Plant Archives Vol. 25, No. 1, 2025 pp. 269-274 e-ISSN:2581-6063 (online), ISSN:0972-5210

Plant Archives

Journal homepage: http://www.plantarchives.org

DOI Url : https://doi.org/10.51470/PLANTARCHIVES.2025.v25.no.1.041

A DECISION TREE APPROACH TO GLYCEMIC INDEX ESTIMATION IN RICE BASED ON HYDROLYSIS PROFILES

C.K. Mohammed Salman, Anil Dahuja, Archana Singh and Veda Krishnan*

Division of Biochemistry, ICAR-Indian Agricultural Research Institute, Pusa, New Delhi, India

*Corresponding author E-mail : veda.krishnan@icar.gov.in, vedakrishnan@iari.res.in

(Date of Receiving-25-11-2024; Date of Acceptance-04-02-2025)

ABSTRACT

Diabetes, a chronic metabolic disorder characterized by elevated blood glucose levels, poses a significant global health challenge. This study explores the application of machine learning techniques to predict the glycemic index (GI) of rice, a staple food with substantial implications for diabetes management. Leveraging a comprehensive data-set of 53 rice accessions, the research develops a decision tree regression model to estimate the predictive glycemic index (pGI) using *in vitro* starch hydrolysis (SH) data. The research methodology involved analyzing starch hydrolysis percentages at multiple time points and calculating area under the curve (AUC) values across various intervals. The decision tree regressor was trained on 80% of the data-set and evaluated using performance metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). The model demonstrated exceptional predictive capabilities, with an R^2 value of 0.9914, explaining approximately 99.14% of variance in pGI data. Notably, feature importance analysis revealed that the AUC for the initial time period was the most influential predictor, contributing 99.30% to the model's predictive power. This emphasizes the critical role of early-stage starch hydrolysis dynamics in determining glycemic response. The study highlights the potential of ML techniques in nutritional biochemistry, providing a robust, interpretable approach for predicting GI. By integrating computational modeling with biochemical analysis, the research offers a scalable framework for high-throughput screening of functional foods and supports the development of personalized dietary interventions.

Key words : Glycemic Index Estimation, Hydrolysis profile, Diabetes management.

Introduction

Diabetes represents a chronic metabolic syndrome characterized by dysregulated glucose metabolism, manifesting through two primary pathophysiological mechanisms: impaired insulin production (type 1 diabetes) or compromised insulin utilization (type 2 diabetes). The escalating global prevalence of this chronic condition has prompted intensified research into predictive and preventive strategies. Contemporary machine learning approaches have emerged as sophisticated tools for risk stratification, integrating multifaceted clinical and demographic variables to develop predictive models for type 2 diabetes onset (Bonsembiante *et al*., 2021; Krishnan *et al*., 2021; Mondal *et al*., 2024). Concurrently, nutritional epidemiology has illuminated the pivotal role of dietary factors, with particular emphasis on the glycemic index (GI) as a critical determinant of metabolic health. Rice, a dietary staple across numerous global regions, has been identified as a nutritional component with significant metabolic implications. Its elevated glycemic index demonstrates potential contributors to type 2 diabetes pathogenesis (Chang *et al*., 2014; Salman *et al*., 2025; Krishnan *et al*., 2021). The glycemic response of rice is intricately modulated by its molecular starch architecture, encompassing crystalline structural configurations, amylose-to-amylopectin proportions, and the complex molecular interactions within amylopectin polymers (Li *et al*., 2023).

Previous studies have demonstrated both negative and positive correlation between amylose content and GI, with low-amylose rice varieties showing higher GI values (Durmus, 2024; Frei *et al*., 2003; Srikaeo and

Sangkhiaw, 2014). Additionally, the physical characteristics of starch, such as granule size and crystallinity, also impact digestibility, as higher crystallinity and limited swelling power are associated with increased resistance to enzymatic hydrolysis. Factors like gelatinization and pasting properties further influence GI, as higher gelatinization temperatures and resistant gel structures reduce enzymatic degradation rates. Alongside starch properties, components such as dietary fiber, proteins, lipids, and bioactive compounds like phytic acid contribute to modulating GI by delaying digestion or inhibiting enzymatic activity (Hernandez-Jaimes *et al*, 2015; Li *et al*, 2023; Mondal *et al*, 2021). Cooking and processing techniques further alter GI by modifying starch structures. While these relationships are well-documented, the application of ML offers a promising approach to unraveling the complex interplay of these factors. Advanced ML algorithms, including XGBoost, Random Forest, and CatBoost, provide powerful tools for predicting GI with high accuracy by analyzing diverse input variables, such as amylose content, crystallinity, and gelatinization parameters (Durmus, 2024). These algorithms leverage large datasets, uncovering subtle patterns to enhance the understanding and optimization of rice products with tailored GI values. Therefore, understanding the factors that influence the GI of rice and developing accurate predictive models using ML techniques can have significant implications for the prevention and management of diabetes. A basic outline on various ML tools are provided in Fig. 1.

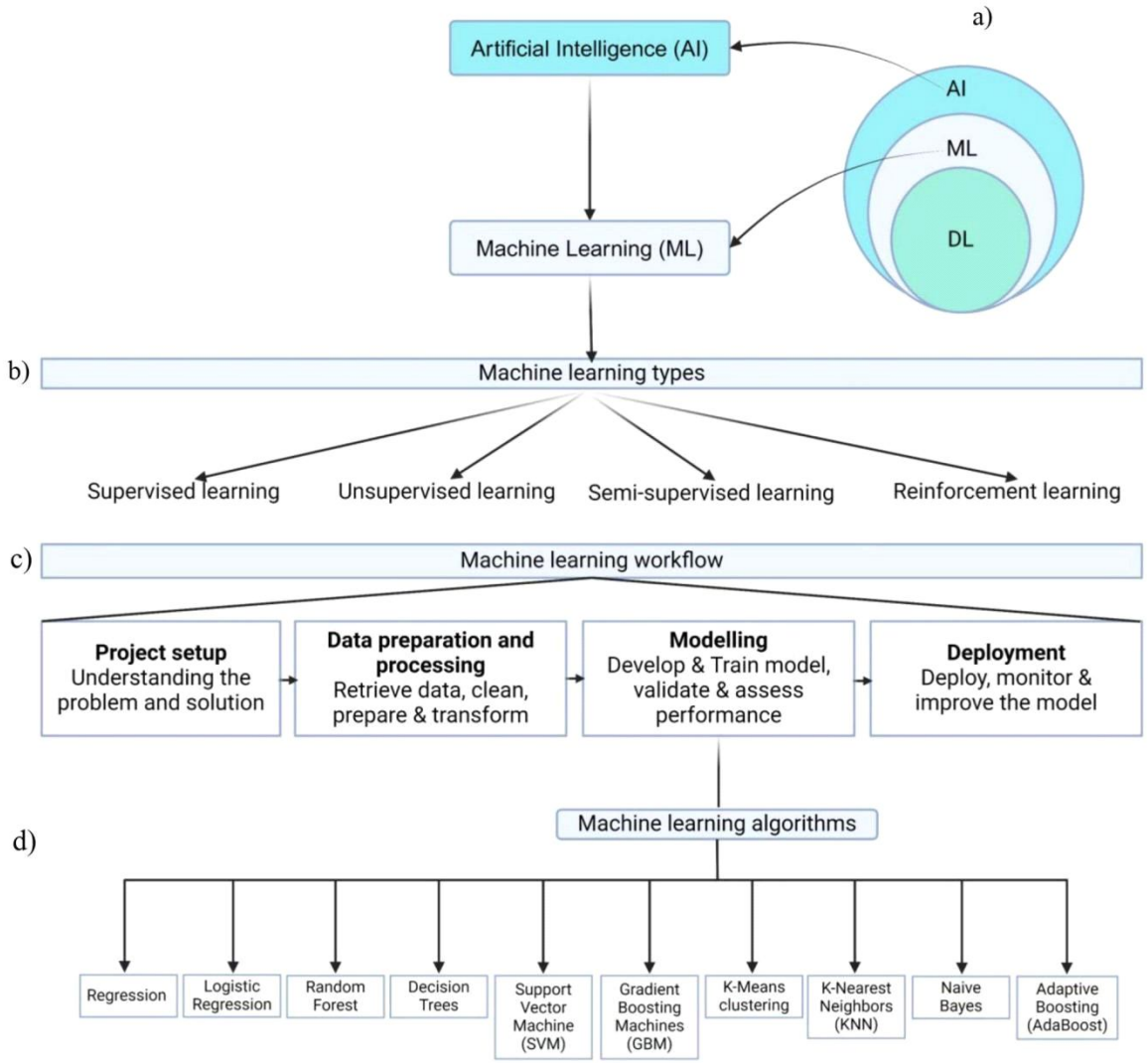

Fig. 1 : Schematic representation of machine learning: types, workflow and key algorithms. a) The image illustrates the hierarchy of Artificial Intelligence (AI), with Machine Learning (ML) as a subset and Deep Learning (DL) as a further specialization within ML. b) It highlights four types of ML: supervised, unsupervised, semi-supervised, and reinforcement learning. c) The workflow includes project setup, data preparation, modeling, and deployment. d) Key ML algorithms such as regression, decision trees, Support Vector Machine (SVM), random forests, and boosting methods like AdaBoost, emphasizing the diverse tools used in ML for various problem-solving approaches.

This current study developed a decision tree model system for accurate predictive glycemic index (pGI) estimation using *in vitro* starch hydrolysis (SH) data from 53 rice accessions, as described by Salman et al., 2025. The model aims to streamline the process of screening rice accessions, offering improved accuracy and efficiency compared to conventional methods. This approach not only facilitates targeted crop breeding for rice varieties with desirable GI characteristics but also supports the rapid identification of promising accessions for further research and development. By integrating ML with biochemical analysis, the study provides a robust tool for advancing the precision and speed of GI prediction in rice breeding programs (Salman *et al*., 2025).

Materials and Methods

Data for analysis

The dataset utilized for ML analysis was derived from an in-house study conducted in our laboratory, building on the findings of Salman *et al*. (2025). This data-set comprised SH data collected at 10 time points (SH0, SH5, SH10, SH20, SH30, SH45, SH60, SH90, SH120, SH180, eg: SH5 represents starch hydrolysis at 5 minutes) for 53 rice accessions. Alongside the SH data, additional parameters such as pGI, total starch (TS) content and inherent glycemic potential were included to enhance the analysis.

To further capture the dynamic nature of SH, the area under the curve (AUC) was calculated for various time intervals using GraphPad Prism (version 10.1.1). These intervals included both cumulative and segment-specific combinations:

- **Cumulative AUCs:** (0–5), (0–10), (0–20), (0–30), (0–45), (0–60), (0–90), (0–120), (0–180)
- **Segment-specific AUCs:** (5–10), (5–20), (5–30), (5–45), (5–60), (5–90), (5–120), (5–180), (10–20), (10–30), (10–45), (10–60), (10–90), (10–120), (10–180), (20–30), (20–45), (20–60), (20–90), (20–120), (20–180), (30–45), (30–60), (30–90), (30–120), (30–180), (45–60), (45–90), (45–120), (45–180), (60–90), (60–120), (60–180), (90–120), (90–180), and (120–180).

The calculated AUC values, in combination with SH data and IGP, provided a comprehensive framework for training ML models. This high-resolution data-set allowed for precise pattern recognition and predictive modeling of glycemic responses in rice accessions, enabling accurate and robust pGI estimation.

Decision tree regressor model development

To capture the non-linear relationships within the data-set, we employed a decision tree regression algorithm, a ML technique that iteratively partitions the data into subsets based on the most informative features. This approach is well-suited for modeling complex interactions between variables that may elude traditional linear regression methods.

Model Training and Evaluation

The data-set was divided into training and testing subsets, with 80 % allocated to training and 20 % to testing. A fixed random seed (random state: 42) was used to ensure reproducibility. The decision tree regression model was trained on the training subset and subsequently evaluated using multiple performance metrics. R-squared (R^2) was calculated to determine the proportion of variance in the target variable (pGI) that the model could explain. The Mean Squared Error (MSE) was used to assess the average squared difference between the observed and predicted values, while the Root Mean Squared Error (RMSE) provided an interpretable measure of prediction error expressed in the same units as the target variable.

GraphPad Prism 10.1.1 was employed for calculating the AUC and other statistical analyses. Google Colab was used for implementing decision tree models and conducting data analysis, providing an interactive and collaborative computational environment (Google Colab, n.d., https://research.google.com/colaboratory/faq.html).

Feature importance Analysis

To quantify the contribution of each feature in predicting the pGI, the importance scores of all predictor variables were extracted from the trained model. These scores provide insights into the relative significance of features in the model's decision-making process, highlighting the key parameters driving GI prediction.

Visualizations

The decision tree model was visualized to provide an intuitive representation of the data partitions and the predictive hierarchy of variables. Additionally, a scatter plot of observed versus predicted pGI values was created to evaluate the model's performance. A regression line was overlaid on the plot to compare the predicted values against a perfect prediction scenario, thereby visually assessing the model's accuracy.

Results and Discussion

Decision tree model analysis for pGI prediction

The decision tree regressor was employed to predict the pGI based on 55 features, including starch hydrolysis percentages (SH%), AUC, TS and IGP. The decision

tree algorithm was chosen for its ability to handle non-linear relationships and interactions between variables, as well as its inherent interpretability. By recursively partitioning the data, the model constructs a tree structure where each node represents a decision rule based on feature thresholds, and the terminal leaf nodes represent predicted pGI values. This hierarchical structure enables the model to learn complex patterns and interactions, making it particularly suited for modeling the intricate relationships observed in *in vitro* digestion datasets.

The model operates by minimizing the variance in the target variable (pGI) within each partition, ensuring that the splits are optimal for prediction accuracy. This approach also allows for an intuitive understanding of the decision-making process, as each branch of the tree can be visualized to trace the pathway leading to specific predictions. By leveraging this capability, the decision tree model serves as a robust and interpretable tool for predicting pGI from biochemical datasets.

Model performance

The data-set was split into training and testing subsets, allocating 80% for training and 20% for testing, with a fixed random state of 42 to ensure reproducibility. The model's performance was assessed using the following metrics:

1. Mean Squared Error (MSE): 0.9610

This metric quantifies the average squared difference between the observed and predicted pGI values, with a lower value indicating better predictive accuracy.

2. Root Mean Squared Error (RMSE): 0.9803

By taking the square root of the MSE, the RMSE provides a more interpretable measure of prediction error in the same units as the target variable.

3. R-squared (R^2): 0.9914

This value indicates that the model explains approximately 99.14% of the variance in the pGI data, demonstrating its ability to capture the underlying patterns effectively.

The low MSE and RMSE values, combined with the high R^2, highlight the model's accuracy and robustness in predicting pGI from the provided features.

Feature Importance Analysis

To gain insights into the model's decision-making process, feature importance was analyzed. The results revealed that AUC (0-5) was the most influential feature, contributing 99.30% to the model's predictive power. This finding underscores the critical role of earlySH dynamics in determining the glycemic response.

In contrast, features such as AUC (0-45) and AUC (30-45) contributed minimally, with respective importance values of 0.48% and 0.07% (Table 1). These differences reflect the diminishing relevance of later time points in predicting pGI, likely due to the plateauing behavior of SH after the initial phase. The stark contrast in feature importance highlights the dominance of early hydrolysis events in driving pGI predictions and provides valuable insights for future experimental designs focusing on the early digestion phase.

Model Visualization

To elucidate the decision-making process, the decision tree structure was visualized. The tree illustrates how feature thresholds are used at each node to split the data, ultimately leading to predictions at the leaf nodes. This graphical representation aids in interpreting the criteria used for predictions and provides a clear understanding of the relationships between features and pGI. Model of the tree developed is provided in Fig. 2.

Prediction analysis and interpretation

A scatter plot comparing predicted mGI (ML based GI) values (*via* the decision tree regressor) against experimental pGI values (*via in vitro* digestion)

Table 1 : Feature Importances in Decision Tree Model for Predicted Glycemic Index (pGI) Prediction. The table lists the relative importance of each feature in predicting pGI using the decision tree regressor model. The abbreviation AUC refers to the Area Under the Curve of starch hydrolysis at different time intervals (e.g., AUC (0-5) indicates the area under the curve from 0 to 5 minutes). TS denotes Total Starch, and SH60 represents Starch Hydrolysis at 60 minutes. The feature importance values indicate the contribution of each feature to the predictive accuracy of the model, with AUC (0-5) being the most significant predictor, accounting for 99.30% of the model's overall predictive power.

Features	Feature importance
AUC (0-5)	0.9930
AUC (0-45)	0.0048
AUC (30-45)	0.0007
AUC (0-10)	0.0003
AUC (90-180)	0.0002
AUC (10-20)	0.0001
AUC (5-45)	0.0001
AUC (30-60)	0.0001
TS	0.0001
AUC (0-90)	0.0001
SH60	0.0001
AUC (5-180):	0.0001

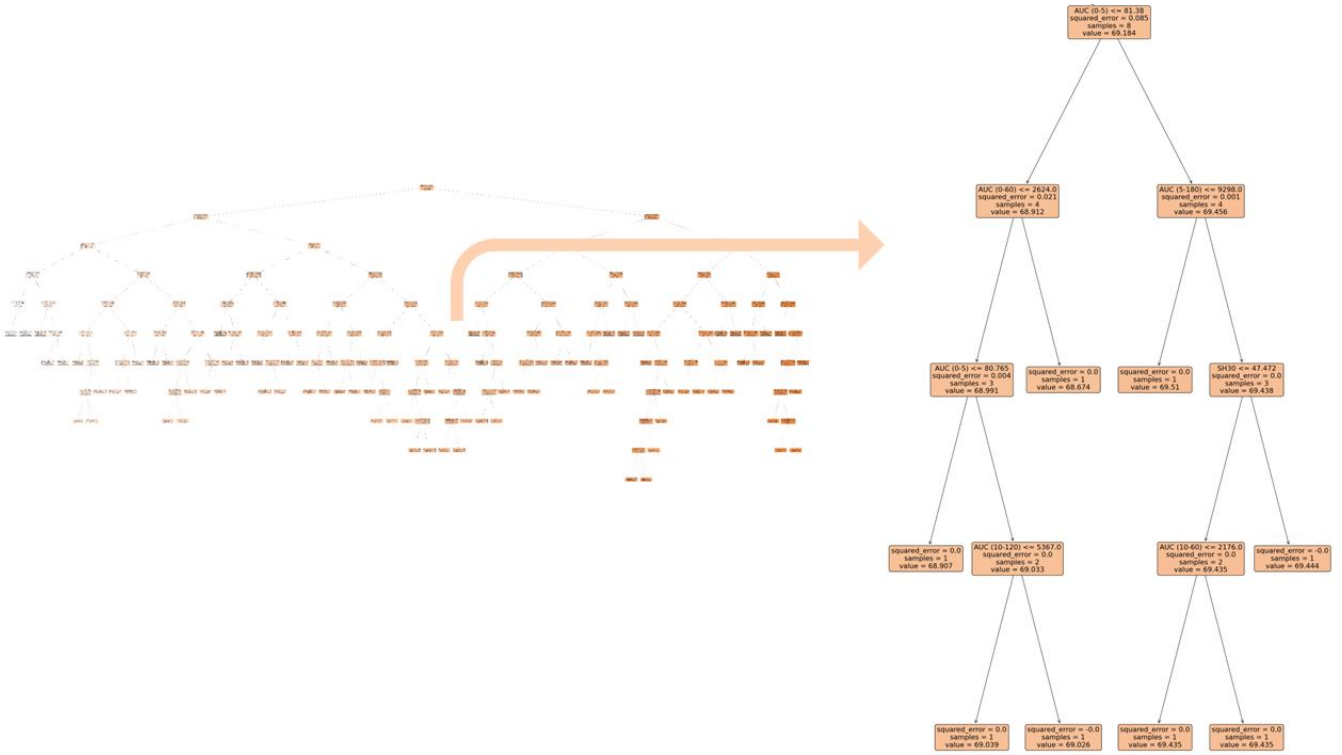

Fig. 2 : Decision tree for estimating predictive Glycemic Index (pGI). a) Decision tree derived the from the data set b) Enlarged portion of the above tree: The node splits based on the AUC (0-5) feature, with the squared error, sample number, and corresponding pGI value displayed. This split illustrates how the model uses AUC (0-5) as a critical decision point to minimize prediction error, guiding the tree in classifying samples into specific pGI categories (AUC-Area Under the Curve).

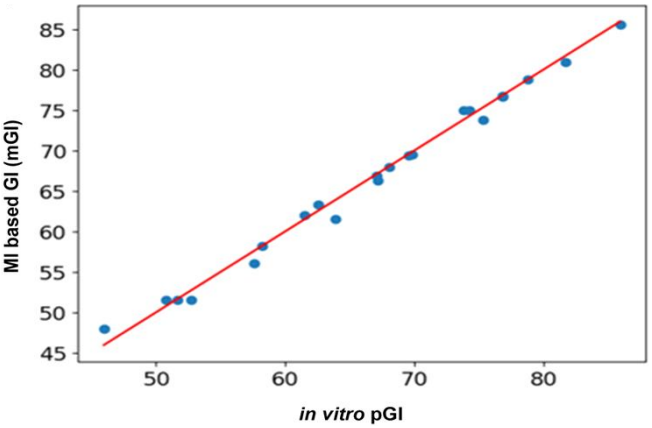

Fig. 3 : Scatter plot analysis for decision tree regressor model (R^2: 0.9914).

demonstrated a strong alignment between predicted and observed data. The regression line, plotted in red, closely follows the data points, reflecting the model's accuracy and reliability, R^2: 0.9914 (Fig. 3). Thus, decision tree regressor provides a robust and interpretable model for predicting pGI. The analysis highlights the importance of specific features and validates the model's accuracy through various performance metrics.

The dominance of early time point AUC in the feature importance analysis aligns with the physiological relevance of early glucose release in determining glycemic response. This reinforces the importance of capturing rapid SH dynamics in predicting GI. The minimal contribution of later AUCs suggests that the saturation phase of hydrolysis plays a negligible role in influencing pGI outcomes, aligning with prior observations of plateauing behavior in enzymatic starch breakdown. These findings aligns with the previous study by Salman *et al*. (2025).

Conclusion

This study demonstrates the efficacy of a decision tree regressor for predicting the pGI using an extensive data-set derived from *in vitro* digestion models. The model exhibited exceptional predictive performance, with high R^2 values (0.9914) and low error metrics, underscoring its capacity to capture the complex, non-linear relationships inherent in glycemic response data. Notably, the feature importance analysis identified AUC (0-5) as the dominant variable driving pGI predictions, aligning with the critical role of early-stage SH in glycemic outcomes. By visualizing the tree structure and conducting a comprehensive feature analysis, this study bridges the gap between computational modeling and biochemical interpretation, advancing the understanding of GI dynamics. Integrating advanced ensemble techniques or

explainable AI tools could further refine the predictive framework, enhancing its precision and transparency. This research underscores the transformative potential of ML in nutritional biochemistry, offering a robust, scalable, and interpretable approach for predicting GI. The findings not only pave the way for high-throughput screening of functional foods but also provide a foundation for the development of integrative, AI-driven frameworks in nutritional science, ultimately contributing to the advancement of personalized dietary interventions and public health initiatives.

References

Bonsembiante, L., Targher G. and Maffeis C. (2021). Type 2 Diabetes and Dietary Carbohydrate intake of Adolescents and Young adults: What is the Impact of Different Choices? *Nutrients*, **13(10)**, Article 10. https://doi.org/10.3390/nu13103344

Chang, U.J., Hong Y.H., Jung E.Y. and Suh H.J. (2014). Chapter 27—Rice and the Glycemic Index. In: Watson, R.R., Preedy V. R. and Zibadi S. (eds.), *Wheat and Rice in Disease Prevention and Health* (pp. 357–363). Academic Press. https://doi.org/10.1016/B978-0-12-401716-0.00027-1

Durmus, Y. (2024). High accuracy prediction of Thai rice glycemic index using machine learning. *Cogent Food & Agriculture*, **10(1)**, 2411032. https://doi.org/10.1080/23311932.2024.2411032

Frei, M., Siddhuraju P. and Becker K. (2003). Studies on the *in vitro* starch digestibility and the glycemic index of six different indigenous rice cultivars from the Philippines. *Food Chem.*, **83(3)**, 395–402. https://doi.org/10.1016/S0308-8146(03)00101-8

Hernandez-Jaimes, C., Lobato-Calleros C., Sosa E., Bello-Pérez L.A., Vernon-Carter E.J. and Alvarez-Ramirez J. (2015). Electrochemical characterization of gelatinized starch dispersions: Voltammetry and electrochemical impedance spectroscopy on platinum surface. *Carbohydrate Polymers*, **124**, 8–16. https://doi.org/10.1016/j.carbpol.2015.02.002

Krishnan, V., Awana M., Singh A., Goswami S., Vinutha T., Kumar R.R., Singh S.P., Sathyavathi T., Sachdev A. and Praveen S. (2021). Starch molecular configuration and starch-sugar homeostasis: Key determinants of sweet sensory perception and starch hydrolysis in pearl millet (*Pennisetum glaucum*). *Int. J. Biological Macromolecules*, **183**, 1087–1095. https://doi.org/10.1016/j.ijbiomac.2021.05.004

Krishnan, V., Mondal D., Thomas B., Singh A. and Praveen S. (2021). Starch-lipid interaction alters the molecular structure and ultimate starch bioavailability: A comprehensive review. *Int. J. Biological Macromolecules*, **182**, 626–638. https://doi.org/10.1016/j.ijbiomac.2021.04.030

Li, C., Hu Y., Li S., Yi X., Shao S., Yu W. and Li E. (2023). Biological factors controlling starch digestibility in human digestive system. *Food Science and Human Wellness*, **12(2)**, 351–358. https://doi.org/10.1016/j.fshw.2022.07.037

Mondal, D., Awana M., Mandal S., Pandit K., Singh A., Syeunda C.O., Thandapilly S.J. and Krishnan V. (2024). Functional foods with a tailored glycemic response based on food matrix and its interactions: Can it be a reality? *Food Chem. X*, **22**, 101358. https://doi.org/10.1016/j.fochx.2024.101358

Mondal, D., Awana M., Aggarwal S., Das D., Thomas B., Singh S.P., Satyavathi C.T., Sundaram R.M., Anand A., Singh A., Sachdev A., Praveen S. and Krishnan V. (2022). Microstructure, matrix interactions and molecular structure are the key determinants of inherent glycemic potential in pearl millet (*Pennisetum glaucum*). *Food Hydrocolloids*, **127**, 107481. https://doi.org/10.1016/j.foodhyd.2022.107481

Salman, C.K.M., Beura M., Singh A., Dahuja A., Kamble V.B., Shukla R.P., Thandapilly S.J. and Krishnan V. (2025). Biomimic models for *in vitro* glycemic index: Scope of sensor integration and artificial intelligence. *Food Chem.: X*, **25**, 102132. https://doi.org/10.1016/j.fochx.2024.102132

Salman, C.K.M., Bollinedi H., Anand A., Singh A., Sundaram R.M., Prathibha K. and Krishnan V. (2025). *In vitro* Glycemic Profiling of Rice: A Dual-Index Approach using Predictive Glycemic Index and Inherent Glycemic Potential. *J. Food Composition and Analysis*, 107229. https://doi.org/10.1016/j.jfca.2025.107229

Srikaeo, K. and Sangkhiaw J. (2014). Effects of amylose and resistant starch on glycaemic index of rice noodles. *LWT - Food Sci. Technol.*, **59(2, Part 1)**, 1129–1135. https://doi.org/10.1016/j.lwt.2014.06.012