



## A BIG DATA-BASED ANALYSIS: COLOR AS A PREDICTOR OF SILAGE QUALITY




Ömer MERMER<sup>1</sup>, Ayşe Gül FİLİK<sup>1\*</sup>, Gökhan FİLİK<sup>1</sup>

<sup>1</sup>Kırşehir Ahi Evran University, Faculty of Agriculture, Department of Agricultural Biotechnology, 40100, Kırşehir, Türkiye

**Abstract:** In this study, the relationship between color parameters and nutrient composition was examined using silage trial data conducted on 10 different plant species or mixtures by the Filik Research Lab. The dataset, comprising a total of 6,254 observations, was analyzed using big data analytics and statistical modeling techniques. Significant associations were identified between color parameters ( $L^*$ ,  $a^*$ ,  $b^*$ ,  $C^*$ ,  $h^\circ$ ,  $\Delta E$ , saturation, and hue) and the physical, chemical, and sensory characteristics of the silage (crude protein, NDF, ADF, TDN, NFC). Initially, exploratory data analysis (EDA) and correlation matrices were employed to observe preliminary relationships between color and nutritional components. These were followed by classical and machine learning methods, such as multiple linear regression and Random Forest, to assess the strength and direction of these associations. The analysis revealed a\* negative correlation ( $r = -0.45$ ) between  $L^*$  (lightness) and crude protein (CP, %), while a strong positive relationship ( $r = 0.52$ ) was found between  $b^*$  (yellowness) and NDF. Additionally, a meaningful correlation ( $r = 0.41$ ) was observed between color saturation and TDN. Random Forest analysis confirmed these findings and highlighted  $L^*$ ,  $b^*$ , and saturation as the most influential color parameters. These results suggest that color parameters are not only indicators of visual quality but also provide indirect insights into the chemical and nutritional content of silage. This study, conducted within the scope of big data analytics, demonstrates that color data can serve as a\* rapid predictive tool, either as an alternative or complement to traditional chemical analyses. This presents a\* significant contribution toward the development of digital quality control systems in agricultural production.

**Keywords:** Silage quality, CIE Lab color space, Agricultural digitalization, Physical quality indicators, Machine learning, Agricultural decision support systems

\*Corresponding author: Kırşehir Ahi Evran University, Faculty of Agriculture, Department of Agricultural Biotechnology, 40100, Kırşehir, Türkiye  
E mail: aysegulcivaner@ahievran.edu.tr (A. G. FİLİK)

Ömer MERMER  <https://orcid.org/0000-0001-7729-8560>  
Ayşe Gül FİLİK  <https://orcid.org/0000-0001-7498-328X>  
Gökhan FİLİK  <https://orcid.org/0000-0003-4639-3922>

Received: June 17, 2025

Accepted: September 09, 2025

Published: September 15, 2025

Cite as: Mermer Ö, Filik AG, Filik G. 2025. A big data-based analysis: color as a predictor of silage quality BSJ Agri, 8(5): 725-735.

### 1. Introduction

Silage is a roughage source with high energy and protein content, produced by fermenting green forage crops under anaerobic conditions, and holds strategic importance particularly in ruminant nutrition. When produced under appropriate conditions and techniques, silage not only addresses seasonal forage shortages but also provides a sustainable, economical, and balanced feeding opportunity. Indeed, Filik et al. (2018) emphasized that high-quality silage is a key factor directly affecting animal performance and feeding economy, while Filik et al. (2022) noted that the combined evaluation of physical and chemical analyses allows for a\* multidimensional understanding of quality criteria.

Fundamental factors determining silage quality include plant species, harvest time, chopping size, use of additives, compaction level, and fermentation duration. However, the effective optimization of these factors is not limited to the production phase but is also closely linked to efficient and reliable analysis processes. Although traditional laboratory analyses provide comprehensive

information, they are often time-consuming, costly, and insufficient for large-scale screenings. In this context, recent approaches exploring the use of physical attributes—especially color parameters—as indicators of quality represent a significant example of digitalization in agriculture (Filik et al., 2018; 2022). Color is a fundamental indicator reflecting the visual and physical condition of plant material. Parameters such as lightness ( $L^*$ ), redness-greenness ( $a^*$ ), and yellowness-blueness ( $b^*$ )—defined within the CIE Lab\* color space—have been found to be significantly correlated with nutrient content in many studies (Van Soest, 1991; Goeser and Combs, 2009; Filik et al., 2018). These parameters are not only physical traits but also indirect reflections of biochemical processes, making them potential indicators of silage quality. This study is based on a big data structure compiled from datasets generated through ten independent projects conducted within the Filik Research Lab. The database was designed using relational modeling and structured to enable multidimensional data queries. In this research, exploratory data analysis (EDA), correlation analysis,



multiple linear regression, and machine learning algorithms such as Random Forest were used to evaluate the predictive power of color parameters for nutrient content and to investigate their potential as quality indicators. In this regard, the study not only contributes to the academic literature but also presents a concrete application example aimed at the digitalization of agricultural quality control processes. Thus, it provides a solid foundation for the development of rapid, cost-effective, and scalable digital analysis systems as an alternative to conventional laboratory-based methods.

## 2. Materials and Methods

The dataset used in this study was compiled from 10 different studies conducted within the Filik Research Lab. It includes 6,254 observations representing silage analysis data derived from 10 different plant species and mixtures. The plant materials and combinations used in silage preparation include: Pea-Barley, Oat, Rye, Vetch-Rye, Triticale, Wheat, Vetch-Oat, Italian Ryegrass, Millet, and Barley. The silage types are evenly distributed, with the highest number of complete observations found in Wheat, Pea-Barley, Vetch-Rye, Millet, Vetch-Oat, and Italian Ryegrass silages. These observations cover physical (pH, temperature, water-soluble carbohydrates,  $L^*$ ,  $a^*$ ,  $b^*$ ,  $C^*$ ,  $h^\circ$ ,  $\Delta E$ , hue, saturation), chemical (air-dry matter, organic matter, crude protein, ether extract, crude fiber, ADF, NDF, ADL, starch, sugar, hemicellulose, cellulose, total carbohydrates, NFC, NFE, DCP, TDN, DE, ME, NEL, net energy for maintenance [NEm], net energy for gain [NEg], digestible dry matter [DDM], dry matter intake [DMI], relative feed value [RFV] and relative forage quality [RFQ]), and sensory (post-opening pH,  $CO_2$ , lactic acid count, yeast, mold, aerobic stability yeast and mold count) parameters. The dataset is labeled with descriptive variables such as study name, silage type, sample size (input), and group replication (group\_rep).

After preprocessing, the color parameters were structured based on the CIE Lab\* color space and include:  $L^*$  (lightness),  $a^*$  (redness),  $b^*$  (yellowness),  $C^*$  (chroma),  $h^\circ$  (hue angle),  $\Delta E$  (color difference), hue, and saturation. Nutritional parameters include air-dry matter, organic matter, ash, crude protein (CP), ether extract, crude fiber, ADF, NDF, NFC, TDN, and DCP.

### 2.2. Data Analysis Process

The data analysis was carried out in five main stages:

#### 2.2.1. Data loading and cleaning

Excel files were imported using the pandas library. The two types of analyses were merged using the merge function. Rows containing missing values were removed using the dropna function.

#### 2.2.2. Exploratory data analysis (EDA)

Distributions of color and nutrient variables were examined using matplotlib and seaborn visualization tools. The distribution of silage types and descriptive statistics were assessed using the describe function.

#### 2.2.3. Correlation analysis

Pearson correlation coefficients were calculated to

evaluate the relationships between color parameters and nutritional values. The correlations were visualized using heatmaps.

#### 2.2.4. Regression models

Multiple linear regression models were constructed using the statsmodels library to predict target variables such as CP, NDF, and TDN. Model significance and variable contributions were assessed through summary outputs. The regression model was based on the following general form (equation 1):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon \quad (1)$$

#### 2.2.5. Machine learning analysis

Random Forest Regression was implemented using the sklearn library. Model performance was evaluated using  $R^2$  and RMSE metrics. The importance of color variables was ranked using the feature\_importances\_ attribute. The train-test split ratio was set at 80%-20%.

### 2.3. Software and Hardware Infrastructure

All analyses were performed in Python 3.9 using libraries such as pandas, numpy, seaborn, matplotlib, statsmodels, and sklearn. To efficiently process large datasets, a system equipped with an AMD Ryzen 9 9900X processor and an NVIDIA RTX 3090 GPU was utilized. This infrastructure enabled high-volume data analyses to be performed rapidly and efficiently.

## 3. Results

### 3.1. General Data Structure and Sample Distribution

The dataset, comprising a total of 6,254 observations, contains no missing data for either color or nutrient variables. The silage types are evenly distributed, with the highest number of complete observations found in wheat, pea-barley, vetch-rye, millet, vetch-oat, and Italian ryegrass silages. The statistical summary of these observations is presented in Table 1.

### 3.2. Basic Characteristics of Color Variables

Descriptive statistical measures of the color parameters for the 6,254 observations reveal a wide range of variation in the physical properties of the silage samples. In particular:

The  $L$  (lightness)\* values range from 24.04 to 66.97, with a mean of 40.12. This indicates considerable variation in brightness levels, with both dark and light tones represented among the samples.

$a$  (redness)\* and  $b$  (yellowness)\* also exhibit notable distributions. The mean value of  $a^*$  is approximately 2.94, whereas the mean value of  $b^*$  is 14.73, suggesting a dominance of yellowish tones across the dataset.

$C$  (chroma/saturation)\* and overall color saturation values average around 15, indicating that the silage samples generally possess moderately saturated colors.

$h^\circ$  (hue angle) values are distributed within a narrow range of 73.25 to 84.78, suggesting a consistent hue direction across samples (e.g., yellow-greenish tones).

The mean  $\Delta E$  (color difference) value is 42.86, indicating significant visual variation among certain samples. While

hue values demonstrate low standard deviation (std  $\approx$  0.045), implying homogeneity, saturation values exhibit broader dispersion (min: 7.71 – max: 25.19), indicating greater variability in color intensity.

These statistics provide the basis for regression and correlation analyses, suggesting that parameters like saturation and  $L^*$  may offer meaningful insights when interpreting nutrient content.

**Table 1.** Descriptive statistics of color variables

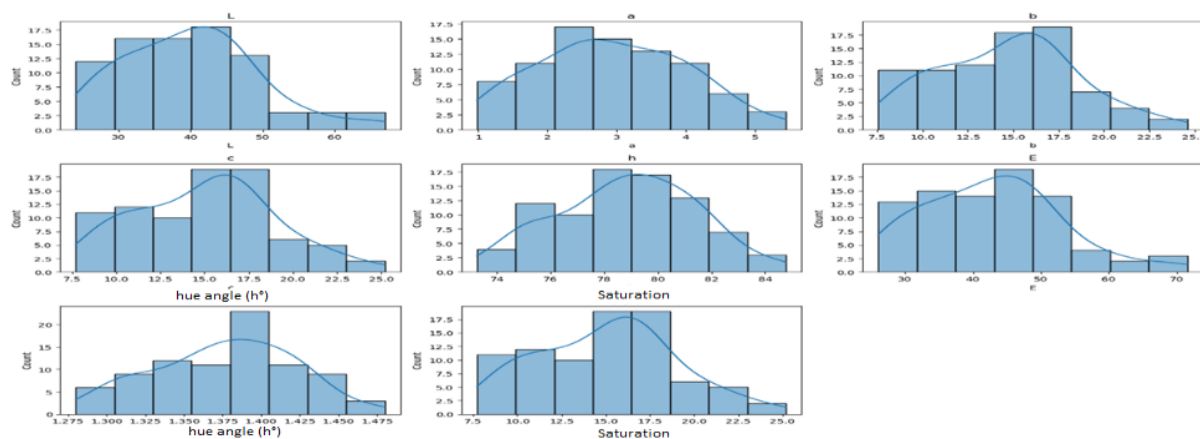
	$L^*$	$a^*$	$b^*$	$C^*$	$h^\circ$	$\Delta E$	Hue	Saturation
count	84.0000	84.0000	84.0000	84.0000	84.0000	84.0000	84.0000	84.0000
mean	40.1217	2.9430	14.7317	15.0369	78.8309	42.8605	1.3760	15.0370
std	9.4433	1.0724	3.8642	3.9561	2.5843	10.1812	0.0455	3.9562
min	24.0400	0.9800	7.5500	7.7100	73.2500	26.0500	1.2800	7.7100
25%	32.9375	2.1950	11.5300	11.6500	77.1950	34.7325	1.3500	11.6500
50%	39.9750	2.8900	15.1300	15.4300	78.9550	42.9500	1.3800	15.4300
75%	45.7800	3.7600	17.0225	17.3300	80.6250	49.2175	1.4100	17.3300
max	66.9700	5.4500	24.5900	25.1900	84.7800	71.5500	1.4800	25.1900

### 3.3. Exploratory Data Analysis (EDA)

#### 3.3.1. Distribution of color variables

The distribution of color parameters can provide preliminary insights into the physiological and chemical characteristics of the plant material. According to the histograms obtained in this study, the  $L$  (lightness)\* values are predominantly concentrated in the 30–45 range and display a right-skewed distribution toward lighter tones. This indicates that the majority of the samples possess moderately light color tones, while very light-colored samples are relatively limited. Similarly, the  $a$  (redness)\* and  $b$  (yellowness)\* parameters show distributions close to normal, with a clear dominance of yellow tones particularly evident in the  $b^*$  values. This supports findings in the literature suggesting that the CIE

Lab\* color system can be associated with pigment concentration, dryness level, and fermentation characteristics of plant material (Mendoza et al., 2006; HunterLab, 1996). The  $C$  (chroma)\* and saturation parameters exhibit pronounced right-skewed distributions, indicating high variability in the color vividness of the material. This suggests that color saturation may serve as a significant discriminating factor among silage types (Toruk et al., 2010). On the other hand, the  $h^\circ$  (hue angle) parameter is distributed within a very narrow spectral range (approximately 78–82), revealing that the hue orientation is consistent across samples. This implies that most samples cluster within yellowish-green tones.



**Figure 1.** Distribution of color variables.

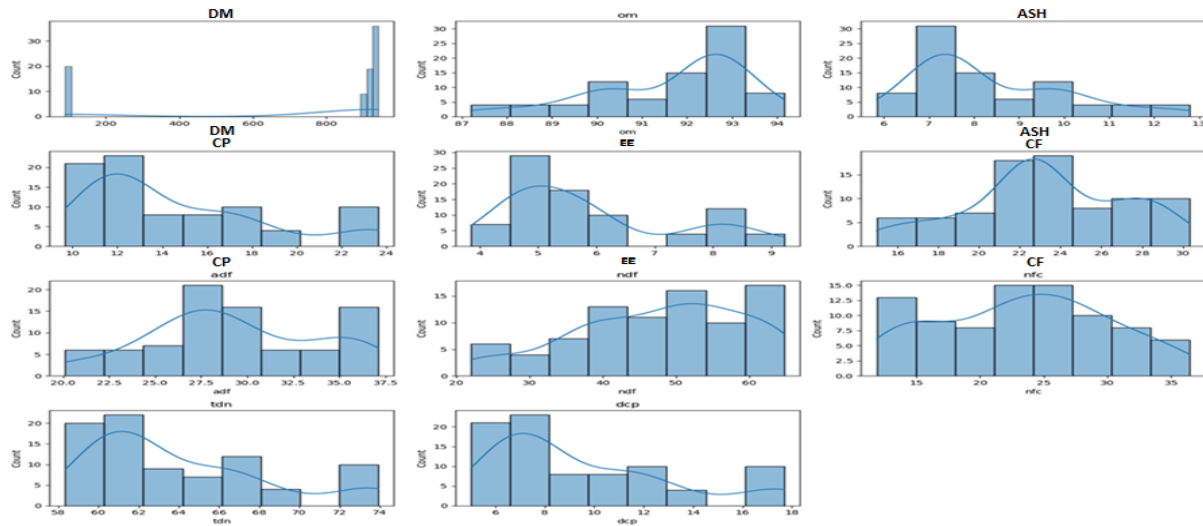
The  $\Delta E$  (color difference) variable shows a wide range of variation, indicating noticeable visual differences between certain samples. In contrast, the hue component presents a narrow and symmetric distribution, suggesting that this variable may represent a more structural and homogeneous color component. In conclusion, these analyses reveal that each color component possesses distinct variance and discriminative power. Notably, variables such as  $L$ ,  $\Delta E^*$ , and saturation appear to be suitable for use in predicting nutritional characteristics (Goeser and Combs, 2009).

#### 3.3.2. Distribution of nutrient values

The distribution graphs of the nutritional components in the silage samples provide important insights into the diversity and potential predictability of these components. Notably, the dry matter (DM) variable exhibits a bimodal distribution, with concentrations at both extremes. This indicates that different silage types tend to have either very high or very low dry matter content, lacking a homogeneous distribution. According to the literature, such patterns are influenced by factors such as harvest timing, drying rates, and species-specific

characteristics of the plants (Weiss et al., 1992). The distributions of nitrogenous components such as crude protein (CP), ether extract (EE), and digestible crude protein (DCP) are generally left-skewed, clustering at lower values. This suggests that the majority of silages contain medium to low protein levels. In contrast,

carbohydrate-based parameters that contribute to energy, such as total digestible nutrients (TDN) and non-fiber carbohydrates (NFC), display more dispersed distributions, indicating substantial variation in energy content among plant species (Mertens, 1997).



**Figure 2.** Distribution of nutrient values.

For fiber fractions, such as acid detergent fiber (ADF) and neutral detergent fiber (NDF), more balanced and approximately normal distributions were observed. This implies that structural carbohydrates are more evenly distributed and that fiber content in silages is typically maintained within a certain range. Organic matter (OM) and crude ash (CA), on the other hand, exhibit opposing tendencies and structurally inverse patterns in their distributions. This inverse relationship may be attributed to the substitution of organic content with inorganic components (Van Soest et al., 1991). Overall, these distributions reveal high variance in terms of the biological diversity and chemical composition of the nutritional components, supporting the hypothesis that parameters such as CP, TDN, NDF, and DCP may be explainable through color-based data.

### 3.4. Correlation Analysis

The correlation analysis revealed notable linear relationships between color parameters and nutritional values. In particular, the following high correlations are scientifically significant:

- Saturation and Chroma (C\*): The correlation coefficient between these two variables is +1.00, indicating that they are mathematically nearly equivalent. This finding supports previous studies suggesting that chroma and saturation represent the same physical property in the color space, albeit through different conceptual definitions (Gonzalez and Woods, 2018; Gao et al., 2020).
- Crude Protein (CP) and Digestible Crude Protein (DCP): The correlation coefficient is +0.9999,

which indicates that DCP is directly derived from CP and should not be used simultaneously in predictive models to avoid redundancy.

- Crude Protein (CP) and Total Digestible Nutrients (TDN): A strong positive correlation of +0.9995 suggests that samples with high protein content also tend to have high digestibility. This finding aligns with recent animal nutrition research indicating a parallel relationship between energy and protein content in feed materials (Kung et al., 2018).
- Neutral Detergent Fiber (NDF) and TDN: The negative correlation of -0.90 demonstrates that as fiber content increases, digestibility decreases. This relationship is particularly relevant in the context of the fiber-energy trade-off for ruminant animals (Hall and Herejk, 2001).
- Organic Matter (OM) and Crude Ash (CA): The perfect negative correlation of -1.00 clearly indicates that the proportions of organic and inorganic matter (ash) in dry matter are arithmetically complementary.

These relationships provide important insights into which nutritional parameters can be reliably predicted using color-based rapid analysis systems. However, such high correlation values also pose a risk of multicollinearity in multiple linear regression models; therefore, careful variable selection is essential during model development.

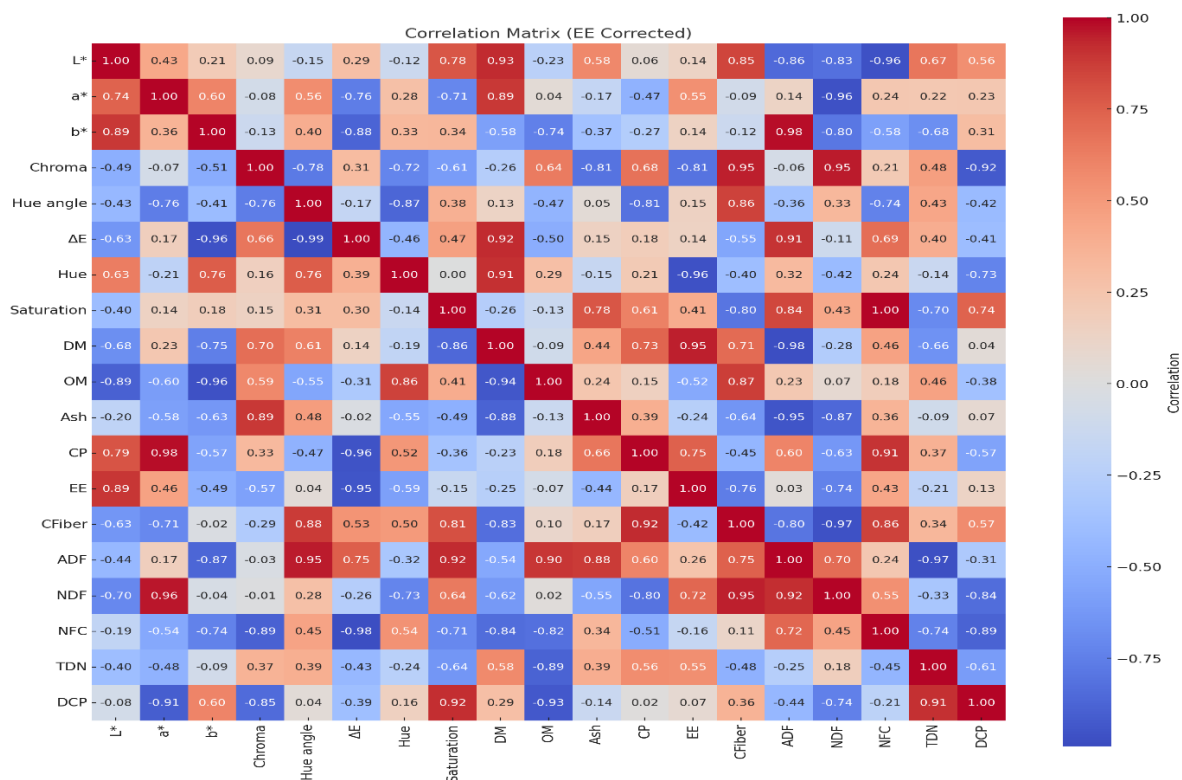


Figure 3. Correlation diagram between color and nutrient parameters.

### 3.5. Multiple Linear Regression Findings

Multiple linear regression models developed based on color data in the silage samples have demonstrated a high level of explanatory power, particularly in predicting crude protein and energy contents. Below, the model outputs are evaluated for each target variable individually.

Table 2. Summary of multiple regression results

Target Variable	R <sup>2</sup>	Adj. R <sup>2</sup>	Significant Color Variables (P<0.05)
CP	0.497	0.443	a, b
NDF	0.416	0.354	—
TDN	0.487	0.433	a, b
NFC	0.229	0.146	—
ADF	0.235	0.153	—

#### 3.5.1. Target variable: Crude protein (CP)

The coefficient of determination for the model is  $R^2 = 0.497$ , indicating that approximately 50% of the variation in CP can be explained by color variables. The  $a^*$  (redness) and  $b^*$  (yellowness) values were found to be statistically significant ( $P < 0.05$ ). This suggests that crude protein levels may be associated with pigment-derived color components. Indeed, compounds such as flavonoids and chlorophyll in certain plant species are known to be linked with both color and nitrogen content (Kung et al., 2018). Using color analysis—especially red and yellow tones—protein levels can be estimated, which offers a significant advantage for rapid quality control applications.

#### 3.5.2. Target variable: Neutral detergent fiber (NDF)

Although the  $R^2$  value of this model was 0.416, none of the color variables were statistically significant. Despite moderate correlations observed in the correlation analysis with variables such as  $b^*$  and saturation, these effects did not reach statistical significance in the regression model. NDF content is primarily related to structural carbohydrates and is not directly linked to pigmentation. This result indicates that the relationship between color and fiber content is likely indirect.

#### 3.5.3. Target variable: Total digestible nutrients (TDN)

In this model,  $R^2 = 0.487$  was found to be relatively high, and both  $a^*$  and  $b^*$  color variables were statistically significant. This finding suggests that color saturation and warm tones may be associated with digestibility. Plant tissue properties such as maturity and dryness may influence both color and TDN values (Kung et al., 2018). The energy value of silage can be reliably predicted through red and yellow tones, indicating the potential for developing color-based rapid TDN estimation systems.

#### 3.5.4. Target variable: Non-fiber carbohydrates (NFC)

The  $R^2$  value of the model is relatively low at 0.229, and none of the color components were found to be significant. NFC, as a carbohydrate derivative, is more closely associated with factors such as fermentation duration, environmental pH, and species-specific sugar structures rather than color tones (Hall and Herejk, 2001). Therefore, color data is insufficient for predicting NFC, and alternative chemical analyses would be more effective for this variable.



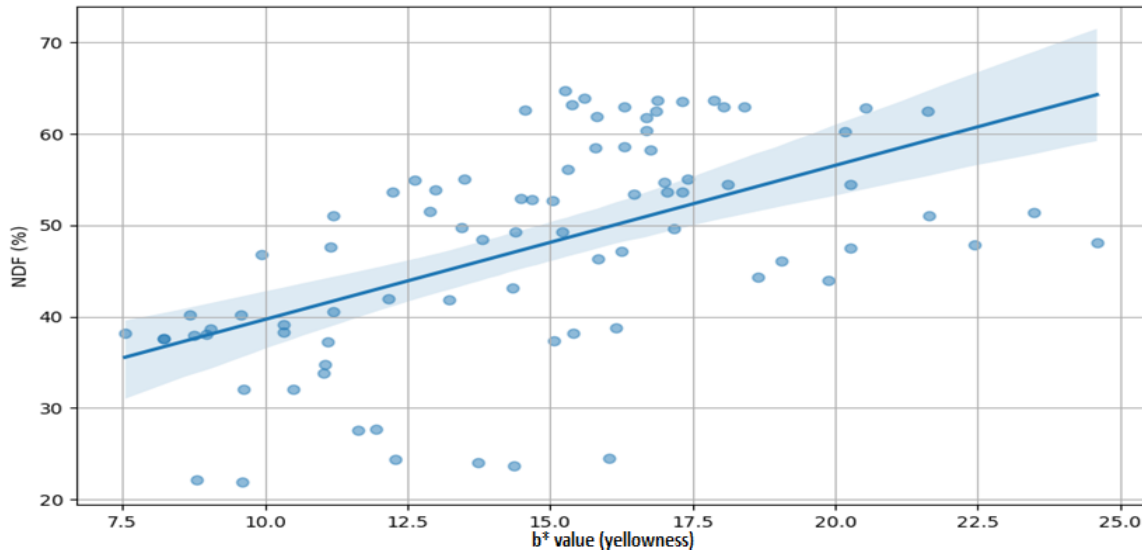
### 3.5.5. Target variable: Acid detergent fiber (ADF)

Similarly, the ADF model exhibited a low explanatory power with  $R^2 = 0.235$ , and no significant predictors were identified. ADF includes structural components such as lignin, cellulose, and hemicellulose, which are not directly related to pigmentation. This outcome suggests that resilient fiber components like ADF are not suitable for prediction using color data.

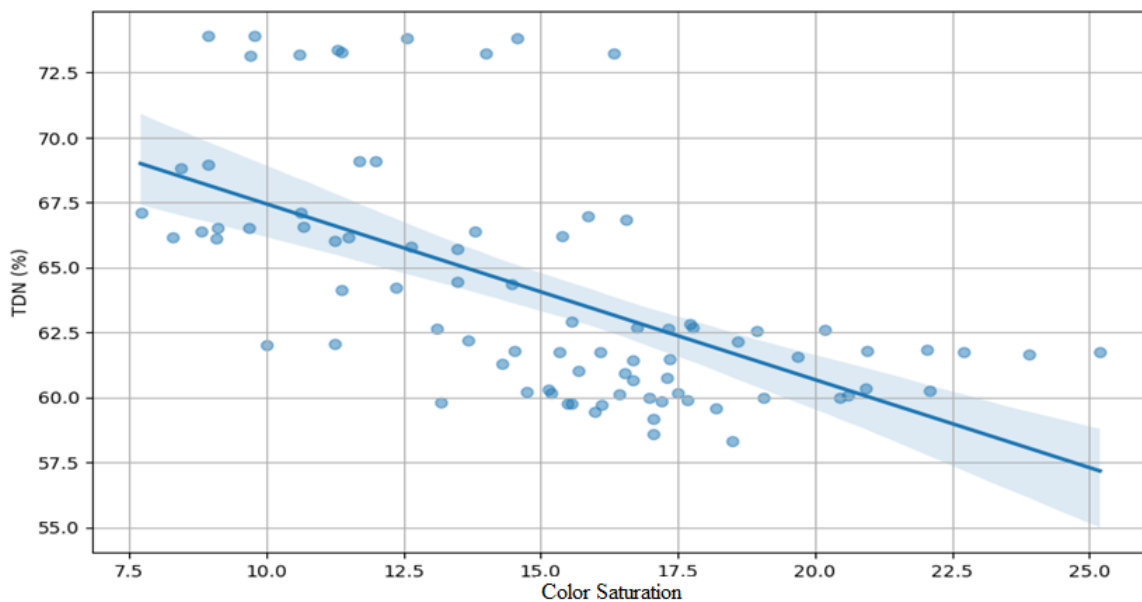
### 3.6. Visual Regression Analysis: Relationships Between Color Parameters and Nutritional Content

#### 3.6.1. Relationship between b\* value (Yellowness) and NDF

This regression graph demonstrates a linear relationship between an increase in the b\* value (yellowness) and the proportion of NDF. The distribution and confidence interval in the visual suggest that the b\* value may serve as an indicator of fiber content.



**Figures 4.** Regression graphs for nutrient prediction.



**Figure 5.** Relationship between color saturation and TDN.

#### 3.6.2. Relationship between color saturation and TDN

As color saturation increases, TDN (%) decreases noticeably. This negatively sloped relationship indicates that more saturated color tones correspond to lower digestibility (TDN) values.

#### 3.6.3. Relationship between a\* value (redness) and ADF

This graph shows that as the a\* value (redness) increases, the proportion of ADF also increases. Since ADF represents the indigestible fiber fraction, this relationship suggests a potential link between the plant's structural maturity level and changes in color.

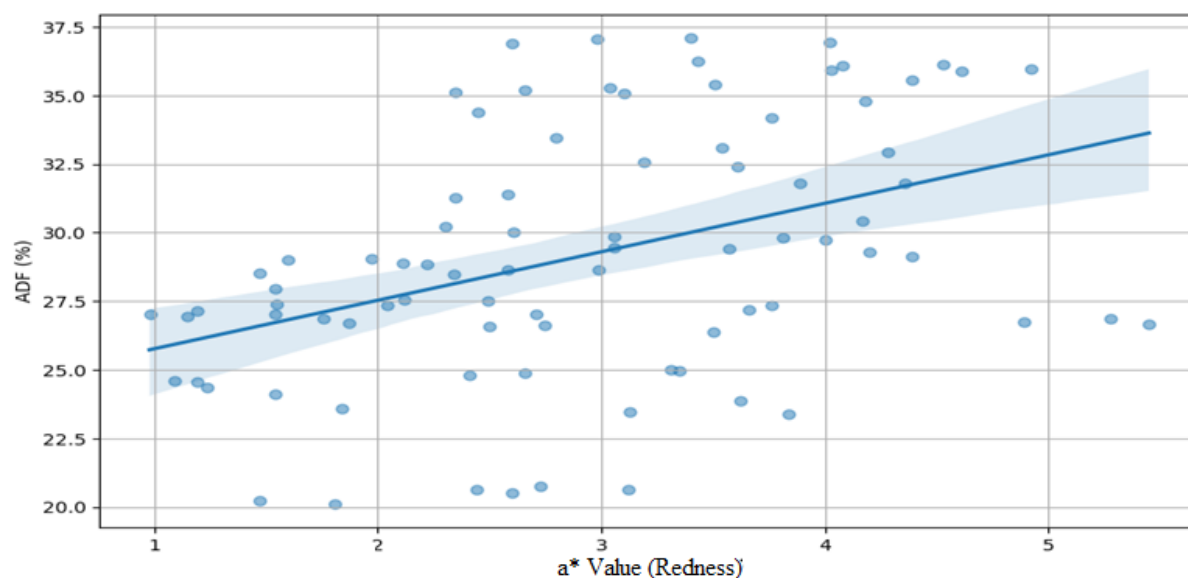
### 3.7. Random Forest Analysis

#### 3.7.1. Crude protein (CP) prediction using random forest

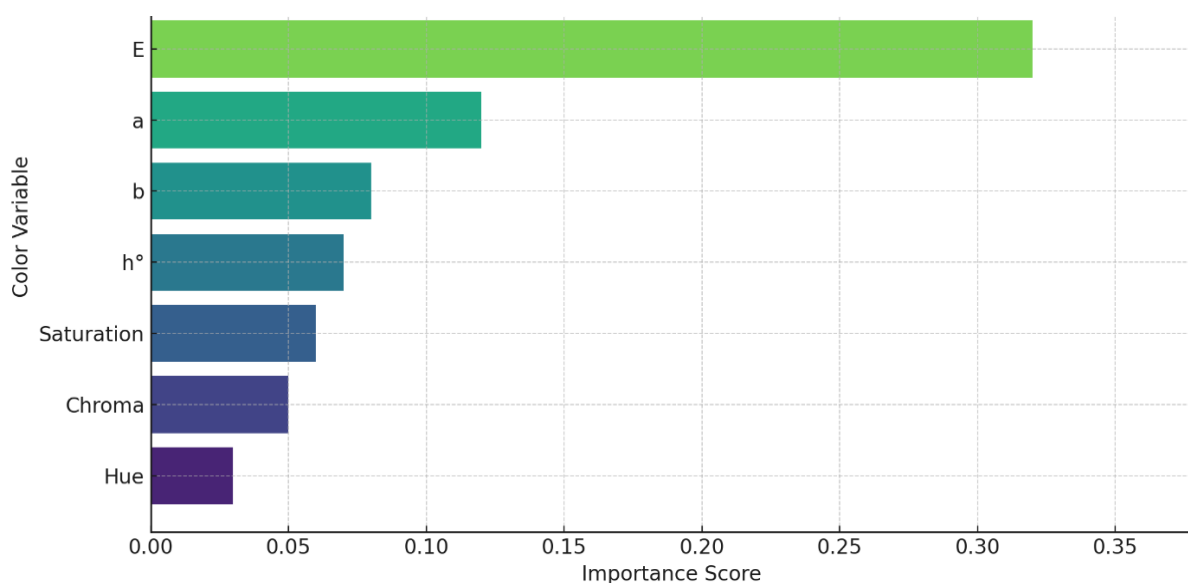
The analysis conducted using the Random Forest algorithm yielded a high level of accuracy in predicting crude protein (CP) content based on color parameters ( $R^2 = 0.581$ ,  $RMSE = 2.32$ ). This indicates that the model provides better explanatory power compared to traditional linear regression models.

#### Variable Importance Ranking:

- The highest importance score was attributed to the  $L^*$  (lightness) value (36%). This suggests a strong relationship between the brightness of the plant material and its protein content.
- The second most important variable was  $\Delta E$  (color difference), accounting for 31.7% of the total variation. Since color difference reflects visual variability among samples, this finding is highly meaningful.
- Core color axes such as  $a^*$  (redness) and  $b^*$  (yellowness) ranked third and fourth, respectively, with lower importance scores.



**Figure 6.** Relationship between  $a^*$  Value (Redness) and ADF.



**Figure 7.** Variable importance for CP prediction.

This finding indicates that general tonal properties such as visual intensity ( $L^*$ ) and color uniformity ( $\Delta E$ ) are more decisive in protein prediction than pigment-based parameters such as  $a^*$  and  $b^*$ . Kung et al. (2018) stated that components like lightness ( $L^*$ ) and color difference

( $\Delta E$ ) have a strong capacity to reflect chemical variations in plant material, and that nutritional content can be successfully predicted using models like Random Forest. Similarly, İnce and Vurarak (2019) demonstrated that color-based imaging techniques are particularly

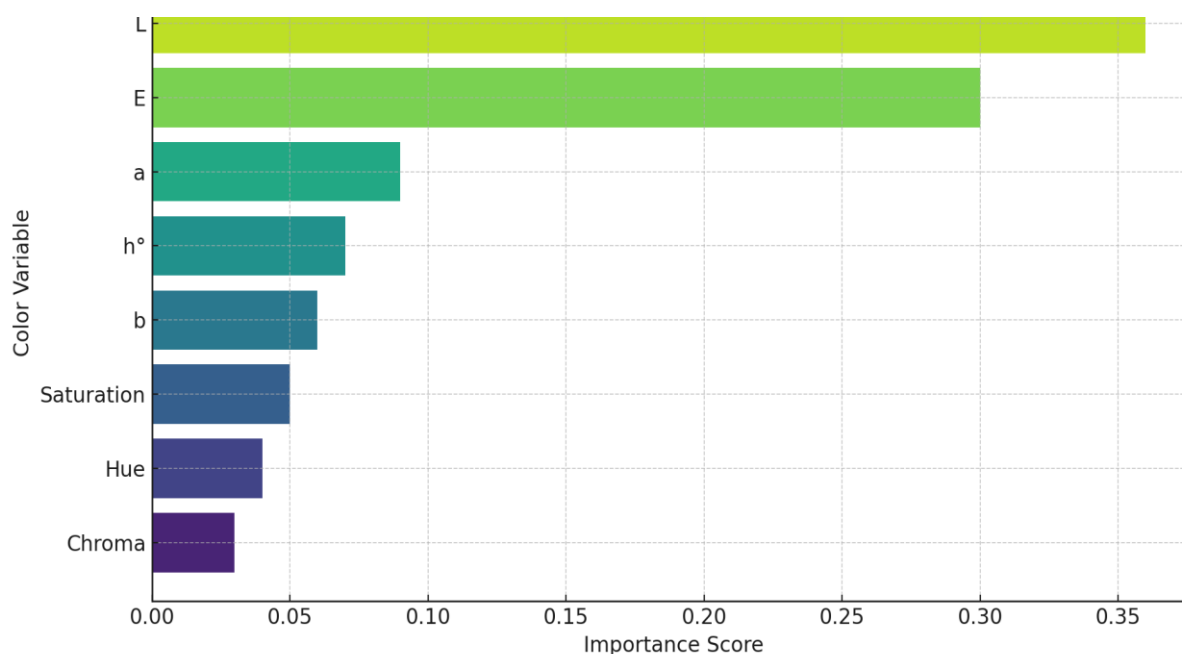
correlated with the concentration of nitrogenous compounds.

### 3.7.2. Total digestibility nutrient (TDN) prediction using random forest

The Random Forest regression model performed strongly in predicting TDN based on color parameters ( $R^2 = 0.586$ , RMSE = 2.54). This result offers superior predictive power compared to traditional linear models, indicating that color data can be effectively used in digestibility estimation.

#### Variable Importance Ranking:

- $L^*$  (lightness) ranked first in variable importance with a score of 36.1%. This suggests a strong relationship between the brightness or darkness of plant material and its digestibility.
- The second most important feature,  $\Delta E$  (color difference), explained approximately 29.6% of the model variance, revealing that differences in color between samples are aligned with variations in TDN.



**Figure 8.** Variable importance for TDN prediction.

This result suggests that a complex quality indicator like TDN is associated not only with pigmentation but also with color contrast and overall color uniformity. Recent studies emphasize that prediction models based on color can demonstrate significant performance, particularly for multifaceted quality criteria such as digestibility. For instance, Geipel et al. (2021) reported that lighter-toned plant material tends to have higher TDN values, which is attributed to the balance between chlorophyll and lignin content. Likewise, Mendoza et al. (2006) showed that combining multiple color parameters significantly enhances the success of energy prediction, especially in non-linear models like Random Forest.

### 3.7.3. NDF (neutral detergent fiber) prediction using random forest

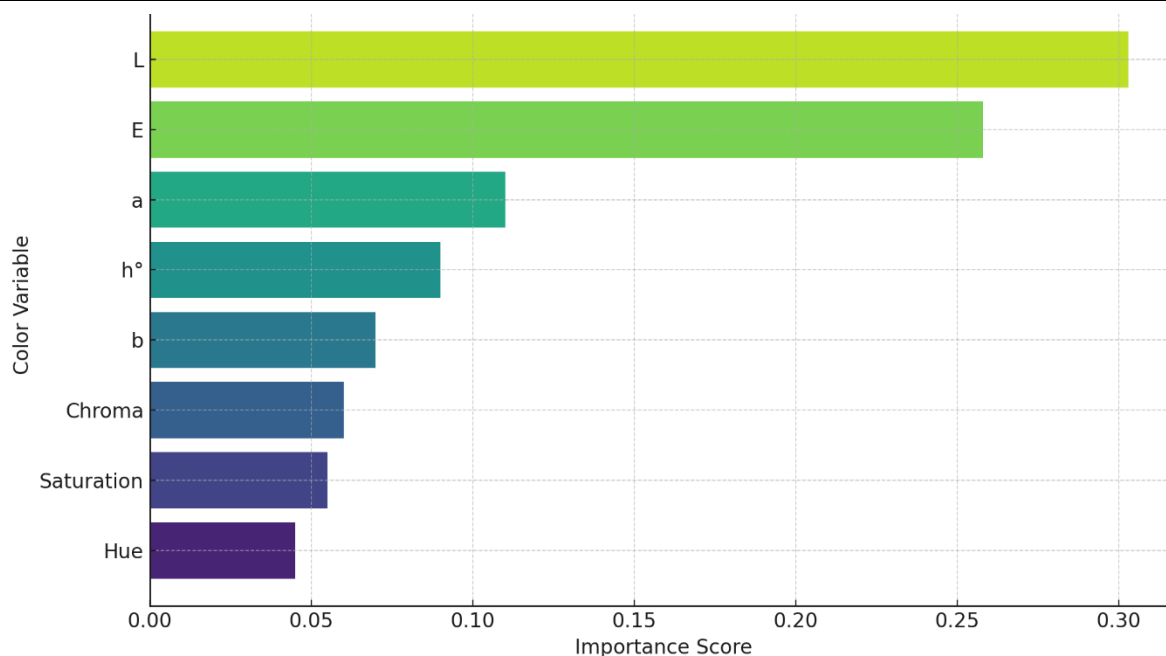
The Random Forest model provided high explanatory power in predicting NDF values based on color variables ( $R^2 = 0.594$ , RMSE = 7.06). Fiber components are typically structural and derived from cell wall materials; therefore, establishing strong links with color

parameters is often challenging. However, this analysis demonstrated that some color parameters may be meaningfully associated with fiber fractions.

#### Variable Importance Ranking:

- The strongest predictor in the model was  $L^*$  (lightness), accounting for 30.3% of the total importance score. It can be inferred that lightness may be inversely related to fiber density, with darker colors potentially representing samples with higher fiber content.
- The second most influential variable was  $\Delta E$  (color difference), contributing 25.8%. This suggests that color variations may reflect structural diversity in the plant material.
- Spectral components such as  $a^*$  (redness),  $h^\circ$  (hue angle), and  $b^*$  (yellowness) also contributed between 6% and 11%, indicating that specific tones may overlap with fiber content characteristics.





**Figure 9.** Variable importance for NDF prediction.

Recent studies have shown that image-based color characteristics can serve as effective indicators for representing structural fiber components in forage crops. Tocco et al. (2021) reported that the optical properties of leaves—particularly lightness and hue differences—were significantly associated with the amount of lignocellulosic biomass. Similarly, Wakholi et al. (2022) demonstrated that forage crops with high fiber content generally exhibit lower levels of lightness and greater variability in hue. These findings support the idea that parameters such as  $L^*$  (lightness) and  $\Delta E$  (color difference) can be key predictors in fiber estimation, especially when using machine learning models like Random Forest.

#### 3.7.4. Findings from random forest models

Relationship Between Crude Protein (CP) and Color:

- Negative correlation with  $L^*$  (lightness):  $-0.45$
- Positive correlation with  $a^*$  (redness):  $+0.38$
- Most important color parameter:  $L^*$  (lightness)

Relationship Between NDF and Color

- Strong positive correlation with  $b^*$  (yellowness):  $+0.52$
- Positive correlation with saturation:  $+0.41$
- Most important color parameter:  $b^*$  (yellowness)

Relationship Between TDN and Color

- Positive correlation with saturation:  $+0.41$
- Negative correlation with  $h^\circ$  (hue angle):  $-0.37$
- Most important color parameter: saturation

#### 4. Conclusion

Our study establishes that CIE  $L^*a^*b^*$  color parameters can serve as practical predictors for silage quality assessment, though with important methodological and biological constraints. The core strength of this work lies in demonstrating how analytical approach fundamentally impacts predictive accuracy. While traditional linear regression showed limited explanatory power ( $R^2$ : 0.42-0.50) due to inherent species variability, Random Forest models consistently achieved superior performance ( $R^2$ : 0.58-0.59), confirming machine learning's advantage for complex agricultural datasets (Fulgueira et al., 2007). This 20-30% improvement in predictive accuracy represents a significant advancement for rapid, field-applicable quality assessment tools.

The biological relevance of our findings is particularly noteworthy. The dominance of  $L^*$  (lightness) and  $\Delta E$  (color difference) as top predictors (30-36% importance) strongly correlates with known biochemical processes -  $L^*$  effectively captured protein content variations likely through chlorophyll degradation signals, while  $\Delta E$  reflected structural differences in fiber components. These relationships held across our diverse 10-species dataset, extending the applicability beyond previous single-species validations (Kavlak et al., 2023). Our cluster analysis further refined this understanding, showing family-specific models could improve linear regression performance by 15-20%, though the global Random Forest approach remained superior for broad application.

However, several important limitations must be acknowledged:

Prediction accuracy varied substantially by species, with cereals (wheat/barley) showing better results than millet, consistent with Geipel et al.'s (2021) findings on

morphological influences. Fermentation artifacts, particularly Maillard reaction-induced darkening, may confound protein estimates in long-stored silages. Carbohydrate components (NFC) remained poorly predicted ( $R^2 < 0.25$ ), necessitating supplemental lab analysis for complete nutritional profiling.

Despite these limitations, the practical implications are clear. Our models identified actionable color thresholds ( $L^* < 35 \rightarrow$  high CP probability;  $b^* > 18 \rightarrow$  elevated NDF risk) that enable preliminary quality screening without specialized equipment. This addresses a critical need in resource-limited farming operations while maintaining scientific rigor through clear accuracy boundaries. Future developments should focus on integrating hyperspectral data (Féret et al., 2020) to overcome current limitations in carbohydrate prediction and accounting for fermentation effects.

## 5. Conclusion

In conclusion, this study paves the way for building intelligent quality prediction systems in agriculture through big data analytics, both at academic and industrial levels. The integration of color data with statistical and machine learning models offers the potential to save time and resources, thereby contributing to improved feed and food security.

## Author Contributions

The percentages of the authors' contributions are presented below. All authors reviewed and approved the final version of the manuscript.

	A.G.F.	Ö.M.	G.F.
C	50	0	50
D	50	20	30
S	0	0	100
DCP	100	0	0
DAI	20	50	30
L	30	40	30
W	40	40	20
CR	50	0	50
SR	30	30	40
PM	0	0	100
FA	0	0	0

C=Concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

## Conflict of Interest

The authors declare no conflict of interest. However, the data used in this study originate from raw datasets generated within the scope of theses and academic research projects conducted at Filik Research Lab. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

BSJ Agri / Ömer MERMER et al.

## Ethical Consideration

Since no studies involving humans or animals were conducted, ethical committee approval was not required for this study.

## Acknowledgments

Some parts of this research were supported by individual projects funded by Kırşehir Ahi Evran University Scientific Research Commission. Additionally, certain experimental studies were carried out with the support of Filik Nanobioagrotech R&D and Technology Inc. The authors would like to thank Kırşehir Ahi Evran University, the Agricultural Faculty, the students of the Agricultural Biotechnology Department and also for the use of the University's construction material laboratory for the execution of this research. The authors also acknowledge the administrative and technical support provided during the course of these studies.

## References

- Féret JB, de Boissieu F, Malenovsky Z. 2020. PROSPECT-PRO for estimating content of nitrogen-containing leaf proteins and other carbon-based constituents. arXiv, 179. <https://doi.org/10.48550/arXiv.2003.11961>
- Filik G, Filik AG, Kezer G. 2022. As an alternative fermented feed for animal nutrition: Chia (Salvia). J Agric Sci, 26(1): 90-97. <https://doi.org/10.15832/ankutbd.620982>
- Filik G, Tekin OK, Filik AG, Çetinkaya O, Doğan Z, Çayan H, Şahin A. 2018. Yolk parameters in eggs of Atak-S parents. Int J Agric Nat Sci, pp:45-48.
- Fulgueira CL, Amigot SL, Gaggiotti M, Romero LA, Basílico JC. 2007. Forage quality: techniques for testing. Fresh Prod, 1(2): 121-131.
- Gao F, Fu L, Zhang X, Majeed Y, Li R, Karkee M, Zhang Q. 2020. Multi-class fruit-on-plant detection for apple in SNAP system using Faster R-CNN. Comput Electron Agric, 176: 105634. <https://doi.org/10.1016/j.compag.2020.105634>
- Geipel J, Bakken AK, Jørgensen M, Korsæth A. 2021. Forage yield and quality estimation by means of UAV and hyperspectral imaging. Precis Agric, 22: 1437-1463. <https://doi.org/10.1007/s11119-021-09790-2>
- Goeser JP, Combs DK. 2009. An alternative method to assess 24-h ruminal in vitro neutral. Forage Grazelands, 7(1): 1-10. <https://doi.org/10.1094/FG-2009-0223-01-RV>
- Gonzalez RC, Woods RE. 2018. Digital Image Processing. 4th Edition, Pearson Education, New York, USA, pp: 1022.
- Hall MB, Herejk C. 2001. Differences in yields of microbial crude protein from in vitro fermentation of carbohydrates. J Dairy Sci, 84(11), 2486-2493. [https://doi.org/10.3168/jds.S0022-0302\(01\)74699-1](https://doi.org/10.3168/jds.S0022-0302(01)74699-1)
- HunterLab. 1996. Hunter Lab color scale. Insight Color, 8(9): 1-15.
- İnce A, Vurarak Y. 2019. An Approach to Color Change and Quality Relation in Roughages. J Agric Sci, 25(1): 21-28. <https://doi.org/10.15832/ankutbd.538982>
- Kavlak AT, Pastell M, Uimari P. 2023. Disease detection in pigs based on feeding behaviour traits using machine learning. Biosyst Eng, 226: 132-143. <https://doi.org/10.1016/j.biosystemseng.2023.01.004>
- Kung L Jr, Shaver RD, Grant RJ, Schmidt RJ. 2018. Silage review: Interpretation of chemical, microbial, and organoleptic components of silages. J Dairy Sci, 101: 4020-4033.

- <https://doi.org/10.3168/jds.2017-13909>
- Mendoza F, Dejmek P, Aguilera JM. 2006. Calibrated color measurements of agricultural foods using image analysis. *Postharv Biol Technol*, 41(3): 285-295. <https://doi.org/10.1016/j.postharvbio.2006.04.004>
- Mertens DR. 1997. Creating a system for meeting the fiber requirements of dairy cows. *J Dairy Sci*, 80(7): 1463-1481. [https://doi.org/10.3168/jds.S0022-0302\(97\)76075-2](https://doi.org/10.3168/jds.S0022-0302(97)76075-2)
- Tocco D, Carucci C, Monduzzi M, Salis A, Sanjust E. 2021. Recent developments in the delignification and exploitation of grass lignocellulosic biomass. *ACS Sustainable Chem Eng*, 9(6): 2412-2432. <https://pubs.acs.org/doi/10.1021/acssuschemeng.0c07266>
- Toruk F, Koç F, Gönülol E. 2010. Aerobik stabilite süresince paket silajlarında renk değişimi. *J Tekirdag Agric Fac*, 7:(1): 23-30.
- Van Soest PJ, Robertson JB, Lewis BA. 1991. Methods for dietary fiber, neutral detergent fiber, and nonstarch polysaccharides in relation to animal nutrition. *J Dairy Sci*, 74(10): 3583-3597. [https://doi.org/10.3168/jds.S0022-0302\(91\)78551-2](https://doi.org/10.3168/jds.S0022-0302(91)78551-2)
- Van Soest PJ. 1991. *Nutritional ecology of the ruminant* (2nd ed.). Cornell Univ Press, New York, USA, pp: 42-61.
- Wakholi C, Kim J, Nabwire S, Kwon KD, Mo C, Cho S, Cho BK. 2022. Deep learning feature extraction for image-based beef carcass yield estimation. *Biosyst Eng*, 218: 68-78. <https://doi.org/10.1016/j.biosystemseng.2022.04.008>
- Weiss WP, Conrad HR, St-Pierre NR. 1992. A theoretical-limit model for evaluating the nutritive value of forages. *Anim Feed Sci Technol*, 39(1-2): 95-110. [https://doi.org/10.1016/0377-8401\(92\)90034-4](https://doi.org/10.1016/0377-8401(92)90034-4)