Anticipating Food Structure of Meat Products from Mastication Physics applying Machine Learning

Dominic Oppen[1], Tabea Attig[2], Jochen Weiss[1*], Christian Krupitzer[2†]

[1]Department of Food Material Science, Institute of Food Science and Biotechnology, University of Hohenheim, Garbenstraße 25, 70599 Stuttgart, Germany

[2]Department of Food Informatics, Institute of Food Science and Biotechnology, and Computational Science Hub, University of Hohenheim, Fruwirthstraße 21, 70599 Stuttgart, Germany

[*] Jochen Weiss, mail: j.weiss@uni-hohenheim.de, phone: +49 711 459 24415
[†] Christian Krupitzer, mail: christian.krupitzer@uni.hohenheim.de, phone: +49 711 23664

Abstract

Alternatives to animal-based products are becoming more relevant. Most of those products rely at some stage on a structuring process; hence researchers are developing techniques to measure the goodness of the structured material. Conventionally, a typical sensory study or texture analysis by measuring deformation forces would be applied to test the produced material for its texture. However, meat alternatives and meat differ in more points than just the texture, making it hard to extract the isolated texture impression. To objectively obtain qualitative and quantitative differences between different food structures, evaluation of oral processing features is an upcoming technology which qualifies as promising addon to existing technologies. The kinematic data of the jaw and exerted forces regarding muscle activities are recorded during mastication. Resulting datasets are high in dimensionality, covering thousands of individual chews described by often more than ten features. Evaluating such a dataset could benefit from applying computational evaluation strategies designed for large datasets, such as machine learning and neural networks. The aim of this work was to assess the performance of machine learning algorithms such as *Support Vector Machines* and *Artificial Neural Networks* or ensemble learning algorithms like *Extra Trees Classifier* or *Extreme Gradient Boosting*. We evaluated different pre-processing techniques and various machine algorithms for learning models with regard to their performance measured with established benchmark values (Accuracy, Area under Receiver-Operating Curve score, F1 score, precision-recall Curve, *Matthews Correlation Coefficient* (MCC)). Results show remarkable performance of classification of each single chew between isotropic and anisotropic material (MCC up to 0.966). According to the feature importance, the lateral jaw movement was the most important feature for classification; however, all features were necessary for an optimal learning process.

Keywords: Machine Learning, Food Structure, Oral Processing, Mastication Physics, Data Science

Abbreviations:

AdaBoost	-	Adaptive Boosting Classifier
ANN	-	Artificial Neural Network
AUC	-	Area Under Curve
AUROC	-	Area Under Receiver-Operating Curve
DT	-	Decision Tree
GNB	-	Gaussian naïve Bayes Classifier
GradBoost	-	Gradient Boosting Classifier
kNN	-	k-Nearest Neighbor
LR	-	Logistic Regression
MCC	-	Matthews Correlation Coefficient
PCA	-	Principal Component Analysis
RF	-	Random Forest
SMOTE	-	Synthetic Minority Oversampling Technique
SVM	-	Support Vector Machine
XGBoost	-	Extreme Gradient Boosting Classifier

Introduction

Food structure is critical in moving food production toward a CO_2-neutral range. As meat production presents one of the components of food production that emits the most greenhouse gases, companies, researchers, and consumers have a common goal: to structure alternative proteins from renewable and climate-neutral sources, with the aim to produce products with the well-accepted and sought-after sensory properties of meat (Dekkers, Boom, & van der Goot, 2018; Grossmann & Weiss, 2021; Ilic, Van Den Berg, & Oosterlinck, 2021).

Besides aromatic, taste, and visual sensory aspects, understanding the structure-texture dependency of meat is part of this. It is known that muscle has an anisotropic structure that extends over broad length scales, from the nanoscale to the macroscale, with many hierarchical levels (Biga et al., 2019; Listrat et al., 2016). In contrast to sensory aspects, which can be easier copied with aroma substitutes, colorants, and spices, understanding how this structure determines how a food product is perceived is unclear. Hence, it cannot be easily copied yet, and further analysis is required (Oppen, Grossmann, & Weiss, 2022).

Therefore, *food science* researchers use medical science knowledge and try to find connections between food structure and mastication physics. This interdisciplinary field is called food oral processing (J. Chen, 2009; Devezeaux De Lavergne, Young, Engmann, & Hartmann, 2021; Foegeding, Vinyard, Essick, Guest, & Campbell, 2015). In this field, several approaches exist to analyze mastication characteristics from the first bite to swallowing and beyond. For example, bolus particle sizes of meat and meat analogues have been collected (Ilić, Djekic, Tomasevic, Oosterlinck, & Berg, 2022), or jaw movements and muscle activities during chewing on different gel-like structures have been recorded (Koç et al., 2014). The common approach to evaluate the dynamic data of jaw movement and muscle activities is to calculate mean values over the whole sequence of mastication or for certain stages of the sequence (Braxton, Dauchel, & Brown, 1996; Brown et al., 1996; Çakir et al., 2012; Kohyama & Mioche, 2004; Le Révérend, Saucy, Moser, &

Loret, 2016). Conducting an ANOVA on the calculated mean values would in theory enable to group the variables into homogenous subgroups. Yet, even large mechanical and rheological differences did not consistently lead to significant differences in the evaluated oral processing features (Melito, Daubert, & Foegeding, 2013). For example, Melito et al. (2013) investigated three types of cheese (Mozzarella, Cheddar and American) on their oral processing behavior calculating differences between the mean values, but none of the features could identify significant differences between all three cheese products.

Recently, work has been conducted to find out how anisotropic structures such as grown meat differ from isotropic gel-like structures by recording jaw movements and muscle activity during chewing: Oppen, Young, Piepho, and Weiss (2023) were able to show relationships between particle size and anisotropy over the course of mastication by modeling features calculated for every signle chew over time of mastication using a linear mixed model. This is of immense value in understanding specific differences in any of the mastication features (such as velocities, amplitude, or muscle activity). It provides the possibility to model the dynamic change of specific features (e.g. the muscle activity per chew) over the course of mastication. A discrimination from e.g. sample A to sample B is not possible following this evaluation strategy.

From an industry's point of view, the assessment described by Oppen et al. (2023) is too specific. It is not of immense value to enable food manufacturers to track the dynamic change of jaw muscle activity while eating procucts. Being able to assign the product to a certain group, based on all features combined, could however be of great benefit. This would for example enable a pass or fail qualification for food quality, or open the possibility to assign meat alternative products to the categories "consumed like meat". Machine learning algorithms have the potential to find particular patterns in datasets that are not linearly interdependent, making it a more flexible tool for big datasets (Khan, Sablani, Nayak, & Gu, 2022; Krupitzer & Stein, 2021). Opposed to the before presented oral processing studies, which make use of linearly modified food model products, real-world scenarios are more complex and the to be evaluated features are not necessarily linearly

dependent. The use of machine learning has already been discussed in a closely related work by Kircali Ata et al. (2023), in which the mechanical properties of meat analogues were correlated with their composition (ash, carbohydrates, fat, protein) by various algorithms. However there is no prior work known to us, evaluating dynamic oral processing data utilizing a comparable evaluation strategy as presented in the current work.

This work aims to assess common machine learning algorithms' suitability to classify a food oral processing dataset of samples with different structures and particle sizes. We applied diverse pre-processing methods for optimizing the dataset and compared a set of established machine learning algorithms concerning their performance using established metrics from the machine learning domain. The remainder of this paper is structured as follows. First, the approach is explained. Second, the pipeline structure is presented in the implementation section. Third, the classification performance of the two different approaches is evaluated in the results and discussion section. Lastly, practical remarks, an outlook, and threats to validity are discussed.

Approach

The approach of this work is to construct a machine learning pipeline based on mastication data for identifying food samples, which differ in particle size and fibrousness. The implementation is conducted in Python (Pilgrim & Willison, 2009).

Dataset

The used dataset was generated in previous work by Oppen et al. (2023), where it is precisely described. Briefly summarized, the researchers attempted to explain the changes of each presented oral processing feature in dependence on the particle size, anisotropy, and progress of mastication by a linear mixed model. The dynamics of the masticatory apparatus of 11 different panelists (4 females, 7 males, 34.5 ± 11.7 years) were measured to generate the mastication data. Isotropic and anisotropic samples of 4 different particle sizes were given to the panelists, and their jaw movement and muscle activities were recorded during mastication. Each panelist consumed

every sample four times, where the first repetition was always discarded in order to exclude previously described exploratory effects coming from the sensory discovery as described by Le Révérend et al. (2016). The four repetitions of one kind were consumed by the panelists subsequently, whereas the different samples were presented in a randomized order.

Anisotropy and isotropy of the samples was simulated by processing of the meat. Anisotropic samples were comprised of whole heat processed muscle tissue whereas isotropic samples were prior to heat processing heavily processed by cutting, resulting in destruction of the muscle fiber cells, therefore loss of fibrous structure. Chemical composition of all samples was the same. The particle sizes were adjusted after cooking by cutting previously diced samples in a bowl chopper for different amount of time. The reached particle sizes d_{90} were for isotropic and anisotropic samples 6.08, 14.01 and 19.49 mm, and 7.87, 16.30 and 23.37 mm, respectively. The uncomminuted sample was determined on a d_{90} of 29.2 mm (Oppen et al., 2023).

Features were calculated for each single chew of the mastication sequence. The measurement was started after placing the sample in the oral cavity and stopped, when the food bolus was swallowed (Oppen et al., 2023). More specifically, 7980 rows corresponding to single chews, and 23 columns describing the features are included in the dataset. This dataset contains features shown in Table **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.1, which only have integer or float values. The interested reader is referred to Oppen et al. (2023) for details regarding the dataset's creation.

Table Fehler! Kein Text mit angegebener Formatvorlage im Dokument**..1: Dataset Features**

Variable *(abbreviation)*	**Explanation**	**Value/Unit**
"P"	Unique ID of each panelist	Integer between 1 and 11
"S"	ID for sample of each structure	Integer between 1 and 4
"D"	ID for structure, isotropic or anisotropic	Logical
"specific_sample"	Specific ID for each unique food sample	Integer between 0 and 7
"R"	Repeated measurement	Integer between 1 and 3

"C"	Contiuous chews in one sequence	Integer between 1 and n, where n = max number of chews
"Rel_chew"	Relative progress of mastication, 0 as start and 1 as end of sequence	Float numbers between 0 and 1
"part_size"	Optically determined Ferret diameter d_{90} of samples	[6.1, 7.8, 14, 16.3, 19.5, 23.3, 29.2]
"Time" $Time_{chew}$	Time from maximal z- to maximal z-value	(s)
"lateral_movement" Lat_{mean}	Mean of absolute y-values over one cycle	(mm)
"max_lateral_movement" Lat_{max}	Max of absolute y-values over one cycle (mm)	(mm)
"lat_side" Lat_{amp}	Mean of y-values (positive = right, negative = left) over one cycle	(mm)
"Vertical_amplitude" $Vert_{amp}$	Difference of max. z-value before downstroke to min. z-value after downstroke	(mm)
"Downward_velocity" V_{down}	Maximal value of the numerical derivative between two opening positions	(mm/s)
"Upward_velocity" V_{up}	Minimal value of the numerical derivative between two opening positions	(mm/s)
"Occlusal_duration" $Time_{occ}$	Duration between two opening positions with a velocity of less than 15 mm/s	(s)
"EMG_Integral" EMG_{AUC}	Numerical integration of rectified, filtered EMG signal of each cycle	(µV*s)
"EMG_max" EMG_{max}	Maxima of rectified, filtered EMG signal of each cycle	(µV)
"EMG_maxint" $EMG^{max}/_{AUC}$	Quotient of **EMG max** and **EMG AUC** of each cycle	-
"EMG_powerstroke" $EMG_{powerstroke}$	Numerical integration of rectified, filtered EMG signal (µV*s) during one cycle where the derivative of the movement was lower than 15 mm/s	(µV*s)
"EMG_occ_quot" $EMG^{powerstroke}/_{AUC}$	Quotient of **EMG power stroke** and **EMG AUC** of each cycle	-

Exploratory Data Analysis

We conducted a conventional exploratory data analysis before applying machine learning algorithms to the dataset, including calculating the feature's correlation, outlier detection, and checking the dataset for missing values. The dataset was further analyzed for class imbalance and Gaussian distribution of the values. Learning on imbalanced data would, without precaution,

result in better prediction values for the larger class. Gaussian distribution of data is often assumed by basic learning algortihms such as Logistic Regression or the Gaussian Naive Bayes algorithm. Based on the findings, random oversampling of data by synthetic minority oversampling technique (see Section Kapitel 1889733088) was applied, eliminating the class imbalance. Further, ensemble learning algorithms which are generally more robust against imbalance and non-gaussian distributed data were applied. The correlation analysis of the features proposed that all features hold essential information and are not redundant, wherefore none of the features were excluded at this stage.

Machine Learning Approach

This work applies two groups of learning algorithms: (i) Classical machine learning algorithms and (ii) ensemble learning classifiers, listed in Table **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.2. Ensemble learning classifiers combine different algorithms, which all learn a model. A meta-learner aggregated the different models' outcomes into one single output value.

We applied a machine learning pipeline, i.e., a well-defined order of actions defined by an algorithm-like structure, to structure the coding operations. In principle, the pipeline might be re-used for other datasets with the same variables or, slightly adjusted, with datasets following another structure. We describe the details of the pipeline's implementation in Section "Pipeline Structure".

Generally, the dataset provides three different variables that can be used as targets for the analyses: particle sizes (4 manifestations), food structures (2 manifestations), and the different samples (8 manifestations embodied by the 4 particle sizes times two different food structures). The goal of our analysis is a classification analysis, i.e., the machine learner analyzes the different variables and builds a model that explains the relations between the manifestations of the variables, called features, and the manifestations of the target variable, the so-called class. Hence,

the machine learning model can be used to predict the value for the target class based on the observed data pattern for the other variables. This work only focused on two target variables: each individual sample and the food structure. Accordingly, we repeated our machine learning pipeline - composed of data examination, pre-processing, classification, and performance evaluation - for each target (shown in Figure **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.1).

Table Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.2: Applied machine learning algorithms with representative source**

Classical Algorithms	Abbreviation	Source
Support Vector Machines	SVM	(Mammone, Turchi, & Cristianini, 2009)
Decision Tree	DT	(de Ville, 2013)
Gaussian Naive Bayes	GNB	(Bayes & Price, 1763; Breese, Heckerman, & Kadie, 2013)
k-nearest Neighbors	kNN	(Altman, 1992)
Logistic Regression	LR	(Cox, 1959)
Artificial Neural Network Classifier	ANN	(Murtagh, 1991)
Ensemble learning Algortihms		
Random Forest	RF	(Breiman, 2001)
Extra Trees Classifier	ExtraTrees	(Geurts, Ernst, & Wehenkel, 2006)
Extreme Gradient Boosting Classifier	XGBoost	(T. Chen & Guestrin, 2016)
Adaptive Boosting Classifier	AdaBoost	(Freund, Schapire, & Abe, 1999; Freund & Schapire, 1997)
Gradient Boosting Classifier	GradBoost	(Friedman, 2002)

Figure Fehler! Kein Text mit angegebener Formatvorlage im Dokument**..1: Flowchart describing the in this work conducted data processing and evaluation steps.**

Implementation

This section first provides an overview of the pipeline structure (shown in Figure **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.1). Afterward, it describes the implementation of the different activities of the machine learning pipeline. In consecutive order, we present the data pre-processing, model learning, and model evaluation.

Pipeline Structure

The machine learning pipeline is divided into six different pipelines for each algorithm. An example of the pipeline structure is described in the section below: The first pipeline does only implement the basic algorithm with default parameters. The second pipeline standardizes the data before training the algorithm. Through standardization, better comparability across the data of different subjects is possible. Following the same argumentation, the third pipeline normalizes the values before training the algorithm instead of standardizing it. In pipeline four, data is first standardized, as this method showed improvements in the performance. Afterward, the principal component analysis (PCA) is applied to reduce dimensionality. Pipeline five uses the ANOVA F-test with standardized data for feature selection before training the algorithms (data not shown). Features are the variables of the data that are included in the machine learning model. Determining the inclusion and exclusion of data variables is part of the learning process. It depends on whether the machine learning algorithm identifies that the values of a variable can contribute to the prediction of the corresponding class. The last pipeline applies standardization and further Synthetic Minority Oversampling Technique (SMOTE), which is a common data augmentation method. As in general, the results of machine learning are improved with an increase in the amount of data, often data augmentation methods are applied to increase the amount of data. We assume that the analyzed dataset might benefit from those techniques as it is relatively small and, hence, tested them.

These different pipelines aim to compare the usefulness of the applied pre-processing techniques and find the best-fitting pipeline for each classification approach. Overall, 11 different machine learning algorithms were trained (Table **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.2). As the pipelines integrate the data preparation techniques mentioned before training, each pipeline must be applied with each machine learning algorithm. Therefore, the generated pipeline structure has a total number of 6x11=66 pipelines.

Data Pre-Processing

In the following this section presents the applied pre-processing techniques. We used dimensionality reduction by a PCA and an ANOVA F-test for feature selection in addition to those presented below. Further, the outlier detection method *iForest* was applied (Cheng, Zou, & Dong, 2019). However, those additional pre-processing techniques did not improve the algorithms and are therefore not further discussed.

Data Transformation

We used the *MinMaxScaler* and *StandardScaler* functions of the commonly applied scikit-learn Python package for machine learning as scaling techniques. Two pipelines of the basic algorithm and each scaling technique were implemented to visualize which scaling method performs the best based on the individual algorithms. The *StandardScaler* showed significant performance improvements in preliminary examinations and was later used as a pre-processing technique for data augmentation.

Data Augmentation

We applied the commonly used oversampling technique SMOTE. This technique has different subtypes, which can be applied by importing the “imblearn.over_sampling” package from scikit-learn. To simplify the code and to be able to compare the results, the common subtype "SMOTE" is used.

Model Learning

Two methods, the *train-test split,* and *k-fold cross-validation,* were used to split the dataset. Both methods have the advantage of preventing over-fitting by separating the training and testing data. Since the dataset includes a low amount of data compared to other datasets used in machine learning, an 80/20 train-test split was applied. This means that 80 % of the data points are used for training the machine learning model, and the remaining 20 % are used for analyzing the model's performance. To enable reproducibility, we configured the split in a way that allows to have the same data split when restarting the code.

To prevent over-fitting and to increase robustness, *k-fold cross-validation* has been performed as a second method for splitting the data into training and testing data. In this case, 10 was chosen as k, i.e., 10 different models were built, and each model was trained on nine folds and tested on the remaining tenth fold. Performances of the different models are measured and averaged on different evaluation metrics automatically by the used procedure. *k-fold cross-validation* has the benefit of using the whole dataset for training and testing. Further, as the final model is an aggregation from best-fitting models of the *k* generated sub-models, the final model is more robust against variations in the data. The results presented in this work were all conducted applying the *k-fold cross-validation* since it showed improved performance compared to the train-test split.

Model Evaluation

This section focuses on the process of model learning and evaluating its performance. Therefore, it presents the selected performance evaluation metrics and determination of the feature importance, i.e., the contribution to the model's explainability of each individual feature in the training process. The testing subset is used to validate the trained algorithms' ability to classify the output feature correctly (fiber structure or each sample). The performance of each algorithm was calculated by using different evaluation metrics: Accuracy, Area under Receiver-Operating Curve (AUROC) score, F1 score, precision and recall, *Matthews Correlation Coefficient* (MCC), ROC

curve and precision-recall curve (Baldi, Brunak, Chauvin, Andersen, & Nielsen, 2000; Fawcett, 2006; Hripcsak & Rothschild, 2005; Keilwagen, Grosse, & Grau, 2014). Different metrics have to be chosen based on the type of machine learning task classification, regression, or clustering for balanced or imbalanced data. The MCC was used as the primary decision criterion in the present work since it is a robust metric. The MCC considers all four confusion matrix categories (true positive, false positive, true negative, and false negative) and only results in high scores if the algorithm shows high scores for all four (Chicco & Jurman, 2020). For the best performing algorithm of the structure classification, the confusion matrix is shown additionally to the MCC. To get more information about the relevance of each feature, when training a specific model, the feature importance was evaluated where applicable – for *Support Vector Machine (SVM), k-nearest-neighbor (kNN), Logistic Regression (LR),* and *Artificial Neural Network (ANN),* the feature importance module is not available.

Results and Discussion

This section describes the results of our analysis. First, we present the results of the exploratory data analysis. Second, we show the analysis and evaluation results of the applied machine learning algorithms.

Exploratory Data Analysis

According to the correlation heatmap shown in Figure **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.2, features with no correlation are colored black, positive linear correlation is colored in orange, and features with negative linear correlation are colored in blue. The analysis aims to identify strong linearly correlated features, as multicollinearity can impact the performance of machine learning, causing unstable regression cefficients. Since no features show direct linear correlation (1), none were removed from the database. Severe multicollinearity is assumed above correlation of 0.7, which was evident for $\mathbf{Time_{Occ}}$ with both, **Time** and $\mathbf{EMG^{powerstroke}/_{AUC}}$. The features $\mathbf{EMG_{AUC}}$, $\mathbf{EMG_{powerstroke}}$, and $\mathbf{EMG^{powerstroke}/_{AUC}}$ are all based on

the muscle activity during one chew and thus show moderate correlations. It can also be noticed that **Lat_{mean}** and the fiber structure ID “D” show a moderate negative correlation. It was however decided that the correlation is not as strong to impact the classification performance, and rather add more information to the classification task. To further account for the potential multicollinearity problem, algorithms which are due to their structure immune to multicollinearity (e.g. XGBoost and Random Forest) were applied.

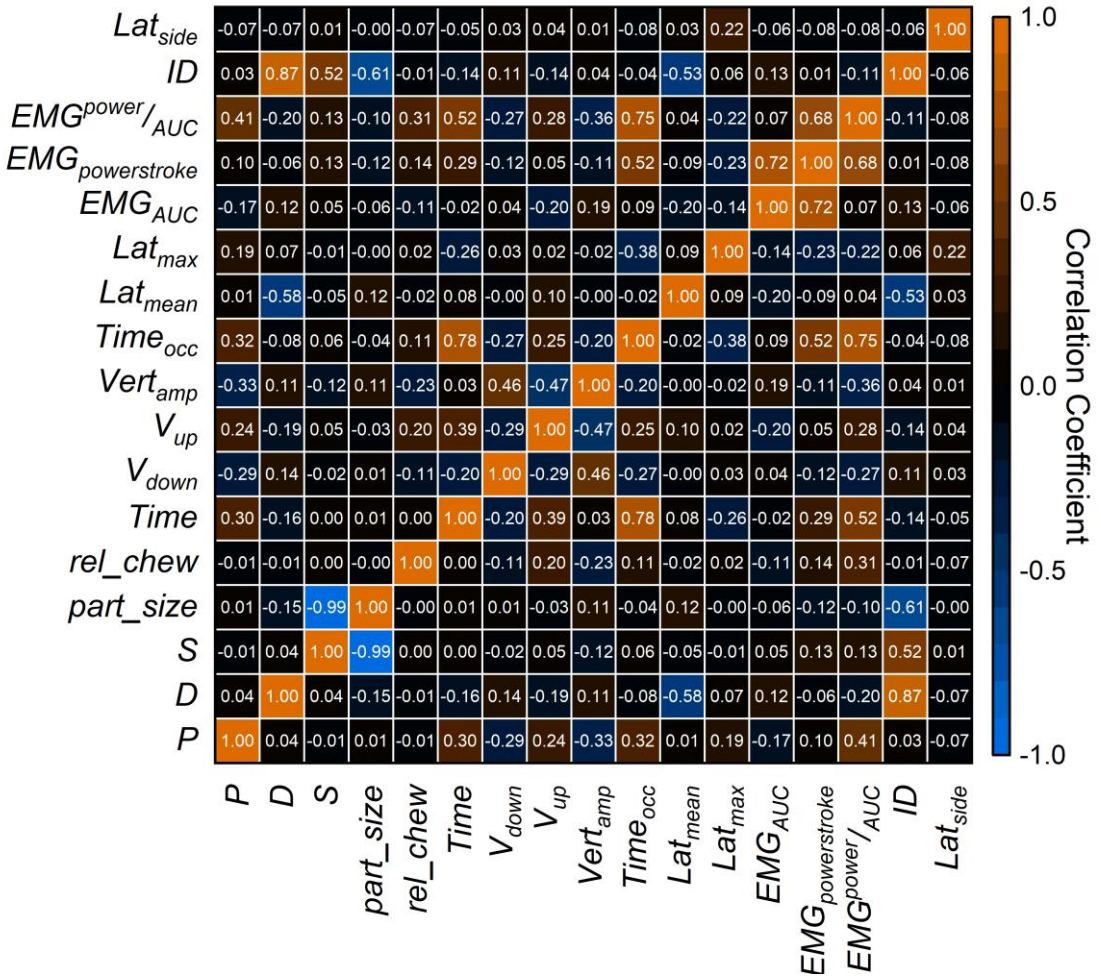

Figure Fehler! Kein Text mit angegebener Formatvorlage im Dokument**..2: Heatmap of the feature correlation analysis conducted as part of the exploratory data analysis. Features and abbreviations are explained in Table Fehler!** Kein Text mit angegebener Formatvorlage im Dokument**..1.**

Data curation regarding outlier detection and deletion of missing values is essential to enable a robust and well-performing machine learning algorithm. However, calculated outliers can also be

misunderstood, especially when dealing with skewed data or small datasets. Therefore, it is essential to know the type of data for deciding whether dropping outliers is suitable. In this case, statistical outliers according to the *iForest* method could be detected; however, this has likely to be attributed to the high inter- and intraindividual variations in the mastication physics over the term of mastication. However, they are considered to convey important information and cannot be neglected. For example, the first bite of a sequence involves jaw movements where the sample is placed between the teeth. Therefore, it most likely has a significantly larger vertical movement than the other bites, which marks it as an outlier. We suspected that eliminating the outliers would lead to a loss of information. More specifically, this could have resulted in the elimination of a particular phase of chewing, such as the first or last bite in a sequence. Eliminating these phases would result in a significant loss of information and poorer model performance. Hence, those identified exceptional values were not removed from the dataset. Missing values could not be detected.

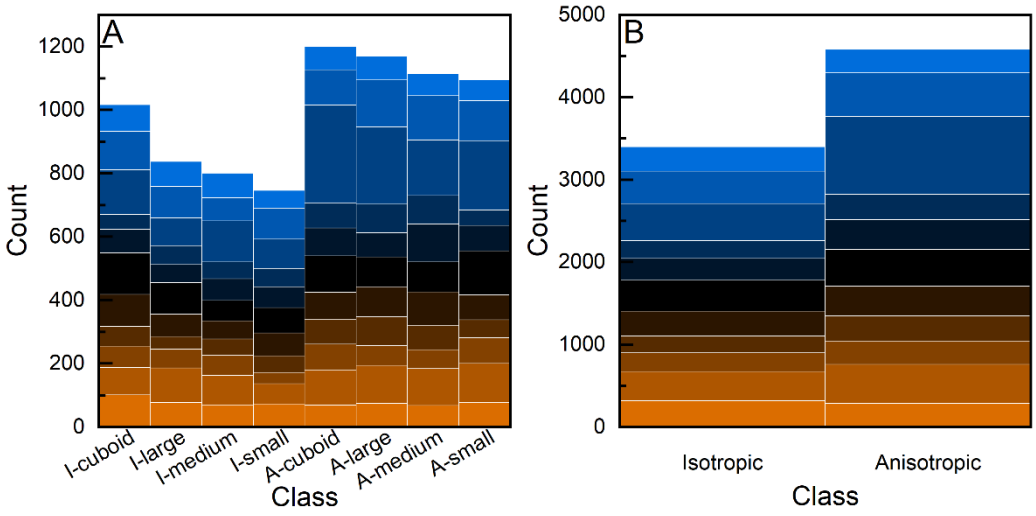

Figure Fehler! Kein Text mit angegebener Formatvorlage im Dokument**..3: Amount distribution of datapoints split by each individual Sample (A) and sample structure (B). Data is shown as cumulative stacks for all subjects. Class description descriptions are given as I: isotropic and A: anistropic with particle sizes: cuboid (not comminuted), large, medium and small**
each color indicates an individual subject (n=11)

Figure **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.3a shows the distribution for eight different samples. The distribution is slightly affected by imbalance. For example, class three of the specific samples only has 746 observations, while class four has 1201 observations. Further, it was noticed that the class of fiber structure is also affected by imbalance, see Figure **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.3b. Since there is no strict definition of when a feature is imbalanced or balanced, the machine learning approach will include techniques for balancing the dataset, as an imbalance in the data might lead to overfitting to a specific aspect.

From Figure **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.3, the difference between the samples can already be estimated immediately: **Samples I-cuboid, -large, -medium, and -small** (see Figure **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.3a) correspond to the isotropic samples, **Samples A-cuboid, -large, -medium, and -small** represent the anisotropic samples, whereas the particle size is decreasing from cuboid (1 “particle”) over large and medium to small. In Figure **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.3b, it can be seen that the dataset includes more measured bites for the anisotropic sample, indicating that these samples were across all panelists masticated with a higher number of chews. A similar effect can be observed for particle size: **I-** and **A-cuboid** had the largest particle size; hence those samples required more bites and are overrepresented in the dataset, while **Samples I**-and **A-small** were the finest samples, therefore, requiring the least number of chews.

Sumarizing, the exploratory data analysis could show that the features are not critically correlated with each other. Outliers could be detected, but were attributed to the complex nature of mastication sequences and therefore not excluded. It was further discovered that the dataset is slightly imbalanced, reasoned by different structure of the samples which need less or more chews until swallowing. The imbalance was not rated as critical, yet methods that can cope with imbalanced datasets were applied.

Machine Learning Evaluation

This section focuses on a thorough analysis and evaluation of the results from various machine learning algorithms listed in Table **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.2, which we used to predict the structure or mixed effect of meat products' structure and particle size. All classifications were further performed with and without respect to the person ID to resolve if subject-dependent effects have to be considered or if the effects observed for one group of subjects are transferrable to another. In the evaluation section, only data for the person ID are shown. The differences are discussed at a later point.

Evaluation by Sample

The sample classification output feature represents each meat sample, which is built out of a combination of sample size (cuboid, large, medium, small) and fiber structure (anisotropic, isotropic). The goal of this multi-class classification is to be able to classify all eight samples with the highest possible performance.

Algorithm Comparison

The accuracy score is the most common metric for evaluating the algorithm's performance. This score shows how many input values are correctly classified out of all input values. An accuracy of more than 0.7 represents a well-performing algorithm. An accuracy of more than 0.9 means the classification performance is very good. Orange data points in Figure **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.4 show the algorithm performance of meat sample classification using *k-fold cross-validation*. Figure **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.4 shows a comparison of the accuracy scores of the applied basic algorithms (Basic), of algorithms with standardized data (Standardized), and of the algorithms where data has been standardized with *StandardScaler* and augmented with SMOTE (Augmented). Accuracies are shown as the mean value of the *k-fold cross-validation* with standard deviation. The ensemble learning algorithms perform better than the classical algorithms (e.g., SVM, *Gaussian naïve Bayes* GNB, kNN, or LR) without data standardization or augmentation,

except for the adaptive boosting classifier. However, the decision tree algorithm performs with an accuracy of 0.3337±0.0138 much better than the other classical algorithms (accuracy =0.1998 to 0.2778). It even performs better than the ANN classifier (accuracy =0.3049). The SVM classifier shows with 0.0107 the smallest standard deviation of accuracy scores over the *k-fold cross-validation* out of all classifiers. The ANN classifier has the largest standard deviation (0.0274), which shrinks to 0.0128 when scaling the data. Moreover, when standardizing the data, SVM, kNN, and the ANN classifiers perform much better than without standardization. The other algorithms are not profiting visibly from standardized data. Applying standardization and data augmentation has no visible effect on the algo'ithm's performance. Only the adaptive boosting algorithm shows little improvement when standardizing and augmenting the data. Overall, the ExtraTrees classifier is worth further optimization and study, as it shows the highest accuracy score out of all algorithms.

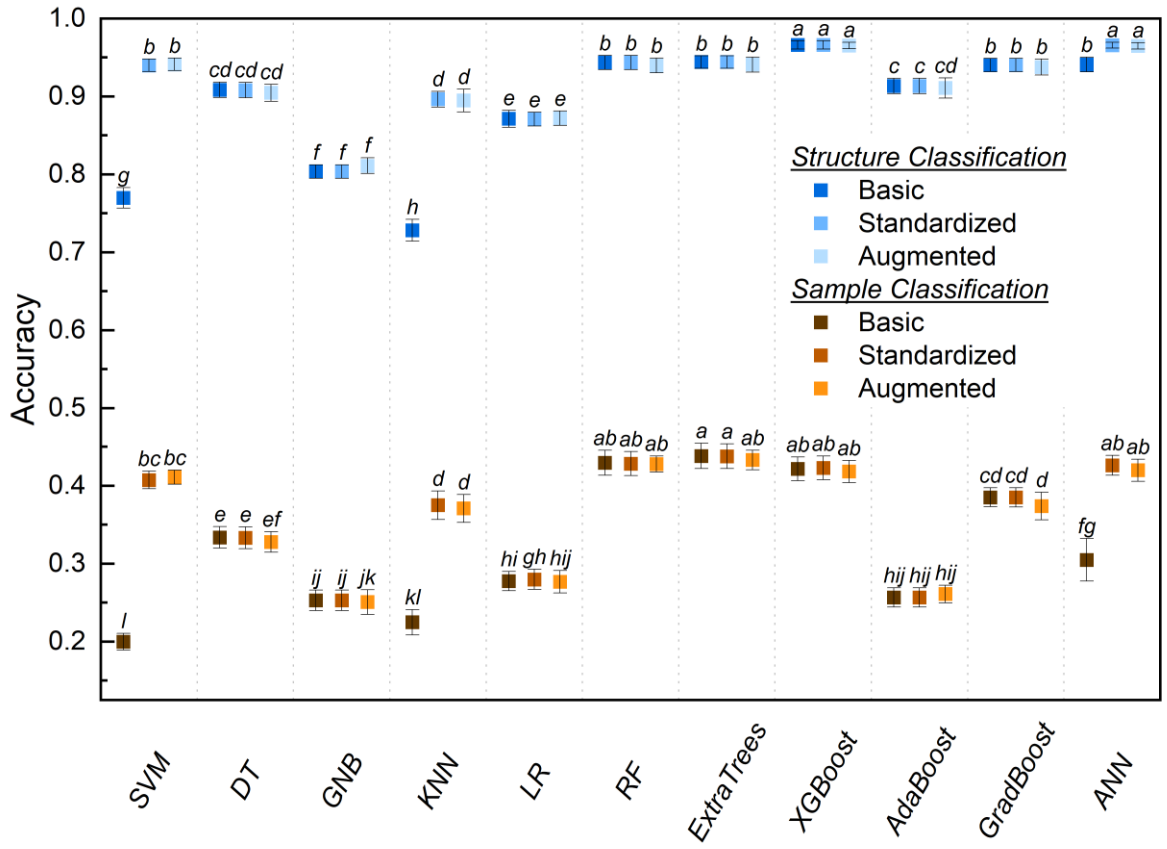

Figure Fehler! Kein Text mit angegebener Formatvorlage im Dokument**..4: Algorithm accuracy comparison of the two classification approaches (By Sample: ■■■; By Structure: ■■■) for three different pre-processing approaches: (i) Basic (■■): Without pre-processing, (ii) Standardized (■■): with standard scaling and (iii) Augmented (■■): with data augmentation after scaling. Values are displayed as mean values with standard deviation of the tenfold cross validation. Algorithm description and abbreviations can be found in Table Fehler!** Kein Text mit angegebener Formatvorlage im Dokument**..2. Accuracy starts from 0.125 on, as this is approximately the random probability for the sample classification. Subscript letters next to datapoints indicate statistically homogenous groups in each classification according to a Tukey test (α=0.05).**

Performance Assessment

When taking all metrics comparatively into account, the best-performing algorithm for sample classification is the ExtraTrees classifier without any pre-processing methods like standardization or data augmentation. However, the performance is not high enough for a practically helpful application. The application of classifying every individual chew of a sequence does not require a very high accuracy, since some of the jaw movements in a sequence will always be chaotic and not to be classified correctly. Although, we suggest that the accuracy should be for at least more

than two thirds of the chews correct. Different evaluation metrics of this classifier are shown in Table **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.3. First, an accuracy of 0.4384 does not indicate a great performance when a score of 1 means perfect classification. Yet, it is still better than random guessing, which would have an accuracy of around 0.125 in the case of perfectly balanced classes. Second, regarding the AUROC score with a value of 0.8541, this result can easily mislead into thinking that the classifier performs well. The AUROC score is calculated as the area under the Receiver-Operation Curve, which illustrates a classifier's diagnostic ability, more specifically, the true positive against the false positive rate of an algorithm as its discrimination threshold is varied (Fawcett, 2006). Third, a *Matthews Correlation Coefficient* (MCC) of 0.3561, which is considered more robust since it takes into account true positive, true negative, false positive, and false negative, is not far away from random guessing (MCC=0) (Chicco & Jurman, 2020). Fourth, the precision, recall, and F1 score are also far from perfect prediction, represented by a score of 1. Overall, it can be concluded that this classification task still requires improvement in algorithm performance or even data transformation to be further implemented.

Reasons for the bad performance of each single class (combination of structure and particle size) might root from having not enough data for training the algorithms. Compared to the structure classification, the dataset is split up into four classes for each structure, resulting in only one fourth of datapoints for each class. A higher number of observations is known to increase model performance, due to a more Gaussian-like distribution, generally more data to train the algorithms robustly, and lower importance of subject-dependent characteristics. Further, small differences in the mastication behavior between the particle sizes of the samples could be a reason for bad performance. This is in doubt, since other researchers already showed that the particle size of samples has a significant influence on the mastication behavior (Kim et al., 2015; Koç, Vinyard, Essick, & Foegeding, 2013).

Table Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.3: Evaluation metrics of the individual sample classification applying "ExtraTrees" classifier without pre-processing.**

Evaluation Metric	Score
Accuracy score	0.43835
AUROC score	0.85407
F1 score	0.43497
Matthews Correlation Coefficient	0.35609
Precision score	0.44053
Recall score	0.43835

The results were further evaluated by plotting Receiver-Operating Curve (ROC) and precision-recall curves and calculating the characteristic area under the curve (AUC). The recall and precision scores were calculated based on different thresholds, while a higher AUC score means better performance. A random classifier would show a horizontal-lined precision-recall curve with a precision based on the positive rate (Keilwagen et al., 2014). The AUC for each of the 8 sample classes and the micro-average precision-recall AUC are shown in Table **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.4. Class **I-cuboid** has the highest AUC score of 0.570. Furthermore, classes **I-small** and **A-cuboid** have an AUC score higher than the micro-average AUC of 0.440. The predictions for classes **I-medium** and **A-large** show the worst performance, with an AUC of 0.368 for class **I-medium** and an AUC of 0.367 for class **A-large**.

Table Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.4: Precision-recall and ROC areas under curve (AUC) for "ExtraTrees" Classifier calculated with scikit-plot for every individual sample class.**

Class	ROC AUC	precision-recall AUC
I-cuboid	0.570	0.88
I-large	0.428	0.87
I-medium	0.368	0.86
I-small	0.493	0.89
A-cuboid	0.483	0.84
A-large	0.367	0.80
A-medium	0.416	0.82
A-small	0.437	0.87
Micro- (Macro-) average	0.440	0.86 (0.86)

Class description descriptions are given as I: isotropic and A: anistropic with particle sizes: cuboid (not comminuted), large, medium and small

According to the AUROC values, the predictions for class **I-small** show the best performance of all classes with an AUROC of 0.89. Furthermore, classes **I-cuboid, -large, medium** and **A-small** are predicted better than the micro- and macro-average ROC (AUROC=0.86). Classes **A-cuboid, -large** and **-medium** are predicted worse than the micro- and macro-average ROC, while class **A-large** has the lowest AUROC value of 0.80. The micro- and macro averages have equal AUROC scores.

Evaluation by Structure

This section looks at the evaluation results of fiber structure classification. Compared to the multi-class classification of samples, this approach is a binary classification. Either anisotropic or isotropic meat structure can be predicted.

Algorithm Comparison

Figure **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.4 visualizes the accuracy scores of *k-fold cross-validation*. First, the plots show much better accuracy scores than the plots of sample classification. Once more, the ensemble learning algorithms and the ANN classifier perform better than the classical algorithms. kNN performs the worst, with an accuracy score of 0.7283. Contrarily, *Extreme Gradient Boosting* (XGBoost) has the best performance out of all fundamental algorithms, with an accuracy score of 0.9662. When standardizing the data, SVM, KNN, and the ANN classifier show visible performance improvement, while the other algorithms' accuracy score does not change visibly. Applying SMOTE to the dataset does not visibly improve any algorithm's performance. It even slightly decreases the performance of the gradient boosting classifier from 0.9405 to 0.9378. *Random Forest* (RF), the *Extra Trees Classifier* (ExtraTrees), and the *Gradient Boosting Classifier* (GradBoost) have similar accuracy scores. It can be noticed that the GNB classifier does not show improvements when applying the pre-processing techniques *StandardScaler* and SMOTE. The accuracy scores of *k-fold cross-*

validation further show a small standard deviation between 0.0037 and 0.0246, which describes that most values are close to the mean accuracy.

Matthews Correlation Coefficient

Table **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.5 shows Matthew's Correlation Coefficients (MCC) of each pipeline and machine learning algorithms since it is considered a robust evaluation metric (Chicco & Jurman, 2020). The classical algorithms SVM, GNB, KNN, and LR show the best MCC when transforming the dataset using the data augmentation technique SMOTE. Furthermore, the decision tree algorithm works best with normalized data. The best performance for the ensemble learning algorithms is achieved without additional pre-processing methods (Standardization and Augmentation) or by standardization. Moreover, standardizing the dataset shows the same evaluation results as only implementing the basic algorithms *Extra Trees Classifier*, XGBoost, AdaBoost, and GradBoost classifier. The best MCC score of 0.9308 is achieved using XGBoost indifferent of the pre-processing (none, Standardization or Augmentation).

Table Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.5: Matthews correlation coefficient of the sample structure classification for three different pre-processing approaches: (i) Basic: Without pre-processing, (ii) Standardized: with standard scaling and (iii) Augmented: with data augmentation after scaling. Algorithm description and abbreviations can be found in Table** Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.2.**

Model	Basic	Standardized	Augmented
SVM	0.52532	0.87854	0.88183
DT	0.81259	0.81234	0.80559
GNB	0.59942	0.59942	0.62193
KNN	0.43893	0.78905	0.78943
LR	0.74280	0.74094	0.74751
RF	0.88646	0.88646	0.87996
ExtraTrees	0.88767	0.88767	0.88283
XGBoost	**0.93080**	**0.93080**	**0.92934**
AdaBoost	0.82505	0.82505	0.82162
GradBoost	0.87906	0.87906	0.87485
ANN	0.88037	0.93023	0.92861

Feature Importance

Table **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.6 shows the feature importance of the fiber structure classification with *k-fold cross-validation*. It is important to note that **Lat_{mean}** (mean absolute lateral jaw displacement per bite) is the top feature in all the presented algorithms, besides the adaptive boosting classifier. The A*daptive Boosting Classifier* is the only algorithm that shows the largest importance of **Lat_{side}** (mean lateral jaw displacement for the side, where positive and negative values describe left and right). However, this feature is calculated with the same values of lateral movement as **Lat_{mean}**. This also explains the importance of **Lat_{mean}** when training the *Adaptive Boosting Classifier*. The same results of **Lat_{mean}** and **Lat_{side}** being the key features is achieved when classifying the eight samples.

Table Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.6: Feature Importance of the classification of sample structure of the best six best performing algorithms without pre-processing. The sum of each column equals 1 and expresses the weight that each feature influenced the model performance.** ***Features explanation and abbreviations can be found in*** **Table** Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.1.**

Feature*	DT	RF	ExtraTree	XGBoost	AdaBoost	GradBoost
Lat_{mean}	0.53737	0.45115	0.35011	0.39205	0.28	0.66126
Lat_{side}	0.15022	0.09156	0.05926	0.12480	0.34	0.17427
$EMG^{power}/_{AUC}$	0.05826	0.05752	0.06858	0.08038	0.04	0.03980
Lat_{max}	0.06280	0.05549	0.05555	0.07464	0.12	0.05665
P	0.02252	0.03810	0.07227	0.07339	0.14	0.02272
Time	0.01988	0.04317	0.05691	0.04731	0.02	0.01259
V_{down}	0.02533	0.03810	0.04274	0.03514	0.02	0.01508
V_{up}	0.02458	0.04094	0.04792	0.03150	0.02	0.00482
EMG_{AUC}	0.02761	0.04670	0.05387	0.03032	0.02	0.00588
$Vert_{amp}$	0.02020	0.03243	0.04244	0.02639	0.00	0.00085
$Time_{occ}$	0.01293	0.02757	0.04247	0.02286	0.00	0.00230
S	0.00882	0.01200	0.03000	0.02226	0.00	0.00115
rel_chew	0.01764	0.02880	0.03644	0.02134	0.00	0.00165
$EMG_{powerstroke}$	0.01183	0.03647	0.04143	0.01762	0.00	0.00098

** Features are listed with decreasing feature importance according to the best performing algorithm, XGBoost*

Furthermore, AdaBoost does not show any feature importance of **S**, **rel_chew**, **$Vert_{amp}$**, **$Time_{occ}$**, and **$EMG_{powertstroke}$** for the classification with person ID. Overall, it can be concluded that the feature **Lat_{mean}** is important in the fiber structure classification approach. With a feature importance of around 0.35 to 0.67, this feature is significantly involved in the training process. The other features are also important but have much smaller feature importance. It can further be seen that all three features related to lateral movement (**Lat_{max}**, **Lat_{mean}**, and **Lat_{side}**) are under the top four important features. Hence, we claim that the main difference between chewing anisotropic and isotropic samples is the lateral movement of the jaw.

Performance Assessment

The best-performing classifier for fiber structure classification is XGBoost. XGBoost shows the same results without additional pre-processing and when standardizing the data. Because of simplicity, the pipeline of XGBoost without scaling is shown below. Table **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.7 presents the achieved scores of selected evaluation metrics. With an accuracy of 0.9662, the fiber structure is classified correctly into anisotropic and isotropic samples. The MCC score of 0.9308 represents an outstanding classification performance, which is not far from ideal (=1). Also, considering the AUROC score of 0.9956, this algorithm performs excellently. Both metrics demonstrate that the algorithm has a high true positive rate without a high error rate of false positives. The F1, precision, and recall scores are also above 0.9, strengthening the previous conclusion. Recall, precision, and the F1 score all together show that the precision of the algorithm (returning only relevant results) and the recall of the algorithm (returning all relevant results) are on a very high level. This presents an outstanding applicability to real world scenarios since oral processing tasks for product development do not have the requirement of a classification with accuracies beyond 99%. In Section „Evaluation by Sample” we stated that even an accuracy of more than 0.66 is not necessarily bad and could be applied to distinguish if, e.g., texturized vegetable protein samples, often applied as meat substitute, could be told apart from real meat structures based on the

mastication physics. The low requirements are because it is not about high risk and sensitive classification tasks, but rather about quality and indications if for example product development is taking steps in the correct direction.

Table Fehler! Kein Text mit angegebener Formatvorlage im Dokument**..7: Evaluation metrics of the sample structure classification applying XGBoost without pre-processing.**

Evaluation Metric	Score
Accuracy score	0.96617
AUROC score	0.99559
F1 score	0.97049
Matthews Correlation Coefficient	0.9308
Precision score	0.97429
Recall score	0.96678

The precision-recall curve, shown in Figure **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.5 (right), plots the precision on the y-axis and the recall on the x-axis. In this figure, the folds of *k-fold cross-validation* are shown separately and as mean values. It can be noticed that all scores show an AUC score above 0.995, which is excellent for a classifier and far away from random guessing (AUC=0.5). Also, when looking at the different folds, it can be noticed that the values have little standard deviation. The ROC curve, shown in Figure **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.5 (left), affirms the results of the precision-recall curve. Being pushed in the left corner, the mean ROC curve visualizes a well-performing classifier. Furthermore, the ten folds show AUROC scores between 0.9936 and 0.9970, demonstrating a small standard deviation.

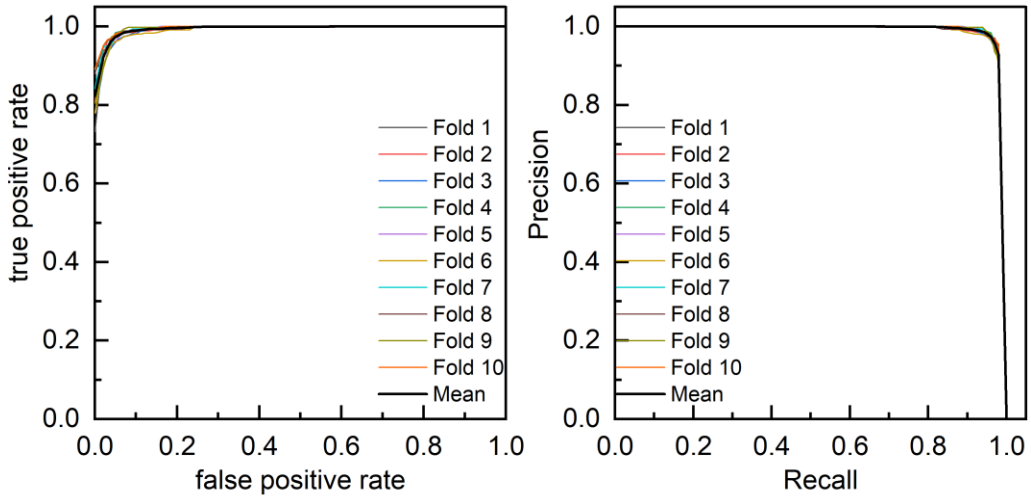

Figure Fehler! Kein Text mit angegebener Formatvorlage im Dokument**..5: ROC- and precision recall curve for the classification by sample structure applying XGBoost without pre-processing. Results are displayed as the ten individual folds of the cross validation and mean value (thick black line) of all folds.**

Visualizing the true and false positive and negative rate, which are all considered in the MCC, again proves that the classification conducted with XGBoost is promising (see Figure **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.6). With 4426 out of 4579 chews correctly identified as anisotropic food and only 117 out of 3401 chews faulty assigned to anisotropic food, the error is kept minimal.

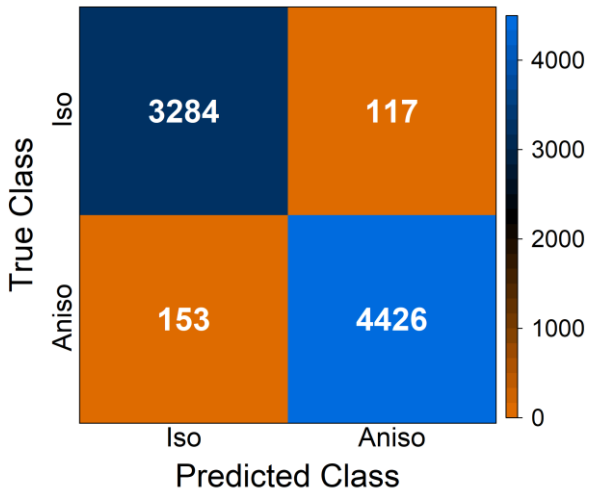

Figure Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.6: True and false predicted classes of the XGBoost classifier displayed as confusion matrix. Values were calculated for the 10-fold cross validation of the sample structure classification. Aniso refers to samples with anisotropic grown meat structure and Iso refers to isotropic sausage like structured products.**

Discussion of Practical Implications

One objective of this work was to determine subject-dependent effects on classification accuracy. It was hypothesized that the unique person ID is essential for correctly classifying the food oral processing dataset, as each person has individual chewing patterns and habits. At least, a strong impact on oral processing parameters by the subjects body weight, salivary flow rate, dental status, or age was proposed by Ketel, de Wijk, de Graaf, and Stieger (2020). Therefore, the model should become more accurate if the data can be adjusted for individuals. However, we demonstrated that the influence of the individuals influence "P" is minimal. Each pipeline was calculated with and without "P" taken into account, but no significant differences were found. This is also supported by the low feature importance of "P", as shown in Table **Fehler! Kein Text mit angegebener Formatvorlage im Dokument.**.6. At first sight that appears to be in contrast with previous findings stating that oral processing characteristics vary at an absolute level depending on the above factors. This work could however show that for specific applications and with correct data pre-processing, the subject-related effects are reduced. It is rather assumed, that the extend of the measured feature differs stronger or weaker in dependence on the subject, which can be eliminated by standardization or choosing of an appropriate algorithm. The mechanism how the mastication process is adapted to the food material however is always of similar kind. Hence, we propose that there exist food material specific oral processing characteristics that are elicted upon consumption, which do not only apply to a specific group of people; but there is evidence for a general dependency between chewing behavior and structure that can be found in every healthy adult human being. For practical applications of oral processing studies, this is of great value, since it proves that it is not only possible to record oral processing parameters for individuals and

bring them in context to the food structure. It is moreover possible to contextualize specific structure and sensory terms with oral processing features, enabling to read for example the fibrousness of food from oral processing data. This could serve as additional tool to the present days conventional texture analysis by deformation and a sensory panel. Conventional sensory studies have the drawback that the panelsists are distracted by more aspects of the product than only the texture, such as appearance aroma and flavor. Benefits against texture analysis by deformation are given by the fact that the food is in the case of oral processing masticated in a real-world scenario, mixed with saliva, brought to bodytemperature, crushed, cut and ground by teeth, and broken down by enzymes. All the before mentioned temporal effects can to date not be simulated in one device. Oral processing however lacks true material specific values, aromatic and flavor perception, and the individual impression of the panelists. Hence it is important to bring multiple methods in combination for a promising and holistic product characterization.

This work could show that pre-processing of the data is beneficial for some algorithms to achieve good accuracies. However, only standardization of the values proved very effective. Other pre-processing techniques, such as ANOVA or PCA, showed no significant effects or even reduced accuracy. The earlier discussed issue of multicollinearity of some features describing time and muscle activity characteristics of mastication sequence might have lead to the worse performance of classic algorithms (e.g. KNN, LR and GNB) in comparison to algorithms which are immune to multicollinearity like XGBoost and Random Forest.

A direct comparison of this works relusts to our previous work (Oppen et al., 2023) applying a linear mixed model to the given dataset is not possible because the scope of the analysis was different. While this study attempted to find data patterns that would allow the samples to be differentiated into classes, the original work aimed to find effects in isolated jaw movement or muscle activity features, understanding which structures and particle sizes result in what kind of differences in jaw movement and muscle activities. The original method did not allow for classification but found statistically significant effects of the isolated features depending on the

progress of mastication, particle size, and state of anisotropy further taking into account individual errors of each individual subject. More specifically, the evaluation utilizing a linear mixed model enabled to model single features, e.g. the peak muscle activity with the fixed variables progress of mastication, particle size and anisotropy. Following, one could exactly see how, at which mastication stage, and to which extent the variation of mentioned variables influenced the peak muscle activity. The dependencies in Oppen et al. (2023) are expressed as functions, which include a coefficient for every fixed effect and a random statement. The model allows to calculate the significance of every effect, providing information if the effect of particle size is significantly influencing e.g. the lateral jaw movement. Differences between specific samples (e.g. small particle size, anisotropic vs. isotropic structure) could not be evaluated with the chosen method of evaluation.

In contrast, this work used machine learning to classify the dataset, assigning a class to each individual chew in the dataset. Either the structure of the sample or the combination of structure and particle size was chosen as the class. The machine learning algorithms thus did not create models based on only one of the oral processing features, but considered them all simultaneously. In this way, it can also be shown which features significantly influence the assignment of a data point to a certain class. In summary, the approach described in Oppen et al. (2023) examined detailed food influences on specific oral processing characteristics, whereas the present work aims to use the differences described to classify model food systems together based on all oral processing characteristics.Classifying the eight food samples with machine learning could show potential, reaching accuracy scores of up to 0.4384 and an AUROC of 0.8541 for the *Extra Trees Classifier*. Against the random probability of 0.125 for eight samples, we could demonstrate that specific patterns could be recognized but not of sufficient precision for an application. It is hypothesized that a higher number of panelists would, due to (i) a more Gaussian-like distribution, (ii) generally more data to train the algorithms robustly, and (iii) lower importance of subject-dependent characteristics, result in better accuracies. For future work, the single bites could be

brought in context to their original sequence again, enabling to enhance the precision by not taking every single bite by itself but taking into account the whole sequence.

For the classification of anisotropy, this work resulted in excellent model performance, showing that the data of each bite can be assigned to the correct structural parameter "isotropic" or "anisotropic" with an accuracy of 0.9662. The feature importance analysis showed that the lateral movement of the jaw contributed strongly to the algorithm's performance. The application of a mixed linear model by Oppen et al. (2023) could not show a significant isolated effect of anisotropy on lateral movement; still, the interaction of masticatory process, particle size, and anisotropy was highly significant (Oppen et al., 2023).

Threats To Validity

This paper aimed to build a machine learning approach that uses existing algorithms to classify the mastication data of jaw movement and muscle activity. However, the goal was not to generate more data with experiments or to build novel algorithms. This work has limitations in some respects, which we will summarize in the following. First, the amount of data or, more precisely, the number of panelists in the dataset is comparably small and might not be representative. Further, robustness might threaten validity since oral processing data are highly individual and not perfectly Gaussian distributed. For further works, using person-independent learning by applying the “leave one subject out” cross-validation would also be suggested. This work covered many learning classifiers from ensemble learning algorithms; however, only one artificial neural network could be implemented in this study. With steady progression in deep learning, long-short term memory neural networks, recurrent neural networks, or feed-forward artificial neural networks, for instance, could be added to the machine learning pipeline. Further, we investigated the use of hyperparameter tuning of all investigated algorithms, i.e., additionally optimizing the parameters of the machine learning algorithms. Evaluated hyperparameters were for example the learning rate, number of estimators, maximal depth and the hidden layer size, depending on the

investigated algorithm. The limited approach of hyperparameter tuning in this work did not improve classification accuracy; however, enlarging the grid and applying it to different algorithms should still be considered in future work. Lastly, more versatile pre-processing methods could be applied, for example, a transformation of the dataset to a Gaussian distribution, which was impossible due to the limited number of panelists.

Conclusions

The findings in this work can be seen as a first step towards establishing machine learning in food science data evaluation. More specifically, an additional tool to analyze high-dimensional data as the present dataset about mastication features of panelists consuming food, for example of anisotropic and isotropic structure, is provided. The different pre-processing techniques like scaling, dimensional reduction, or feature selection did not improve model performance in most cases. Only the application of standardization could enhance the accuracy of SVM, KNN, and ANN significantly. Differentiating between the eight individual samples with different macroscopic (coarse to fine food particles in four variations) and microscopic (grown meat fiber structure and meat protein gel) structure did not result in high accuracy, but ensemble learning classifiers could classify the samples significantly better as compared to random guessing. It was shown that the binary classification of food microstructure is very promising. With an accuracy of 0.9662 and AUROC values of up to 0.9956, the algorithm XGBoost performed best in classifying the dataset for the microstructure of the sample. As a key outcome, it was shown that machine learning algorithms can classify food into anisotropic meat structures and isotropic meat protein gel structures based on masticating patterns. In the field of oral processing, great efforts have been made in the past to find slight differences in individual features of oral processing and to understand which food structure causes what kind of change in the chewing process. The present work makes use of this knowledge and transforms the findings into an application-relevant tool. The conducted classification method could for example be applied to evaluate if structured plant protein samples classify, based on their mastication pattern, as isotropic or anisotropic meat like

structure. This could reveal potential candidates for meat substitute materialA similar approach could be taken in the meat production industry: For example, it could be tested whether novel meat processing techniques or adapted recipes are to be classified as different from the reference method. This would show whether consumer perception could be changed or whether the changes remain undetected. Machine learning algorithms based on this work can be used for other food oral processing studies that measure various features to describe the kinematics of jaw movements and electromyographic data of the jaw muscles. Further, as future work, deep learning algorithms could be studied after enlarging the dataset with more individuals. Additionally, the first works study how to use machine learning approaches to model a digital food twin (e.g., Henrichs et al. (2022), Krupitzer and Stein (2021) and Krupitzer, Noack, and Borsum (2022)). Using the applied machine learning algorithms for oral food processing in such a digital twin might be beneficial for calculating the mastication of an adjusted receipt based on historical data.

Ethical Statement

This research was approved by the ethical committee of Hohenheim on 30th of March 2021 (no case number available). Participants gave informed consent via the statement "I am aware that my responses are confidential, and I agree to participate in this survey" where an affirmative reply was required to enter the session. They were able to withdraw from the session at any time without giving a reason. The products assessed were safe for consumption

Acknowledgements

The Authors would like to thank all panel members, which participated in the oral processing experiments.

Declaration of interests

None.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial,

or not-for-profit sectors.

Publication Bibliography

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician, 46* (3), 175-185.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics, 16* (5), 412-424.

Bayes, T., & Price. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London, 53*, 370-418.

Biga, L. M., Dawson, S., Harwell, A., Hopkins, R., Kaufmann, J., LeMaster, M., Matern, P., Morrison-Graham, K., Quick, D., & Runyeon, J. (2019). Anatomy and physiology. In: OpenStax/Oregon State University

Braxton, D., Dauchel, C., & Brown, W. E. (1996). Association between chewing efficiency and mastication patterns for meat, and influence on tenderness perception. *Food quality and preference, 7* (3-4), 217–223.

Breese, J. S., Heckerman, D., & Kadie, C. (2013). Empirical analysis of predictive algorithms for collaborative filtering. *arXiv preprint arXiv:1301.7363*.

Breiman, L. (2001). Random Forests. *Machine Learning, 45* (1), 5-32.

Brown, W. E., Langley, K. R., Mioche, L., Marie, S., Gérault, S., & Braxton, D. (1996). Individuality of understanding and assessment of sensory attributes of foods, in particular, tenderness of meat. *Food quality and preference, 7* (3-4), 205–216.

Çakir, E., Koç, H., Vinyard, C. J., Essick, G., Daubert, C. R., Drake, M., & Foegeding, E. A. (2012). Evaluation of texture changes due to compositional differences using oral processing. *Journal of Texture Studies, 43* (4), 257-267.

Chen, J. (2009). Food oral processing—A review. *Food hydrocolloids, 23* (1), 1–25.

Chen, T., & Guestrin, C. (2016). XGboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). San Francisco, California, USA.

Cheng, Z., Zou, C., & Dong, J. (2019). Outlier detection using isolation forest and local outlier factor. In *Proceedings of the Conference on Research in Adaptive and Convergent Systems* (pp. 161–168). Chongqing, China: Association for Computing Machinery.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21* (1), 6.

Cox, D. R. (1959). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B (Methodological), 21* (1), 238-238.

de Ville, B. (2013). Decision trees. *WIREs Computational Statistics, 5* (6), 448-455.

Dekkers, B. L., Boom, R. M., & van der Goot, A. J. (2018). Structuring processes for meat analogues. *Trends in food science and technology, 81*, 25–36.

Devezeaux De Lavergne, M., Young, A. K., Engmann, J., & Hartmann, C. (2021). Food oral processing—an industry perspective. *Frontiers in Nutrition, 8*, 634410.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27* (8), 861-874.

Foegeding, E. A., Vinyard, C. J., Essick, G., Guest, S., & Campbell, C. (2015). Transforming structural breakdown into sensory perception of texture. *Journal of Texture Studies, 46* (3), 152-170.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence, 14* (771-780), 1612.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences, 55* (1), 119-139.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38* (4), 367-378.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63* (1), 3-42.

Grossmann, L., & Weiss, J. (2021). Alternative protein sources as technofunctional food ingredients. *Annual review of food science and technology, 12* (1), 93-117.

Henrichs, E., Noack, T., Pinzon Piedrahita, A. M., Salem, M. A., Stolz, J., & Krupitzer, C. (2022). Can a Byte Improve Our Bite? An Analysis of Digital Twins in the Food Industry. *Sensors, 22* (1), 115.

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association, 12* (3), 296-298.

Ilić, J., Djekic, I., Tomasevic, I., Oosterlinck, F., & Berg, M. A. v. d. (2022). Materials properties, oral processing, and sensory analysis of eating meat and meat analogs. *Annual review of food science and technology, 13* (1), 193-215.

Ilic, J., Van Den Berg, M., & Oosterlinck, F. (2021). How do we eat meat–the role of structure, mechanics, oral processing, and sensory perception in designing meat analogs. *IOP Conference Series: Earth and Environmental Science, 854* (1), 012036.

Keilwagen, J., Grosse, I., & Grau, J. (2014). Area under Precision-Recall Curves for Weighted and Unweighted Data. *PLOS ONE, 9* (3), e92209.

Ketel, E. C., de Wijk, R. A., de Graaf, C., & Stieger, M. (2020). Relating oral physiology and anatomy of consumers varying in age, gender and ethnicity to food oral processing behavior. *Physiology & behavior, 215*, 112766.

Khan, M. I. H., Sablani, S. S., Nayak, R., & Gu, Y. (2022). Machine learning-based modeling in food processing applications: State of the art. *Comprehensive reviews in food science and food safety, 21* (2), 1409-1438.

Kim, E. H.-J., Jakobsen, V. B., Wilson, A. J., Waters, I. R., Motoi, L., Hedderley, D. I., & Morgenstern, M. P. (2015). Oral Processing of Mixtures of Food Particles. *Journal of Texture Studies, 46* (6), 487-498.

Kircali Ata, S., Shi, J. K., Yao, X., Hua, X. Y., Haldar, S., Chiang, J. H., & Wu, M. (2023). Predicting the textural properties of plant-based meat analogs with machine learning. *Foods, 12* (2), 344.

Koç, H., Çakir, E., Vinyard, C. J., Essick, G., Daubert, C. R., Drake, M. A., Osborne, J., & Foegeding, E. A. (2014). Adaptation of oral processing to the fracture properties of soft solids. *Journal of Texture Studies, 45* (1), 47–61.

Koç, H., Vinyard, C. J., Essick, G. K., & Foegeding, E. A. (2013). Food oral processing: conversion of food structure to textural perception. *Annual review of food science and technology, 4*, 237–266.

Kohyama, K., & Mioche, L. (2004). Chewing behavior observed at different stages of mastication for six foods, studied by electromyography and jaw kinematics in young and elderly subjects. *Journal of Texture Studies, 35* (4), 395–414.

Krupitzer, C., Noack, T., & Borsum, C. (2022). Digital Food Twins Combining Data Science and Food Science: System Model, Applications, and Challenges. *Processes, 10* (9), 1781.

Krupitzer, C., & Stein, A. (2021). Food informatics: Review of the current state-of-the-art, revised definition, and classification into the research landscape. *Foods, 10* (11), 2889.

Le Révérend, B., Saucy, F., Moser, M., & Loret, C. (2016). Adaptation of mastication mechanics and eating behaviour to small differences in food texture. *Physiology and Behavior, 165*, 136-145.

Listrat, A., Lebret, B., Louveau, I., Astruc, T., Bonnet, M., Lefaucheur, L., Picard, B., & Bugeon, J. (2016). How muscle structure and composition influence meat and flesh quality. *The Scientific World Journal, 2016:3182746*.

Mammone, A., Turchi, M., & Cristianini, N. (2009). Support vector machines. *WIREs Computational Statistics, 1* (3), 283-289.

Melito, H. S., Daubert, C. R., & Foegeding, E. A. (2013). Relationships between Nonlinear Viscoelastic Behavior and Rheological, Sensory and Oral Processing Behavior of Commercial Cheese. *Journal of Texture Studies, 44* (4), 253-288.

Murtagh, F. (1991). Multilayer perceptrons for classification and regression. *Neurocomputing, 2* (5), 183-197.

Oppen, D., Grossmann, L., & Weiss, J. (2022). Insights into characterizing and producing anisotropic food structures. *Critical reviews in food science and nutrition*, 1-19.

Oppen, D., Young, A. K., Piepho, H.-P., & Weiss, J. (2023). Fibrous food and particle size influence electromyography and the kinematics of oral processing. *Food research international, 165*, 112564.

Pilgrim, M., & Willison, S. (2009). *Dive into python 3* (Vol. 2): Springer.