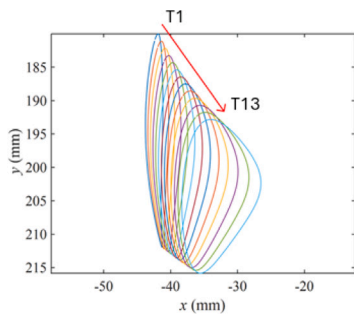


Fig. 2. T1 and T13 trajectories (Chen et al., 2022).

2.2. Dataset description

The dataset consists of segmented images of food boluses generated during controlled robotic chewing experiments. Artificial saliva was replicated using an enzyme-active mouthwash (Oral7, 2025), introduced at a controlled flow rate of 1 ml/min using a syringe pump and a chewing rate of 3s. The use of such saliva substitutes has been explored in previous work to assess their role in altering the mineral content of dental enamel (Kielbassa et al., 2001). The robotic chewing platform enables precise and repeatable control of chewing parameters, ensuring consistency across trials and allowing systematic analysis of how different motion patterns influence particle fragmentation, size reduction, and bolus formation. A standardised food sample of 1 g biscuit (Maliban Biscuit Manufactories (Pvt) Ltd., 2026), as shown in Fig. 1, was selected due to its uniform initial structure and repeatability across trials. Two distinct chewing trajectories were examined to capture variations in chewing kinematics (Chen et al., 2022): Trajectory T1, which involves purely vertical motion and represents simple compression-based chewing, and Trajectory T13, which combines vertical motion with shear forces to represent more complex grinding mastication patterns as shown in Fig. 2. For each trajectory, 512 images were collected across 32 replicate chewing trials, each conducted with a fresh 1 g biscuit sample and a fully reset artificial saliva delivery system. Each trial contributed 16 images corresponding to the unchewed state and 15 successive chewing cycles. Since each image originates from a distinct independent trial, the assumption underlying 5-fold grouped-stratified cross-validation is satisfied. All 512 images were used in model evaluation through 5-fold group-stratified cross-validation, where predictions were generated for every image exactly once when it appeared in the held-out fold, ensuring that performance metrics are computed over the

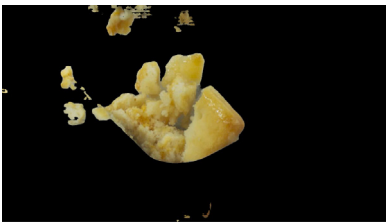


Fig. 3. A segmented image using SAM2 to isolate food bolus.

Table 1
Three-phase chewing interpretation.

Chew count class	Operational description	Nominal cumulative saliva (ml)
Unchewed	High structural integrity, no lubrication	0
Initial/Mid (1 to 10)	Early fracture and progressive fragmentation with increasing lubrication	0.05–0.50
Final (11 to 15)	Bolus formation	0.55–0.75

full dataset. Image pre-processing was performed to standardise input data and prepare images for feature extraction. The pre-processing pipeline included image segmentation, where images were segmented using the Segment Anything Model 2 (SAM2) (Ravi et al., 2024) to isolate the food bolus from background elements as shown in Fig. 3. SAM2 was selected for bolus segmentation due to its state-of-the-art segmentation capability, validated across diverse object categories in the original publication. In this setup, the food bolus presents as a visually distinct object against the blue silicone oral cavity background, representing a well-defined foreground–background contrast that is well-suited to SAM2’s segmentation approach. All segmented images were visually inspected, and any frames exhibiting failed or incomplete segmentations were excluded from the dataset prior to feature extraction. This step ensured feature extraction focused solely on the food material, reducing noise from the robotic system components.

These images were initially grouped into three. The three-class phase is defined as the primary classification: (1) the pre-processing state, where the food retains full structural integrity and no lubrication has occurred; (2) the active breakdown state, where mechanical fragmentation dominates and saliva is being progressively incorporated; and (3) the final ready state, where structural disintegration is complete and lubrication is sufficient for safe swallow. These categories describe controlled chewing progression and do not imply experimentally validated physiological transitions. The three-class phases is summarised in Table 1. To further investigate the discriminability, a finer granularity of four-class partition was defined using equal-interval groupings of five chews each across the 15-chew cycle robotic sequence as summarised in Table 2. At a cycle duration of 3 s and a delivery rate of 1 mL/min, the nominal cumulative volume delivered by cycle n is 0.05n mL. Representative images from each chewing phase of the four chew phases for T1, T13 trajectories are shown in Fig. 4.

2.3. Feature extraction

A hybrid feature extraction approach was employed, combining deep learning-based features and classical image descriptors to capture both high-level structural patterns and low-level texture characteristics of the chewed food bolus. Deep features were extracted using a convolutional neural network to represent complex visual patterns associated with bolus morphology, while classical features quantified statistical texture and intensity variations related to surface roughness and saliva distribution. All features were extracted exclusively from the

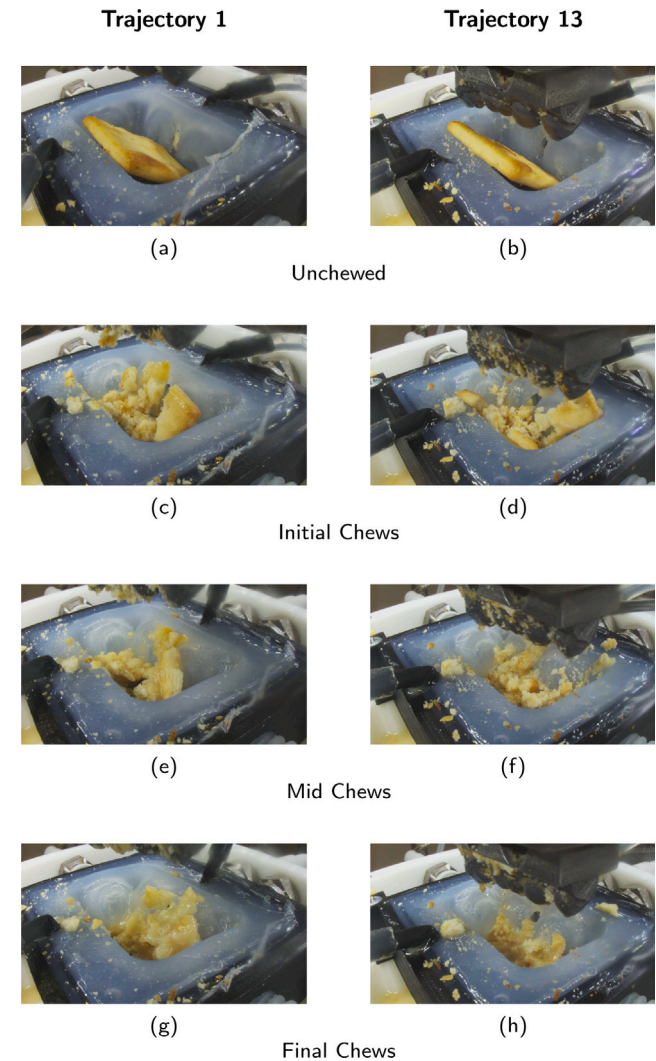


Fig. 4. Progression of food bolus deformation during robotic chewing under Trajectory 1 and Trajectory 13, from unchewed state to swallowable bolus.

Table 2
Four-phase chewing: based on equal chew-count intervals for finer analysis of bolus transformation.

Chew count class	Operational description	Nominal cumulative saliva (ml)
Unchewed	High structural integrity, no lubrication	0
Initial (1 to 5)	Early fracture and particle generation	0.05–0.25
Mid (6 to 10)	Progressive fragmentation with increasing lubrication and particle aggregation	0.30–0.50
Final (11 to 15)	Bolus formation	0.55–0.75

segmented food bolus region to minimise background influence and improve classification accuracy.

2.3.1. Deep learning features

Critically, image-based classification of food bolus transformation during in-vitro robotic mastication represents a novel and emerging research area, with very limited prior literature providing guidance on optimal deep learning backbone selection for this specific task.

However, VGG16 was selected as a representative pretrained CNN and for several domain specific and practical reasons. First, its uniform stack of 3×3 convolutional filters captures fine-grained local texture patterns particularly effectively, unlike architectures such as ResNet or EfficientNet, whose residual connections and compound scaling are optimised for object-level semantic discrimination rather than low-level texture encoding (Rani et al., 2025). Second, the fully connected layers of VGG16 produce a fixed 512-dimensional feature vector that is straightforward to concatenate with low-dimensional classical descriptors (GLCM, lightness, variance), making it particularly compatible with the hybrid fusion pipeline employed here; by contrast, ResNet and EfficientNet require custom global average pooling strategies to yield equivalent fixed-length representations (Addisu et al., 2025). Third, VGG16 has been shown to be a reliable and high-performing feature extractor for small, domain-specific image datasets under transfer learning, where its relatively simple and interpretable architecture reduces the risk of overfitting compared to deeper architectures with more complex connectivity (Bakasa and Viriri, 2023). This is an important consideration given that the present dataset comprises 512 images per trajectory.

VGG16 (Simonyan and Zisserman, 2014), pre-trained on ImageNet, was used to extract deep features from the images. The segmented images were resized to 224×224 pixels using an aspect-ratio-preserving strategy, where each image was first scaled such that its longest dimension was resized to 224 pixels while maintaining the original width-to-height ratio, with zero-padding applied symmetrically to the shorter dimension only. This ensures the bolus occupies the maximum possible proportion of the input canvas, with padding confined to narrow symmetric borders. Since SAM2 produces tightly cropped segmented images with background pixels already masked to zero, the bolus occupies the dominant portion of the resized canvas across all chewing phases, minimising the extent of black background introduced. The extracted VGG16 features were L2-normalised prior to fusion with classical descriptors, which standardises feature vector magnitude across samples.

2.3.2. Classical image features

To complement deep learning features, three categories of classical descriptors were extracted. First, texture properties were captured using GLCM analysis, computing three key metrics (Zubair and Alo, 2024): Energy, which reflects texture uniformity; Homogeneity, indicating the closeness of element distribution to the diagonal; and Dissimilarity, measuring variation between pixel pairs. Second, mean lightness values were derived from the L channel of the LAB colour space, providing insights into overall brightness, which is a factor correlated with moisture content and particle size (Akarawita et al., 2024b). Finally, intensity variance was calculated to quantify texture heterogeneity, revealing the degree of particle size variation and mixing within the bolus. The extracted features were concatenated to form a 517-dimensional feature vector, as shown in Fig. 5. This fusion strategy combines the hierarchical representations from deep learning with interpretable texture descriptors from classical methods.

2.4. Feature normalisation and ablation study

Before classification, features were standardised using a Standard-Scaler, which centres each feature to zero mean and scales it to unit variance across the dataset. This normalisation ensures that features from different sources and scales contribute equally to subsequent analysis. Principal Component Analysis (PCA) was then applied to reduce dimensionality while retaining 95% of the total variance, reducing computational complexity and helping mitigate overfitting, particularly important when dealing with relatively small datasets (Bian et al., 2022). Model evaluation was performed using trial-grouped stratified cross-validation, where all images originating from the same chewing trial were assigned to the same fold. The overall methodology is represented as a flow chart in Fig. 6.

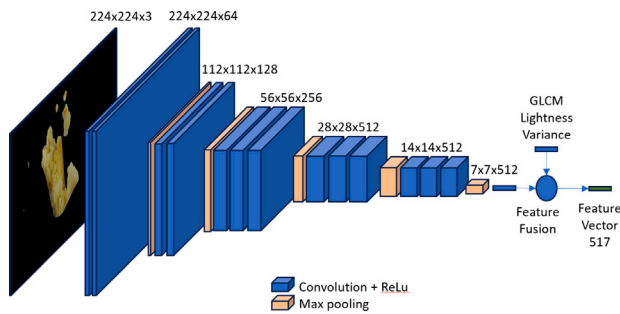


Fig. 5. VGG16-based feature extraction fused with classical features.

To evaluate the contribution of each feature group to classification performance, a formal ablation study was conducted comparing three feature configurations: (1) Classical-only, comprising GLCM energy, homogeneity and dissimilarity, mean lightness, and intensity variance (5-D); (2) VGG16-only, comprising the deep features extracted from the penultimate VGG16 layer (512-D); and (3) Full Hybrid, the concatenation of both feature groups (517-D). Each configuration was processed and evaluated through the same pipeline. This controlled design ensures that observed accuracy differences are attributable solely to the feature content, not to differences in pipeline design or hyperparameter tuning.

2.5. Classification models

Classification of sequential or progressive phases is a common challenge in food science and robotics. Traditional machine learning approaches, including SVM, Linear Discriminant Analysis (LDA), and Random Forest (RF) methods, have been successfully applied to food texture classification tasks (Jiménez-Carvelo et al., 2019). These three classifiers were selected to represent a complementary range of decision boundary strategies: linear (LDA), kernel-based (SVM), and ensemble tree-based (RF), which is a standard practice in machine learning benchmarks for multi-class image classification problems involving high-dimensional features (Syarif and Setiawan, 2020).

2.6. Hyperparameter tuning

To ensure a fair and rigorous comparison across classifiers and kernel functions, hyperparameters for SVM, LDA, and RF were tuned using a systematic nested cross-validation strategy. An inner 5-fold stratified grouped cross-validation loop was applied within the full hybrid feature space (517-D) to search over the predefined hyperparameter grid. The configuration yielding the highest mean inner-validation accuracy was selected and then evaluated on the held-out trial groups in the outer cross-validation loop. For SVM, three kernel functions were included in the search space simultaneously, ensuring that kernel selection itself was data-driven rather than assumed. The parameter grid is summarised in Table 3. For RF, the search covered the number of estimators, tree depth, minimum samples per split and leaf, and the maximum feature subset considered at each split, as summarised in Table 4. The class weight parameter was fixed to balanced for both SVM and RF throughout all experiments to account for class imbalance. For LDA, the solver, tolerance, and shrinkage parameters were included in the hyperparameter search space as in Table 5.

2.7. Evaluation methodology

Model evaluation was performed using a 5-fold stratified grouped cross-validation framework. In this approach, all images from the same chewing trial were kept within the same fold, ensuring that samples from a given trial did not appear simultaneously in the training and

Table 3

Hyperparameter search space for SVM classification.

Hyperparameter	Values tested
Kernel	rbf, polynomial, sigmoid
Regularisation C	0.01, 0.1, 1, 5, 10
Gamma	scale, auto, 0.0001, 0.001, 0.01
Polynomial degree	2, 3
Coefficient	-1.0, 0.0, 0.5, 1.0, 2.0
Class weight	balanced

Table 4

Hyperparameter search space for RF classification.

Hyperparameter	Values tested
Number of trees, n_estimators	100, 200, 300
Maximum tree depth, max_depth	5, 10, 20
Minimum samples for split	2, 5, 10
Minimum samples per leaf	4, 8, 16
Maximum features	sqrt, log2, 0.5
Class weight	balanced

Table 5

Hyperparameter search space for LDA classification.

Hyperparameter	Values tested
Solver	svd, lsqr, eigen
Tolerance, tol	10^{-4} , 10^{-3} , 10^{-2}
Shrinkage	None, auto, 0.01, 0.1, 0.3, 0.5, 0.7, 0.9

validation sets. This provided a stricter assessment of model generalisation to unseen chewing trials while maintaining class balance across folds where possible.

To prevent data leakage, all preprocessing steps, including standardisation and PCA-based dimensionality reduction, were embedded within scikit-learn Pipelines and fitted only on the training data of each fold. Hyperparameter tuning for LDA, SVM, and RF was performed using an inner 5-fold stratified grouped cross-validation loop, forming a nested cross-validation setup. The best hyperparameter configuration was selected based on the mean inner-validation balanced accuracy and then evaluated on the held-out trial groups in the outer fold.

Overall accuracy was presented as mean $\pm$ standard deviation, together with the 95% confidence interval. Confusion matrices and class-wise precision, recall, and F1-scores were computed from the held-out predictions across all folds. For visual interpretation only, LDA was used to project the feature space into two dimensions to examine class separability and misclassification patterns.

3. Results

3.1. 3-Phase classification performance

The 3-phase classification results for trajectories T1 and T13 are summarised in Table 6. Following the trial-grouped cross-validation evaluation, LDA achieved the highest mean accuracy for both trajectories, with 75.9% for T1 and 82.3% for T13. SVM showed comparable performance, particularly for T1, while RF produced the lowest mean accuracy among the three classifiers. Across all models, T13 achieved higher accuracy than T1, suggesting that the additional shear component in the T13 trajectory improved feature separability between chewing phases.

Detailed class-wise performance for the best-performing 3-phase model is presented in Table 7. For T1, the Unchewed class is clearly differentiated (precision, recall, and F1-score of 1.000), while moderate confusion is observed between the Initial/Mid and Final classes, reflected in F1-scores. The macro-average F1-score of 0.8861 indicates balanced performance across classes, whereas the weighted average 0.8373 reflects the influence of class imbalance.

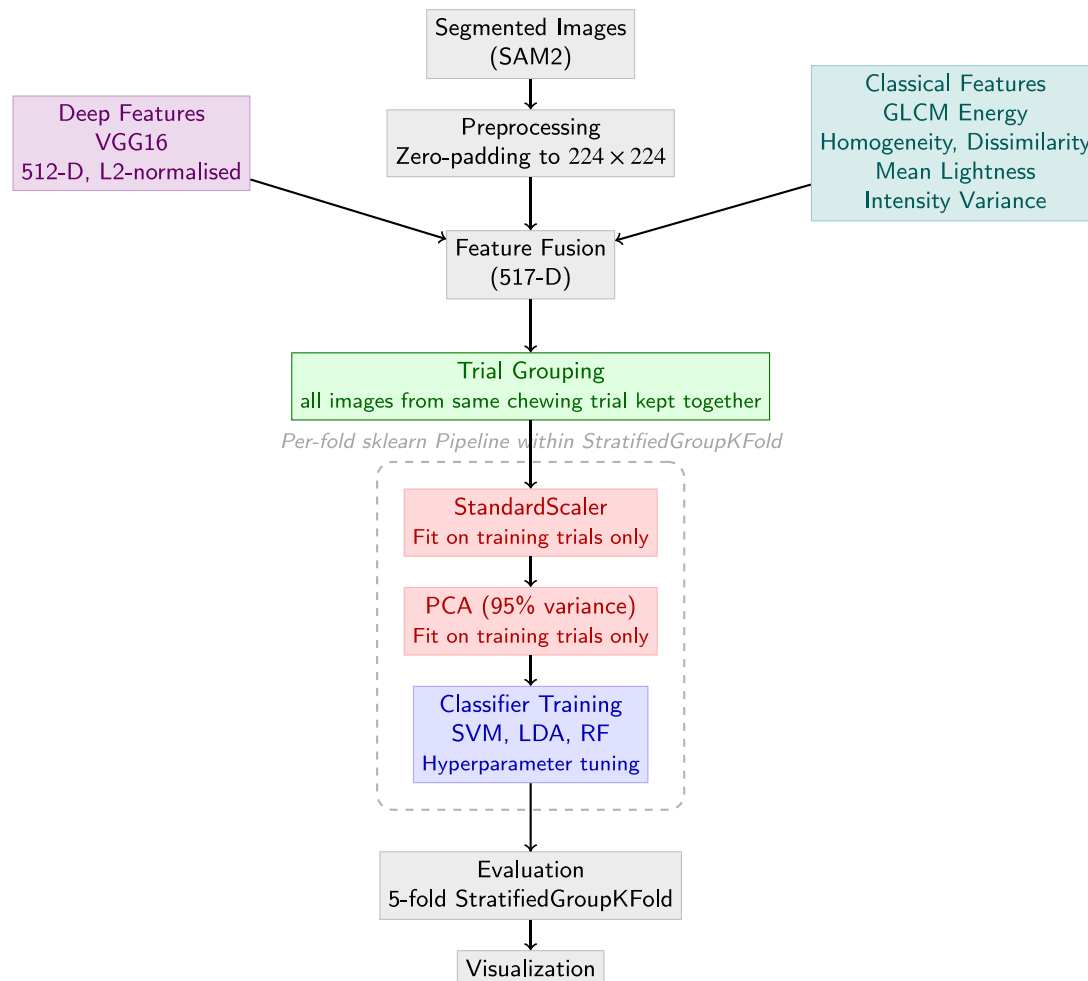


Fig. 6. Methodology flowchart illustrating feature extraction, feature fusion, trial-grouped cross-validation, dimensionality reduction, and classifier evaluation.

Table 6
Model performance for 3-phase chewing classification.

Model	T1		T13	
	Mean±SD	95% CI	Mean±SD	95% CI
LDA	75.9 ± 5.5%	69%–83%	82.3 ± 1.7%	77%–87%
SVM	75.6 ± 2.7%	72%–79%	81.0 ± 2.1%	73%–89%
RF	73.7 ± 5.5%	67%–81%	78.0 ± 5.4%	73%–83%

For T13, improved separability between classes is observed. Although the Unchewed class shows a slight reduction, the Initial/Mid and Final classes achieve comparable higher F1-scores. The macro-average F1-score increased from 0.8063 for T1 to 0.8517 for T13, while the weighted-average F1-score increased from 0.7588 to 0.8236. These results indicate that the T13 trajectory provides more discriminative visual features for distinguishing progressive chewing phases, although some overlap remains between adjacent chewed classes.

3.2. 3-Phase confusion matrix analysis

Table 8 presents the confusion matrices for the 3-phase chewing phase classification for trajectories T1 and T13. For trajectory T1, the Unchewed class was classified without error, with all 32 samples correctly identified. The main source of misclassification occurred between the Initial/Mid and Final classes. Specifically, 59 Initial/Mid samples were misclassified as Final, while 64 Final samples were misclassified as Initial/Mid. This indicates considerable overlap between the

Table 7
Class-wise evaluation metrics for the best LDA 3-phase model.

(a) Trajectory T1				
Class	Precision	Recall	F1-score	Support
T1Unchewed	1.0000	1.0000	1.0000	32
T1Initial/Mid	0.8031	0.8156	0.8093	320
T1Final	0.6194	0.6000	0.6095	160
Macro Avg	0.8075	0.8052	0.8063	512
Weighted Avg	0.7580	0.7598	0.7588	512

(b) Trajectory T13				
Class	Precision	Recall	F1-score	Support
T13Unchewed	0.9688	0.9688	0.9688	32
T13Initial/Mid	0.8706	0.8406	0.8553	320
T13Final	0.7076	0.7562	0.7311	160
Macro Avg	0.8490	0.8552	0.8517	512
Weighted Avg	0.8258	0.8223	0.8236	512

two chewed phases under the T1 trajectory. For trajectory T13, the Unchewed class also showed strong separability, with 31 out of 32 samples correctly classified. Compared with T1, confusion between the Initial/Mid and Final classes was reduced, with 50 Initial/Mid samples misclassified as Final and 39 Final samples misclassified as Initial/Mid. The number of correctly classified Final samples also increased from 96 for T1 to 121 for T13. These results suggest that the inclusion of shear motion in T13 improves discrimination between progressive chewing phases. Overall, most errors occurred between adjacent chewed

Table 8
Confusion matrices for 3-phase chewing classification.

(a) Trajectory T1				
True/Pred	Unchewed	Initial Mid	Final	
Unchewed	32	0	0	
Initial	0	261	59	
Final	0	64	96	

(b) Trajectory T13				
True/Pred	Unchewed	Initial Mid	Final	
Unchewed	31	1	0	
Initial	1	269	50	
Final	0	39	121	

Table 9
Model performance for 4-phase chewing phase classification.

Model	T1		T13	
	Mean±SD	95% CI	Mean±SD	95% CI
LDA	64.7 ± 4.6%	59%–70%	68.2 ± 1.5%	66%–70%
SVM	62.8 ± 4.1%	58%–68%	66.8 ± 3.9%	62%–72%
RF	58.1 ± 5.6%	51%–65%	61.5 ± 3.4%	57%–66%

classes, rather than between non-adjacent, reflecting the continuous and transitional nature of bolus transformation during mastication.

3.3. 4-Phase classification performance

The classification performance for the 4-phase chewing model is presented in Table 9. Compared with the 3-phase formulation, all models showed lower accuracy, reflecting the increased difficulty of separating chewing progression into finer phase categories. For both trajectories, LDA achieved the highest mean accuracy, with 64.7% for T1 and 68.2% for T13. SVM showed comparable performance, while RF produced the lowest accuracy for T1.

Consistent with the 3-phase results, trajectory T13 generally showed better classification performance than T1, suggesting that the inclusion of shear motion improves feature separability. However, the reduced accuracy compared with the 3-phase case indicates greater overlap between adjacent chewing phases when the classification task is defined with finer granularity.

Class-wise evaluation of the best-performing 4-phase model is shown in Table 10. For both trajectories, the Unchewed class remained highly separable, with F1-scores of 1 for T1 and 0.9688 for T13. In contrast, the Initial Chews, Mid Chews, and Final classes showed lower F1-scores, indicating increased confusion among neighbouring chewed phases. The macro-average F1-score increased from 0.7211 for T1 to 0.7421 for T13, confirming a modest improvement for the shear-assisted trajectory. Overall, these results indicate that chewing phase classification becomes more challenging as the number of phase categories increases, mainly due to overlap between adjacent transitional stages.

3.4. 4-Phase confusion matrix analysis

Table 11 presents the confusion matrices for the 4-phase chewing classification for trajectories T1 and T13. For both trajectories, the Unchewed class was classified with minimal error, confirming clear separability from the chewed phases. For trajectory T1, most errors occurred between neighbouring chewed phases. Initial samples were mainly misclassified as Mid, while Mid and Final samples showed mutual confusion. In particular, 47 Final samples were misclassified as Mid, indicating overlap between later chewing stages. For trajectory T13, a similar pattern was observed; however, the Final class showed improved classification, with 111 correctly classified samples compared

Table 10
Class-wise evaluation metrics for the best 4-phase model.

(a) Trajectory T1				
Class	Precision	Recall	F1-score	Support
T1Unchewed	1.0000	1.0000	1.0000	32
T1Initial chews	0.7260	0.6625	0.6928	160
T1Mid chews	0.5160	0.6062	0.5575	160
T1Final	0.6644	0.6062	0.6340	160
Macro Avg	0.7266	0.7188	0.7211	512
Weighted Avg	0.6582	0.6484	0.6513	512

(b) Trajectory T13				
Class	Precision	Recall	F1-score	Support
T13Unchewed	0.9688	0.9688	0.9688	32
T13Initial chews	0.7801	0.6875	0.7309	160
T13Mid chews	0.5272	0.6062	0.5640	160
T13Final	0.7161	0.6937	0.7048	160
Macro Avg	0.7480	0.7391	0.7421	512
Weighted Avg	0.6929	0.6816	0.6854	512

Table 11
Confusion matrices for 4-phase chewing phase classification.

(a) Trajectory T1					
True/Pred	Unchewed	Initial	Mid	Final	
Unchewed	32	0	0	0	
Initial	0	106	44	10	
Mid	0	24	97	39	
Final	0	16	47	97	

(b) Trajectory T13					
True/Pred	Unchewed	Initial	Mid	Final	
Unchewed	31	1	0	0	
Initial	1	110	43	6	
Mid	0	25	97	38	
Final	0	5	44	111	

Table 12
Ablation study: LDA classification accuracy (Mean ± Std, 5-fold CV).

Config	T1 3-Ph	T13 3-Ph	T1 4-Ph	T13 4-Ph
Classical	63 ± 8%	79 ± 3%	56 ± 4%	61 ± 3%
VGG16	75 ± 8%	82 ± 5%	62 ± 7%	65 ± 2%
Hybrid	76 ± 6%	82 ± 5%	65 ± 5%	68 ± 2%

with 97 for T1. Misclassifications remained concentrated between Initial, Mid, and Final phases. Overall, the confusion matrices highlight that errors predominantly occur between neighbouring phases, rather than between non-adjacent classes.

3.5. Ablation study

The ablation study results in Table 12 present the mean classification accuracy and standard deviation for each feature configuration under 5-fold cross-validation. Classical features alone achieved moderate classification performance, classical features contain useful discriminative information. The VGG16-only configuration generally improved performance compared with classical features alone, particularly for the T1 3-phase and T1 4-phase settings. The Hybrid feature configuration achieved the highest or joint-highest accuracy across all configurations. Compared with VGG16-only features, the Hybrid model improved accuracy by 1% for T1 3-phase, matched the T13 3-phase accuracy while reducing variability, and improved the 4-phase classification accuracy by 3% for both T1 and T13. These results suggest that classical image features provide complementary information to deep features, particularly when the classification task becomes more granular.

Table 13

Best LDA hyperparameters identified via grid search.

Setting	Solver	Shrinkage
T1 4-class	lsqr	0.7
T1 3-class	lsqr	0.7
T13 4-class	lsqr	0.1
T13 3-class	lsqr	0.9

3.6. Best hyperparameters

The best LDA hyperparameters identified through grid search are reported in Table 13. Across all trajectory and classification settings, the least-squares solver (lsqr) was selected. The optimal shrinkage value varied between settings and tolerance was not applicable for the selected lsqr solver.

3.7. Dimensionality reduction visualisation

LDA-based two-dimensional visualisations were used to examine class separability patterns for trajectories T1 and T13. Fig. 7 shows the 3-phase classification, while Fig. 8 shows the 4-phase classification. These plots were used for visual interpretation only; model performance was evaluated using the trial-grouped cross-validation results reported in the previous sections. Each point is coloured according to its true chewing-phase label. Coloured decision regions indicate which class the model would predict at each location in the LD1–LD2 plane. For the 3-phase and 4-phase classifications, the Unchewed class formed a clearly separated cluster for both T1 and T13, confirming strong visual distinction between unchewed and chewed bolus states. In 4-phase classifications the Initial/Mid and Final classes showed partial overlap, indicating that the transition between chewed phases is gradual rather than sharply separated. This overlap was lower for T13, suggesting that the addition of shear motion improves feature dispersion between progressive chewing stages. For the 4-phase classification, greater overlap was observed among the Initial, Mid, and Final classes. This agrees with the lower classification accuracy obtained for the finer phase formulation. The Unchewed class remained clearly separated, while most ambiguity occurred among adjacent chewed phases. Overall, the LDA projections support the confusion matrix results, showing that classification errors mainly arise from the continuous and transitional nature of bolus breakdown rather than confusion between non-adjacent phases. This pattern reflects the progression pattern of the Mouth Process Model proposed by Hutchings and Lillford (Hutchings and Lillford, 1988), in which food undergoes continuous structural breakdown and lubrication rather than discrete phase changes. The observed class overlap reflects the gradual and continuous nature of food breakdown, where similar texture patterns make it inherently challenging to achieve clear separation, although it may also be influenced by the way the chew-count-based class labels were defined or the insufficient discriminative information in the dataset. As artificial saliva is introduced at a controlled rate, lubrication evolves systematically alongside mechanical breakdown. This systematic co-evolution of mechanical changes provides a basis for tentatively associating the LDA axes.

4. Discussion

4.1. Phase classification performance

The results show that hybrid image features can discriminate robot-generated bolus states under a trial-grouped cross-validation framework. For the 3-phase classification, LDA achieved the highest mean accuracy for both trajectories, with 75.9% for T1 and 82.3% for T13. The improved performance for T13 in 3-phase, 4-phase, and all classifiers suggests that the additional shear component produces more

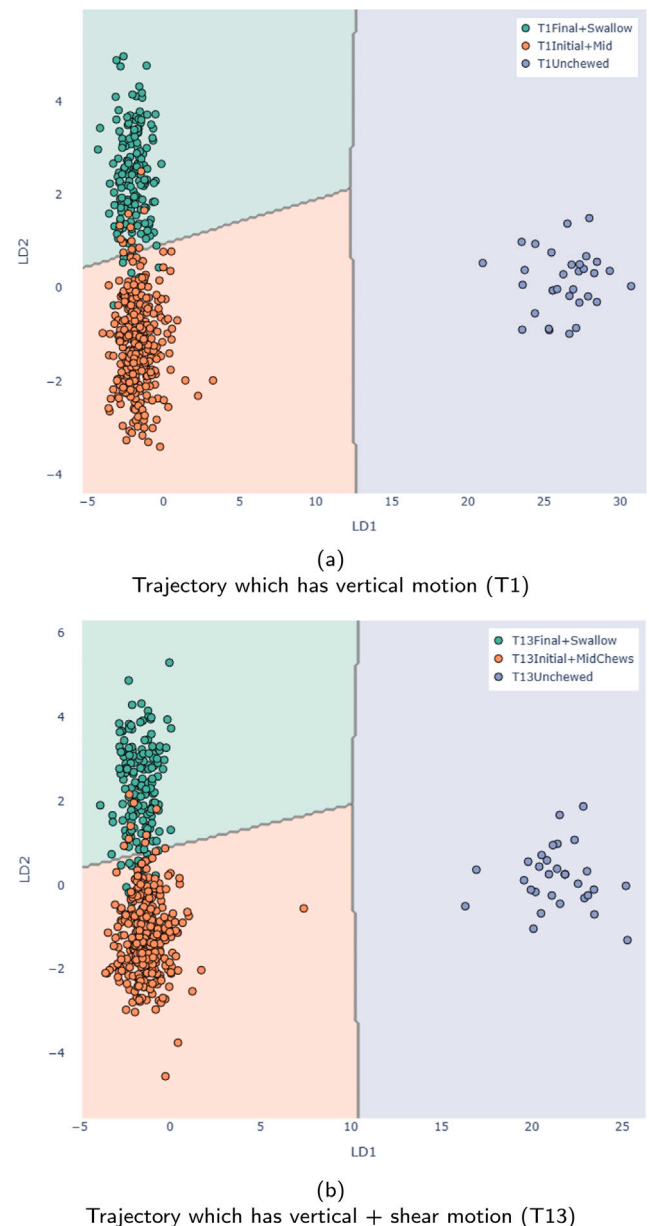


Fig. 7. Plots (a) and (b) represents the LDA projection of segmented bolus images across the three chewing phases.

distinctive visual changes during bolus transformation compared with purely vertical compression.

Across both trajectories, the Unchewed class was clearly separated from the chewed classes, confirming that intact biscuit samples have distinct image characteristics. Most misclassifications occurred between adjacent chewed phases, particularly between Initial/Mid and Final, rather than non adjacent phases. This indicates that the visual transition between later chewing stages is gradual and overlapping, and not sharply separated.

The 4-phase classification showed lower accuracy than the 3-phase formulation, with best accuracies of 64.7% for T1 and 68.2% for T13. This reduction is expected because finer phase divisions increase overlap between neighbouring chewing stages. The Initial, Mid, and Final classes showed greater mutual confusion, while the Unchewed class remained highly separable. Therefore, the 4-phase results suggest that the proposed framework captures meaningful image-based progression

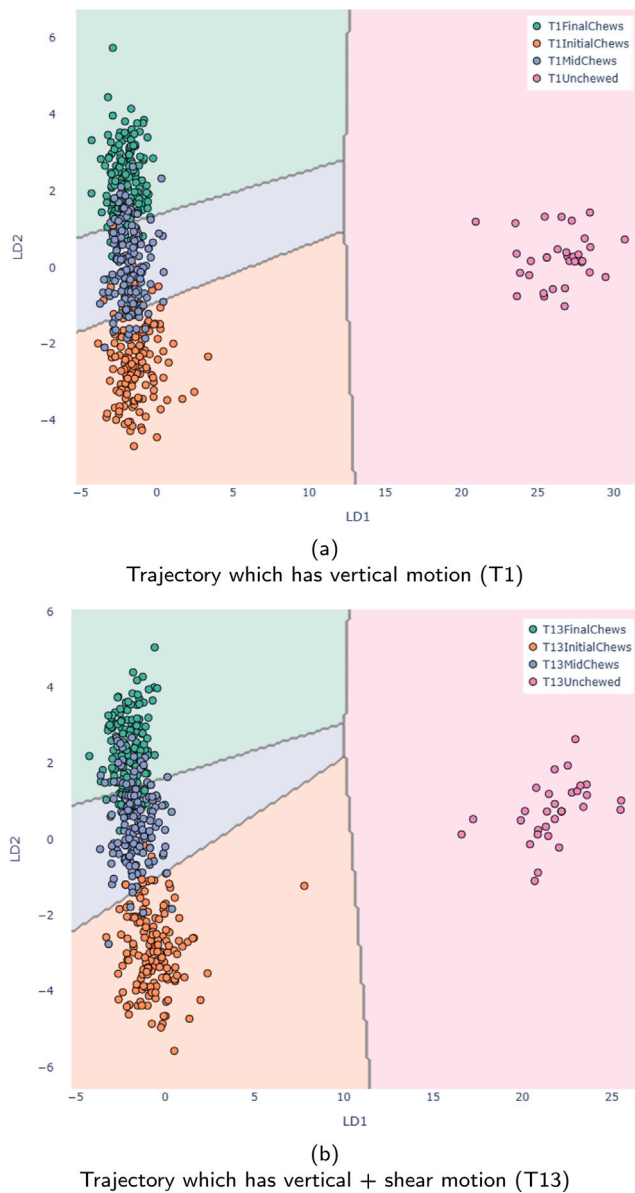


Fig. 8. Plots (a) and (b) represents the LDA projection of segmented bolus images across the five chewing phases. The distribution conceptually reflects the progression consistent with the Mouth Process Model proposed by Hutchings and Lillford, where oral processing involve gradual structural breakdown and increasing lubrication towards swallowability (Hutchings and Lillford, 1988).

during robotic mastication, but also highlight the difficulty of assigning discrete labels to a continuous bolus transformation process.

The LDA visualisations support these findings by showing clear separation between unchewed and chewed samples, together with overlap among adjacent chewed phases. However, the LDA axes are interpreted only as mathematical projections of the image-derived feature space and not as direct physiological variables. Future work should validate these image-based phase distinctions against independent measurements of bolus mechanical properties, lubrication, or swallowability.

4.2. Effect of chewing trajectory

Across all classifiers and both phase-count configurations, trajectory T13 consistently outperforms T1, with improved mean accuracy and reduced standard deviation. This pattern strongly suggests that

the inclusion of shear motion in T13 generates more discriminative textural changes in the bolus compared to purely vertical compression in T1. Shear forces introduce additional mechanical actions with lateral spreading, grinding, and particle redistribution, that produce more distinct morphological and surface texture transformations at each chewing stage. The LDA decision boundaries overlaid in Figs. 7 and 8 provide a visual illustration of the classifier's separation strategy in the projected two-dimensional space. It should be noted that these overlaid boundaries serve as an illustrative purpose and not as a quantitative evidence beyond what is already reported in the accuracy metrics. Across both 3-phase and 5-phase classifications, T13 exhibits broader class dispersion and more distinct cluster formation, particularly for the Final grouping. The emergence of a well-separated final cluster under T13 suggests that shear-enhanced mastication produces a bolus state with sufficiently distinctive textural characteristics to be reliably identified by the classifier.

4.3. Contribution of hybrid feature fusion

The ablation study shown in Table 12 provides direct and systematic evidence for the complementary roles of deep and classical features. Classical features alone achieve substantial standalone accuracy, confirming that GLCM texture metrics, mean lightness, and intensity variance encode physiologically meaningful information about bolus state. This is consistent with prior work Akarawita et al. (2024a,b) which established these descriptors as effective indicators of bolus texture and moisture content in vitro chewing experiments of a different food type. VGG16 features carry the dominant discriminative signal, reflecting the network's capacity to encode complex global structural patterns associated with bolus fragmentation, deformation, and surface morphology under transfer learning. Critically, the full hybrid configuration consistently and systematically outperforms VGG16 or classical features alone, across both trajectories and both phase-count settings in ways that improve classification reliability, particularly in the more challenging 4-phase setting.

4.4. Classifier behaviour

The trial-grouped nested cross-validation results showed that LDA achieved the best overall performance across both trajectories and both phase formulations. This suggests that, after standardisation and PCA-based dimensionality reduction, the main separation between chewing phases can be captured effectively using linear discriminant boundaries. The selected LDA configurations consistently used the lsqr solver with shrinkage, indicating that regularisation was beneficial for stabilising classification in the relatively high-dimensional feature space. SVM produced comparable results but did not exceed LDA, while RF generally showed lower accuracy, suggesting that tree-based ensemble modelling was less effective for the available dataset and feature representation. Overall, the results indicate that a regularised linear classifier provides an interpretable approach for chewing phase classification under the grouped cross-validation framework.

5. Conclusion

This study developed a non-contact, hybrid vision-based framework for classifying chew count-based bolus states during bioinspired robotic mastication. The framework used a dataset of 1024 images and trial-level grouped nested cross-validation to compare LDA, SVM, and RF classifiers under the same hybrid feature space. Using the 3-phase schema conceptually grounded in the Hutchings and Lillford Mouth Process Model, the best-performing LDA achieved accuracies of 75.9% for T1 and 82.3% for T13, and 4-phase accuracies of 64.7% for T1 and 68.2% for T13. Across both phase formulations and all three classifiers, the T13 generally improved classification performance compared with the T1, suggesting that combined vertical-lateral motion

produced more discriminative image features during bolus transformation. The Unchewed class was consistently well separated from chewed samples, while most errors occurred between adjacent chewed phases rather than non adjacent phases. This indicate that finer chew-count-based phase separation remains challenging. The ablation results further showed that combining VGG16-derived features with classical descriptors provided the joint-best performance across the tested settings. This suggests that the two feature groups provide complementary information for image-based bolus monitoring. Overall, the proposed framework provides a reproducible framework for non-contact monitoring of bolus transformation under controlled robotic mastication. Future work should extend the framework to larger experimental datasets, additional food types, real-time implementation, alternative deep learning backbones, and independent rheological measurements to strengthen the link between image-derived predictions and physical bolus properties.

CRedit authorship contribution statement

Isurie Akarawita: Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Bangxiang Chen:** Writing – review & editing, Data curation, Conceptualization. **Jaspreet Singh Dhupia:** Writing – review & editing, Supervision. **Martin Stommel:** Writing – review & editing, Supervision. **Weiliang Xu:** Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

All authors confirm that there are no conflicts of interest associated with this publication.

Acknowledgments

We are pleased to acknowledge the financial support of Riddet Institute, A New Zealand Centre of Research Excellence (CoRE), for providing the PhD scholarship to the first author.

Data availability

Data will be made available on request.

References

- Addisu, E.G., Yirga, T.G., Yirga, H.G., Yehuala, A.D., 2025. Transfer learning-based hybrid VGG16-machine learning approach for heart disease detection with explainable artificial intelligence. *Front. Artif. Intell.* 8, 1504281.
- Akarawita, I., Chen, B., Dhupia, J.S., Stommel, M., Xu, W., 2024a. A novel approach to evaluate robotic in vitro chewing effect on food bolus formation using the glm image analysis technique. *IEEE Open J. Ind. Electron. Soc.* 5, 682–691.
- Akarawita, I., Chen, B., Dhupia, J.S., Stommel, M., Xu, W., 2025. ChewNet: A multimodal dataset for in vivo and in vitro beef and plant-based burger patty boluses with images, texture, and force profiles. *Data Brief* 111890.
- Akarawita, I., Chen, B., Dhupia, J., Xu, W., Stommel, M., 2024b. Estimation of the moisture content of in vitro chewed food boluses by image processing. In: 2024 30th International Conference on Mechatronics and Machine Vision in Practice. M2VIP, IEEE, pp. 1–4.
- Bakasa, W., Viriri, S., 2023. Vgg16 feature extractor with extreme gradient boost classifier for pancreas cancer prediction. *J. Imaging* 9 (7), 138.
- Bian, J., Zhao, D., Nie, F., Wang, R., Li, X., 2022. Robust and sparse principal component analysis with adaptive loss minimization for feature selection. *IEEE Trans. Neural Networks Learn. Syst.* 35 (3), 3601–3614.
- Bu, L., Hu, C., Zhang, X., 2024. Recognition of food images based on transfer learning and ensemble learning. *PLoS One* 19 (1), e0296789.
- Chen, B., Dhupia, J.S., Morgenstern, M.P., Bronlund, J.E., Xu, W., 2022. Development of a biomimetic masticating robot for food texture analysis. *J. Mech. Robot.* 14 (2), 021012.
- Chen, B., Dhupia, J.S., Morgenstern, M.P., Zhang, F., Xu, W., 2023. Opportunities and challenges in biomimetic robotic simulation for in vitro testing of chewable drugs. In: ASME International Mechanical Engineering Congress and Exposition, vol. 87639, American Society of Mechanical Engineers, V006T07A032.
- Hassan, N., Miah, A.S.M., Okuyama, Y., Shin, J., 2025. Neurological disorder recognition via comprehensive feature fusion by integrating deep learning and texture analysis. *IEEE Open J. Comput. Soc.*
- Hu, Q., Lu, W., Luo, H., Dai, D., Deng, Y., Song, A., 2023. Noncontact detection method for food texture assessment based on air puff combined with structured light imaging. *IEEE Trans. Instrum. Meas.* 72, 1–13.
- Hutchings, J.B., Lillford, P., 1988. The perception of food texture—the philosophy of the breakdown path. *J. Texture Stud.* 19 (2), 103–115.
- Ishaq, H.M., Siddiqui, M.N., Rafiq, M.U., Raza, M.H., Niazi, A.K., Yaseen, W., Qadri, S., Rehman, M.A.U., 2025. Applying VGG-16 for classification of rice varieties incorporating color and texture features. *Lahore Garrison Univ. Res. J. Comput. Sci. Inf. Technol.* 9 (1), 11–19.
- Jiménez-Carvelo, A.M., González-Casado, A., Bagur-González, M.G., Cuadros-Rodríguez, L., 2019. Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity—A review. *Food Res. Int.* 122, 25–39.
- Kielbassa, A.M., Shohadai, S.P., Schulte-Mönting, J., 2001. Effect of saliva substitutes on mineral content of demineralized and sound dental enamel. *Support. Care Cancer* 9 (1), 40–47.
- Li, B., Thomas, G., Williams, D., 2017. Detection of ice on power cables based on image texture features. *IEEE Trans. Instrum. Meas.* 67 (3), 497–504.
- Maliban Biscuit Manufacturers (Pvt) Ltd., 2026. Cheesebits. <https://www.malibangroup.com/cheesebits/>. (Accessed 28 April 2026).
- Marconati, M., Ramaioli, M., 2020. The role of extensional rheology in the oral phase of swallowing: an in vitro study. *Food Funct.* 11 (5), 4363–4375.
- Mishellany-Dutour, A., Peyron, M.-A., Croze, J., François, O., Hartmann, C., Alric, M., Woda, A., 2011. Comparison of food boluses prepared in vivo and by the AM2 mastication simulator. *Food Qual. Pref.* 22 (4), 326–331.
- Nishi, K., Shibata, A., Nagahama, Y., Kimura, K., Inoue, M., Higashimori, M., 2019. Mastication class estimation for food bolus by using convolutional neural network. In: 2019 IEEE International Conference on Systems, Man and Cybernetics. SMC, IEEE, pp. 3377–3382.
- Oral7, 2025. Oral7® moisturising mouthwash. Product information and description available. Online URL <https://www.oral7.com/product/oral7-moisturising-mouthwash-copy/>.
- Priya, K., Maheswari, P.U., et al., 2021. Deep learnt features and machine learning classifier for texture classification. In: *Journal of Physics: Conference Series*, vol. 2070, IOP Publishing, 012108.
- Qadri, S.A.A., Huang, N.-F., Wani, T.M., Bhat, S.A., 2024. Advances and challenges in computer vision for image-based plant disease detection: a comprehensive survey of machine and deep learning approaches. *IEEE Trans. Autom. Sci. Eng.* 22, 2639–2670.
- Rani, R., Bharany, S., Elkamchouchi, D.H., Ur Rehman, A., Singh, R., Hussien, S., 2025. VGG-EffAttnNet: Hybrid deep learning model for automated chili plant disease classification using VGG16 and EfficientNetB0 with attention mechanism. *Food Sci. Nutr.* 13 (7), e70653.
- Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al., 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.
- Ren, S., Chen, B., Dhupia, J.S., Stommel, M., Xu, W., 2025. Analyzing food bolus formation during chewing through image segmentation. *Eng. Appl. Artif. Intell.* 160, 111884.
- Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Syarief, M., Setiawan, W., 2020. Convolutional neural network for maize leaf disease image classification. *Telkomnika (Telecommunication Comput. Electron. Control)* 18 (3), 1376–1381.
- Tournier, C., Grass, M., Zope, D., Salles, C., Bertrand, D., 2012. Characterization of bread breakdown during mastication by image texture analysis. *J. Food Eng.* 113 (4), 615–622.
- Wang, X., Chen, B., Dhupia, J., Bronlund, J., Han, J., Xu, W., 2025. Conceptualization and mechanical design of a cam-follower based chewing robot with three oral chambers. *IEEE Access*.
- Wang, X., Xu, W., Dhupia, J., Bronlund, J., 2022. New concept and challenges for a next-generation chewing robot. In: 2022 28th International Conference on Mechatronics and Machine Vision in Practice. M2VIP, IEEE, pp. 1–4.
- Xu, Y., Lv, B., Wu, P., Chen, X.D., 2024. Creating similar food boluses as that in vivo using a novel in vitro bio-inspired oral mastication simulator (iBOMS-III): The cases with cooked rice and roasted peanuts. *Food Res. Int.* 190, 114630.
- Zhou, X., Yu, J., 2022. Development and validation of a chewing robot for mimicking human food oral processing and producing food bolus. *J. Texture Stud.* 53 (4), 419–429.
- Zubair, A.R., Alo, O.A., 2024. Grey level co-occurrence matrix (GLCM) based second order statistics for image texture analysis. *arXiv preprint arXiv:2403.04038*.