Hungarian University of Agriculture and Life Sciences

Doctoral School of Plant Science

Ph.D. Dissertation

Integrative Approaches in Enhancing Processing Tomato (*Solanum lycopersicum L.*) Cultivation: From Root Development and Machine Learning Predictions to Genetic Diversity Evaluation

DOI: 10.54598/004840

By

Oussama M'hamdi

Gödöllő, Hungary

2024

The Doctoral School: Doctoral School of Plant Sciences

Name: Oussama M'hamdi

Discipline: Agricultural Sciences

Head: *Professor Dr. Lajos Helyes*
Head of Doctoral schools of the Hungarian University of Agricultural and Life Sciences
Director of Institute of Horticultural Sciences
Hungarian University of Agriculture and Life Sciences

Supervisors: *Professor Dr. Zoltán Pek*
Institute of Horticultural Sciences
Hungarian University of Agriculture and Life Sciences

Dr. Sándor Takács
Institute of Horticultural Sciences
Hungarian University of Agriculture and Life Sciences

……

Approval of the Head of Doctoral School

…… ……

Approval of the Supervisor Approval of the Supervisor

1 INTRODUCTION AND OBJECTIVES

Tomatoes (Solanum lycopersicum L.) are integral to global diets and play a crucial role in the agricultural economy, with extensive cultivation due to their nutritional value and economic importance (Agbemafle et al., 2014). In 2021, global tomato production reached approximately 186 million metric tons across 4.9 million hectares, contributing to an industry valued at around 181.74 billion US dollars (FAO, 2022). This scale of production underscores the tomato's prominence in both fresh and processed forms across diverse cultures and cuisines, where it serves as a fundamental ingredient in various dishes such as soups, sauces, and juices (Bergougnoux, 2014; J. Liu et al., 2021). In addition to its culinary versatility, the tomato is recognized for its rich nutritional profile, being a source of essential vitamins, minerals, and bioactive compounds like lycopene, which has been linked to a reduced risk of prostate cancer and cardiovascular diseases (Giovannucci, 1999; Burton-Freeman & Sesso, 2014). The integration of advanced agricultural technologies and sustainable farming practices has been instrumental in increasing yields and improving the quality of tomatoes, further solidifying their position as a staple crop worldwide (Nemeskéri et al., 2019; X. Wang et al., 2019). Research also highlights the significant health benefits of tomatoes, including their contribution to immune function, blood pressure regulation, and cognitive health, with studies showing their potential in mitigating the risk of neurodegenerative diseases (Slavin & Lloyd, 2012; Meeusen, 2014). Additionally, the processing of tomatoes can enhance the bioavailability of certain nutrients, particularly lycopene, making processed tomato products like sauces and juices beneficial dietary components (Basu & Imrhan, 2007).

The cultivation of processing tomatoes, faces numerous challenges. These challenges encompass a broad array of environmental stressors, notably the variability in water availability, such fluctuations in water supply significantly affect the quality and yield of tomatoes, impacting each phenological stage of the plant's growth in distinct ways (Nemeskéri & Helyes, 2019; Takács et al., 2020). Adding to these challenges is the necessity to consistently adhere to stringent fruit quality standards, a task that becomes increasingly challenging under diverse climatic conditions (Giuliani et al., 2019). Furthermore, the ripening process of tomatoes is yet another critical aspect, marked by significant alterations in metabolic pathways (Zhu et al., 2022). These changes are crucial in determining the fruit's external appearance, internal quality parameters such as Brix value and Lycopene content, and the distinctive colour indicative of ripeness. Therefore, predicting quality in diverse climatic conditions becomes paramount, as uniformity in quality and appearance is essential for consumer acceptance. Additionally, the genetic diversity inherent in different tomato genotypes presents both challenges and opportunities. While this diversity

demands adaptation in cultivation practices for each genotype, it also provides significant opportunities for agricultural advancements and the breeding of varieties more resilient to specific environmental stressors (Udriște et al., 2022). Therefore, the cultivation of processing tomatoes encompasses a wide spectrum of scientific and practical considerations, each playing a crucial role in ensuring the sustained production and availability of this globally important crop.

Current research reveals a significant gap in comprehending the full impact of environmental factors on tomato root development and fruit quality. The potential of utilizing advanced predictive tools, such as machine learning, to assess fruit quality in relation to varying environmental and genetic influences remains largely unexplored. This thesis explores the complex relationship between environmental conditions and root development, delves into the application of machine learning techniques for predicting fruit quality attributes, and examines the interaction between genetic composition and environmental factors.

Objectives to achieve

The primary aim of this research is to comprehensively understand the growth, development, and quality of processing tomato plants in response to various environmental conditions and genetic factors. Specifically, this study seeks to achieve the following objectives:

To Investigate Root Development in Processing Tomato Plants under Different Water Supply Levels

Investigate the impact of differential water supply on the root system architecture of processing tomato plants (Solanum lycopersicum). This includes monitoring changes in root count, length, and overall development using non-destructive methods. The goal is to understand how water stress or abundance affects root growth patterns, potentially impacting overall plant health and yield

To Perform a Comparative Analysis of Machine Learning Models in Predicting Tomato Fruit Quality

Utilize two advanced machine learning techniques, eXtreme Gradient Boosting (XGBoost) and Artificial Neural Network (ANN), to predict key quality attributes of processing tomato fruits. These attributes include Brix, Lycopene content, and a/b ratio. The data for this analysis comprises variables like different cultivars, planting locations, years, and climatic factors, aiming to establish robust predictive models for fruit quality assessment.

To Evaluate the Genetic Resources of Processing Tomato Plants in Diverse Environments

Conduct a comprehensive assessment of the genetic diversity among different processing tomato genotypes and their response to various environmental conditions using GGE biplot analysis.

This objective focuses on understanding how genetic variation influences important quality traits like Brix and Lycopene content across different years and locations, providing insights into genotype-environment interactions.

2 MATERIALS AND METHODS

2.1 Part 1: Root Development Monitoring under Different Water Supply Levels

2.1.1 Plant Material and Experimental Set-Up

The experiment was conducted at the Horticultural Experimental Farm of the Hungarian University of Agriculture and Life Sciences in Gödöllő, Hungary, using the processing tomato hybrid H1015, a determinate variety commonly grown in the region. Plants were spaced 140 cm between rows and 20 cm between plants, resulting in a density of 3.57 plants per square meter. Seedlings were transplanted on 14 May 2020 and 15 May 2021. Three irrigation treatments were applied: 100% of crop evapotranspiration (I100), 50% of I100 (I50), and a non-irrigated control, with each treatment replicated three times. In 2020, irrigation began on 8 June, continued on 29 June, and ended on 3 August. In 2021, it started on 3 June and ended on 2 August.

2.1.2 Image Acquisition

Images of roots were captured using a CI-600 In-Situ Root Imager (CID Bio-Science, USA), which uses a rotating scanner inside a transparent tube inserted into the soil. The images, taken at 300 dpi resolution, were collected at three depths: 10–30 cm, 30–50 cm, and 50–70 cm. Photos were taken weekly from June 25 to August 18, 2020, and from June 9 to July 14, 2021, for 3 plants per irrigation treatment. Scanner tubes were installed near randomly selected plants after transplanting. Root mapping was done manually using RootSnap 1.4 software, and the data were exported to a spreadsheet.

2.1.3 Relative Chlorophyll Content and Photosynthetic Activity

Measurements were conducted on randomly selected plants within each treatment at around 12:00 on each measurement date. The SPAD index was measured using a SPAD 502 chlorophyll meter. A PAM 2500 fluorometer device was used to measure chlorophyll fluorescence. Data were acquired from the device with PamWin-4 4.01 software. In total, 16 measurements per treatment were taken, which included 4 measurements per treatment in each repeated block. All measurements were carried out non-destructively on healthy, fully developed leaves.

2.2 Part 2: A Comparative Analysis of XGBoost and Neural Network Models for Predicting Tomato Fruit Quality

2.2.1 Dataset Description

A comprehensive dataset was utilized encompassing physicochemical characteristics and environmental factors across a diverse selection of tomato cultivars over five consecutive growing seasons from 2017 to 2021. The dataset included observations of 48 cultivars and 28 locations (Loc) within Hungary. The number of cultivars and locations varied each year, with measurements taken after harvest to assess key quality traits: Brix, lycopene content, and fruit colour (a/b ratio). In total, 28,474 measurements were recorded for these three main variables.

To understand the impact of meteorological factors on tomato cultivation, data was collected from the Operational Drought and Water Scarcity Management System in Hungary over multiple growing seasons (30 May to 30 August). Key factors included the number of days with optimal temperatures (21°C to 27°C), total precipitation, rainy days, and average relative humidity. Days with ideal humidity levels (40% to 70%) and instances of high humidity (over 90%) were also tracked. Additionally, soil types at each location were classified based on the USDA system.

2.2.2 Measurement of tomato quality traits

Tomato physicochemical properties were assessed using advanced automated stations. Brix was measured by the Maselli SV01 system, which processed the tomatoes into juice and performed automatic refractometric analysis, with a range of 0 to 10 Brix and accuracy of ±0.15, following the nD/Bx ICUMSA standard. Lycopene content was determined through automated spectrophotometry, with a range of 0 to 80 mg/100 g and accuracy of ±0.5 mg/100 g. Fruit color was also measured spectrophotometrically, using colorimetric coordinates (L, a, b) to derive the chromaticity ratio, ensuring accurate colour balance evaluation.

2.2.3 Data Preprocessing

The dataset underwent several preprocessing steps to ensure quality and support analysis. Categorical attributes like 'Loc', 'Cultivar', and 'SoilTyp' were one-hot encoded for compatibility with machine learning algorithms (Goodfellow et al., 2016; Lecun et al., 2015). Missing values in numerical columns were imputed with the mean, while categorical columns were filled with the mode, preserving data distribution. After cleanup, variable relationships were explored through a correlation matrix visualized with a seaborn heatmap.

2.2.4 Machine learning models

2.2.4.1 XGBoost Model

The XGBoost (eXtreme Gradient Boosting) model, known for handling missing values and evaluating feature importance, was used for prediction (Chen & Guestrin, 2016). Lag features (1 to 3-time steps) and a rolling mean (3-time window) were engineered for 'Predicted Variable' columns (Brix, Lycopene, a/b ratio) to capture temporal patterns (Box et al., 2015). The dataset was split using a 5-fold Time Series Split to maintain temporal sequence integrity. Features were standardized using StandardScaler, and the XGBoost Regression model was optimized via grid search and 3-fold cross-validation. Model performance was evaluated using R-squared, RMSE, and MRE.

2.2.4.2 ANN Model

Artificial Neural Networks (ANNs), designed to model complex non-linear relationships, were used for prediction (Goodfellow et al., 2016). Data was sorted chronologically by 'Year', with lag features and a rolling average (three-time point) generated to capture temporal patterns. Hyperparameter tuning determined the architecture, including the number of neurons, dropout rates, and learning rates (Bergstra et al., 2012). The model had two hidden layers, dropout regularization, and an output layer. Random search with early stopping helped prevent overfitting. The dataset was split using 5-fold Time Series Split, and both training and test sets were standardized. The ANN's performance was evaluated using R-squared, RMSE, and MRE.

2.2.5 Feature Importance Analysis with SHAP

The SHAP (SHapley Additive exPlanations) analysis (Lundberg & Lee, 2017) was used to explain the impact of individual features on the predictions of both XGBoost and ANN models. SHAP values measure each feature's contribution by evaluating their marginal effects across all possible feature combinations. For the XGBoost model, SHAP values were computed for features like 'Loc', 'Cultivar', 'SoilTyp', 'AvgT', and others, after data standardization. In the ANN model, the GradientExplainer method was used for SHAP value computation. SHAP summary plots visualized feature importance, showing the magnitude and direction of each feature's influence.

2.3 Part 3: Evaluation of Tomato Plant Genetic Resources for Brix and Lycopene in Different Environments

2.3.1 Plant Material

In our study, we evaluated six commercial tomato varieties, each with unique traits. The H1015 variety, developed by Heinz, is known for Extended Field Storage (EFS™), disease resistance, and adaptability to various climates, with a second-early maturity. The N6416 hybrid is early-

maturing, resistant to Tomato Spotted Wilt Virus (TSW), and has high acidity, ideal for industrial use. Prestomech F1 matures very early, with square/round fruits, high sugar content, and resistance to overripening. UG11227 and UG812J are industrial varieties, the latter known for resistance to mechanical damage. Ussar stands out for its mild, juicy flavour and versatility. These varieties were chosen for their diverse attributes to analyse performance in various environments.

2.3.2 Environments

This study assessed the performance of six tomato varieties across three locations in Hungary over five years (2017–2021). Trials were conducted annually in Szarvas, which has nutrient-rich meadow chernozem soil, in Mezöberény in 2018 and 2019 on fertile casting meadow soil, and in Kocsér in 2020 on sandy to humus sand soil. These diverse environments allowed for a thorough evaluation of the varieties' resilience and productivity under varying agro-climatic conditions, providing valuable insights for GGE biplot analysis.

2.3.3 Instrumental Measurements

The physicochemical properties of tomatoes were assessed using automated stations. Brix value (sugar content) was measured with the Maselli SV01 device, which converted tomatoes to juice and performed automated refractometric analysis, displaying results from 0 to 10 Brix with an accuracy of ±0.15 Brix. Lycopene concentration was determined via spectrophotometry, with a measurement range of 0 to 80 mg/100 g, accuracy of ±0.5 mg/100 g, and repeatability of ±0.25 mg/100 g.

2.3.4 GGE biplot

The GGE biplot was constructed by plotting the first (PC1) and second (PC2) principal component scores of genotypes and environments, derived from the singular value decomposition (SVD) of environment-centred data (Yan et al., 2000). This method visualizes genotype-environment interactions, showing the 'which-won-where' pattern, and ranks genotypes by performance and stability. The GGE biplot was analyzed using Genstat.v12 software with row-metric preserving settings (SVP = 2), no transformation, no scaling, and environment-centering (Centring = 2).

3 RESULTS AND DISCUSSION

3.1 Part 1: Root Development Monitoring under Different Water Supply Levels

3.1.1 General Results Regarding Root Count and Root Length

In 2020, a statistical analysis showed that the full irrigation treatment (I100) resulted in a smaller number of roots with less total length than in the water-stressed treatments, meaning 45% less

root per plant and 40% less total length compared to the I50 treatment and control, respectively, with no significant difference between the stressed treatments. The results of the 2021 analysis showed that plants under the control treatment produced the highest number of roots with the highest total length, followed by the I100 and then the I50 treatments. No significant difference was found between the two irrigated treatments in 2021. Overall, fewer roots were captured in 2021 than in 2020. The reason for this difference can be attributed to the different periods when a long-term irrigation deficit could develop and the different irrigation treatments that were applied.

According to our results in 2020, plants generally grew more roots with a greater total root length in the middle and bottom layers. In contrast, roots in the top layer did not exceed 71 roots with a 2081 mm total length. The 2021 results reinforced this pattern, indicating that root density in the soil increased with depth.

3.1.2 Evaluation of the Time Scale for the Monitored Root Zone

In 2020, the plants under the mild stress treatment (I50) exhibited significantly more roots with longer total lengths by the end of the monitoring period compared to the control and I100 treatments. Although the initial data suggested that the I50 and I100 treatments started on similar grounds, by the second week of monitoring, the rapid growth rate of the I50-treated plants led to a high root count comparable to the control. This observation could suggest that mild stress conditions stimulate the plants to develop more roots to absorb available water.

In 2021 growing season demonstrated a reduced number of roots and total root length in all the treatment groups. Notably, the plants in the control group showed the most extensive root growth. The irrigated treatments produced similar root counts during the experiment, and the three treatments barely differed in the final two weeks of the monitoring period in root length, while the higher number of roots in the control was continuous from the second week of the monitoring period.

3.1.3 Evaluation of the Layer Scale for the Monitored Root Zone

In the 2020, the distribution of roots was not uniform in the 10–70 cm rooting depth. The top layer developed a smaller number of roots with the least total length under all treatments. Regarding the top layer, the highest number and length were captured in the control. The plants that received full irrigation developed the highest number and longest roots in the middle layer compared to the other two soil layers. The 2020 growing season data revealed no significant difference in either root number or total length between the middle and bottom layers in the mild stress treatment.

In 2021, the deeper the soil layer was, the more and longer roots were developed, and the highest counts were observed in the control treatment in the 50–70 cm layer. However, the total length of roots was the lowest in the I50 treatment in the 10–30 cm layer, and the number of roots was equally low in the irrigated treatments in this layer. Comparing the number and length of roots, the data revealed that the plants in the control group developed significantly more roots in each layer compared to the irrigated treatments due to severe water stress. The results for the I50 treatment were inconsistent in the two growing seasons since the same level of root number was detected as in the I100 treatment.

3.1.4 Comparison of the Root Development in the Two Years

The comparative results demonstrate that the tomato plants cultivated in 2020 exhibited more substantial root growth and lengthier roots compared to those grown in 2021, regardless of the treatment applied. Consequently, the highest quantity and length of roots were observed in 2020 under the I50 treatment. The highest root count was recorded at 128 and 69, with corresponding total lengths of 4313 mm and 2607 mm for the years 2020 and 2021, respectively. Meanwhile, the minimum root count was observed in the 10–30 cm soil layer, with 70 and 41 roots and total lengths of 228 mm and 1610 mm, respectively, for 2020 and 2021. In 2020, the root counts were nearly equal in the 30–50 and 50–70 cm layers, whereas in 2021, both exhibited a consistent increase towards the bottom layer. The differences between treatments were less explicit in 2021.

The statistical analysis of the interaction effects between the year of measurement and the water treatment demonstrated a significant effect on both root count and root length, suggesting that the effectiveness of water treatments on root development varies depending on the year. On the other hand, the interaction between year and layer on root count shows significance (p-value = 0.0164), while it isn't significant for Root Length (p-value = 0.115), indicating that the influence of soil layer on root length is consistent across different years.

3.1.5 Effect of Different Treatments on Relative Chlorophyll Content (SPAD) and Chlorophyll Fluorescence (Fv/Fm)

The available data facilitate a comparison of the SPAD values of the tomato plants under the different irrigation treatments on each measurement date. In 2020, on 8 July, the I100 treatment displayed a lower SPAD value compared to both the I50 treatment and the control treatment. From 15 July to 29 July, the I50 treatment’s values generally surpassed those of the I100 treatment but fell short of the control treatment’s values. The SPAD values for all three treatments diminished during this period. In 2021, the SPAD values of the control treatment group were significantly higher than the I100 and I50 groups during the whole measurement period. By 14 July, the differences between the treatments became more pronounced and

significant. The I100 treatment showed a slight increase in SPAD value to 57.4, while the I50 and the control treatments exhibited a larger increase to 61.4 and 70.2, respectively, indicating a greater chlorophyll content.

The data show the chlorophyll fluorescence values of tomato plants under the different irrigation treatments on each measurement date. In 2020, the I100 plants initially exhibited lower values compared to the I50 and control plants. However, over time, the chlorophyll fluorescence values for the I100 plants gradually increased and eventually surpassed the values of the I50 and control plants by the time of the maturity period, after the irrigation had ended. The statistical analysis indicated that, except for the measurement taken on 29 July, there were no significant differences observed between the treatments. A similar observation was recorded in 2021, when the statistical analysis revealed that there were no significant differences between the treatments on all measurement dates, except for 29 July, where the control treatment had a higher value than the I50 treatment, which reported the lowest value. Both irrigated treatments displayed very similar values.

3.2 Part 2: A Comparative Analysis of XGBoost and Neural Network Models for Predicting Tomato Fruit Quality

3.2.1 Correlation Heatmap

The correlation heatmap provides a clear view of the linear relationships between climatic variables, Brix, Lycopene, and the a/b ratio. It uses a color spectrum from blue (negative correlations) to red (positive), as validated by Waskom (2021). Key insights include a strong positive correlation between 'AvgT' and 'T21_27', indicating that higher average temperatures correlate with more days in the optimal range for growth. 'TotPrecip' and 'RainDays' are closely aligned, highlighting the link between rainfall and precipitation. In contrast, 'AvgRH' negatively correlates with 'RH40_70', showing that higher humidity reduces the ideal humidity days for cultivation. Additionally, 'RH_90+' has a strong positive correlation with 'Brix', suggesting that high humidity may influence sugar concentration in fruits. The 'a/b ratio' also correlates with various climatic factors. These eight meteorological variables were used as independent factors in predictive models to analyze their impact on fruit quality and yield.

3.2.2 Model performance on Brix prediction

The developed algorithms exhibited a high degree of accuracy when estimating the Brix values. The XGBoost model yields an impressively a robust R^2 value of 0.98 and low RMSE of 0.07. Such results not only vouch for the XGBoost algorithm's capability but also highlight the significance of the chosen features in predicting Brix values from other climatic and quality variables. On the other hand, the ANN model resulted in an R^2 of 0.89 and RMSE of 0.17,

marking its good performance in intricate predictive modelling scenarios. The presented scatter plots from the two distinct models provide insights into their performance efficacy in predicting Brix values. Both plots display a significant concentration of data points around the black line representing x=y, highlighting the commendable accuracy of both models. For the XGBoost model and the ANN respectively, the percentage of predictions deviating less than 5% are 97% and 89%. It's noteworthy that a predominant cluster of data points signifying that the models' predictions are not only accurate but also consistent. These statistics underscore the models' competence in closely estimating the actual Lycopene content, despite some error margins which are to be expected in predictive modelling.

The MRE for the XGBoost model was as low as approximately 0.25% in some intervals, indicating high predictive accuracy, but it reached upwards of 2% in others, suggesting a reasonable predictive performance overall. On the other hand, the MRE for the ANN model, which varied significantly, ranging from approximately 0.5% to nearly 7%. While both models showed areas of agreement between actual and predicted Brix values, the ANN model exhibited higher variability in prediction accuracy.

3.2.3 Model performance on Lycopene prediction

The XGBoost model yielded an R^2 value of 0.87 and an RMSE value of 0.61, accounting for 87% of the variance in observed Lycopene content. In contrast, the ANN model had an R^2 of 0.84 and an RMSE of 0.86, attesting to its substantial explanatory capability. While both models exhibited commendable accuracy in predicting Lycopene content, minor inconsistencies were observed. The line representing ideal prediction, where predicted values coincide with actual measurements, serves as a benchmark for accuracy. It was revealed, that a significant proportion of predictions from both models lied within the 10% deviation margin, underscoring their precision. More specifically, for the XGBoost model and the ANN respectively, the percentage of predictions deviating less than 5% were 84.55% and 86.45% respectively.

Regarding the Mean Relative Error (MRE) the XGBoost model demonstrated a more stable performance, with most data groups maintaining an MRE below 4%, suggesting generally robust predictive accuracy. On the other hand, the ANN model, exhibited higher variability in its MRE, oscillating across different values and suggesting varying degrees of predictive accuracy. Notably, some segments exhibited a relatively high MRE, peaking just below 6%. The XGBoost model presented slightly superior performance in terms of consistency and reduced error.

3.2.4 Model performance on a/b ratio

The XGBoost model had demonstrated a high degree accuracy, achieving an R^2 value of 0.93 and an RMSE of 0.03, indicating a strong fit to the data. In contrast, the ANN model had yielded a higher RMSE of 0.138. While this suggested a reasonable proximity of predictions to actual observations, the model's negative R^2 value of -0.35 indicated a poor fit to the dataset. This finding suggested that either the current ANN model was not optimal for this dataset, or there were underlying issues with the dataset or its processing. In terms of prediction deviation, for the XGBoost model, 99.45% of predictions had been within 5% of the actual values. This indicated a high level of accuracy for most predictions. Notably, the ANN model had displayed significant deviations beyond the ±5% and ±10% margins, suggesting areas of unreliability. It is worth noting that despite the moderate correlation observed in the ANN model, indicating a positive linear relationship between observed and predicted values, the negative R^2 value pointed to its failure in adequately fitting the variance in the data. This discrepancy underscored the importance of comprehensive evaluation metrics in model assessment. The RMSE of 0.138, while seemingly small, was significant if the dependent variable in the dataset exhibited low variability. This magnitude of RMSE reflected that the ANN model's predictions were, on average, 0.138 units away from the actual values, leading to consistent and notable inaccuracies. Thus, the practical utility of the ANN model in this context was limited, as evidenced by its negative R^2 value, despite a moderate correlation.

In our analysis, the XGBoost model demonstrated satisfactory predictive performance. Its MRE fluctuated but remained relatively low, peaking slightly above 0.8%. In contrast, the ANN model exhibited significantly greater variability in its predictions. The MRE of the ANN model reached as high as approximately 12%, indicating that, on average, its predictions deviated by a maximum of 12% from the actual values.

3.2.5 SHAP

Brix

The most important difference between the SHAP plots of the two-machine learning model was that positive feature values contributed to mainly positive SHAP values in the ANN model but they were sorted differently for the XGBoost. The 'Cultivar' feature was paramount in the XGBoost model, displaying a broad range of SHAP values that are both positive and negative values, indicating a robust association between certain cultivars and elevated Brix levels. This suggested the significance of genetic attributes in enhancing water soluble solids content. The features related to humidity, such as 'RH40_70' and 'AvgRH' showed a substantial spread of SHAP values across the x-axis, suggesting variable effects on Brix prediction, where both low

and high relative humidity levels could either positively or negatively impact the accumulation of water-soluble solids in fruits, contingent upon other interacting variables. In contrast, in the ANN model the plot revealed a consistent pattern: higher feature values are invariably associated with positive SHAP values, while lower feature values correspond to negative SHAP values. This suggests a monotonic behavior where the magnitude of a feature's value is directly proportional to its impact on the model's output. The 'Cultivar' feature demonstrated a more uniform effect across the entire dataset, with a tendency toward positive contributions, reflecting its significant and consistent influence on the model's prediction of the Brix. Similarly, the SHAP values for 'Loc' and 'SoilTyp' indicate that geographical location and soil type are influential factors in predicting Brix levels, with higher and lower values of these features consistently impacting the model's output. The variable 'Year' also emerged as a significant temporal factor in the ANN model, potentially capturing the effects of varying climatic conditions across years, indicative of the model's capability to assimilate temporal dynamics into its predictive mechanism. The SHAP analysis showed that the XGBoost model attributed more importance to ‘AvgT’ than to ‘TotPrecip’, by contrast, the effect of ‘TotPrecip’ on the prediction of Brix was important in the ANN model. However, the ways in which these factors influenced Brix predictions in each model differed, possibly reflecting inherent differences in data assumptions and the models' strategies for integrating features.

Lycopene

The analysis of the XGBoost model revealed that the 'Cultivar' and 'RH40_70' features had a significant impact on the model's predictions of Lycopene content. The 'Cultivar' feature, in particular, showed a wide spread of SHAP values, indicating that different cultivars had varying levels of influence on the Lycopene content prediction. This suggested a complex, potentially non-linear relationship with the target variable. 'RH40_70' showed a more concentrated range of SHAP values, suggesting a consistent but less influential effect on the model's predictions. Other features were represented with SHAP values clustered closer to the center, implying a more moderate impact on the Lycopene content prediction. For the ANN, the 'Cultivar' feature exhibited the most substantial influence on the model's output with a broad spread of dots, indicating that the influence was more positive than negative. This implied a complex interplay where certain cultivars could have had a substantial impact, either augmenting or diminishing the potential Lycopene content determined by genetic background. Although the general directionality of feature values and their impact on the model’s predictions might have suggested a monotonic pattern, the spread and distribution of the SHAP values did not necessarily imply a linear relationship but rather implied a consistent pattern recognized by the neural network where

certain features were favorable for Lycopene production. The colour gradient added another layer of interpretability. For instance, the XGBoost plot showed that both high and low values of 'AvgT' did not exhibit simple linear relationships with Lycopene content. Instead, its impact was nuanced, with both high and low values influencing predictions in both positive and negative directions. This complexity may have mirrored how biological processes formed agricultural crops in response to environmental factors. Additionally, temporal trends reflected in the 'Year' feature's SHAP values could have pointed to evolving agricultural practices or climatic shifts over time, further highlighting the multifaceted nature of Lycopene biosynthesis.

a/b Ratio

The 'Year' feature in the XGBoost model had exhibited a high distancing of SHAP values, with clusters on both the positive and negative sides of the zero line, indicating a variable influence on the model's prediction, with some years contributing to an increase and others to a decrease in the predicted a/b ratio. The 'Cultivar' feature exhibited a unidirectional effect, with a pronounced aggregation of its SHAP values on the positive side, indicating a uniform contribution to the increase in the model's predicted a/b ratio. Notably, this increase is predominantly associated with the lower encoded values of 'Cultivar,' as indicated by the abundance of blue points. Conversely, 'TotPrecip' was predominantly associated with decreases in the a/b ratio, suggesting a positive relationship. For the ANN model, interpreting the SHAP values became more challenging due to the negative R^2 score. The model had predominantly exhibited negative SHAP values for features such as 'Cultivar', 'SoilTyp', and 'RH40_70'. These consistently downward predictions indicated that these features often reduced the predicted value compared to the model's baseline. The dominance of negative SHAP values and the lack of variation in SHAP value direction, unlike the variability observed in the XGBoost model, raised concerns about potential overfitting, insufficient feature representation, or inadequate network architecture to capture the complexities of the dataset. Furthermore, the ANN's poor performance metric, as highlighted by the negative R^2 score, had implied that the model was less informative than a simple average of the target variable, suggesting that the model's internal representations and learned weights did not generalize well to the data's underlying structure.

3.3 Part 3: Evaluation of Tomato Plant Genetic Resources for Brix and Lycopene in Different Environments

3.3.1 GGE biplot analysis

The first two principal PCs explain 77.10% (PC1 = 50.12%, PC2 = 26.97%) and 82.71% (PC1 = 67.42%, PC2 = 15.28%) of the total variation of the GGE model respectively for the Brix and

Lycopene values. PC1 defines the mean performance of the genotype, while PC2 shows the GEI of each variety, which is a measure of variability (stability).

Focusing on individual genotype performance, UG812J and Ussar, with PC1 values above zero, demonstrate high Brix and good adaptability. Prestomech, uniquely positioned near the biplot origin along PC2, demonstrates consistent stability across various conditions, although it exhibits average Brix content. Conversely, genotypes with PC1 values below zero show the opposite trend. The biplot's lack of clustering highlights significant environmental and genotype variation. Moreover, the acute angles observed between Szarvas2017 and Szarvas2021 indicate a positive correlation between each of these environments. This pattern repeats between the environments Szarvas2018, Szarvas2019, Szarvas2020, Mezobereny2018, Mezobereny2019. However, obtuse angles observed between Szarvas2017 and Szarvas2021 one side, and Szarvas2019 and Mezobereny2018 on the other side indicate a negative correlation between these environments. Same thing was observed when comparing Szarvas2020 with Szarvas2021 and Szarvas2019. A right angle between Szarvas2017 and Szarvas2018 indicates no correlation between these environments.

Regarding the Lycopene content, only H1015 and Ussar outperform the average and show good adaptability. The biplot also suggests environmental similarities, with Szarvas2017 and Mezobereny2018 clustering together, indicating shared attributes. This pattern is repeated for Szarvas2019, Szarvas2020, and Mezobereny2019. All environments except for Szarvas2017 exhibit an acute angle between each pair of environments, indicating a positive correlation between them. However, obtuse angles between Szarvas2017 and the trio of Szarvas2019, Szarvas2020, and Mezobereny2019 suggest negative correlations between these environments.

3.3.2 The Which-Won-Where patterns

The biplot evaluating the Brix values reveals the existence of three distinct mega-environments. The first mega-environment comprises Szarvas2017, whereas the second one includes Kocser2020 and Szarvas2021, and the third one all the rest of the environments. Notably, genotypes located in the same sector with a particular environment are the best performers in that environment. genotypes Prestomech is located in the same sector with environment as Kocser2020 and Szarvas2021, therefore, we would expect it to have the highest Brix values in these environments. Same thing goes for Ussar which is located in the same sector with Szarvas2018, Szarvas2019, Szarvas2020, Mezobereny2018 and Mezobereny2019.

The biplot evaluating the Lycopene values of identifies two mega-environments in the data, the first is consisting of Szarvas2017 and Kocser2020, the second of all the rest environments.

Among these, genotype H1015 is expected to exhibit the best Lycopene in the environments Szarvas2018, Szarvas2019, Szarvas2020, Szarvas2021, Mezobereny2018 and Mezobereny2019 since it is located in the same sector.

3.3.3 Ranking biplot

In the biplot analysis assessing Brix and Lycopene content, distinct patterns in genotype performance emerged. H1015 exhibited the highest mean Brix value, followed by N6416 and Prestomech, whereas UG11227 and UG812J recorded the lowest, falling below average. Notably, Prestomech demonstrated remarkable stability in Brix performance. In terms of Lycopene content, H1015 and Ussar showed higher mean values. Contrarily, Prestomech, alongside Ussar, UG812, UG1227, and N6416, exhibited lower mean Lycopene values. Ussar and UG812J were notably stable in their Lycopene performance. It is worth pointing out that according to Yan & Tinker (2006) if the biplot accounts for only a small fraction of the overall variation, it's possible that some genotypes which appear stable might not be genuinely stable. This is because their variability may not be completely captured in the biplot.

3.3.4 Comparison Biplot

The “ideal” genotype at the center of concentric circles, characterized by a position on the AEA (Absolutely Stable Axis) in the positive direction and a vector length matching the longest vectors of genotypes on the AEA's positive side, indicative of the highest mean performance. Therefore, the smaller the circle containing a genotype the more attributes it shares with the “ideal genotype”, which makes it more desirable than others. In this context, N6416 surpasses Prestomech in terms of desirable high Brix values, while UG1127 ranks as the least desirable. Conversely, for Lycopene content, H1015 leads in desirability, followed by Ussar, with Prestomech exhibiting the poorest performance across all environments.

Szarvas2018 and Mezobereny2019 emerge as the most favorable environments for achieving high mean Brix values combined with genotype stability. Similarly, Szarvas2021 and Mezobereny2018 are identified as optimal for high mean Lycopene values and genotype stability.

Results depicts a crucial idea related to "stability". The term "high stability" is favorable only when linked with the mean performance (Yan & Tinker, 2006). This criterion reveals that while Prestomech and Ussar are deemed 'highly stable', they exhibit lower Brix values compared to the less stable genotype, N6416.

4 CONCLUSIONS AND RECOMMENDATIONS

4.1 Conclusions

4.1.1 Root Development Monitoring under Different Water Supply Levels

Our study highlights the adaptability of tomato plants in response to varying water supply levels. The development of deeper roots under water stress, as observed in our findings, emphasizes plants' inherent strategies to counter water deficits and optimize water uptake. According to our results, the root system expansion to layers with higher soil moisture levels can happen quickly (<one week). The data suggested that root length could triple in 8 days. However, tomato plants that are irrigated regularly with sufficient water quantities develop shorter roots during the intensive root development phase.

Our findings also shed light on the impact of water supply on root system efficacy, with lower irrigation rates and water quantity levels stimulating more intensive root development. The observed variances in root growth over the two consecutive years, influenced by factors such as irrigation water levels and temperature variations, underscore the multifaceted nature of plant responses to environmental conditions.

The relationship between relative chlorophyll content and root development is stronger during the intensive root development period. The consistency in chlorophyll fluorescence across treatments, despite varying water conditions, suggests robust plant mechanisms that maintain photosynthetic efficiency under stress, even if the relative chlorophyll content is affected.

Our research contributes valuable insights into the adaptive strategies of plants under drought stress. This knowledge could inform plant breeding efforts aimed at developing cultivars that are more effectively adapted to water-deficient conditions. It is also pertinent to irrigation professionals seeking to enhance the use of soil layers and improve the effectiveness of root zones.

4.1.2 Comparative Analysis of XGBoost and Neural Network Models for Predicting Tomato Fruit Quality

Our study offers a detailed analysis of Brix, Lycopene, and a/b ratio predictions using XGBoost and ANN models. For Brix prediction, the XGBoost model proved to be highly effective, explaining approximately 98% of the variance in actual Brix values, compared to about 89% by the ANN model. In Lycopene content prediction, the XGBoost model demonstrated high efficacy with an 87% variance explanation, marginally outperforming the ANN model, which accounted for 84%. However, in predicting the a/b ratio, the XGBoost model maintained strong

performance with 93% of the variance explained, while the ANN model was notably less effective, indicated by a negative R^2 value of -0.35.

These findings underscore the superior predictive capabilities of the XGBoost model in these scenarios and reveal limitations of the ANN model, especially in predicting the a/b ratio. The SHAP summary plot analysis shows that both models effectively predict Brix values and Lycopene content in tomatoes, but with different focal points. XGBoost emphasized the genetic makeup of cultivars and their interaction with environmental factors, whereas the ANN model captures complex genetic interactions and direct feature relationships. Additionally, our results highlighted the significant influence of temporal factors, particularly 'Year', on the a/b chromaticity ratio, suggesting a complex interplay with climatic conditions and agricultural practices. The limitations of the ANN model in this aspect, as evidenced by its negative SHAP values and R^2 score, underline the necessity of meticulous model selection, optimization, and validation in precision agriculture.

4.1.3 Evaluation of Genetic Resources for Brix and Lycopene in Different Environments

The comprehensive analysis utilizing GGE biplot methodology has revealed distinct patterns of adaptability and performance across different tomato genotypes in varied environmental contexts, demonstrating the complexity and multidimensional nature of these interactions. The identification of mega-environments and their corresponding well-suited genotypes for particular quality traits provides a strategic framework for targeted breeding programs and cultivation practices aimed at enhancing tomato quality. The genotype H1015, for instance, emerged as a notable performer with high mean values in both Brix and Lycopene, suggesting its potential as a cornerstone in breeding programs focused on improving nutritional quality and taste.

Moreover, the variability in performance among genotypes across different environments underscores the essential role of GEI in determining the phenotypic expression of principal quality traits. This variability presents both challenges and opportunities for breeders in selecting and developing genotypes that can deliver consistent performance across diverse environmental conditions. The study also highlights the importance of maintaining genetic diversity in breeding programs, as different genotypes exhibit varied responses to environmental factors, thus enabling the cultivation of tomatoes that meet specific quality standards in different environments.

The limitations inherent in biplot analysis, such as the potential for deceptive stability in cases where a small fraction of the total variation is accounted for, emphasize the need for further research. Comprehensive datasets encompassing a wider range of environmental variability are essential for developing a more nuanced understanding of GEI and its implications for tomato quality trait improvement.

Together, these studies illustrate a comprehensive picture of the factors influencing tomato plant production and fruit quality. The adaptive strategies of plants to water stress, the predictive power of machine learning models, and the critical role of genetic diversity in shaping quality traits are all interconnected facets of a larger agricultural ecosystem. This research underscores the potential for a holistic approach to crop management, one that leverages advanced technologies and genetic insights to foster sustainable and efficient agricultural practices.

4.2 Recommendations

Water Management: Adaptive water management strategies must be developed to align with root architecture, optimizing water use efficiency and sustainability based on plant responses to water availability.

Machine Learning Adoption: Advanced machine learning models, like XGBoost, should be used to enhance decision-making in agriculture, allowing precise predictions of fruit quality and improving system resilience and productivity.

Breeding Programs: Breeding efforts should focus on selecting genotypes with high-quality traits and adaptability to various environments, helping agriculture withstand climate challenges.

Further Research: More research is needed to integrate environmental, genetic, and technological data to improve predictive models, address climate impacts, and enhance sustainable agricultural practices.

5 NEW SCIENTIFIC RESULTS

- ❖ Enhanced Root Development under Water Stress

Water-stressed tomato plants (I50 and control) exhibited significantly more and longer roots compared to fully irrigated plants (I100), particularly in deeper soil layers (30-70 cm). While full irrigation promotes initial root development primarily in the upper soil layers with eventual expansion to deeper layers, limited irrigation (I50 treatment) encourages deeper root growth from the outset as plants seek available water sources in the subsoil.

- ❖ Yearly Variations in Root Growth

Tomato plants exhibited significantly greater root growth and longer roots in 2020 compared to 2021, with the I50 treatment showing the most substantial root development (128 roots, 4313 mm in 2020 vs. 45 roots, 2058 mm in 2021); these variations were attributed to temperature and precipitation patterns, which impacted the development and distribution of root systems across different soil depths.

❖ Superior Performance of XGBoost Model

XGBoost consistently outperformed ANN, achieving high accuracy in predicting Brix (R^2 = 0.98, RMSE = 0.07) and lycopene content (R^2 = 0.87, RMSE = 0.61), and excelling in colour prediction (a/b ratio) with a R^2 of 0.93 and RMSE of 0.03. ANN lagged behind particularly in colour prediction, showing a negative R^2 value of −0.35.

❖ Importance of Specific Features in Prediction Models

SHAP value analysis further highlights the critical role of specific features such as 'Cultivar', relative humidity, and soil type, underscoring the complex interplay between genetic makeup, environmental conditions, and tomato quality.

❖ Differences in Model Interpretability

SHAP analysis reveals distinct differences in feature importance and model interpretability between XGBoost and ANN models, offering nuanced understanding of how genetic factors like 'Cultivar', environmental variables, and even temporal dynamics influence tomato quality traits such as Brix values, Lycopene content, and a/b ratio.

❖ Genotype and Environment Interaction

The interaction between genotype and environment is significant, contributing to 29.59% of the variance in Brix and 18.74% in lycopene. This indicates that the effect of genotype on these parameters is influenced by environmental factors and vice versa.

❖ Identification of Superior Genotypes

Genotypes such as UG812J and Ussar exhibit high Brix values and adaptability across diverse environments, while genotypes like H1015 and Ussar are highlighted for their superior Lycopene content, making them ideal candidates for breeding programs focused on nutritional quality enhancement.

6 SCIENTIFIC PUBLICATIONS

Published in impact factor journals

M'hamdi, O., Égei, M., Pék, Z., Ilahy, R., Nemeskéri, E., Helyes, L., & Takács, S. (2023). Root Development Monitoring under Different Water Supply Levels in Processing Tomato Plants. Plants, 12(20), 3517. https://doi.org/10.3390/plants12203517

M'hamdi, O., Takács, S., Palotás, G., Ilahy, R., Helyes, L., & Pék, Z. (2024). A Comparative Analysis of XGBoost and Neural Network Models for Predicting Some Tomato Fruit Quality Traits from Environmental and Meteorological Data. Plants, 13(5), 746. https://doi.org/10.3390/plants13050746

Publications published in non-Impact factor journals

M'HAMDI OUSSAMA, MÁRTON ÉGEI, ZOLTÁN PÉK, SÁNDOR TAKÁCS (2023). Effect of different water supply levels on the root system of industrial tomatoes. KERTGAZDASÁG, 55(1), 50-64. https://budaicampus.uni-mate.hu/documents/54944/7554750/50-64__zoldseg_Mhamdi_et+al_.pdf/129f46a5-a145-abea-552e-7b9550b2ea84?t=1678280426867

Submitted and accepted to Acta Horticulturae

M'hamdi, O., Égei, M., Pék, Z., Ilahy, R., Nemeskéri, E., Helyes, L., & Takács, S. (2024). Adaptive Responses of Tomato Plants to Varying Irrigation Levels: Insights into Root Development Efficiency

M'hamdi, O., Takács, S., Palotás, G., Ilahy, R., Helyes, L., & Pék, Z. (2024). Prediction of Tomato Quality Traits Utilizing Machine Learning Models

Posters

- Adaptive Responses of Tomato Plants to Varying Irrigation Levels: Insights into Root Development Efficiency
- Prediction of Tomato Quality Traits Utilizing Machine Learning Models

Conferences

Participation: 16th World Processing Tomato Congress and 18th ISHS Symposium on Processing Tomatoes (07/06/2026 - 10/06/2026).

Award: Winner of the ISHS Young Minds Award for Best Poster Presentation at the XVII International Symposium on Processing Tomatoes, Budapest, June 2024, for the paper titled "Prediction of Tomato Quality Traits Utilizing Machine Learning Models."