Quality Evaluation of Tea using Chemical Sensors and NIR Spectroscopy: Machine Learning Approach

Thesis submitted

By

Angiras Modak

Doctor of Philosophy (Engineering)

Department of Instrumentation and Electronics Engineering
Faculty Council of Engineering & Technology
Jadavpur University
Kolkata, India

2025

JADAVPUR UNIVERSITY

KOLKATA- 700032

INDIA

INDEX NO. 212/19/E

Title of the Thesis **Quality Evaluation of Tea using Chemical Sensors and NIR Spectroscopy: Machine Learning Approach**

Professor
Dept. of Instrumentation & Electronics Engg
Jadavpur University
Saltlake, 2nd Campus
Kolkata-700 106

Name, Designation & Institution of the Supervisor

Dr. Runu Banerjee Roy
Professor
Dept. of Instrumentation and Electronics Engineering
Jadavpur University
Salt Lake Campus, Sector- III, Block- LB, Plot-8, Kolkata- 700106, India.
Phone: 08240235742
Fax: 03323357254

List of Publications

Journals

[1] **A. Modak**, D. Das, T. N. Chatterjee, B. Tudu and R. Banerjee Roy, "Regression-Based EGCG Detection in Green Tea Employing MIP Electrodes," in *IEEE Sensors Journal*, vol. 24, no. 11, pp. 18090-18097, 1 June 1, 2024, doi: https://doi.org/10.1109/JSEN.2024.3390438.

[2] Modak A, Moulick M, Das D, Roy RB. Effective tannic acid prediction in black tea using regression approach: Fusion of electrochemical and spectroscopic responses. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. 2026 Apr 22:127958. doi: https://doi.org/10.1016/j.saa.2026.127958.

Conference Proceedings

[1] **A. Modak**, T. N. Chatterjee, S. Nag, R. B. Roy, B. Tudu and R. Bandyopadhyay, "Linear Regression Modelling on Epigallocatechin-3-gallate Sensor Data for Green Tea," 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, India, 2018, pp. 112-117, doi: https://doi.org/10.1109/ICRCICN.2018.8718696.

[2] **A. Modak**, D. Das and R. B. Roy, "Classification of Green Tea Samples Based on Epigallocatechin-3-Gallate Content Using Molecular Imprinting Polymerization Technique," 2024 IEEE 3rd International Conference on Control, Instrumentation, Energy & Communication (CIEC), Kolkata, India, 2024, pp. 355-360, doi: https://doi.org/10.1109/CIEC59440.2024.10468449.

[3] **Modak, A.**, Moulick, M., Roy, R.B. (2025). Efficient Near Infrared Spectroscopy Based Feature Selection of Tannic Acid for Black Tea Evaluation. In: Singh, J.P., Singh, M.P., Singh, A.K., Mukhopadhyay, S., Mandal, J.K., Dutta, P. (eds) Computational Intelligence in Communications and Business Analytics. CICBA 2024. Communications in Computer and Information Science, vol 2367. Springer, Cham. doi: https://doi.org/10.1007/978-3-031-81339-9_13.

Statement of Originality

I, ***Angiras Modak****, registered on* ***10th July, 2019****, do hereby declare that this thesis entitled "****Quality Evaluation of Tea using Chemical Sensors and NIR Spectroscopy: Machine Learning Approach****" contains a literature survey and original research work done by the undersigned candidate as part of Doctoral studies.*

All information in this thesis has been obtained and presented in accordance with existing academic rules and ethical conduct. I declare that, as required by these rules and conduct, I have fully cited and referred to all materials and results that are not original to this work.

I also declare that I have checked this thesis as per the "Policy on Anti-Plagiarism, Jadavpur University, 2019", and the level of similarity as checked by iThenticate software is **3%**.

Angiras Modak

Angiras Modak
File No. 212/19/E
Date: 08.06.2026

Dr. Runu Banerjee Roy
Professor
Dept. of Instrumentation and Electronics Engineering
Jadavpur University
Salt Lake Campus, Kolkata

Professor
Dept. of Instrumentation & Electronics Engg
Jadavpur University
Saltlake, 2nd Campus
Kolkata-700 106

Certificate from the Supervisor

Date: 08.06.2026

*This is to certify that the thesis entitled "**Quality Evaluation of Tea using Chemical Sensors and NIR Spectroscopy: Machine Learning Approach**", submitted by Angiras Modak, who got his name registered on 10th July, 2019, for the award of Ph.D. (Engineering) degree of Jadavpur University, is absolutely based upon his own work under supervision of Prof. Runu Banerjee Roy and that neither his thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award anywhere before.*

Dr. Runu Banerjee Roy

Professor

Dept. of Instrumentation and Electronics Engineering

Jadavpur University

Salt Lake Campus, Kolkata

This thesis is dedicated to

my beloved parents,

my dearest wife,

and my precious daughter

Thank you for your endless love, patience, unwavering support, guidance, and sacrifice.

Every step forward was made possible by your unfaltering belief in my pursuits.

This accomplishment is as much yours as it is mine.

I am forever grateful for your presence in my life.

Acknowledgement

Any success requires the effort of many, and this is no exception. First and foremost, I would like to extend my sincere gratitude to my supervisor, Dr. Runu Banerjee Roy, Professor at Jadavpur University. Without her able guidance, support, and constant encouragement, this thesis would not have materialised. Her technical expertise and motivation have helped me overcome the difficulties in this course of work. Her kind assistance and her patience with my knowledge gaps in this area have been unparalleled. It was a great privilege and honour to work and study under her guidance.

I would also like to extend my profound gratitude and heartfelt thanks to Prof. Rajib Bandyopadhyay, Dept. of IEE, Jadavpur University, for his invaluable advice at critical junctures of this course. His lectures have left a strong impression on me, and I have always cherished the positive memories of our interactions. I am grateful to him for being the foundation upon which this laboratory, and all the work and emotions within it, has been built.

Furthermore, I would like to extend my gratitude to Prof. Bipan Tudu, Dept. of IEE, Jadavpur University. His innovative teaching methods and problem-solving capabilities have consistently inspired me during challenging times in this work.

I would also like to thank Dr. Debangana Das, NIT Agartala, India, for being a constant scientific collaborator, motivator, and advisor throughout my research duration. I thank my research partners, Dr. Trishita Nandi Chatterjee, Dr. Mahuya Banerjee, Dr. Sreya Nag, Mrs. Madhurima Moulick, Mr. Dipan Bandyopadhyay, Mr. Deepam Ganguly, Mr. Sumit Kundu, Mr. Mohasin Ali Miah, and Dr. Srikanta Acharya for creating a cheerful atmosphere and providing support.

Finally, I would like to extend my love and gratitude to my parents, wife, and daughter, whose unwavering support, vigilance, and constant encouragement enabled this dissertation to blossom from a mere bud into a beautiful flower. Their

understanding of the stress I experienced during my research, along with their tactful handling of my mood fluctuations, was immensely gratifying. Every time I was ready to quit, they would not let me, and for that, I will remain forever grateful. This thesis stands as a testament to their genuine love, compassion, and support throughout the course of this work. Lastly, I express my gratitude to the Almighty God for His abundant blessings.

Abstract

Tea has been one of the most widely consumed non-alcoholic beverages worldwide since ancient times due to its distinctive flavour and health benefits. It includes components such as polyphenols, alkaloids, tannic acids and amino acids that contribute to its taste and aroma. Tea is classified into six distinct types, which are determined by the diverse processing techniques employed. These types include dark, black, oolong, yellow, green, and white teas. Among them, black and green tea are highly consumed by many people globally. However, variations in processing and manufacturing practices may result in lower-quality tea products, affecting both market reputation and consumer interests. Assessing the quality of tea is essential for both producers and consumers; however, it poses significant challenges due to the multitude of compounds and their varying impacts on tea quality.

Typically, the determination of tea quality is conducted through human sensory analysis, which offers direct and comprehensive measurements of different characteristics. The classification of tea is derived from the ratings given by seasoned tea tasters, who evaluate the perceived intensities of appearance, aroma, and flavour through their sensory perception. Though human sensory evaluation is inherently subjective, even the same individual may yield different assessments across various trials, and sensory panels can exhibit inconsistencies due to individual differences. Additionally, numerous emotional and physical factors (such as mental or bodily states) can further compromise the accuracy of human sensory evaluations. Furthermore, conventional tea quality assessment methods are often time-consuming and may lack consistency.

To address these issues of subjective sensory evaluation, objective instrumental methods are used. Near-Infrared (NIR) spectroscopy provides a rapid chemical fingerprint, while Molecularly Imprinted Polymers (MIPs) are a sensor

development technique used to create highly selective electrochemical sensors. Both approaches offer reliable, consistent alternatives to human panels for tea quality assessment. As a result, the present research aims to enhance the tea quality prediction, particularly for black and green tea, using NIR and MIP electrochemical methods and implements four proposed models using machine learning (ML) and deep learning (DL) methods.

The first proposed model employs an ML approach to assess the quality of green tea by analysing its EGCG content. Evaluation of EGCG is conducted using an MIP electrode, employing a three-electrode configuration to capture the voltammetric response. Following data pre-processing, the UMAP algorithm is applied to enhance the output by selecting features. Then, K-means clustering is used to differentiate between samples. The classification task is performed using the XGBoost algorithm, renowned for its quick convergence and minimal error. Additionally, the Epigallocatechin-3-gallate (EGCG) levels are measured using High-performance liquid chromatography (HPLC) data for each green tea sample. The effectiveness of the model is evaluated through various metrics, demonstrating its potential of classifying green tea accurately.

The second proposed model introduces a novel Particle Swarm Optimisation-based Random Forest (PSRF) regression model. The modified PSO, obtained by hyper-tuning the traditional PSO, is combined with the RF Model to perform the regression process. First, the study collects the data by incorporating multiple MIP sensors to measure Tannic Acid (TAN), Theophylline (THEO), and Caffeine (CAFF). The collected data are fused and prepared as a dataset for quality analysis. Then, a greedy-based genetic algorithm (Greedy-GA) is employed to perform feature selection, selecting the enriched feature set from the input data. The proposed PSRF is to conduct the regression process. The efficacy of the sensor fusion has been evaluated using standard metrics and demonstrates enhanced accuracy with a lower error rate in predicting black tea quality.

The third proposed framework focuses on differentiating black tea samples using NIR spectroscopy, specifically analysing the TAN content, as it is a significant factor in tea quality. The samples are examined in absorbance mode over a wavelength range of 900 to 1700 nm. Data preprocessing is conducted with the standard scalar method to standardise the responses, which are then subjected to feature selection. Data reduction techniques, including Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP), and Kernel PCA (KPCA), are utilised to further analyse the selected features. The study's effectiveness is assessed using criteria such as the Silhouette, Calinski-Harabasz, and Davies-Bouldin indices, with KPCA yielding a satisfactory Silhouette coefficient of 0.64. This indicates that KPCA effectively aids in predicting TAN content and serves as a reliable method for evaluating the quality of black tea.

The final proposed model develops an MIP electrochemical sensor to predict tannic acid content in black tea samples, enhancing prediction accuracy by integrating electrochemical responses with Near-Infrared (NIR) spectroscopic data. A regression model based on a Siamese Neural Network-Long Short-Term Memory (SNN-LSTM) architecture is employed, optimised with a Particle Swarm Optimisation-Dragonfly Optimisation (PSO-DFO) algorithm for effective feature selection and a Convolutional Neural Network (CNN) for feature extraction. Individual system responses are analysed through Canonical Correlation Analysis (CCA) and UMAP techniques to achieve effective data fusion. CCA computes canonical weights for feature projection, while UMAP assists in dimensionality reduction. The SNN-LSTM model successfully identifies complex relationships among compounds in black tea and achieves strong predictive capability through CCA, indicating strong correlation and improved performance in quantifying tannic acid content.

Overall, this thesis makes a significant contribution to the field of the tea industry by assessing the enhanced prediction of tea quality based on the chemical composition of black and green tea samples. The integration of MIP electrochemical, NIR spectroscopy, and AI techniques significantly improves the accuracy and reliability of tea quality predictions, providing producers and consumers with a robust framework for distinguishing high-quality tea, ultimately enhancing market reputation and consumer confidence.

Table of Contents

List of Tables

List of Figures

Abbreviation

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
NIR	Near-Infrared Spectroscopy
MIP	Molecularly Imprinted Polymers
EGCG	Epigallocatechin-3-Gallate
TAN	Tannic Acid
E-Nose	Electronic Nose
E-Tongue	Electronic Tongue
E-Eye	Electronic Eye
PCA	Principal Component Analysis
UMAP	Uniform Manifold Approximation and Projection
SVM	Support Vector Machine
RF	Random Forest
PSO	Particle Swarm Optimisation
ACO	Ant Colony Optimisation
DFO	Dragonfly Optimisation
CCA	Canonical Correlation Analysis
LSTM	Long Short-Term Memory
THEO	Theophylline
CAFF	Caffeine
DT	Decision Tree
SNN	Siamese Neural Network
PLSR	Partial Least Squares Regression

CHAPTER 1

INTRODUCTION AND SCOPE OF THESIS

This chapter starts with an introduction about the tea, its nutrient benefits, and the importance of quality assessment of tea, particularly for black and green tea, followed by an overview of tea chemistry. In the tea chemistry section, the focus is on tea processing and a brief discussion of the chemical compounds responsible for its colour, aroma, and taste. It also discusses the traditional methods of the tea quality prediction approach, such as human tea tasting and chemical analysis techniques. In addition, it covers the details of the pattern recognition method involved in this present study. Then, it provides summarized details of the existing research in the field of tea quality assessment. Finally, the chapter ends with the aim and objective and the scope and significance of the study.

LIST OF SECTION

- Introduction
- Tea and Its Chemistry
- Tea Quality Evaluation by Instrumental Methods
- Quality Evaluation with Electronic Sensor System
- Brief Description of an Electronic Tongue
- Pattern Recognition Methods
- Literature Review
- Objective of the Work
- Scope of the Thesis
- Thesis Organisation
- Conclusion

CHAPTER 1

INTRODUCTION AND SCOPE OF THE THESIS

1.1. Introduction

The increasing complexity of product quality assessment across industries has made data analysis an important component for ensuring consistency, reliability, and process optimisation. In particular, the application of machine learning (ML) techniques has gained significant traction in fields such as tea manufacturing, where precise quality evaluation and process optimisation are paramount. Reported high classification performance in tea blending assessment and grade classification demonstrates the potential of these data-driven methodologies. Beyond classification, advanced ML approaches, such as reinforcement learning and predictive modelling, are being leveraged to optimise key manufacturing processes, including fermentation. This enables real-time monitoring and adjustment, ensuring not only consistent product quality but also reduced waste. Deep convolutional neural networks (DCNNs), for example, have demonstrated near-perfect accuracy in classifying fermentation levels by analysing colour and texture features. This integration of DL with traditional analytical techniques enhances the precision of tea quality evaluation while simultaneously supporting more sustainable practices within the industry. By automating specific processes and reducing dependence on human expertise, these technologies ensure strict quality control and meet the growing demand for high-quality tea products. As research in this area continues to evolve, the fusion of ML methodologies with advanced sensing technologies promises to transform the tea industry, offering innovative solutions for both quality assurance and process optimisation. This synergy enhances the reliability of evaluations and ensures that consumers can enjoy consistently high-quality tea, supported by scientific rigour. The ongoing development of more sophisticated methods will undoubtedly enrich the global tea experience, blending tradition with cutting-edge technology to meet the demands for authenticity and health-conscious products.

Tea is considered a rich beverage that people have consumed since ancient times because of its esteemed flavours and health benefits [1]. Embedded in ancient traditions and rituals, these practices are deeply integrated into the social and cultural structures of numerous societies worldwide. According to delicate records, from the subtle taste of green tea to the robust flavour of black tea, each variety offers a unique tasting experience that promotes appreciation

among enthusiasts [2]. The global tea market is expected to exceed US$1,481.6 billion by 2027, growing at an annual growth rate of 6.4%. Therefore, the importance of ensuring high-quality tea cannot be overstated [3].

Similarly, nutritional research continues to highlight the health benefits of tea, primarily due to its biologically active compounds. The most common bioactive polyphenols, the prominent catechins and flavonoids, contribute significantly to their effects on preventive health. Research indicates that these compounds can help combat free radicals, reduce inflammation, and prevent cancer, which is crucial for lowering the risk of long-term health issues such as heart disease, obesity, and certain types of cancer. It is there. Additionally, components such as L-theanine and caffeine improve cognitive function and provide a calming effect, which also promotes tea consumption for mental health [4].

Several factors, including tea varieties, geographical diseases, processing methods, and storage, influence the complex chemistry of tea. For example, tea processing involves essential steps such as reduction, rolling, oxidation, drying, etc. [5]. Each step influences the chemical composition and sensory properties of the final product, ranging from its scent to its mouthfeel. Oxidation levels are significant. For example, black tea undergoes complete oxidation, resulting in a robust taste profile, whereas green tea remains minimally processed to preserve its fresh and fruity notes. Understanding this process is essential for both manufacturers and consumers who want to appreciate the nuances of tea quality [6].

Although the richness of tea varieties and their flavours present challenges for value assessments, they have traditionally been modelled in a specific way. Traditional approaches mainly depend on the evaluation of sensory input directed by skilled tasters, who measure several features like texture, fragrance, and aroma [7]. Although these approaches introduce subjectivity and variability, they also create potential contradictions in assessing the quality of tea. The fact that many consumers lack the knowledge and expertise to determine tea quality based on sensory characteristics is an urgent need for a more objective approach to quality assessment [8].

Modern advances in analytical technologies, such as high-performance liquid chromatography (HPLC) and Near-Infrared Spectroscopy (NIR) [9], have significantly improved tea quality evaluation [10]. HPLC enables the detailed profiling of key compounds, including catechins, caffeine, and other phytochemicals, by separating, identifying, and quantifying the various components within a tea sample [11]. On the other hand, NIR provides a rapid, non-destructive

method for assessing tea quality by measuring infrared absorption spectra, offering insights into chemical composition and moisture content, making it ideal for routine quality checks in commercial environments [12]. Despite the numerous advantages of the aforementioned precise instruments, their high cost and dependence on expert operators in laboratory settings have led to the development of more affordable, easy-to-use handheld devices for quality measurement. The electronic replication of biological sensing systems has already provided a reliable pathway in this direction. The Electronic-Nose (E-Nose) mimics human odour detection to analyse tea's scent, ensuring consistency in aroma across production batches, while the E-Tongue quantifies taste profiles, such as sweetness, bitterness, and sourness. The Electronic-eye (E-Eye) utilises image processing to evaluate visual characteristics, including colour and shape, which impact market acceptance [13].

Recently, ML and advanced data analysis techniques have emerged as transformative tools in the tea industry, particularly for quality evaluation and assessment. Traditional methods of assessing tea quality, such as sensory analysis and chemical testing, are often subjective, time-consuming, and labour-intensive. The integration of ML with modern technologies like chemical sensors and computer vision has provided a more objective, efficient, and scalable approach to tea quality assessment.

DL algorithms have been widely applied across various stages of the tea production process. For instance, supervised learning techniques such as Support Vector Machines (SVM) and Random Forests (RF) have been used for tasks like chemical composition analysis and flavour profiling [14]. These methods enable rapid prediction of tea quality attributes by analysing large datasets of sensory and chemical parameters. Similarly, deep learning (DL) models, such as Convolutional Neural Networks (CNNs), have demonstrated high accuracy in image-based tea classification tasks, including grading needle-shaped green tea and identifying the quality grades of fresh tea leaves. For example, studies using enhanced YOLOv8x models with attention mechanisms have achieved precision rates exceeding 95% in classifying fresh tea leaves based on their visual features [15].

Computer vision systems combined with ML models have also proved effective in evaluating specific quality indicators like colour and texture. In one study, a decision tree-based AdaBoost (DT-AdaBoost) model achieved a correct discrimination rate of 98.5% for needle-shaped green tea by analysing its colour features. Another notable application involved using ResNet models with attention modules to improve the efficiency and accuracy of tea blending assessments and

grade classifications, achieving over 95% accuracy for oolong and black teas. Beyond classification tasks, ML techniques are being employed to optimise processes in tea manufacturing. Reinforcement learning and predictive modelling facilitate the real-time monitoring of key stages, like fermentation, to ensure consistency in product quality while minimising waste. For example, DCNNs have been used to classify fermentation levels with near-perfect accuracy by analysing colour and texture features.

The integration of ML with traditional analytical methods not only enhances the precision of tea quality evaluation but also supports sustainable practices in the industry. By automating processes and reducing reliance on human expertise, these approaches ensure consistent quality control while meeting the growing demand for high-quality products. As research progresses, the fusion of ML-driven methodologies with advanced sensing technologies promises to revolutionise the tea industry by offering innovative solutions for quality assurance and process optimisation [15].

Moreover, the landscape of evaluating tea quality has evolved through the integration of traditional sensory methods with advanced instrumental technology. This overall approach facilitates strict quality control, ensuring that consumers can enjoy high-quality tea products backed by scientific analysis and reliability [16]. These technologies continue to demonstrate promising potential for objective and reliable tea quality evaluation, significantly enhancing its quality, and promises innovation and improvements in human-rated industries. The consumer's demands to address authenticity will ultimately enrich this advancement in the global tea experience and weave together tradition, science, and appreciation in every cup [17].

In spite of considerable advances made in assessing the quality of tea based on spectroscopy, chemometrics, and DL algorithms, certain deficiencies are observed in the current practices. Most traditional techniques involve independent use of spectroscopy or sensory-based analysis, which hampers robustness and reproducibility of results. In addition, relatively few studies have been conducted to examine the use of MIP-based electrochemistry in conjunction with NIR spectroscopy in data fusion for prediction of tea quality. In addition, most of the current systems do not provide multi-output prediction with optimized DL models. Hence, there exists a need to develop an integrated system that could predict the quality of tea based on sensor fusion and feature optimisation.

1.2. Tea and Its Chemistry

Tea, derived from the Camellia sinensis plant, is a complex beverage rich in a wide range of chemicals that influence its taste, aroma, colour, and health benefits. The composition of tea varies significantly depending on multiple factors, including the tea cultivar, geographical conditions, and the processing methods employed. At its core, tea chemistry is primarily characterised by several key components: polyphenols, amino acids, carbohydrates, proteins, caffeine, and essential oils [18].

Polyphenols are the most studied and abundant compounds in tea, accounting for a significant portion of its health benefits. Among the polyphenols, catechins, specifically epicatechin (EC), epicatechin gallate (ECG), epigallocatechin (EGC), and epicatechin, are notable for their potent antioxidant properties. These compounds contribute to tea's ability to combat oxidative stress, potentially reducing the risk of chronic diseases such as heart disease and cancer [19].

In addition to polyphenols, tea contains a substantial amount of amino acids, with L-theanine being the most prominent. L-theanine is recognised for its calming properties and ability to induce relaxation without causing drowsiness. This unique compound modulates the effects of caffeine, creating a balanced alertness that one may appreciate.

Caffeine, while well known for its stimulating effects, is just one part of tea's intricate chemistry. The caffeine content in tea can vary significantly, typically ranging from 20 to 60 milligrams per serving, depending on the type of tea and the brewing method used. The presence of other compounds modifies caffeine's effect, with L-theanine promoting a calming influence that counters caffeine's more stimulatory properties. The essential oils in tea play a crucial role in creating its unique aroma profiles. These volatile compounds are released during the brewing process, contributing significantly to the sensory experience of consuming tea. The interaction between aromas and flavour components influences the overall perception of quality in teas [20].

The chemical composition of tea is not only valuable for its health benefits but also for the sensory attributes that define different tea varieties [21]. The complex arrangement of these components can lead to diverse taste profiles, ranging from the bitterness of black tea to the delicate grassy notes of green tea and the fragrant floral undertones found in oolong teas.

1.2.1. Black Tea Processing

The processing of tea leaves is a critical step that transforms fresh leaves into the various types of tea available globally, with unique characteristics. The degree of oxidation distinguishes the primary types of tea, namely white, green, oolong, and black. Each processing method involves a series of steps designed as illustrated in Figure 1.1 to create specific flavour profiles and chemical compositions [22]. In addition to the carefully controlled steps that transform freshly picked tea leaves into the diverse varieties of tea enjoyed worldwide, each processing stage significantly impacts the chemical processes involved, including withering, rolling, oxidation, drying, and sometimes ageing.

- **Withering**: The first step in processing tea involves wilting or softening the fresh leaves by subjecting them to controlled airflow and temperature. This step reduced moisture content and made the leaves pliable, which is crucial for subsequent rolling [23]. Withering not only initiates the enzymatic processes that influence flavour but also prepares the leaves for better handling in the following steps. The duration and environmental conditions during withering can significantly impact the concentration of certain compounds, particularly polyphenols.
- **Rolling**: The mechanical process involves crunching or rolling the withered tea leaves to rupture the cell walls and initiate the extraction of essential oils and juices from the leaf cells. Rolling also increases the surface area of the leaves, facilitating oxidation, a crucial process for producing certain types of tea, particularly black tea. The rolling methods can influence the leaf structure, which in turn affects the infusion characteristics and flavour profiles [24].
- **Oxidation (Fermentation):** Oxidation is the most crucial stage determining the final characteristics of tea. During this, polyphenol oxidase enzymes catalyse the oxidation of catechins, transforming them into complex flavonoid compounds. The degree of oxidation defines how we classify tea; fully oxidised leaves result in black tea, partially oxidised leaves yield oolong tea, and minimally oxidised leaves produce green tea. The oxidation time, temperature, and humidity conditions during this process are carefully monitored to achieve the desired taste and appearance [25].
- **Drying**: after oxidation, the leaves must be dried to halt any further enzymatic activity. Drying removes moisture and locks in the flavours developed during processing.

Several methods are employed for drying tea, including hot air drying, pan firing, and sun drying, each contributing distinct qualities to the final product. This process is key to preserving the chemical composition of the tea, enhancing its shelf life, and playing a role in the development of aromatic compounds [26].

- **Ageing:** Some tea, particularly certain varieties of pu-erh tea, benefits from an ageing process. This can involve controlled storage environments that allow chemical reactions to continue over time, resulting in enhanced flavours and aromas. The ageing process can significantly change the tea's taste profile, often smoothing out harsh notes while developing unique earthy undertones.

The intricacies of tea processing underscore its profound impact on the chemical composition of the tea and its sensory properties. Through variation in processing methods and conditions, tea producers can create diverse profiles that cater to different consumer preferences and market demands. Understanding these processes is essential for both the producers aiming to perfect their craft and consumers seeking to appreciate the nuances in their tea experience [27].

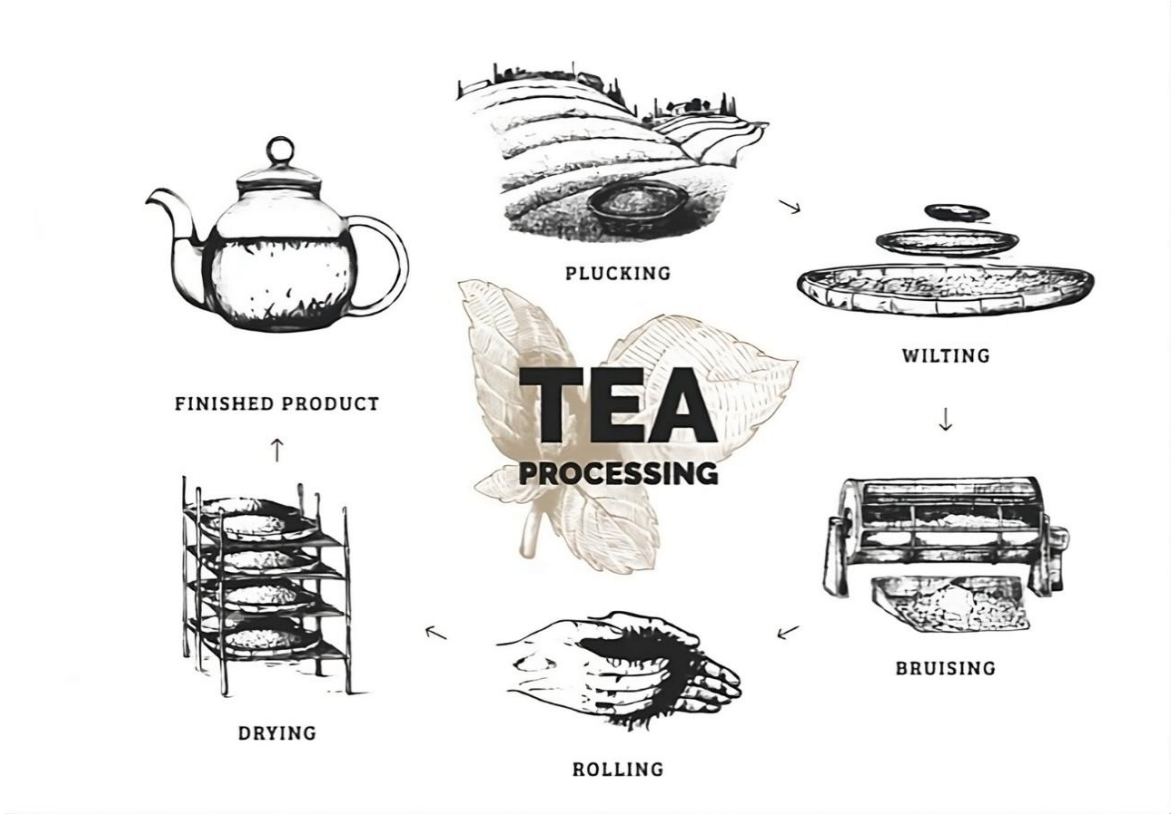


Figure 1.1 Tea Processing

1.2.2. Tea Constituents

Tea chemistry is remarkably complex, with fresh tea leaves containing thousands of chemical compounds. These compounds break down, form complexes, and create new compounds during processing. Steeping tea volatilises compounds, known as the aroma complex, and nonvolatile compounds, stimulating our senses [28]. This complexity makes it difficult to pinpoint specific chemicals responsible for particular tastes; however, tea chemicals are grouped into broad categories to help understand their effects during processing. The most important compounds in fresh tea leaves include polyphenols, amino acids, enzymes, pigments, carbohydrates, methylxanthines, minerals, and volatile flavour and aroma compounds. These components contribute to the appearance, aroma, and taste of teas. During processing, these compounds change to produce the finished teas [29].

1.2.2.1. Polyphenols

Polyphenols are abundant in tea leaves. Flavanols transform into theaflavins and thearubigins during oxidation, contributing to the dark colour and robust flavours of oxidised teas. Flavonols, flavones, isoflavones, and anthocyanins influence the colour and taste of the infusion. Tea polyphenols are significant natural antioxidants, with their antioxidant capacity influenced by their spatial configuration and hydroxyl groups [30].

- **Catechins**: predominant polyphenols in fresh tea leaves, especially EGCG, EGC, EC, and ECG. Catechins are potent antioxidants. In non-oxidised teas, such as green tea, catechins remain essentially unchanged; however, in oxidised teas, like black tea, catechins convert to theaflavins and thearubigins through enzymatic oxidation.

- **Theaflavins**: The orange-red pigments form during catechin oxidation, especially in black tea processing. Theaflavins contribute to the colour, briskness, and astringency of black tea. They form through the enzymatic oxidation of catechins.

- **Thearubigins**: These brown pigments also form during catechin oxidation. Larger and more complex than theaflavins, thearubigins contribute to the colour and body of black tea. They also form through the enzymatic oxidation of catechins [9].

1.2.2.2. Amino Acid

Tea brewing contains considerable amounts of amino acids, including aspartic acid, glutamic acid, arginine, alanine, tyrosine, and theanine. Theanine, a non-protein amino acid unique to tea, has a positive effect on relaxation, cognitive performance, emotional status, sleep quality,

and overall health. Theanine levels can decrease slightly during processing but generally remain stable, contributing to the overall flavour profile [31].

1.2.2.3. Aromatic Compounds

Volatile substances contribute to the flavour and aroma of tea. These compounds easily enter the air from tea leaves or tea liquor, reaching our olfactory system as a vapour. The complex aroma comprises hundreds, possibly thousands, of volatile, flavoured compounds formed during processing rather than existing in leaves. The specific VOC profiles depend on the type of tea and its processing conditions. For example, exlinalool and linalool oxide produce floral-sweet scents. eGeraniol and phenylacetaldehyde produce floral aromas. Nerolidol, benzaldehyde, methyl salicylate and phenyl ethanol produce fruity flavours. -hexenal, cis-3-hexanol, and β-ionone-iononenproducers flavours [32].

1.2.3. Conventional Methods for Tea Quality Evaluation

Evaluating tea quality is crucial for ensuring customer satisfaction and maintaining market standards. Traditional methods primarily depend on sensory assessments by trained experts. However, these methods are subjective, labour-intensive, and often lack consistency. Sensory evaluation involves visual inspection, aroma assessments, taste evaluation, and infusion appearance. Physicochemical analysis provides a more objective approach by measuring moisture content, ash content, water extract, crude fibre, caffeine content, and polyphenol content.

1.2.3.1. Sensory Evaluation

The evaluation uses human senses to assess the quality of tea. It also includes both internal quality and appearance evaluation.

- ***Visual Inspection****:* Experts assess the dry tea leaves for attributes such as colour, size, and uniformity. The colour should align with the type of tea.
- ***Aroma Assessments****:* The aroma of dry and infused tea leaves is evaluated to identify aromatic notes like floral, fruity, vegetal, and smoky.
- ***Taste Evaluation****:* The tea's taste, or liquor, is assessed for briskness, astringency, body, and flavour. The balance of these attributes determines overall quality.
- ***Infusion Appearance****:* The colour and clarity of the tea infusion are evaluated, with the colour expected to be bright and clear, free from cloudiness or sediments.

1.2.3.2. Limitations of Sensory Evaluation

Sensory evaluation has inherent limitations. It is subjective, depending on individual preferences and experience, leading to variations between evaluators and inconsistency for the same evaluator over time. It is also labour-intensive, requiring trained personnel, which increases costs and time. Moreover, it provides qualitative data without quantifying specific chemical compounds [33].

1.2.3.3. Physiochemical Analysis

Physicochemical analysis provides a more objective means of evaluating tea quality through measurements of various chemical properties.

- **Moisture Content**: Excessive moisture can cause mould growth and spoilage. Moisture content is measured using drying ovens or moisture analysers.
- **Ash Content**: Ash content indicates the mineral content and is determined by burying the tea sample and weighing the residue.
- **Water Extract**: This process measures the soluble solids in tea by using hot water to extract the tea and then measuring the dissolved solids.
- **Crude Fib**re: This measures the indigestible plant material in tea by digesting the tea sample with acid and alkali and weighing the residue.

1.3. Tea Quality Evaluation by Instrumental Methods

The assessment of tea quality has undergone significant evolution, transitioning from traditional sensory evaluations to sophisticated instrumental techniques. These methods offer objective, reproducible, and rapid assessments, overcoming those limitations of subjective human evaluations, which are prone to variability and require experienced personnel. The instrumental method provides detailed chemical profiles and physical characteristics, essential for standardising tea quality and meeting consumer expectations in the global market. This section explores several key instrumental methods used in tea quality evaluation [34].

1.3.1. Quality Evaluation with HPLC

HPLC is a cornerstone technique in tea quality analysis, prized for its ability to separate, identify, and quantify various chemical compounds present in tea leaves. HPLC's high resolution and sensitivity make it an indispensable tool for researchers and quality control professionals in the tea industry.

Fundamentals of HPLC: The operators of HPLC separate compounds based on their interaction with a stationary phase and a mobile phase. The stationary phase is typically a solid material packed into a column, while the mobile phase is a liquid solvent that carries the sample through the column. As the sample passes through the column, its components interact differently with the stationary phase, causing them to elute at different times. These separated compounds are then detected and quantified using various detectors, such as UV-Vis detectors, fluorescence detectors, or mass spectrometers. HPLC remains at the forefront of tea research, providing essential chemical insights for quality control, product development, and understanding the health benefits of tea [35].

1.3.2. Quality Evaluation with NIR

NIR spectroscopy provides rapid, non-destructive analysis of tea, making it ideal for real-time monitoring of quality attributes throughout the supply chain. Beyond fundamental components analysis, NIR provides insights into complex structural and physical properties.

Innovative application in tea assessments

- **Tea variety identification**: NIR spectroscopy can differentiate tea varieties based on their spectral fingerprints, supporting breeding programs and varietal authentication.
- **Forecasting sensory characteristics**: NIR models can estimate sensory traits like scent, flavour, and texture by linking spectral data to sensory panel evaluations.
- **Observation of the withering process**: The process used to track moisture reduction and chemical alterations during withering, allowing for accurate management of this critical processing phase.
- **Evaluation of fermentation level:** The spectroscopy aids in evaluating the level of fermentation in black tea production by monitoring alterations in polyphenol makeup.
- **Adulterant identification**: Detect adulterants like foreign plant substances or synthetic colourants in tea samples.

Accurate NIR analysis requires robust calibration models developed using representative tea samples and reference methods. Moreover, in spectral preprocessing, the appropriate technique removes noise and corrects the scattering effects, improving model accuracy. Furthermore, regular instrument calibration and standardisation are necessary to ensure data comparability across different instruments and laboratories [36].

Emerging direction

- **Hyperspectral imaging:** The combination of NIR spectroscopy with spatial information, providing detailed maps of chemical composition across tea leaves or infusions.
- **Portable NIR Devices**: Development of portable NIR spectrometers enables on-site quality assessments in tea plantations, factories, and retail outlets.
- **Integration with IoT**: Linking NIR data to IoT platforms allows for real-time monitoring of tea quality parameters and data-driven decision-making.

NIR spectroscopy is transforming tea quality assessments by providing rapid, non-destructive, and information-rich analysis, supporting enhanced optimisation of quality control processes.

1.3.3. Quality evaluation with thermogravimetric analysis with Fourier Transform-Infrared (TGA-FTIR)

Thermogravimetric analysis coupled with TGA-FTIR provides a comprehensive understanding of tea's thermal behaviour and volatile emissions. Beyond the fundamental decomposition analysis, TGA-FTIR is used to characterise complex thermal degradation pathways and identify key aroma compounds.

Specialised uses in Tea Research

- **Characterising Thermal Stability**: TGA measures the thermal stability of tea components, which is crucial for determining optimal storage and processing conditions.
- **Analysing Volatile Signatures**: FTIR identifies volatile compounds released during heating, providing insights into tea aroma, profile, and how it changes with temperature.
- **Assessing Organic Matter Composition**: TGA-FTIR quantifies the amount and composition of organic matter in tea, indicating purity and quality.
- **Studying Tea Extracts**: This technique is used to analyse tea extracts, identifying compounds responsible for bioactivity and health benefits.

TGA is conducted under controlled atmospheres, such as nitrogen or oxygen, to simulate various processing or storage conditions, thereby revealing oxidative and thermal degradation pathways. Then FTIR identifies and quantifies the gases evolved during TGA, linking specific thermal events to the release of key aroma compounds. Furthermore, kinetic models derived from TGA data provide insights into the thermal degradation mechanism of tea compounds, aiding in the optimisation process.

1.3.4. Advancements in TGA-FTIR

The high-resolution TGA, with improved resolution, allows for more precise measurements of weight changes, enabling detailed analysis of complex tea matrices. Rapid FTIR scanning captures the transient changes in volatile emissions, providing a dynamic view of aroma release during heating. Combining TGA-FTIR with mass spectrometry enhances the compound identification and quantification, providing a more comprehensive analysis of tea volatiles.

TGA-FTIR offers valuable insights into tea composition and thermal behaviour, enabling better control of the process, storage, and product development [37].

1.4. Quality Evaluation with Electronic Sensor System

1.4.1. Quality Evaluation with Electronic Eye

The E-eye is a bionic-inspired diagnostic instrument designed to replicate human visual examination, providing an impartial assessment of the visual characteristics of tea. In contrast to conventional methods that depend on personal human assessment, the E-eye provides a reliable and quantifiable evaluation of tea leaves and brews. In addition to basic leaf grading, it is proficient in identifying visual flaws and automating the sorting procedure, ensuring improved product purity and quality [38]. The E-eye is capable of categorizing various tea samples and assessing product quality based on appearance-associated traits [39].

1.4.2. Quality evaluation with Electronic Nose

The E-Nose is a rapid, objective tool for assessing tea aroma profiles, simulating the human olfactory system. Beyond simple aroma fingerprinting, E-Nose technology is used to monitor subtle aroma changes and detect off-odours, supporting quality control and process optimisation.

It is also used in real-time fermentation monitoring in black tea by tracking aroma changes, enabling precise adjustments to optimise fermentation conditions. Additionally, it detects off-

odours that indicate spoilage or contamination, thereby ensuring the safety and quality of the product. Likewise, in geographical origin authentication, an E-nose can differentiate teas from different regions based on their unique aroma profiles, thereby aiding in authentication [40]. E-nose technology provides rapid, objective aroma assessment, enhancing quality control and process optimisation in the tea industry [41].

1.4.3. Quality Evaluation with Electronic Tongue

The electronic tongue (E-Tongue) is a measurement device created to replicate human taste sensing, providing an unbiased means to evaluate the flavour profile of tea. Rather than relying on personal human taste panels, the E-Tongue utilises a variety of chemical sensors to identify and measure nonvolatile compounds that contribute to the critical flavour of tea, such as sweetness, bitterness, sourness, and umami. This knowledge provides a reliable and measurable assessment of taste, which is crucial for quality assurance and correction within the tea industry [42].

1.4.3.1. Applications in Tea Quality Evaluation

The E-Tongue has diverse applications in tea quality assessments, enabling a more detailed and consistent analysis compared to traditional methods [43].

- **Taste Profiling**: The E-Tongue can differentiate between various types of tea based on their unique taste profiles. This is particularly useful for quality control during tea production and for product development, allowing manufacturers to find a delicate tune flavour to meet specific market demands.
- **Grading and classification**: By integrating the E-tongue data with chemometrics techniques, tea samples can be classified according to their quality grade based on taste attributes. Studies have demonstrated high accuracy in classifying tea quality using methods such as Linear Discriminant Analysis [44].
- **Monitoring Taste Changes during Processing**: The E-Tongue can be used to monitor the changes in taste profiles during tea processing stages, such as fermentation, oxidation, and drying. This allows for real-time adjustments to optimise this process and ensure the desired taste characteristics are achieved.

- The E-Tongue can assess the taste characteristics of tea blends and predict their overall sensory quality. This is valuable for creating new blends with consistent and appealing taste profiles.
- **Astringency Measurement**: E-Tongue can distinguish tea samples with different astringency levels with high accuracy, typically exceeding 85%.

1.4.3.2. Advantages of E-Tongue Technology

The E-Tongue offers several advantages over traditional sensory evaluation methods.

- **Objectivity**: The E-Tongue provides an objective measure of taste, eliminating the subjectivity associated with human sensory panels.
- **Reproducibility**: The E-Tongue delivers consistent and reproducible results, ensuring reliability in taste assessments.
- **Non-Destructive**: E-tongue analysis is non-destructive, preserving the sample for future testing if needed.
- **Mimicking Human Sensory Awareness**: The e-tongue rapidly mimics the sensory awareness of human taste by using the fingerprints of response signals in tea samples.
- **Cost and Time Saving**: E-Tongue reduces the cost and time calculations with preparation and preserving sensory panels [45].

1.4.3.3. Future Trends and Developments

The field of E-Tongue technology is continuously evolving. Current research focuses on improving sensor selectivity, integrating AI for better data processing, and developing more robust and portable E-tongue systems. The integration of triboelectric sensors, motivated by the natural sense of taste, is another field undergoing active advancement. These developments are expected to enhance the precision, dependability, and functionality of E-tongue in assessing tea quality and various sectors within the food industry. By addressing existing constraints and integrating advanced technologies, E-tongues are poised to be a crucial asset for ensuring quality, uniformity, and innovation in the global tea market.

1.4.4. Brief Description of Electronic Tongue

E-tongues provide numerous benefits, such as impartiality, reliability, swift analysis, and savings in both time and cost. They are able to identify tea samples with varying astringency

levels with outstanding precision. They also quickly replicate the sensory perception of human taste and lower the expenses and time involved in training and sustaining sensory panels [46].

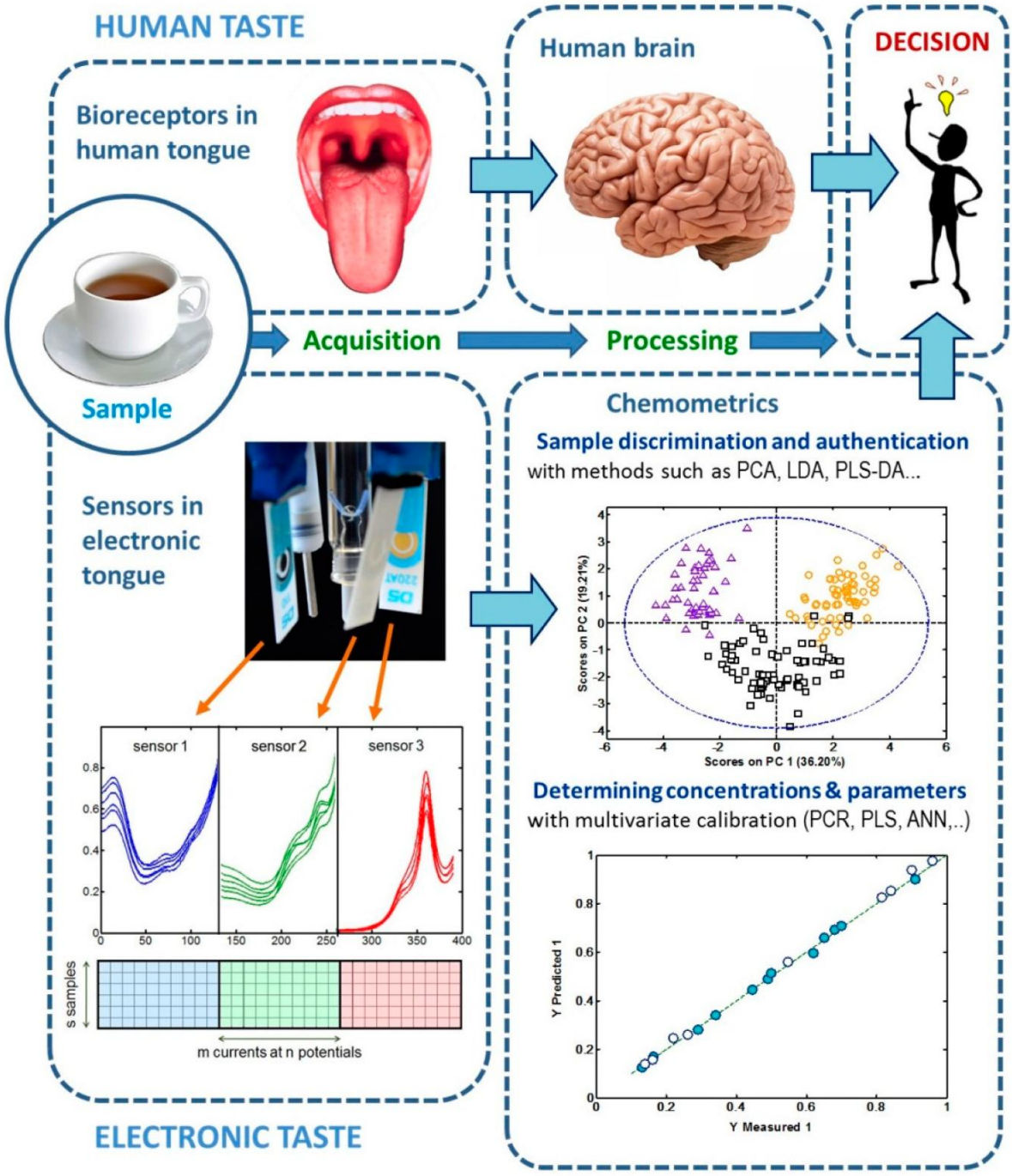


Figure 1.2 E-Tongue [48]

The E-tongue is composed of an electrochemical cell, a sensor array, and a pattern recognition system. The electrochemical cell consists of working electrodes, a reference electrode, and a counter electrode [47]. The first generation e-tongue array consists of widely tuned, non-specific potentiometric electrodes made of noble metals. These sensors interact with flavour

compounds within a sample, resulting in changes in their electrical characteristics. These alterations are subsequently examined through pattern recognition systems and AI algorithms to determine and measure the taste profile.

The data generated by the e-tongue is complex and requires sophisticated data processing to extract meaningful information as shown in Figure 1.2. Chemometrics is a scientific field that applies mathematical, statistical, and additional techniques, utilising formal reasoning, to develop or select the best measurement methods and experiments, and to extract the most chemical information through the analysis of chemical data. These procedures are designed to identify patterns in the sensor data and associate them with specific taste characteristics or quality metrics.

A significant benefit of the e-tongue is its capacity to replicate the scores of tea tasters through electronic methods. It provides a more objective and reliable evaluation compared to human sensory panels, which can be prone to fatigue and inconsistencies. The e-tongue enhances both speed and user-friendliness. A complete tea testing procedure can be executed with just one mouse click, yielding results within minutes. This renders it a valuable resource for consistent quality assessment in the tea sector.

Furthermore, utilising an e-tongue can help avoid a decline in the accuracy of sensory panellist ratings, which may result from their exhaustion and neural fatigue. The e-tongue can accurately identify tea samples with varying levels of astringency, achieving an accuracy rate exceeding 85%. E-tongues, when combined with additional electronic sensing systems, have the potential to establish new standards for bioactive products that meet global market needs.

1.4.5. Sensor Technology

The development of effective electronic tongues for tea quality assessment relies heavily on the selection of appropriate sensor materials, each chosen to address specific challenges in analysing such a complex beverage. Noble metal electrodes, such as gold and platinum, are often employed for their excellent conductivity and chemical stability, ensuring reliable and repeatable measurements in diverse tea matrices. As a cost-effective alternative, graphite electrodes are also widely used, and their performance can be significantly enhanced through modification with nanoparticles or conductive polymers to improve sensitivity towards key tea compounds. A particularly innovative approach involves Molecularly Imprinted Polymers (MIPs), which are not merely a material but a powerful sensor development technique. MIPs are used to create synthetic receptors with binding sites specifically tailored to target marker

molecules in tea, such as catechins or theaflavins. By integrating these highly selective MIPs onto an electrode surface, sensors can precisely identify and quantify these quality-defining compounds, thereby significantly advancing the objectivity and accuracy of electronic taste analysis for tea grading and authentication.

MIPs represent an innovative approach for the selective identification and extraction of key compounds in tea, particularly in advanced methods for detecting and evaluating tea quality. These specifically designed polymers function as artificial receptors. Crafted with unique binding sites that demonstrate a strong attraction to target molecule increases the sensitivity and selectivity of the sensor.

The development of MIPs requires a template molecule, which is a particular compound of interest present in tea, such as caffeine, catechins, Theaflavins. This template directs the creation of binding sites throughout the polymerisation process. Functional monomers, selected for their capacity to engage with the template via hydrogen bonding, hydrophobic interactions, or other chemical forces, are combined with it. A cross-linking agent is subsequently added to form a firm, three-dimensional polymer structure. After polymerisation, the extraction of the template molecule creates accurately formed cavities in the polymer matrix that correspond to the target analyte. The resulting MIP is now prepared to selectively attach to and isolate the desired compound from a tea sample. Nonetheless, it is crucial to recognise that the imprinting procedure can occasionally retain leftover template molecules in the polymer, potentially interfering with the analysis [49].

In the realm of tea quality assessment, MIPs provide numerous significant benefits. Their elevated selectivity enables the separation and pre-concentration of particular compounds, even amid various other interfering substances typically present in tea extracts. This increased selectivity results in improved accuracy and sensitivity in subsequent analytical measurements, such as HPLC, Gas Chromatography-Mass Spectrometry **(**GC-MS), or electrochemical detection.

The Figure 1.3, illustrates the process of MIP development, beginning with a template molecule and functional monomers that form a template monomer complex. This complex undergoes polymerization with the addition of cross-linkers to create a binding site specific to the template molecule. Following this, the template is removed, leaving a cavity that retains the shape of the template, which can subsequently bind the target analyte. Alongside this, the schematic for the

MIP-EGCG development represents the overall procedure for synthesizing these polymers tailored for specific applications, ensuring effective molecular recognition.

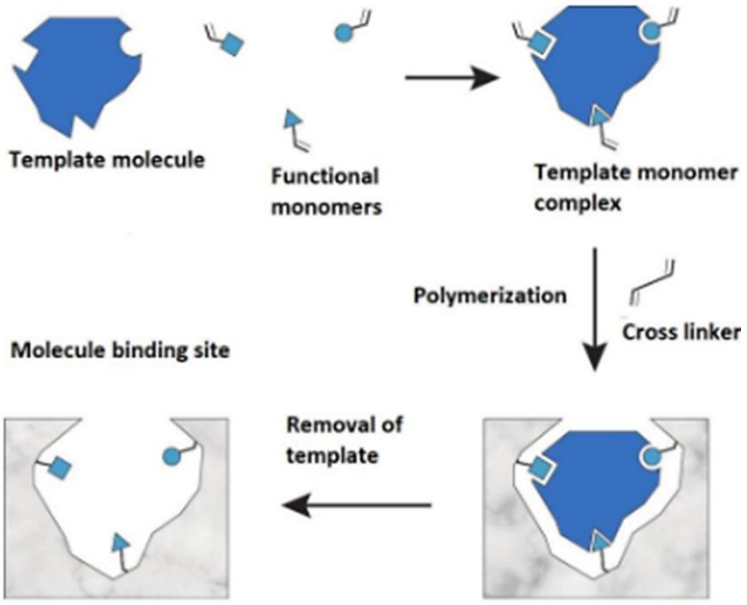


Figure 1.3 MIP Development

The electronic tongue was initially conceived to mimic human taste, offering a solution to the subjectivity of sensory panels. It has since evolved from simple sensor arrays into a sophisticated analytical instrument. This evolution incorporates advanced materials and intelligent data processing, enabling highly objective and detailed taste evaluation.

1.4.6. Electronic Tongue Measurement Methods

The sense of taste is considered to be the biological process which permits humans to observe a larger range of flavours, a capability that distinguishes the larger range of tastes with the comparatively smaller and few taste buds which are enabled with the incorporation of signals provided at the taste buds and the adjacent examinations and the gratitude from the NN (Neural Networks) in the brain. The E-Tongue utilises non-selective sensors to replicate taste receptors and employs an Artificial Intelligence (AI) algorithm to simulate brain examination operations, based on stimuli derived from natural taste observation methods.

In contrast to the natural taste, artificial taste offers a broader range of options and provides the advantages of impartiality and efficiency. Those projecting features facilitate the integration of E-tongues in different sectors. For example, the E-tongues are utilised for contrasting the quality of red wine and beer to identify productive areas for rice and ginger. In the

environmental monitoring sector, E-tongues are used for the detection of heavy metal ions. In the field of pharmaceuticals, devices are utilised to modify the taste of medications. The conventional E-tongues are created using electrochemical methods, including voltammetry and potentiometry [50].

1.4.6.1. Potentiometric Method

Potentiometry depends on polymeric membrane Ion-Selective Electrodes (ISEs), a well-established analytical technology integrated into the physiological testing of key electrolytes. The potentiometric sensors offer several advantages, including being tiny, fast-responding, easy to use, and cost-effective, as well as being resilient to colour and turbid interventions.

Figure 1.4 depicts the visual representation of potentiometric sensors. Additionally, the ISEs have some individual characteristics that generate information related to the activities of ions, which is considered to be different from other examination techniques that provide complete concentration. They are independent of the sample volume, allowing for a significant reduction in sample volume without compromising the detection range. Those properties result in the ISEs for a unique indicator conductor [51].

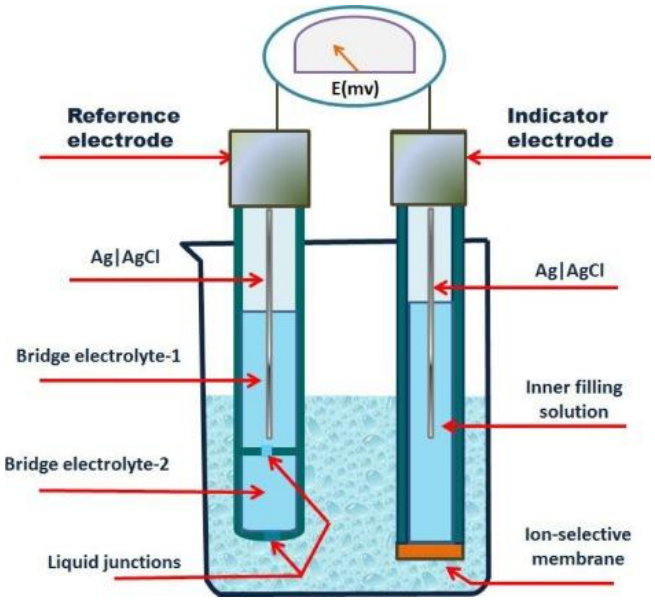


Figure 1.4 Potentiometric sensors [52]

1.4.6.2. Voltammetric Method

In voltammetry, the potential is integrated to the working electrode, which results in the current extracted during the reduction of redox-active species on the electrode plane being calculated. The utilisation of voltammetry involves certain advantages such as being sensitive, adaptable,

comfortable and robust. In contrast to potentiometry, voltammetry is less impacted by electrical disturbances and contains a more favourable signal-to-noise ratio. It provides numerous diverse examination methods, including cyclic, stripping, and pulse voltammetry.

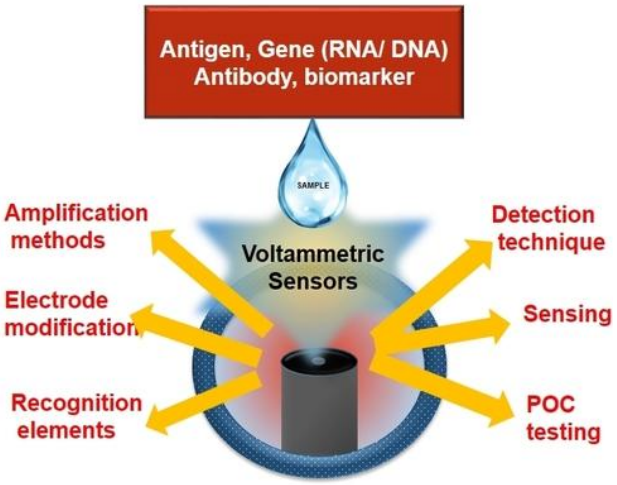


Figure 1.5 Voltammetric Sensors [53]

Figure 1.5 illustrates the visual representation of voltammetric sensors. Additionally, various types of information are obtained using this technique. Notably, the redox-active species have been calculated at a fixed potential using pulse voltammetry. The voltammetric E-tongue is utilised in various applications, including the detection of milk deterioration caused by microbial growth, which correlates with Colony Forming Units (CFU). Different types of teas are categorised into various groups. The voltammetric E-Tongue is used for monitoring the quality of drinking water [54].

1.4.6.3. Amperometric sensors

Amperometric sensors [55] are electrochemical tools for measuring the current generated by a redox reaction, making them considerably valuable for identifying specific substances in tea, such as sugars, antioxidants, or contaminants. These sensors work by measuring the potential difference between the electrodes, which triggers a chemical reaction in the presence of the target substance. The current produced is directly proportional to the concentration of the substance being measured, permitting for precise, real-time monitoring of tea quality.

Figure 1.6 depicts the Amperometric sensors that can be used to monitor parameters such as the concentration of polyphenols, which are dynamic for determining the antioxidant properties and flavour of tea. These sensors can also identify the presence of unwanted substances, such as pesticides or heavy metals, ensuring the tea meets safety standards. Their high sensitivity

allows for the detection of even trace amounts of substances, making them an invaluable tool for quality control [57].

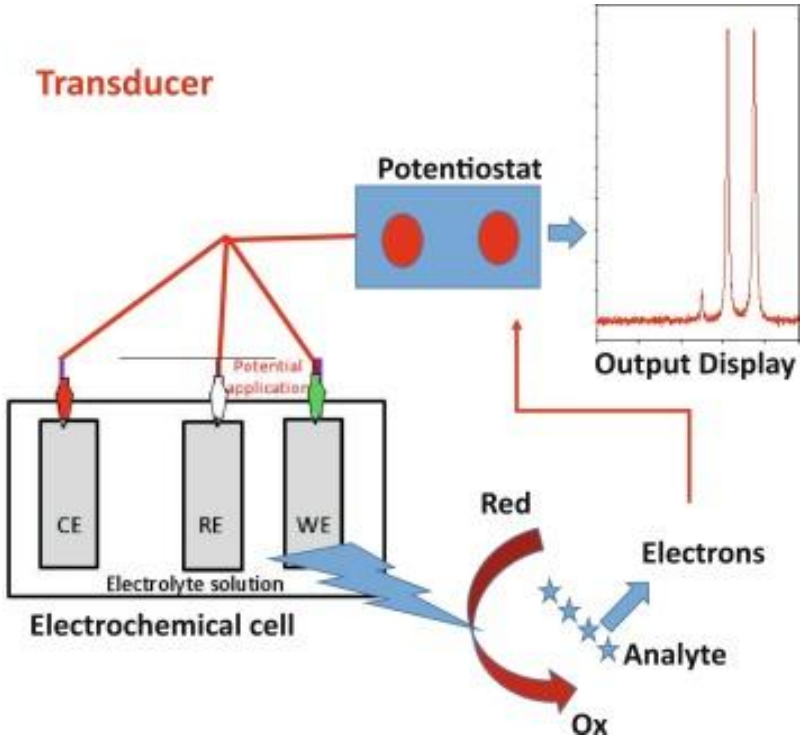


Figure 1.6 Amperometric sensors [56]

Additionally, offers advantages in terms of portability, speed, and cost-effectiveness compared to traditional analytical methods. By integrating these sensors into the production line, tea manufacturers can perform continuous, in-situ monitoring, improving efficiency and ensuring consistent quality. The result is a safer, more reliable tea product for consumers.

1.4.6.4. Capacitive sensors

The capacitive sensors are utilised in the E-Tongue. The measures alterations in capacitation for the identification and quantification of diverse materials in liquids or gases. E-Tongues are tools designed to simulate the human sense of taste. The array of chemical sensors with broad selectivity is contrasted with the use of recent mathematical processes for signal processing, pattern recognition, and multivariate data analysis [58].

Figure 1.7 depicts the capacitive sensors, which measure the substance's ability to hold an electric charge by calculating the difference between two electrodes separated from the substance. When the composition of the substance changes, its dielectric properties also change, resulting in altered capacitance, which is then related to the concentration of the

substance. It is incorporated into various fields, including environmental monitoring, food quality control, and biomedical diagnostics. In the e-tongues, it is utilised for examining complex solutions such as beverages, bio-samples and pharmaceuticals [59].

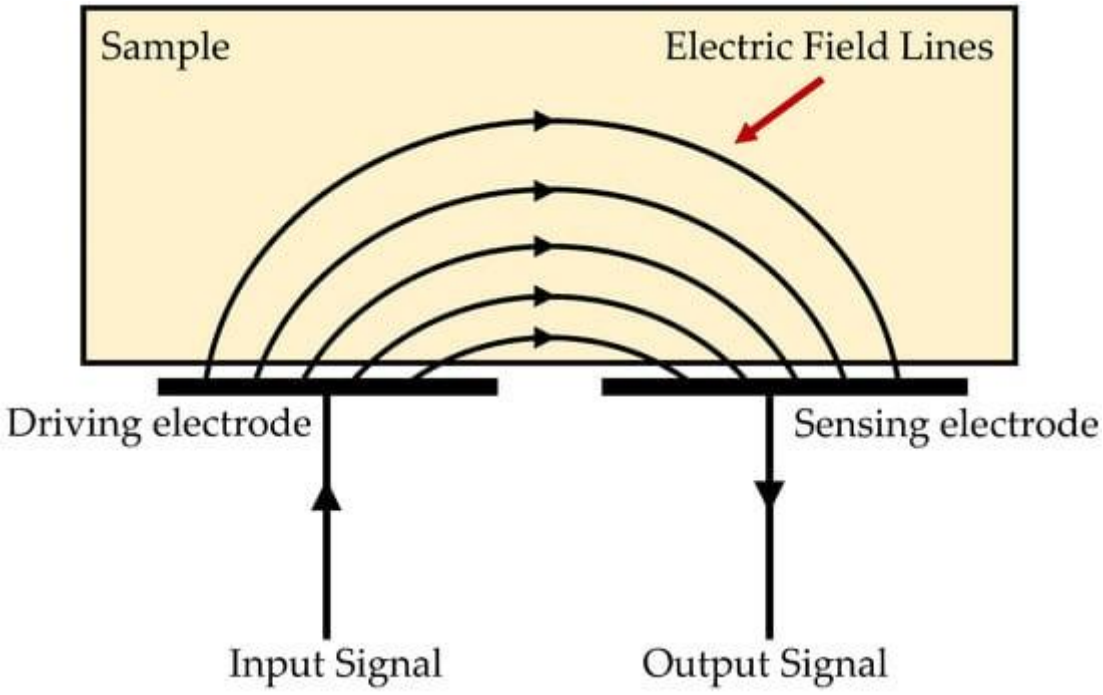


Figure 1.7 Capacitive sensors [44]

The advantages of capacitive sensors include higher sensitivity, lower power usage, and the capability of operating at room temperature. It has been downsized and integrated into a portable device, making it suitable for on-site evaluations. The materials employed include polymers, composites and metal oxide.

1.4.6.5. Impedance sensors

It holds importance in the tea sector for assessing the quality of tea. These sensors measure the impedance of the tea sample, providing information about its chemical composition and overall quality.

Figure 1.8 depicts the visual representation of impedance sensors. The sensor identifies variations in resistance by using an Alternating Current (AC) signal [61], which reflects diverse characteristics such as moisture levels, texture, and density. These sensors are especially advantageous in identifying various tea types or recognising adulteration. They provide quick, non-invasive, and precise measurements, making perfect quality control in tea manufacturing [62]. The capacity to provide real-time information facilities for effective decision-making in

tea processing, enhancing the dependability and quality. The demand for premium tea plays a crucial role in maintaining the quality of the products.

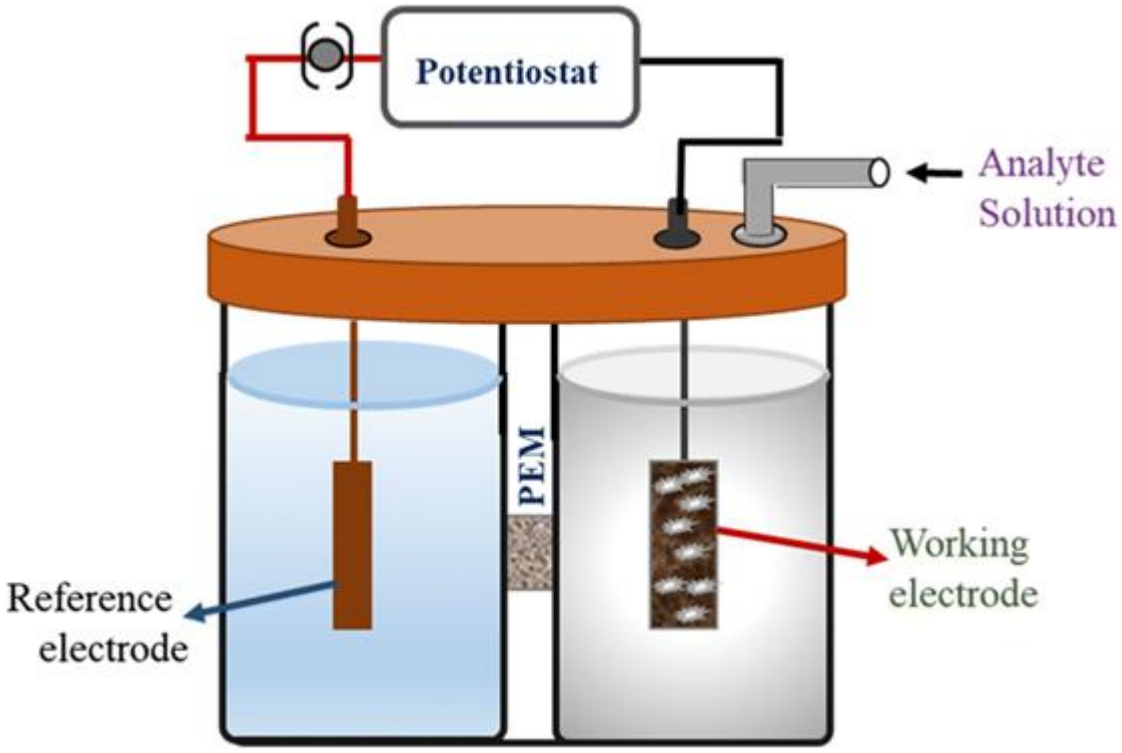


Figure 1.8 Impedance sensors [60]

1.5. Pattern Recognition Methods

It is considered the crucial part in the e-tongue response analysis process. It includes the algorithms for classifying and explaining the complex data provided by the sensor array. As the e-tongues depend on arrays of wider, selectively tuned sensors, the raw data is produced with difficulty and requires thorough examination for the extraction of meaningful information. The primary aim of pattern recognition is to identify patterns present in the associations of sensor data, which are used for the classification of samples and the prediction of properties of specified compounds.

Diverse pattern recognition methods are utilised in the e-tongue data examinations, which include:

- Principal Component Analysis (PCA) is a robust technique for reducing dimensionality, which aids in finding the essential components within high-dimensional datasets. It operates by converting the initial features into a fresh collection of orthogonal variables, referred to as principal components, which encapsulate the

highest variance in the data. PCA simplifies the visualisation and interpretation of complex datasets by decreasing the number of dimensions. This decrease enables a more manageable analysis, thereby improving classification, clustering, and pattern recognition. PCA is primarily advantageous when the data exhibits strong correlations, as it highlights the most essential features while reducing noise. Accordingly, it simplifies improved decision-making and insight creation. PCA is also frequently used in fields such as ML, image processing, and bioinformatics for data analysis and feature extraction.

- Linear Discriminant Analysis (LDA) [44] is a supervised classification technique that aims to identify a linear collection of features that enhances class separations. It achieves this by finding the linear integration that most effectively distinguishes between various classes in the dataset. LDA aims to improve the variance between classes while minimising modification within classes. In contrast to PCA, which is unsupervised and emphasises variance, LDA employs class labels to determine the most effective way to differentiate between classes. The approach lowers dimensionality while preserving class-separate information. LDA is commonly applied in tasks involving pattern recognition and classification, including face recognition and medical diagnostics. It works best when the data follows a standard delivery, and the class exhibits similar covariance.

- SVM is a supervised classification method that seeks to identify the ideal hyperplane that most effectively divides data into various classes within a high-dimensional space. It functions by choosing a hyperplane that maximises the distance between the nearest data points of every class, referred to as support vectors. This margin optimisation ensures that the classifier has the highest generality capability and can exactly categorise new data. SVM is capable of addressing both linear and non-linear classification issues by applying kernel functions to convert data into higher-dimensional spaces. When the data is not linearly separable, SVM can utilise the kernel trick to identify a separating hyperplane in a modified space. SVM is recognised for its strength and efficiency, particularly in high-dimensional data conditions. It is frequently used in fields such as image recognition, text categorisation, and bioinformatics.

- Artificial Neural Networks (ANN) are computational systems modelled after the organisation and operation of the human brain. They are composed of layers of linked nodes, or neurons, that analyse and learn from information in a manner akin to how the brain recognises patterns and relationships. ANN can identify complex patterns and connections in data, making it suitable for tasks like image recognition, speech analysis, and understanding natural languages. The network undergoes training via back propagation, which sends the error backwards through the network to modify the weights and improve precision. ANNs are excellent at handling non-linear relationships and can generalise weights and enhance accuracy. Able to learn from labelled and unlabelled data, they are extensively utilised in DL applications for tasks that demand significant computational resources.

- Cluster Analysis is an unverified technique employed to categorise samples according to their resemblance without any prior label information. It detects patterns in the data, allowing for the creation of clusters in which samples of the same type are more similar to each other than to those of different types. Frequently used methods in cluster examination encompass hierarchical clustering, which creates a tree-like diagram to depict nested groups, and k-means clustering, which splits data into a set number of clusters by reducing the variance within each cluster. These methods are commonly applied in areas such as data mining, image segmentation, and market division to uncover natural clusters in data, including customer categories or sensor outputs. Cluster analysis facilitates the structuring and consideration of extensive datasets by leveraging similarities in the characteristics of the models.

Distinguishing pattern recognition plays a crucial role in e-tongue skill by enabling the extraction of valuable insights from sensor data. The process involves several key stages:

1. **Data Acquisition:** The first step involves collecting sensor data from the e-tongue system. This data is typically generated by multiple sensors that mimic the human sense of taste, and it is crucial for capturing a diverse set of samples from the environment or the substance being analysed.

2. **Data Pre-processing:** This step focuses on cleaning and transforming the raw data to enhance its quality and reliability for further analysis. It includes processes such as

stabilisation, baseline correction, and noise reduction to eliminate any irrelevant variations or errors, ensuring that the data is suitable for accurate pattern recognition.

3. **Feature Extraction:** Feature extraction is essential for identifying the most informative and discriminative features from the processed data. This step focuses on reducing dimensionality while retaining the key characteristics that are most relevant for distinguishing between different classes or categories of samples.

4. **Model Training:** In this phase, a pattern recognition model is trained using a subset of data with known class labels. This training procedure enables the model to learn the relationships between the features and the corresponding labels, allowing it to recognise patterns and relations in future, unseen data.

5. **Model Validation:** After training, the model's performance is estimated using an independent validation set. This step tests the prototypical exactness, value of precision, and generalisation abilities, ensuring it can make reliable predictions on new, unseen data. Adequate model validation is crucial to prevent overfitting and assess the model's real-world applicability.

6. **Classification or Prediction:** The last stage includes using the trained model to classify new, unidentified samples or predict their properties. This enables the e-tongue system to make real-time decisions based on the patterns it has learned, such as recognising the taste profile of a beverage or detecting a product's authenticity.

In conclusion, pattern recognition is crucial for e-tongue technology, which enables the extraction of expressive information from complex sensor data. This enables practical applications such as quality control, product verification, taste analysis, and sensory evaluations, transforming industries like food and beverage, pharmaceuticals, and environmental observation.

1.6. Literature Review

In tea production, antioxidant activity plays a vital role in measuring the quality and effectiveness of tea products. The information gathered from these sensors is often complex and includes a significant amount of background noise. Therefore, it is essential to focus on getting accurate information from the sample being tested. The way the electronic tongue

sensors react depends on the ions and compounds in the test liquid. This means that each different solution can cause the sensors to respond in their individual, unique way.

Typically, unsupervised pattern recognition methods, including PCA, are valuable tools for visualising data distribution, commonly applied to data assessment and discovery. Table 1.1 provides a concise overview of the literature review on various pattern recognition techniques for evaluating tea quality using an electronic tongue.

Table 1.1 Outline of Literature Review

Sl. No.	Measurement Type	Objective	Feature transformation technique	Machine learning technique	Refs
1.	Cyclic voltammetry	To evaluate the antioxidant activity of green tea products during storage	normalisation or dimensionality reduction (e.g., PCA)	regression or classification algorithms	[63]
2.	Fluorescence signalling	The goal is to develop ratiometric molecularly imprinted particle probes for the reliable detection of carboxylate-containing molecules.	Ratiometric analysis		[64]
3.	Quantitative determination of pyrethroid insecticide residues.	The goal is to achieve sensitive detection of pyrethroid insecticide	normalisation or scaling of sensor data	SVM, ANN, PCA	[65]

Sl. No.	Measurement Type	Objective	Feature transformation technique	Machine learning technique	Refs
		residues in tea samples.			
4.	Fourier Transform Infrared (FTIR) and Near Infrared Spectroscopy	Geographical origin identification of black tea samples	Spectroscopic fingerprinting	k-Nearest Neighbours (k-NN) and SVM	[66]
5.	Near-Infrared Spectroscopy	Assurance of tea quality	DL-based feature extraction (TeaNet architecture)	Convolutional Neural Networks (CNN)	[67]
6.	Hyperspectral Imaging	Non-destructive testing and visualisation of catechin content during black tea fermentation	PCA, Multiplicative Scatter Correction (MSC), Standard Normal Variate (SNV)	Extreme Learning Machine (ELM) and Support Vector Regression (SVR)	[68]
7.	Near-Infrared Reflectance Spectroscopy (NIRS) and Computer Vision Sensors	Quality evaluation of Keemun black tea	Mid-level data fusion combining spectral, shape, colour, and aroma data	Least-Squares SVM (LS-SVM), (SVR), ELM, and Partial Least Squares Discriminant Analysis (PLS-DA	[69]
8.	Micro-NIR(NIRS), Computer	Rapid and comprehensive grade evaluation	Mid-level data fusion combining spectral, shape,	SVM, Least-Squares-SVM (LS-SVM),	[70]

Sl. No.	Measurement Type	Objective	Feature transformation technique	Machine learning technique	Refs
	Vision, and Colourimetric Sensor Array	of Keemun black tea	colour, and aroma data	ELM, PLS-DA, and SVR	
9.	Surface-Enhanced Raman Spectroscopy (SERS)	Quality assessment of black tea during its fermentation process	Chemometric preprocessing techniques	Partial Least Squares Regression (PLSR)	[71]
10.	Chemical indices of green and black teas (e.g., pigments, flavonol glycosides, catechins ratios)	Predicting the sensory quality of green and black teas	Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) and variable selection using Random Forest	SVM	[72]
11.	Micro-NIR(NIRS) and Computer Vision System (CVS)	Rapid and real-time detection of black tea fermentation quality	Mid-level data fusion combining colour, spectral, and sensory evaluation data	SVM for fermentation degree discrimination and theaflavins content prediction	[73]
12.	Fluorescence Hyperspectral Imaging (FHSI)	Reliable identification and classification of	Spectral preprocessing using Standard Normal Variate (SNV), MSC,	This text discusses SVM) and (PLS-DA)	[74]

Sl. No.	Measurement Type	Objective	Feature transformation technique	Machine learning technique	Refs
		oolong tea species	PCA, and Variable Importance in Projection (VIP) for feature selection		
13.	UV-Visible Spectroscopy and Multispectral Omics	Exploration and identification of critical colourant compounds in black tea infusion	Metabolite analysis and spectral preprocessing using untargeted metabolomics (UHPLC-Q-TOF-MS)	PLSR	[75]
14.	NIRS	Tea grade discrimination	Data fusion strategies combining spectral, shape, colour, and possibly aroma data	SVM, Random Forest (RF)	[76]
15.	Smartphone Imaging coupled with Micro-Near-Infrared Spectrometer (micro-NIRS)	Evaluation of black tea quality	Quantisation of HSV colour channels for image features and spectral preprocessing for NIRS data	SVM is used for classification	[77]
16.	Lab-made Computer	Evaluation of Congou black	Extraction of morphological	SVM	[78]

Sl. No.	Measurement Type	Objective	Feature transformation technique	Machine learning technique	Refs
	Vision System (CVS)	tea quality based on morphological features	parameters and chemometric preprocessing		
17.	PCA and PLS applied to spectroscopic data.	Authentication of black tea from narrow-geographic origins	PCA for dimensionality reduction and PLS for predictive modelling	LDA and SVM classifiers	[79]
18.	NIR	Non-invasive identification of commercial green tea blends	PCA for dimensionality reduction	SVM	[80]
19.	RGB Imaging and Hyperspectral Imaging	Rapid classification of tea coal disease	Pre-processing using Standard Normal Variate (SNV), second derivative (2-D), Savitzky-Golay (S-G) smoothing, and feature selection with Competitive Adaptive Reweighted Sampling (CARS) and Uninformative Variable	ResNet18, VGG16, and AlexNet	[81]

Sl. No.	Measurement Type	Objective	Feature transformation technique	Machine learning technique	Refs
			Elimination (UVE) algorithms		
20.	NIRS	Prediction of tea polyphenol content	Deep feature extraction using 1D and 2D CNNs	CNN	[82]
21.	RGB Imaging	Classification of oolong tea varieties based on visual attributes	Data augmentation (image cropping, flipping) and extraction of 13 visual features (colour and texture)	ResNet50	[83]
22.	NIRS	Monitoring the dynamic changes in catechins during black tea drying	Chemometric preprocessing techniques, including Standard Normal Variate (SNV), MSC, and PCA for spectral data optimisation	PLSR	[84]
23.	Fluorescence Hyperspectral Imaging	Classification and adulteration detection of Mengding Mountain green tea varieties	Spectral preprocessing using SNV, MSC, and PCA for feature extraction	SVM and PLS-DA	[85]

Sl. No.	Measurement Type	Objective	Feature transformation technique	Machine learning technique	Refs
24.	Fourier Transform Infrared (FTIR) and Near Infrared [9] Spectroscopy	Geographical origin identification of black tea samples	Spectroscopic fingerprinting	k-Nearest Neighbours (k-NN) and SVM	[86]

1.6.1. Research Gap and Motivation

From the literature reviewed, it is clear that substantial advancement has been made in the field of tea quality assessment based on analytical chemistry, spectroscopy, sensors, and DL methods. However, there exist certain gaps in the current research work.

Primarily, several studies mainly depend on sensory or instrumental analysis alone, which may not provide an overall and objective analysis. Secondly, although electrochemical sensor analysis and NIR spectroscopy individually perform well, their combined use within the framework of intelligent sensors has gained relatively less attention in tea quality prediction. Thirdly, majority of the previous studies have mainly concentrated on classification rather than multi-output regression based modelling, which is required for predicting multiple compounds associated with tea quality.

In addition to this, the use of optimised feature selection, hybrid architectures of AI and data fusion models have received minimal consideration in tea quality estimation. In addition, some of the reported methods involve the use of costly equipment, extensive sample preparation procedures, and are also relatively less scalable. Thus, the present study attempts to overcome such limitations with the help of MIP sensors, NIR spectroscopy, DL, optimization and fusion of data for better tea quality prediction.

1.7. Objective of the Work

This thesis seeks to advance the field of tea quality analysis through the following objectives, which focus on,

Objective 1:

- To apply linear regression modelling on Epigallocatechin-3-Gallate sensor data for green tea.

- To classify green tea samples based on Epigallocatechin-3-Gallate content using the molecular imprinting polymerisation techniques.
- To employ regression-based EGCG detection in green tea utilising MIP electrodes.

Objective 2:

- To assess the black tea sample effectively using a multi-output regression approach.
- To use the modelling of MIP-based fused chemical sensors.

Objective 3:

- To perform efficient Near Infrared Spectroscopy-based feature selection of Tannic Acid for black tea evaluation.
- To predict Tannic Acid effectively in black tea samples using a regression approach
- To fuse the response obtained from the MIP electrochemical sensors and NIR spectroscopy.

This thesis seeks to develop tea quality assessment methods based on objective and data-driven methods using electrochemistry-based sensing techniques, NIR, feature optimisation, and ML and DL algorithms. There has been an effort towards increasing prediction accuracy, sensor fusion for robustness, and scalability of the analysis method for industrial applications of tea.

1.8. Scope of the Thesis

The scope of this thesis is centred on the application and advancement of electronic tongue (E-Tongue) technology for the objective evaluation of tea quality. This work is designed to address the limitations of conventional and instrumental methods by leveraging the unique capabilities of E-Tongue systems to analyse complex chemical compositions and sensory attributes present in tea infusions. The novelty of the present work lies in the integration of MIP-based electrochemical sensing, NIR spectroscopy, optimisation-assisted feature selection, and intelligent fusion-driven predictive modelling within a unified tea quality evaluation framework.

The study systematically investigates various E-Tongue measurement methods, including potentiometric, voltammetric, amperometric, capacitive, impedimetric, optical, biosensor, and acoustic wave sensor techniques, to determine their suitability and effectiveness in differentiating tea samples based on quality parameters. Special emphasis is placed on the design and optimisation of sensor arrays, as well as the development of robust pattern recognition algorithms for accurate classification and prediction of tea quality.

The research is confined to the evaluation of selected tea types, focusing on the chemical and sensory markers most relevant to quality assessment. It does not extend to the fusion of the E-Tongue with other sensor technologies, such as electronic noses or eyes. Furthermore, the thesis examines the integration of MIPs to enhance selectivity in sensor design. The work also includes a comprehensive review of existing literature and a critical comparison of E-Tongue performance with conventional analytical techniques. The ultimate aim is to establish a rapid, reproducible, and non-destructive method for assessing tea quality that can be applied in both research and industry settings, thereby providing a foundation for further development in sensor-based food quality evaluation.

1.9. Thesis Organisation

- Chapter 1 introduces tea quality evaluation, research context, sample overview, method selection, research goals, and scope, and includes a literature review.
- Chapter 2 outlines the data analysis approaches, summarising the ML and DL techniques used.
- Chapter 3 examines single-molecule quality assessment using MIPs, focusing on EGCG analysis in green tea through regression and categorisation methods.
- Chapter 4 evaluates black tea quality using multi-output regression with MIP-based fused sensors, detailing data collection, feature selection, modelling, and performance.
- Chapter 5 focusing on tannic acid feature selection and predicts tannic acid content in black tea by integrating MIP sensor and NIR data, employing a proposed SNN-LSTM regression framework.
- Chapter 6 concludes with key findings, contributions, future research directions, and final remarks.

1.10. Conclusion

In conclusion, the chapter has recognised the foundation for investigating advanced methods in evaluating tea quality. The quality of tea is complex, shaped by its processing method, chemical makeup, and sensory characteristics. Although traditional sensory analysis, which relies on trained tasters, is essential, it has inherent limitations regarding subjectivity and efficiency. To address these limitations, instrumental methods such as HPLC, NIR spectroscopy, TGA-FTIR, electronic noses, electronic tongues, and electronic eyes have

emerged as valuable tools, providing objective and quantitative data regarding various quality parameters of tea.

The chapter presents essential instrumental methods used to assess tea quality, including HPLC for measuring specific compounds, NIR spectroscopy for non-invasive evaluation, and electronic noses and tongues for examining aroma and flavour characteristics. We emphasised the capabilities of electronic tongues and MIPs that function as artificial receptors for specific compounds. These sophisticated methods, combined with pattern recognition algorithms, enable the rapid and reliable evaluation of tea quality.

This thesis clearly outlined its aims and scope, emphasising the comprehensive analysis and comparison of electronic tongues and MIPs with traditional instrumental techniques. This study aims to enhance the development of a more effective and impartial quality control system for the tea sector, enabling real-time tracking of tea quality during the manufacturing process. In summary, this chapter lays the foundation for a more in-depth investigation of sophisticated detection and evaluation methods, creating a pathway for the following chapters that will examine data analysis, separate molecule quality assessment, and the thorough evaluation of tea samples through cutting-edge methods. Through the use of these sophisticated tools, the tea sector can enhance quality assurance, meet consumer needs, and ensure the authenticity and quality of tea products.

REFERENCES

[1] L. Fu and W. Ying, "Advancements in the application of carbon nanotubes for tea quality and safety assessment," *Fullerenes, Nanotubes and Carbon Nanostructures*, vol. 31, no. 11, pp. 1007-1018, 2023, doi: https://doi.org/10.1080/1536383X.2023.2246162.

[2] H. S. A. Huda *et al.*, "Exploring the ancient roots and modern global brews of tea and herbal beverages: A comprehensive review of origins, types, health benefits, market dynamics, and future trends," *Food Science & Nutrition*, vol. 12, no. 10, pp. 6938-6955, 2024, doi: https://doi.org/10.1002/fsn3.4346.

[3] J. Ouyang *et al.*, "Insights into the flavor profiles of different grades of Huangpu black tea using sensory histology techniques and metabolomics," *Food Chemistry: X*, vol. 23, p. 101600, 2024, doi: https://doi.org/10.1016/j.fochx.2024.101600.

[4] M. T. El-Saadony *et al.*, "Chemistry, bioavailability, bioactivity, nutritional aspects and human health benefits of polyphenols: A comprehensive review," *International journal of biological macromolecules*, p. 134223, 2024, doi: https://doi.org/10.1016/j.ijbiomac.2024.134223.

[5] J. Moreira *et al.*, "Tea Quality: An Overview of the Analytical Methods and Sensory Analyses Used in the Most Recent Studies," *Foods*, vol. 13, no. 22, p. 3580, 2024, doi: https://doi.org/10.3390/foods13223580.

[6] I. Ben Alaya, G. Alves, J. Lopes, and L. R. Silva, "Use of Encapsulated Polyphenolic Compounds in Health Promotion and Disease Prevention: Challenges and Opportunities," *Macromol*, vol. 4, no. 4, pp. 805-842, 2024, doi: https://doi.org/10.3390/macromol4040048.

[7] S. S. Rodrigues, L. G. Dias, and A. Teixeira, "Emerging Methods for the Evaluation of Sensory Quality of Food: Technology at Service," *Current Food Science and Technology Reports*, vol. 2, no. 1, pp. 77-90, 2024, doi: https://doi.org/10.1007/s43555-024-00019-7.

[8] A. I. de Almeida Costa, M. J. P. Monteiro, and E. Lamy, *Sensory Evaluation and Consumer Acceptance of New Food Products: Principles and Applications*. Royal Society of Chemistry, 2024, doi: https://doi.org/10.1039/9781839166655.

[9] S. Jakabová, J. Árvay, M. Šnirc, J. Lakatošová, A. Ondejčíková, and J. Golian, "HPLC-DAD method for simultaneous determination of gallic acid, catechins, and methylxanthines and its application in quantitative analysis and origin identification of

green tea," *Heliyon*, vol. 10, no. 16, 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e35819.

[10] I. Ziani, H. Bouakline, A. El Guerraf, A. El Bachiri, M.-L. Fauconnier, and F. Sher, "Integrating AI and advanced spectroscopic techniques for precision food safety and quality control," *Trends in Food Science & Technology*, p. 104850, 2024, doi: https://doi.org/10.1016/j.tifs.2024.104850.

[11] H. R. Tan and W. Zhou, "Metabolomics for tea authentication and fraud detection: Recent applications and future directions," *Trends in Food Science & Technology*, p. 104558, 2024, doi: https://doi.org/10.1016/j.tifs.2024.104558.

[12] J. Zhang, X. Wu, C. He, B. Wu, S. Zhang, and J. Sun, "Near-Infrared Spectroscopy Combined with Fuzzy Improved Direct Linear Discriminant Analysis for Nondestructive Discrimination of Chrysanthemum Tea Varieties," *Foods*, vol. 13, no. 10, p. 1439, 2024, doi: https://doi.org/10.3390/foods13101439.

[13] L. Wang *et al.*, "Evaluation of the quality grade of congou black tea by the fusion of GC-E-nose, E-tongue, and E-eye," *Food Chemistry: X*, p. 101519, 2024, doi: https://doi.org/10.1016/j.fochx.2024.101519.

[14] X. Zhao, L. Cai, P. Huang, and C. J. A. a. S. Cui, "Enzymatic Synthesis and Sensory Evaluation of N-Cinnamoyl-L-Phenylalanine as a Novel Flavor Enhancer: Impact on Taste Perception and Mechanistic Insights," *Food Chemistry*, 2025, doi: https://doi.org/10.1016/j.foodchem.2025.144542.

[15] J. Liang, J. Guo, H. Xia, C. Ma, and X. J. F. C. Qiao, "A black tea quality testing method for scale production using CV and NIRS with TCN for spectral feature extraction," *Food Chemistry*, vol. 464, p. 141567, 2025, doi: https://doi.org/10.1016/j.foodchem.2024.141567.

[16] M. S. Kabir, M. A. Pathan, K. Tanvir, and D. J. Gomes, "Smart Agriculture Using Advancing Tea Leaf Quality Assessment With Deep Learning," in *Intelligent Internet of Everything for Automated and Sustainable Farming*: IGI Global Scientific Publishing, pp. 149-186, 2025, doi: https://doi.org/10.4018/979-8-3373-0020-7.ch005.

[17] U. Yadav, S. Khurana, K. Rawal, B. Arora, and A. Yadav, "Integrating AI with Electronic Tongue (E-Tongue) for Real-Time Detection in the Food Industry," in *Artificial Intelligence in the Food Industry*: CRC Press, pp. 277-296, 2025,.doi: https://doi.org/10.1201/9781032633602-14.

[18] Y. Li *et al.*, "Integration of non-targeted/targeted metabolomics and electronic sensor technology reveals the chemical and sensor variation in 12 representative yellow teas,"

Food Chemistry: X, vol. 21, p. 101093, 2024, doi: https://doi.org/10.1016/j.fochx.2023.101093.

[19] H. Xia *et al.*, "Rapid discrimination of quality grade of black tea based on near-infrared spectroscopy (NIRS), electronic nose (E-nose) and data fusion," *Food Chemistry*, vol. 440, p. 138242, 2024, doi: https://doi.org/10.1016/j.foodchem.2023.138242.

[20] Y. Song *et al.*, "Exploring key aroma differences of three Fenghuang Dancong oolong teas and their perception interactions with caffeine via sensomics approach," *Food Research International*, vol. 202, p. 115665, 2025, doi: https://doi.org/10.1016/j.foodres.2025.115665.

[21] X. Zhao, P. Yu, N. Zhong, H. Huang, and H. Zheng, "Impact of Storage Temperature on Green Tea Quality: Insights from Sensory Analysis and Chemical Composition," *Beverages*, vol. 10, no. 2, p. 35, 2024, doi: https://doi.org/10.3390/beverages10020035.

[22] Y. Wei *et al.*, "Green preparation, safety control and intelligent processing of high-quality tea extract," *Critical reviews in food science and nutrition*, vol. 64, no. 31, pp. 11468-11492, 2024, doi: https://doi.org/10.1080/10408398.2023.2239348.

[23] A. A. Shitandi, "Tea processing and impact on catechins, theaflavin and thearubigin formation," in *Tea in health and disease prevention*: Elsevier, pp. 193-205, 2013, doi: https://doi.org/10.1016/B978-0-12-384937-3.00016-1.

[24] Q. Chen *et al.*, "Re-rolling treatment in the fermentation process improves the taste and liquor color qualities of black tea," *Food Chemistry: X*, vol. 21, p. 101143, 2024, doi: https://doi.org/10.1016/j.fochx.2024.101143.

[25] Z. Ding *et al.*, "Lightweight CNN combined with knowledge distillation for the accurate determination of black tea fermentation degree," *Food Research International*, vol. 194, p. 114929, 2024, doi: https://doi.org/10.1016/j.foodres.2024.114929.

[26] Z. Shu, H. Zhou, L. Chen, Y. Wang, Q. Ji, and W. He, "Effect of Four Different Initial Drying Temperatures on Biochemical Profile and Volatilome of Black Tea," *Metabolites*, vol. 15, no. 2, p. 74, 2025, doi: https://doi.org/10.3390/metabo15020074.

[27] J.-P. Ke *et al.*, "Cordyceps militaris fermentation changes the flavor and chemical profiles of Lu'an GuaPian green tea with fat-lowering and anti-aging activities," *Microchemical Journal*, vol. 201, p. 110676, 2024, doi: https://doi.org/10.1016/j.microc.2024.110676.

[28] X. Guo, L. Wang, X. Huang, and Q. Zhou, "Characterization of the volatile compounds in tea (Camellia sinensis L.) flowers during blooming," *Frontiers in Nutrition*, vol. 11, p. 1531185, 2025, doi: https://doi.org/10.3389/fnut.2024.1531185.

[29] Q. Li *et al.*, "Characterization and exploration of dynamic variation of volatile compounds in vine tea during processing by GC-IMS and HS-SPME/GC–MS combined with machine learning algorithm," *Food Chemistry*, vol. 460, p. 140580, 2024, doi: https://doi.org/10.1016/j.foodchem.2024.140580.

[30] H. Zuo *et al.*, "Tea-Derived Polyphenols Enhance Drought Resistance of Tea Plants (Camellia sinensis) by Alleviating Jasmonate–Isoleucine Pathway and Flavonoid Metabolism Flow," *International Journal of Molecular Sciences*, vol. 25, no. 7, p. 3817, 2024, doi: https://doi.org/10.3390/ijms25073817.

[31] J. Cui, B. Wu, and J. Zhou, "Changes in amino acids, catechins and alkaloids during the storage of oolong tea and their relationship with antibacterial effect," *Scientific Reports*, vol. 14, no. 1, p. 10424, 2024, doi: https://doi.org/10.1038/s41598-024-60951-5.

[32] H. Karami *et al.*, "Advanced evaluation techniques: Gas sensor networks, machine learning, and chemometrics for fraud detection in plant and animal products," *Sensors and Actuators A: Physical*, p. 115192, 2024, doi: https://doi.org/10.1016/j.sna.2024.115192.

[33] E. Bagnulo, G. Strocchi, C. Bicchi, and E. Liberto, "Industrial food quality and consumer choice: Artificial intelligence-based tools in the chemistry of sensory notes in comfort foods (coffee, cocoa and tea)," *Trends in Food Science & Technology*, p. 104415, 2024, doi: https://doi.org/10.1016/j.tifs.2024.104415.

[34] F. Wu *et al.*, "Characteristic aroma screening among green tea varieties and electronic sensory evaluation of green tea wine," *Fermentation*, vol. 10, no. 9, p. 449, 2024, doi: https://doi.org/10.3390/fermentation10090449.

[35] Z. Cai *et al.*, "Quality evaluation of Lonicerae Japonicae Flos and Lonicerae Flos based on simultaneous determination of multiple bioactive constituents combined with multivariate statistical analysis," *Phytochemical Analysis*, vol. 32, no. 2, pp. 129-140, 2021, doi: https://doi.org/10.1002/pca.2882.

[36] W. Zhang *et al.*, "Data-driven optimization of nitrogen fertilization and quality sensing across tea bud varieties using near-infrared spectroscopy and deep learning," *Computers and Electronics in Agriculture*, vol. 222, p. 109071, 2024, doi: https://doi.org/10.1016/j.compag.2024.109071.

[37] C. Călin, E.-E. Sîrbu, M. Tănase, R. Győrgy, D. R. Popovici, and I. Banu, "A Thermogravimetric Analysis of Biomass Conversion to Biochar: Experimental and

Kinetic Modeling," *Applied Sciences*, vol. 14, no. 21, p. 9856, 2024, doi: https://doi.org/10.3390/app14219856.

[38] D. Das, S. Nag, and R. B. Roy, "Machine Vision System: An Emerging Technique for Quality Estimation of Tea," in *Edible Electronics for Smart Technology Solutions*: IGI Global, pp. 445-476, 2025, doi: https://doi.org/10.4018/979-8-3693-5573-2.ch018.

[39] X. Li, A. Sanaeifar, S. Zhang, Z. Zhan, and Y. He, "Recent Technological Advances in Tea Quality and Safety," *Natural Products in Beverages: Botany, Phytochemistry, Pharmacology and Processing*, pp. 83-127, 2024, doi: https://doi.org/10.1007/s44462-025-00021-9.

[40] R. Zhi, L. Zhao, and D. Zhang, "A framework for the multi-level fusion of electronic nose and electronic tongue for tea quality assessment," *Sensors*, vol. 17, no. 5, p. 1007, 2017, doi: https://doi.org/10.3390/s17051007.

[41] Y. Yu, Q. Li, Z. Hua, C. Yin, and Y. Shi, "An effective multi-source information fusion method for electronic nose and hyperspectral to identify the spring tea quality at different harvesting periods," *Measurement*, vol. 243, p. 116452, 2025, doi: https://doi.org/10.1016/j.measurement.2024.116452.

[42] G. Ren, X. Zhang, R. Wu, L. Yin, W. Hu, and Z. Zhang, "Rapid characterization of black tea taste quality using miniature NIR spectroscopy and electronic tongue sensors," *Biosensors*, vol. 13, no. 1, p. 92, 2023, doi: https://doi.org/10.3390/bios13010092.

[43] C. Sheng *et al.*, "Metabolomics and electronic-tongue analysis reveal differences in color and taste quality of large-leaf yellow tea under different roasting methods," *Food Chemistry: X*, vol. 23, p. 101721, 2024, doi: https://doi.org/10.1016/j.fochx.2024.101721.

[44] F. Abdollahi-Mamoudan, C. Ibarra-Castanedo, and X. P. V. Maldague, "Advancements in and Research on Coplanar Capacitive Sensing Techniques for Non-Destructive Testing and Evaluation: A State-of-the-Art Review," *Sensors*, vol. 24, no. 15, p. 4984, 2024, doi: https://doi.org/10.3390/s24154984.

[45] T. Yang *et al.*, "Effect of spreading time on the taste quality of steamed green tea based on E-tongue evaluation and chemometric statistical analysis," *Journal of Food Measurement and Characterization*, pp. 1-13, 2024, doi: https://doi.org/10.1007/s11694-024-02394-0.

[46] B. Aouadi *et al.*, "Historical evolution and food control achievements of near infrared spectroscopy, electronic nose, and electronic tongue—Critical overview," *Sensors*, vol. 20, no. 19, p. 5479, 2020, doi: https://doi.org/10.3390/s20195479.

[47] A. A. Almario, O. P. Calabokis, and E. A. Barrera, "Smart E-Tongue Based on Polypyrrole Sensor Array as Tool for Rapid Analysis of Coffees from Different Varieties," *Foods*, vol. 13, no. 22, p. 3586, 2024, doi: https://doi.org/10.3390/foods13223586.

[48] C. Pérez-Ràfols, N. Serrano, C. Ariño, M. Esteban, and J. M. Díaz-Cruz, "Voltammetric electronic tongues in food analysis," *Sensors*, vol. 19, no. 19, p. 4261, 2019, doi: https://doi.org/10.3390/s19194261.

[49] X. Ni *et al.*, "Research progress of sensors based on molecularly imprinted polymers in analytical and biomedical analysis," *Journal of Pharmaceutical and Biomedical Analysis*, p. 115659, 2023, doi: https://doi.org/10.1016/j.jpba.2023.115659.

[50] J. Liu *et al.*, "Bioinspired integrated triboelectric electronic tongue," *Microsystems & Nanoengineering*, vol. 10, no. 1, p. 57, 2024, doi: https://doi.org/10.1038/s41378-024-00690-9.

[51] J. Ding and W. Qin, "Recent advances in potentiometric biosensors," *TrAC Trends in Analytical Chemistry*, vol. 124, p. 115803, 2020, doi: https://doi.org/10.1016/j.trac.2019.115803.

[52] A. Ma'mun, M. K. Abd El-Rahman, M. J. J. o. p. Abd El-Kawy, and b. analysis, "Real-time potentiometric sensor; an innovative tool for monitoring hydrolysis of chemo/bio-degradable drugs in pharmaceutical sciences," *Journal of pharmaceutical biomedical analysis*, vol. 154, pp. 166-173, 2018, doi: https://doi.org/10.1016/j.jpba.2018.02.007.

[53] S. Nambiar, M. Mohan, and A. J. C. Rosin Jose, "Voltammetric Sensors: A Versatile Tool in COVID-19 Diagnosis and Prognosis," *ChemistrySelect*, vol. 8, no. 6, p. e202204506, 2023, doi: https://doi.org/10.1002/slct.202204506.

[54] T. Zhang *et al.*, "Insights into the Sources, Structure, and Action Mechanisms of Quinones on Diabetes: A Review," *Molecules*, vol. 30, no. 3, p. 665, 2025, doi: https://doi.org/10.3390/molecules30030665.

[55] S. Jayaraman *et al.*, "A smartphone-based tunable tyrosinase functional mimic modulated portable amperometric sensor for the rapid and real-time monitoring of catechol," *Chemical Engineering Journal*, vol. 497, p. 154811, 2024, doi: https://doi.org/10.1016/j.cej.2024.154811.

[56] Z. Yang *et al.*, "Nanozyme-Enhanced Electrochemical Biosensors: Mechanisms and Applications," *Small,* vol. 20, no. 14, p. 2307815, 2024, doi: https://doi.org/10.1002/smll.202307815.

[57] S. Ü. Pektaş, M. Keskin, O. C. Bodur, F. J. J. o. F. C. Arslan, and Analysis, "Green synthesis of silver nanoparticles and designing a new amperometric biosensor to determine glucose levels," *Journal of Food Composition Analysis,* vol. 129, p. 106133, 2024, doi: https://doi.org/10.1016/j.jfca.2024.106133.

[58] F. Shangguan, W. Cui, S. Liang, X. Lin, K. J. H. i. S. Wang, Engineering, and Technology, "A Capacitive Transparent Liquid Concentration Detecting System Based on LabVIEW," *Highlights in Science, Engineering Technology,* vol. 100, pp. 50-59, 2024, doi: https://doi.org/10.54097/8fnswa62.

[59] J. Tang *et al.*, "Electrochemical detection of rutin in black tartary buckwheat tea and related health-care pills with new ionic liquid-based supramolecular hydrogels," *Food Control,* vol. 155, p. 110045, 2024, doi: https://doi.org/10.1016/j.foodcont.2023.110045.

[60] C. S. Kushwaha and S. J. J. o. M. S. Shukla, "Non-enzymatic potentiometric malathion sensing over chitosan-grafted polyaniline hybrid electrode," *Journal of Materials Science,* vol. 54, no. 15, pp. 10846-10855, 2019, doi: https://doi.org/10.1007/s10853-019-03625-2.

[61] S. Yang *et al.*, "Non-targeted metabolomics reveals the effects of different rolling methods on black tea quality," *Foods,* vol. 13, no. 2, p. 325, 2024, doi: https://doi.org/10.3390/foods13020325.

[62] Y. Lin *et al.*, "A newly-discovered tea population variety processed Bai Mu Dan white tea: Flavor characteristics and chemical basis," *Food Chemistry,* vol. 446, p. 138851, 2024. https://doi.org/10.1016/j.foodchem.2024.138851.

[63] L. Jiang and K. Zheng, "Towards the intelligent antioxidant activity evaluation of green tea products during storage: A joint cyclic voltammetry and machine learning study," *Food Control,* vol. 148, p. 109660, 2023, doi: https://doi.org/10.1016/j.foodcont.2023.109660.

[64] Y. Sun, K. Gawlitza, V. Valderrey, B. Bhattacharya, and K. Rurack, "Ratiometric Molecularly Imprinted Particle Probes for Reliable Fluorescence Signaling of Carboxylate-Containing Molecules," *ACS Applied Materials & Interfaces,* vol. 16, no. 37, pp. 49944-49956, 2024, doi: https://doi.org/10.1021/acsami.4c09990.

[65] Y. Duan, D. Wang, Z. Xu, S. Yu, X. Zhang, and Z. Liu, "Sensitive determination of pyrethroid insecticide residues in tea using a molecularly imprinted fiber array based on homemade solid-phase microextraction coatings," *Microchemical Journal*, vol. 182, p. 107897, 2022, doi: https://doi.org/10.1016/j.microc.2022.107897.

[66] Y. Li *et al.*, "Fingerprinting black tea: When spectroscopy meets machine learning a novel workflow for geographical origin identification," *Food Chemistry*, vol. 438, p. 138029, 2024, doi: https://doi.org/10.1016/j.foodchem.2023.138029.

[67] J. Yang *et al.*, "TeaNet: Deep learning on Near-Infrared Spectroscopy (NIR) data for the assurance of tea quality," *Computers and Electronics in Agriculture*, vol. 190, p. 106431, 2021, doi: https://doi.org/10.1016/j.compag.2021.106431.

[68] C. Dong *et al.*, "Nondestructive testing and visualization of catechin content in black tea fermentation using hyperspectral imaging," *Sensors*, vol. 21, no. 23, p. 8051, 2021, doi: https://doi.org/10.3390/s21238051.

[69] Y. Song, X. Wang, H. Xie, L. Li, J. Ning, and Z. Zhang, "Quality evaluation of Keemun black tea by fusing data obtained from near-infrared reflectance spectroscopy and computer vision sensors," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 252, p. 119522, 2021, doi: https://doi.org/10.1016/j.saa.2021.119522.

[70] L. Li *et al.*, "Rapid and comprehensive grade evaluation of Keemun black tea using efficient multidimensional data fusion," *Food Chemistry: X*, vol. 20, p. 100924, 2023, doi: https://doi.org/10.1016/j.fochx.2023.100924.

[71] X. Luo *et al.*, "Using surface-enhanced Raman spectroscopy combined with chemometrics for black tea quality assessment during its fermentation process," *Sensors and Actuators B: Chemical*, vol. 373, p. 132680, 2022, doi: https://doi.org/10.1016/j.snb.2022.132680.

[72] L. Lu *et al.*, "An efficient artificial intelligence algorithm for predicting the sensory quality of green and black teas based on the key chemical indices," *Food Chemistry*, vol. 441, p. 138341, 2024, doi: https://doi.org/10.1016/j.foodchem.2023.138341.

[73] G. Jin *et al.*, "Rapid and real-time detection of black tea fermentation quality by using an inexpensive data fusion system," *Food Chemistry*, vol. 358, p. 129815, 2021, doi: https://doi.org/10.1016/j.foodchem.2021.129815.

[74] Y. Hu, L. Xu, P. Huang, X. Luo, P. Wang, and Z. Kang, "Reliable identification of oolong tea species: nondestructive testing classification based on fluorescence

hyperspectral technology and machine learning," *Agriculture,* vol. 11, no. 11, p. 1106, 2021, doi: https://doi.org/10.3390/agriculture11111106.

[75] P. Long *et al.*, "Discovery of color compounds: Integrated multispectral omics on exploring critical colorant compounds of black tea infusion," *Food Chemistry*, vol. 432, p. 137185, 2024, doi: https://doi.org/10.1016/j.foodchem.2023.137185.

[76] Q. Li *et al.*, "Machine learning technique combined with data fusion strategies: A tea grade discrimination platform," *Industrial Crops and Products,* vol. 203, p. 117127, 2023, doi: https://doi.org/10.1016/j.indcrop.2023.117127.

[77] L. Li *et al.*, "Evaluation of black tea by using smartphone imaging coupled with micro-near-infrared spectrometer," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy,* vol. 246, p. 118991, 2021, doi: https://doi.org/10.1016/j.saa.2020.118991.

[78] G. Ren, N. Gan, Y. Song, J. Ning, and Z. Zhang, "Evaluating Congou black tea quality using a lab-made computer vision system coupled with morphological features and chemometrics," *Microchemical Journal,* vol. 160, p. 105600, 2021, doi: https://doi.org/10.1016/j.microc.2020.105600.

[79] N. Mohammadi, M. Esteki, and J. Simal-Gandara, "Machine learning for authentication of black tea from narrow-geographic origins: combination of PCA and PLS with LDA and SVM classifiers," *LWT*, vol. 203, p. 116401, 2024, doi: https://doi.org/10.1016/j.lwt.2024.116401.

[80] V. G. K. Cardoso and R. J. Poppi, "Non-invasive identification of commercial green tea blends using NIR spectroscopy and support vector machine," *Microchemical Journal,* vol. 164, p. 106052, 2021, doi: https://doi.org/10.1016/j.microc.2021.106052.

[81] Y. Xu *et al.*, "A deep learning model for rapid classification of tea coal disease," *Plant Methods,* vol. 19, no. 1, p. 98, 2023, doi: https://doi.org/10.1186/s13007-023-01074-2.

[82] N. Luo, Y. Li, B. Yang, B. Liu, and Q. Dai, "Prediction model for tea polyphenol content with deep features extracted using 1D and 2D convolutional neural network," *Agriculture,* vol. 12, no. 9, p. 1299, 2022, doi: https://doi.org/10.3390/agriculture12091299.

[83] Y. Zhu *et al.*, "Classification of oolong tea varieties based on computer vision and convolutional neural networks," *Journal of the Science of Food and Agriculture,* vol. 104, no. 3, pp. 1630-1637, 2024, doi: https://doi.org/10.1002/jsfa.13049.

[84] L. Li, X. Sheng, J. Zan, H. Yuan, X. Zong, and Y. Jiang, "Monitoring the dynamic change of catechins in black tea drying by using near-infrared spectroscopy and

chemometrics," *Journal of Food Composition and Analysis*, vol. 119, p. 105266, 2023, doi: https://doi.org/10.1016/j.jfca.2023.105266.

[85] Z. Zou *et al.*, "Classification and adulteration of mengding mountain green tea varieties based on fluorescence hyperspectral image method," *Journal of Food Composition and Analysis*, vol. 117, p. 105141, 2023, doi: https://doi.org/10.1016/j.jfca.2023.105141.

[86] W. Xing *et al.*, "Suitability evaluation of tea cultivation using machine learning technique at town and village scales," *Agronomy*, vol. 12, no. 9, p. 2010, 2022, doi: https://doi.org/10.3390/agronomy12092010.

CHAPTER 2

DATA ANALYSIS

This chapter discusses various data analysis methods that leverage Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP). Besides, it covers details regarding optimisation methods such as the Genetic Algorithm (GA), Swarm Intelligence, Particle Swarm Optimisation (PSO), Ant Colony Optimisation, and Dragonfly Optimisation (DFO). Additionally, this chapter provides information on the performance metrics involved in regression and classification models. Finally, this chapter deliberates an explainable AI approach and its implementation challenges, concluding with a summary of the key details.

LIST OF SECTION

CHAPTER 2

DATA ANALYSIS

2.1. Introduction

The tea industry is currently experiencing a significant transformation, with the application of the disruptive force of Artificial Intelligence (AI) and its subset, Machine Learning (ML). This transformation is particularly evident in the areas of quality evaluation and decision-making. The unique ability of AI is analyse complex data sets across cultivation, harvesting, and processing, which enables precise quality assessment, production optimization, and adaptation to market conditions [1]. ML algorithms are especially valuable in the tea industry. These algorithms, which learn from data to identify patterns and correlations without explicit programming, are enhancing decision-making. This research explores various AI methods for tea quality, including classical approaches such as rule-based and fuzzy logic systems, which are well-suited for handling uncertainty and subjectivity in quality assessment. The discussion also covers ML methods using supervised and unsupervised learning for tasks such as classification, regression, and clustering in the tea sector.

In addition, this chapter describes Deep Learning (DL) technologies, a subset of ML that utilises artificial neural networks to mimic human information processing for tasks like image recognition and natural language processing. The focus will be on applying DL methods, particularly image-based feature extraction, to enhance tea quality evaluation. Beyond algorithmic discussions, the chapter emphasises the crucial role of performance classification metrics and error metrics for regression in validating the effectiveness of AI models. Additionally, the importance of explainable AI (XAI) has been highlighted, introducing techniques like SHAP and LIME to foster trust and responsible AI use by making complex AI decisions more transparent and interpretable for stakeholders. It also explores how AI can address the unique challenges within the tea industry, including fluctuations in raw material quality, environmental changes, and evolving customer preferences, by enabling real-time monitoring and agile management through the integration of diverse data sources. Finally, it can critically examine the hurdles to implementing AI within the tea sector, such as data quality issues, algorithmic bias, and the need for domain-specific expertise, while outlining future research directions to overcome these obstacles and unlock the full potential of AI in enhancing tea quality assessments.

2.2. AI Techniques

This section outlines the wide range of AI techniques in a way that allows us to hierarchically categorise them and relate them graphically through Figure 2.1. The goal is to visually and conceptually connect the algorithms, guiding the reader from fuzzy logic to ensemble and DL methods.

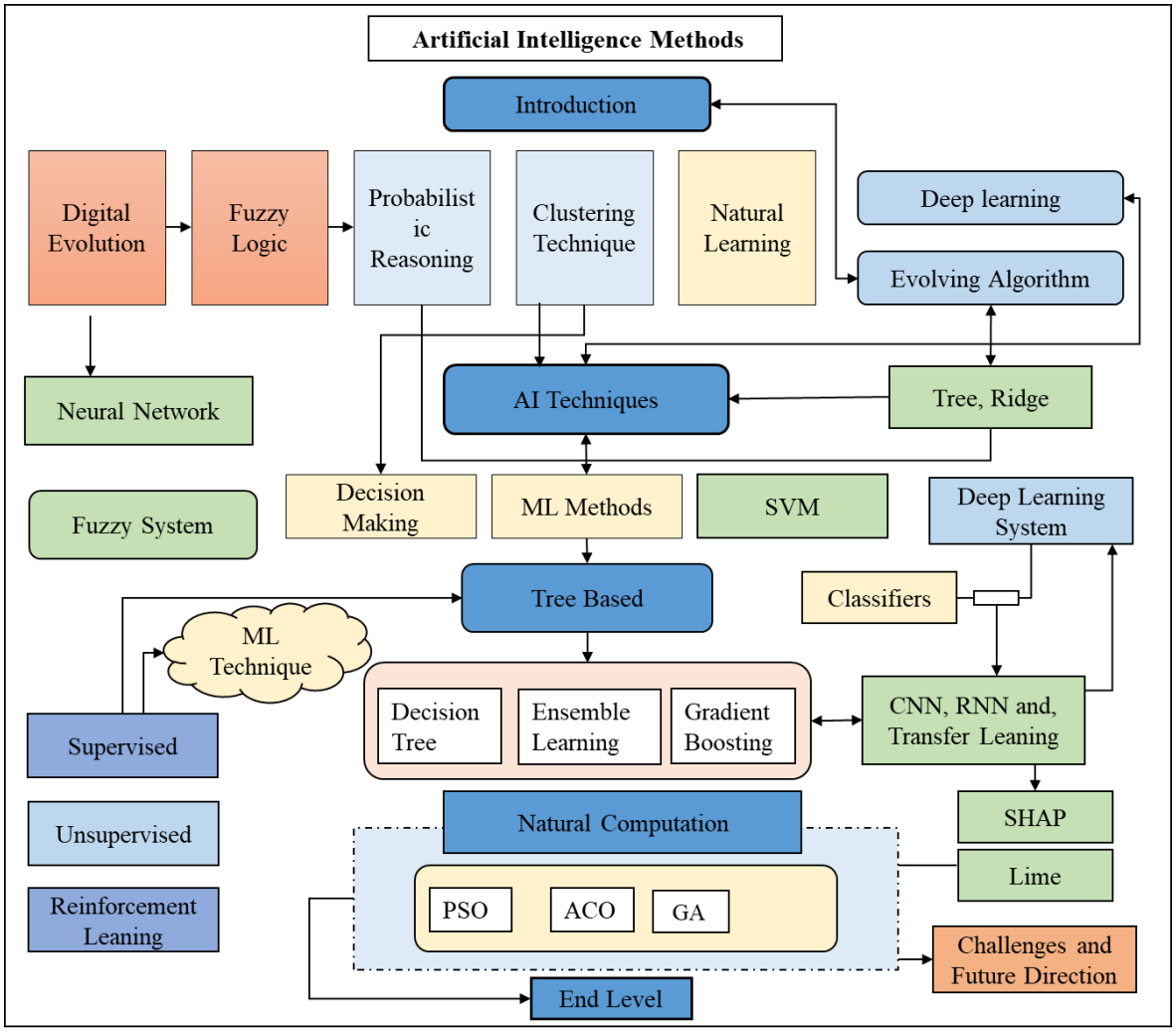


Figure 2.1 Data Analysis of AI Techniques

2.2.1. Decision Making

AI has become a powerful tool for enhancing decision-making across various domains, offering speed, accuracy, and scalability unmatched by traditional methods. AI systems can process massive datasets to identify patterns and correlations that are difficult or impossible for humans to detect, resulting in more informed and efficient decisions.

2.2.1.1. Rule-Based System (Fuzzy Systems)

Decision-making AI techniques encompass a variety of knowledge-based systems that employ reasoning techniques to integrate data and preexisting rules, enabling intelligent decision-making processes. Of these methods, rule-based systems, more specifically, fuzzy systems, are central to approximating human-like reasoning capabilities. Fuzzy systems, in their most prevalent formulation, employ linguistic IF-THEN rules, which enable the integration of words or uncertain information into a rules-based decision-making system. It implies that fuzzy systems can manage real-world complexity, particularly when precise data is unavailable or difficult to estimate.

Several key contributions have significantly advanced the development of fuzzy systems for AI-driven decision-making in uncertain environments.

The Fuzzy Wang Mendel Method is a data-driven approach that extracts fuzzy rules directly from initial datasets, allowing for the systematic incorporation of domain knowledge and enabling precise and robust reasoning for applications such as industrial automation and environmental monitoring [2, 3]. Fuzzy Adaptive Resonance Theory Mapping (Fuzzy ARTMAP) is a supervised neural-fuzzy architecture that combines neural networks and fuzzy logic for adaptable pattern recognition, utilising incremental learning to continuously update its knowledge base. This makes it well-suited for classification in complex and dynamic environments such as image recognition and speech processing [4]. Similarly, Fuzzy Neural Networks (FNNs) integrate fuzzy inference and neural network learning to enhance pattern classification. They utilise fuzzy membership functions to represent input variables and then employ a neural network to process these fuzzy inputs, resulting in increased accuracy in applications such as financial forecasting and medical diagnosis. Finally, the Fuzzy Recogniser of Signal with Time (FRST) addresses signal classification by using a fuzzy linguistic approach to represent the temporal dynamics of sensor data [5]. This enables it to identify patterns based on temporal changes, thereby improving decision-making in areas such as real-time monitoring and smart manufacturing. Together, these advancements highlight the evolving capability of fuzzy systems to address complex decision-making in diverse applications.

2.2.2. Machine Learning Techniques

ML techniques are not just pivotal; they are transformative in leveraging data for sophisticated learning and prediction, surpassing the limitations of traditional coding methods based on

predefined rules. This approach enables systems to continually improve over time by directly acquiring knowledge from data, making them inherently flexible and responsive. The transformative power of ML is not just evident; it's inspiring in diverse applications, such as image identification, natural language processing, and predictive analysis [6].

2.2.2.1. Unsupervised Learning

Unsupervised learning represents an essential category of ML in which algorithms are trained using a dataset lacking labelled outputs. The primary objective of supervised learning is to discern the underlying structure or distribution within the data, thereby uncovering hidden patterns within any prior knowledge of the expected clustering tasks, and dimensionality reduction application [7].

2.2.2.1.1. Clustering

Clustering algorithms are a type of unsupervised learning algorithm that are used to group a set of objects such that the objects in one group (or cluster) are more similar to each other than objects in different groups. This clear explanation should make the audience feel knowledgeable and confident in their understanding of ML. Clustering algorithms have many applications, such as market segmentation, social network analysis, organising computing clusters, and image compression [8].

a. K-means

One of the most common clustering algorithms is K-means, which works by dividing the data set into k unique groups. K-means is an iterative algorithm that assigns each data point to one of the k clusters based on feature similarity. The goal is to achieve low intra-cluster variance and high inter-cluster variance. This is achieved by computing the centroids of the clusters and dynamically updating them as data points are allocated, leading to an increasingly accurate clustering solution [9].

2.2.2.1.2. Dimensionality Reduction

Dimensionality reduction involves the use of algorithms to reduce the number of features in a dataset while retaining its general characteristics. By lowering the dataset, dimensionality reduction can improve model accuracy, decrease complexity, and minimise the effects of the curse of dimensionality, which arises when the feature space becomes sparser as the number

of dimensions increases [10]. Besides, dimensionality reduction methods such as t-Distributed Stochastic Neighbour Embedding (t-SNE) are highly beneficial for visualising high-dimensional tea quality data. By reducing numerous attributes into two or three dimensions, producers can better understand inter-variable relationships, leading to more informed cultivation and processing decisions.

Principal Component Analysis (PCA)

PCA is a linear method that finds directions of maximum variance and projects data onto a lower-dimensional subspace. It's useful for noise reduction and data compression. PCA achieves this by performing Eigen decomposition on the data's covariance matrix, according to GeeksforGeeks [11]. The following provides the steps involved in PCA.

Standardise the Dataset: Centre the data by subtracting the mean of each feature from the dataset.

$$X' = X - \mu \tag{2.1}$$

Where, X is the original dataset, μ is the mean vector of the dataset, and X' is the centered data.

Calculate the Covariance Matrix: Compute the covariance matrix to understand how the features vary in relation to one another.

$$C = \frac{1}{n-1} X^T X \tag{2.2}$$

Where X is the data matrix, and n is the number of observations. The principal components are derived from the eigenvectors corresponding to the largest eigenvalues.

Calculate Eigenvalues and Eigenvectors: Solve the characteristic equation to find the eigenvalues and eigenvectors of the covariance matrix.

Sort Eigenvalues and Eigenvectors: Sort the eigenvalues in descending order and select the top k Eigenvalues. The corresponding eigenvectors form the new feature space. $Sort\ (\lambda i)\ and\ select\ top\ K$.

Form the feature vector: Construct a feature vector by forming a matrix of the selected eigenvectors.

$$W = [e_1, e_2, \dots\dots, e_k] \quad (2.3)$$

Where W Is the feature vector and e_i These are the selected eigenvectors.

Transform the original Dataset: Finally, transform the original into the new feature space using the feature vectors.

$$Z = X'W \quad (2.4)$$

Here Z Is the transformed dataset in the reduced-dimensional space?

a. UMAP

UMAP is a modern non-linear dimensionality reduction approach that preserves both local and global structures. Unlike PCA, which assumes linear relationships among features, UMAP uses manifold learning to produce a low-dimensional representation of high-dimensional data. UMAP helps visualise complex datasets and is quickly becoming popular in genomics, image analysis, and webs of complex data visualisation [12]. UMAP is utilised to optimise the extracted features from the rolling bearing dataset to decrease dimensionality while maintaining the key structure of the data [13].

b. Independent Component Analysis (ICA)

ICA is a complex statistical technique used to disassemble multivariate signals into independent components that are additive. ICA is at its most valuable when the observed data is a mixture of many source signals, and the source signals can be assumed statistically independent [13].

$$X = AS \quad (2.5)$$

Here X is the observed mixed signal, A is the mixing matrix, and S Represent the independent source signals.

ICA is referred to in data analysis for an intelligent agricultural environment monitoring system. ICA separates mixed signals into independent components, thereby isolating and analysing individual environmental factors that are monitored by the wireless sensor network.

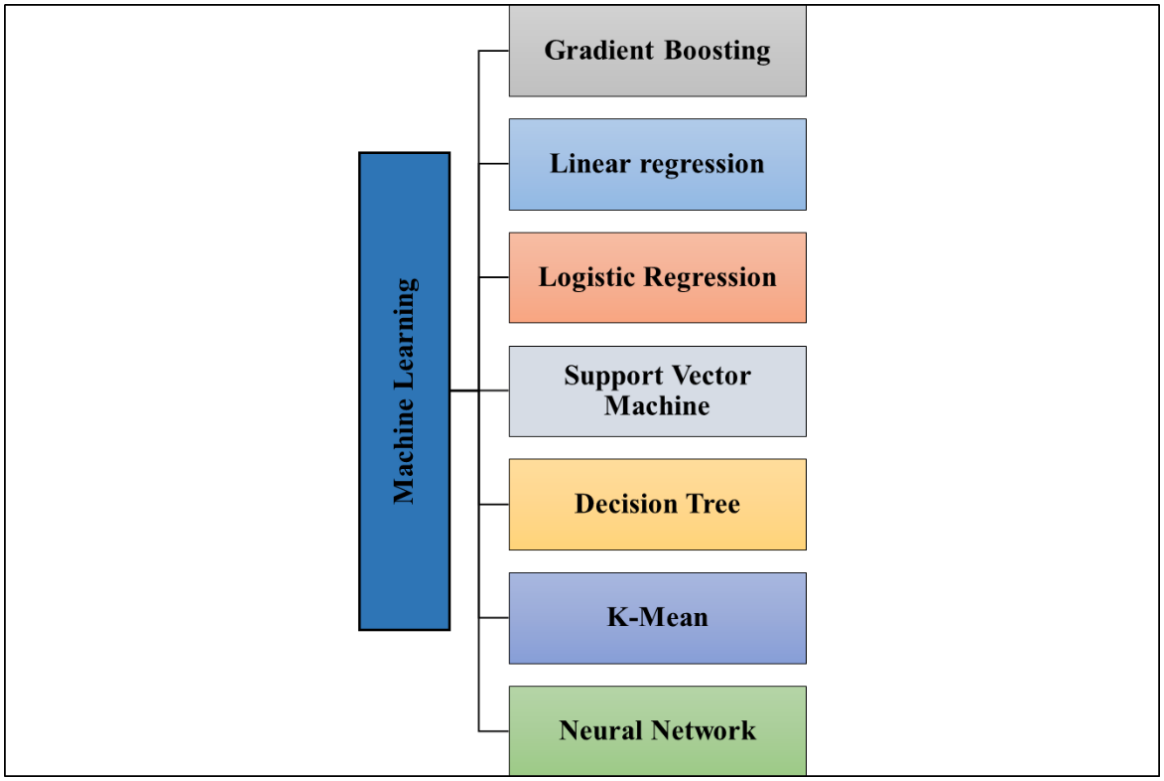


Figure 2.2 Supervised Algorithms

2.2.2.2. Supervised Learning

Supervised learning is an effective paradigm in ML, where models are trained on labelled datasets to make predictions or classifications. For example, when analysing electronic nose and tongue data to determine tea quality, effective supervised learning techniques are necessary to draw valuable conclusions from the multivariate outputs of sensors. The use of supervised learning techniques, as illustrated in Figure 2.2, can significantly enhance the accuracy of estimates regarding a tea quality measurement [14, 15].

2.2.2.2.1. Regression

a. Linear Regression along with Regularisation Techniques

Linear Regression (LR)

LR is a fundamental method of regression analysis. It works on the assumption of finding a linear relationship between one or more input variables (features) and a continuous output variable. Being simple, the model is extremely interpretable, enabling practitioners to see the effect of each feature on the predicted output [16].

Despite the advantages associated with employing linear regression methods, there are also weaknesses. Linear regression can be highly susceptible to the noise or outliers of the data, which can distort the accuracy of results and predictions. Furthermore, linear regression can produce unreliable models when any of the predictor variables are highly correlated with each other (referred to as multi-collinearity), since this can create instability in the coefficient estimates. The mathematical representation of linear regression can be expressed as,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \in \quad (2.6)$$

Where, y is the predicted output, β_0 is the intercept, β_1 Are the coefficients, and x_1 are the input features, $\in$ Is the error term.

b. Regularisation Techniques

Regularisation methods play a pivotal role in ML applications, especially for regression models. Overfitting, sensitivity to noise, and multicollinearity are three significant challenges to the robustness of predictive models. Lasso (L1 regularisation), Ridge (L2 regularisation), and Elastic Net regularisation are the three essential regularisation methods that provide much to model robustness and interpretability.

Lasso Regression

Lasso Regression adds an absolute value penalty to the coefficients, which forces the model to use a sparse representation. As a result, the model shrinks some of the coefficients to zero, effectively performing feature selection. By retaining only the most significant predictors, Lasso reduces the model's complexity, making it easier to interpret while also reducing the risk of overfitting. This is particularly helpful in high-dimensional data, where many features may not make significant contributions to the prediction [17].

The cost function for Lasso regression can be defined as

$$J(\beta) = n1 \sum_{i=1}^{n} (yi - y^i)2 + \lambda \sum_{j=1}^{p} | \beta j | \quad (2.7)$$

Ridge Regression (RR)

In contrast, RR uses a squared penalty on coefficients, which can shrink larger weights but cannot eliminate any features from the model. This is an attractive property. Ridge is a suitable modelling procedure when there is a high correlation among predictor variables, as this

stabilises the estimates and improves performance [18]. Since we retain all the features in the model, Ridge provides a more comprehensive understanding of each feature's contribution to the model's output. For regression, the cost function is,

$$J(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (2.8)$$

c. Elastic Net Regression

Elastic Net surpasses Lasso and Ridge by using both L1 and L2 penalties. By coupling the two, Elastic Net promotes sparsity with coefficient estimate stability. It is beneficial in situations where predictors are correlated, as it surpasses the strength of feature selection by retaining the stability of many correlated variables.

During sensory data analysis, methods like electronic nose and tongue system analysis of tea quality may benefit from the use of regularisation techniques. These statistical methods will go a long way to improving model accuracy. By utilising regularisation, the models we adopt will be better fitted to the data, while also being more interpretable. This enhances our understanding of the underlying variables that affect the tea quality in question [19].

d. Polynomial Regression

Polynomial regression is a regression analysis in which the relationship between the independent variable and dependent variable is represented as an nth-degree polynomial. Simple linear regression, on the other hand, expects linearity in the relationship. Still, polynomial regression permits the establishment of a polynomial equation between the data that represents a nonlinear, higher-order relationship. This adaptability makes it particularly effective in several domains such as economics, biology, and engineering, where the relationship at hand is not necessarily linear. Since polynomial terms are involved, this technique can identify trends within the data that linear models can't, which will ultimately lead to more accurate predictions and conclusions [20].

e. Partial Least Squares (PLS)

PLS is a proper statistical technique commonly used in chemometrics and sensor data analysis. PLS is efficient when the number of predictors is larger than the number of observations, which can lead to issues with collinearity. The PLS technique achieves this by projecting the input features and output responses to a new latent space with maximum covariance, allowing for the construction of regression models from the latent variables.

One key advantage of PLS over traditional approaches, such as PCA, is its dual focus on both input variance and the relationship between inputs and outputs. This characteristic enables PLS to effectively reduce dimensionality while preserving predictive power, making it an ideal choice for fusing electronic nose and tongue data. In such applications, where several correlated sensor signals interact with various superiority parameters, PLS proves to be an invaluable tool. [1].

$$Y = XB + E \tag{2.9}$$

Here, Y Is the matrix of responses, X Is the matrix of predictors, and B Is the matrix of coefficients and E Is the residual Matrix.

f. Support Vector Machines (SVM)

SVMs are a type of sophisticated margin-based classifier and regressor that seeks an optimal hyperplane that successfully separates two classes of data with the maximum margin. Such a process improves generalisation, making SVM valuable in high-dimensional feature spaces [21].

SVMs seek to find an optimal hyperplane that separates different classes in the feature space. The optimisation problem can be expressed as,

$$\min \frac{1}{2} \left|\left|w\right|\right|^2 \tag{2.10}$$

Subject to the constraints,

$$y_i(w^T x_i + b) \geq 1, \forall_i \tag{2.11}$$

Where w the weight vector, b is the bias, and y_i denotes the class labels.

The underlying idea of SVM is to maximise the margin between classes, which leads to better classification. Additionally, the invention of the kernel trick enables mapping input data into higher-dimensional feature spaces using multiple kernel functions, such as linear, polynomial, and radial basis functions, to effectively model nonlinear relationships.

SVMs offer several benefits, not the least of which is their ability to avoid overfitting even in contexts with small sample sizes. The process of maximising a margin, in and of itself, is a level method of protecting against noise and outliers. Additionally, SVM can be considered in

the context of regression with Support Vector Regression (SVR), where it employs epsilon-insensitive loss functions while still having acceptable predictive performance [22, 23]. Table 2.1 provides a summary of the regression models.

Table 2.1 Summary of Key Points of Regression Models

Technique	Purpose	Key Characteristics	Strengths
Linear Regression	Predict continuous output	A linear model sensitive to noise and multicollinearity	Interpretable and straightforward, but prone to overfitting
Lasso (L1)	Regularisation for feature selection	Shrinks coefficients, promoting sparsity	Selects relevant features, combats overfitting
Ridge (L2)	Regularisation for coefficient shrinkage	Penalises large coefficients	Stabilises estimates when predictors are correlated
Elastic Net	Hybrid regularization	Combines L1 and L2 penalties	Balances feature selection and stability
PLS	Dimension reduction + regression	Projects data to latent variables, maximising covariance	Handles collinearity and compresses data effectively
SVM	Classification and regression	Maximises margin; kernel transforms data to nonlinear spaces	Powerful with nonlinear boundaries, robust margins

g. Tree-Based Modelling

Tree-based models have become foundational models in traditional data analysis because they are intuitively interpretable and flexible. For sensor data analysis and classification projects, such as those involving electronic nose and tongue data, tree models enable the recursive partitioning of the feature space into easily interpretable and manageable decision regions [24].

Decision Tree (DT)

A DT is a more structured, flowchart-like model that can be applied to decision-making plans and classification. Here, the internal nodes represent tests on the features, branches represent the outcome of the tests, and leaf nodes represent the prediction (or class label). A decision tree algorithm employs a recursive partitioning approach, where the dataset is split at each node based on the feature value that yields the most significant increase in class purity or reduction in classification error, as determined by the square root method for the cross-entropy criterion. Different criteria can be used for this step, such as Gini impurity and entropy, for classification [25].

For instance, in assessing tea quality, features derived from sensors that measure aroma and taste can be successively split to form distinct groups, each expressing different grades of tea. This approach enhances the understanding of how various attributes contribute to overall tea quality [26].

h. Ensemble Learning

Individual DTs can be reasonably interpretable; however, they can suffer from high bias and variance. Ensemble learning combines multiple models, leveraging their collective power to evaluate the overall performance of the ensemble. There are many ensemble techniques, but the two main ones we are concerned with here are bagging and boosting. Bagging, also known as bootstrap aggregating, aims to reduce variance by averaging predictions from multiple independent models. Boosting improves accuracy sequentially by increasing the weight of the data points that were previously misclassified. This can lead to more accurate and robust predictions [27].

i. Bagging

Bootstrap Aggregating (often referred to as Bagging) is an ensemble learning method designed to reduce variance by averaging the predictions of several base learners. Bagging involves generating bootstrapped samples, which are randomly selected with replacement from the original dataset. The different models are heterogeneous, and Bagging is built around leveraging that heterogeneity [28].

Random Forest (RF)

RF is a robust ensemble learning algorithm capable of handling nonlinear relationships and complex feature interactions [29]. It utilises an ensemble learning method, combining multiple decision trees to produce a single, more accurate prediction. Each decision tree within the forest is trained on a distinct, randomly generated subset of the original dataset, created through bootstrapping (random sampling with replacement). Additionally, at each node within a tree, only a random subset of features is considered for splitting, further enhancing diversity. For classification, the final prediction is based on the majority vote from all trees, while for regression, it's the average of all trees' predictions [30].

j. Boosting

Boosting is a sequential ensemble techniques that transform a set of weak learners (models slightly better than random guessing) into a strong learner by iteratively emphasising hard-to-predict samples. Boosting algorithms enhance performance by additively constructing models that correct the mistakes of earlier models, under the direction of the gradient of a loss function [31, 32].

AdaBoost

AdaBoost (Adaptive Boosting) is a powerful ensemble learning technique that sequentially trains weak classifiers, such as decision stumps, to construct a robust strong classifier. It begins by assigning equal weights to all training instances. In each subsequent iteration, a weak learner is trained, and the weights of misclassified instances are adaptively increased, forcing the next learner to focus on these more challenging cases. This iterative process effectively reduces both bias and variance. Ultimately, AdaBoost combines the outputs of these weak learners through a weighted voting system, where each learner's contribution is proportional to its accuracy, yielding a highly accurate and robust classifier for various classification problems.

Gradient Boost

Gradient Boosting constructs models sequentially, with each subsequent model refining the mistakes of the previous one. It minimises a loss function by reducing residuals via gradient descent, allowing it to handle complex datasets and enhance prediction performance through cautious model fitting. Extends the boosting framework by optimising any differentiable loss function using gradient descent. At each boosting iteration, a new weak learner is fit to the

pseudo-residuals (the gradient of the loss function with respect to current predictions). Improves prediction capability by sequentially adding trees that minimise the residual error. Allows tuning of learning rate and tree depth to control overfitting and convergence speed.

Light Gradient Boosting Machine* (*LightGBM)

LightGBM was designed with efficiency and scalability in mind, particularly for handling massive datasets. LightGBM employs a histogram-based method for accelerated training and reduced memory consumption, as well as the ability to parallelise and utilise GPUs, which is why it works well for high-dimensional datasets.

- An optimised gradient boosting framework designed for enhanced computational efficiency and scalability.
- Utilises histogram-based algorithms that bucket continuous feature values, accelerating split finding.
- Employs leaf-wise tree growth with depth constraints rather than level-wise, often yielding better accuracy with fewer trees.
- Scales well to large datasets and supports parallel and GPU training.
- Handles categorical features natively and reduces memory consumption.

Extreme Gradient Boosting (XGBoost)

XGBoost extends conventional gradient boosting methods by adding regularisation, which prevents overfitting. Its distributed computing and optimisation across memory and speed have made it a popular option in several ML competitions and practical implementations. It is a highly optimised and scalable gradient boosting implementation with regularisation features. It incorporates both L1 and L2 regularisation on tree weights to combat overfitting. Additionally, it supports parallel processing and distributed computing, making it well-suited for handling large and complex datasets. Additionally, it provides automated handling of missing data and customised objective functions [33].

Table 2.2 represents a comparison of tree-based and ensemble methods highlighting the concept and key features along with its pros and cons.

Table 2.2 Tree-Based and Ensemble Methods

Method	Concept	Key Features	Pros	Cons
Decision Tree	Recursive partitioning	Easy interpretability, handles mixed data types	Transparent, simple, fast	Prone to overfitting, unstable
Random Forest	Bagging + feature randomness	Bootstrap sampling, random feature subset splitting	Reduces variance, estimates variable importance	Less interpretable than a single tree
AdaBoost	Sequential weighting of errors	Reweigh samples based on errors	Reduces bias, focuses on complex examples	Sensitive to noise and outliers
GBM	Gradient descent on loss	Fits trees to residual gradients	Flexible loss functions, accurate	Potentially slow training
LightGBM	Efficient Gradient Boosting	Leaf-wise growth, histogram bins, parallelisation	Fast training, scalable, accurate	May overfit if not tuned
XGBoost	Regularised gradient boosting	L1/L2 regularisation, distributed training	Highly optimised, handles large datasets, robust	Complexity in tuning

2.2.3. Deep Learning (DL)

DL, a powerful subset of supervised Learning, has garnered significant attention for its superior aptitude in addressing complex pattern recognition challenges [34].

CNNs are a type of DL architecture that excel in processing spatially correlated data, such as images. Unlike shallow learning, where features are manually designed, CNNs autonomously

learn to extract features hierarchically through layers of convolution, pooling, and non-linear activation. This depth enables them to identify increasingly complex patterns, mirroring aspects of the human visual system. CNNs have transformed image processing applications such as object detection and medical image analysis, achieving high accuracy. They also show promise in analysing sensor data, including taste and volatile patterns, in applications such as black tea quality evaluation, by interpreting spatial and temporal correlations in sensor arrays. One of their key strengths is their efficiency in handling diverse input data, which reduces the need for domain-specific feature engineering common in other models [35, 36]. This adaptability instils confidence in their applicability to a wide range of tasks.

2.2.4. Natural Computing

Natural Computing encompasses a range of computational methods and techniques inspired by biological processes and the collective behaviours observed in nature. These techniques optimise and model complex systems through methods that are designed to mimic biological evolution, neural processes, and swarm intelligence. For example, evolutionary algorithms simulate natural selection, which evolves optimal solutions through mechanisms like mutation, crossover, and selection [37]. Natural Computing approaches are particularly adept at addressing diverse challenges in optimisation, ML, and AI. With approaches influenced by nature, they optimise and model different classes of complex systems as they often require overcoming large solution spaces, evolving in complex environments. AI techniques have demonstrated significant potential in addressing complex tea quality evaluation problems [38].

2.2.4.1. Evolutionary Computation

Evolutionary computation is a collection of population-based, stochastic optimisation techniques based on the process of natural selection acting in biological evolution [39]. Of these, the most well-known is the Genetic Algorithm (GA), which mimics the principle of "survival of the fittest" to improve solutions over generations [40].

Genetic Algorithm

GA maintains a population of candidate solution individuals, each represented by an encoded chromosome. Each individual is measured in terms of fitness with respect to an objective function defined by the problem. The GA applies selection to choose fitter individuals to transmit genetic material; crossover, which merges two or more solutions (parents) to create

new individual solutions (offspring); and mutation, which introduces random perturbations to maintain diversity in the search process and avoid premature convergence [41].

The effectiveness of each solution is evaluated using a fitness function tailored to the specific objectives of the problem. GAs employ three primary genetic operations,

- **Selection:** This process identifies and favours individuals with higher fitness, forming a mating pool from which genetic material is transmitted to create new solutions.
- **Crossover:** By merging two or more parent solutions, crossover generates offspring that inherit characteristics from their parents, facilitating the exploration of novel regions within the solution space.
- **Mutation:** Random perturbations are introduced to offspring chromosomes to maintain genetic diversity, thereby preventing the algorithm from converging prematurely on suboptimal solutions.

This evolutionary process enables GAs to perform a global search in high-dimensional, nonlinear, and complex spaces without gradient information, making them suitable for parameter optimisation in fuzzy classifiers, adjusting sensory data processing pipelines, and feature selection within sensory fusion systems, such as black tea quality assessment [42]. The stochastic nature and population-based search of GAs confer them inherent robustness to local minima, adaptability to changing situations, and flexibility in encoding a diverse range of problem constraints and objectives. This robustness ensures their reliability in finding optimal solutions, providing a sense of security in their application [43].

2.2.4.2. Swarm Intelligence (SI)

SI is a class of optimisation algorithms inspired by the collective behaviour of decentralised, self-organised systems found in nature, such as bird flocks or ant colonies. These methods utilise simple agents that interact locally with their environment and each other to solve complex optimisation problems collaboratively. SI approaches are beneficial for tackling multidimensional, non-linear challenges often encountered in areas such as sensor signal processing and feature extraction. Through exploration and exploitation processes, these methods efficiently navigate vast search spaces at minimal cost. They are applicable in areas like fusion model estimation and parameter optimisation in multi-sensor systems. In essence, SI offers compelling optimisation engines by mimicking natural self-organisation to achieve global solutions without needing a central authority.

Particle Swarm Optimisation (PSO)

PSO is a population-based meta-heuristic based on swarms of animals, such as schools of fish or flocks of birds. A swarm of candidate solutions, called particles, work together to explore the search space. Each particle modifies its position by utilising its own experience and that of its neighbours, balancing personal best and global best positions to converge to optimum solutions [44]. In PSO, every particle has a velocity vector that dictates its motion within the search space. Velocity and position are updated iteratively based on the particle's optimal known position (pbest), which is the best solution found by it so far. The swarm's global best position (gbest), which represents the optimum solution discovered by any particle [45].

Ant Colony Optimisation (ACO)

Ant Colony Optimisation [46] is inspired by the foraging behaviour of real ants, which effectively find the shortest path between their nest and food sources. Artificial ants in ACO construct candidate solutions by probabilistically traversing different states (nodes or edges in a graph) based on two key factors: pheromone trails and heuristic information. Ants deposit pheromones on the paths they take, with the amount being proportional to the quality of the solution found; this makes better paths more attractive to subsequent ants. This positive feedback mechanism reinforces reasonable solutions. However, to prevent premature convergence to suboptimal solutions, pheromones evaporate over time, adding a dynamic element that encourages exploration of alternative paths. The probabilistic nature of path selection enables ACO to balance exploring new routes and exploiting known good ones, resulting in high-quality solutions for complex optimisation problems over repeated iterations [47].

Dragon Fly Optimisation (DFO)

The DFO is a metaheuristic swarm intelligence algorithm inspired by the static (hunting) and dynamic (migratory) swarming behaviours of dragonflies.

It leverages five core behaviours: separation to avoid collisions, alignment to match velocities, cohesion to move towards the centre of the swarm, attraction to food sources, and distraction from enemies. These behaviours, controlled by adjustable weights, facilitate both exploration (searching broadly for solutions) and exploitation (refining promising solutions). DFO works by computing a velocity vector for each artificial dragonfly based on these behaviours,

allowing them to estimate and update their positions in the search space. This balancing act between exploration and exploitation makes DFO effective for optimisation problems, including the optimisation of clustering algorithms, parameter tuning, and sensor calibration in dynamic and uncertain environments. DFO's ability to adapt to changes in data distribution and quality further enhances its utility in real-world scenarios [48]. Table 2.3 summarises the key details of the optimisation algorithm used in tea quality assessment.

Table 2.3 Summary Table for Evolutionary Methods

Method	Base Learner	Key Features	Pros	Cons
GA	Population of candidates	Selection, Crossover, Mutation	Robust against local minima; adaptable; versatile	Can be slow to converge; requires careful parameter tuning
PSO	Particles in a swarm	Velocity and position updates; balance between personal and global bests	Simple to implement; fast convergence; few parameters	May get stuck in local optima; sensitive to parameter settings
ACO	Artificial ants	Pheromone depositing, evaporation, probabilistic path choice	Effective for discrete problems; adaptive to dynamic environments	Computationally intensive; performance can degrade with increasing problem size
DFO	Artificial dragonflies	Separation, Alignment, Cohesion, Attraction, Distraction	Good exploration-exploitation balance; robust in complex environments	Newer method, less tested compared to others; complexity in implementation

2.3. Performance Metrics

Performance metrics are crucial for objectively evaluating the quality of tea and analysing sensor data. These metrics quantitatively assess the efficiency and dependability of models, establishing baselines for comparison between different techniques. For tea quality assessment using sensor data, such as from electronic noses, tongues, and MIP-based chemical sensors, both regression-based and classification-based metrics are employed to ensure a robust assessment of accuracy. This approach addresses the limitations of traditional subjective sensory evaluation methods, which often suffer from inconsistencies and bias. By integrating DL with sensor technologies, it's possible to analyse multivariate and time-series data objectively, enabling accurate predictions and classifications of tea quality factors like flavour, aroma, visual appearance, caffeine, and polyphenol content, thereby revolutionising the field.

2.3.1. Regression-Based Performance Metrics

Regression is typically used to predict continuous variables, such as the levels of critical chemical compounds in tea. The most crucial regression measures used to assess the degree to which predicted values match actual measured values are as follows:

1. R^2 (Coefficient of Determination): R^2 represents the proportion of variance in the observed data explained by the model. The closer the R^2 value is to 1, the better the model's fit and prediction ability, which means that the model more accurately describes the underlying process in the data.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (2.12)$$

Where y_i Is the actual value, $\hat{y}_i$ is the predicted values, and $\bar{y}$ is the mean of actual values

2. Mean Squared Error (MSE): MSE is a measure of the average of the squared differences between predicted values and the actual values. Lower MSE values indicate greater accuracy, and MSE penalises larger errors more than minor errors, which may be more informative in some situations, such as when outliers can substantially skew the results.

$$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 \quad (2.13)$$

Here n is the number of observations

3. Root Mean Squared Error (RMSE): RMSE is the square root of MSE and has the same unit as the predicted variable, but is more interpretable. It is low for good model performance.

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \quad (2.14)$$

4. Mean Absolute Error [49]: The MAE is the average of all absolute errors of the predicted versus actual values. MAE provides an understanding of the average amount of a prediction error, with a smaller MAE indicating better accuracy. Therefore, MAE is advantageous when evaluating models in situations of consistency.

$$MAE = \frac{1}{n}\sum|y_i - \hat{y}_i| \quad (2.15)$$

These regression measures are typically applied in research studies that project determinants of tea quality based on sensor data. Hence, regression approaches can enhance predictive performance and, therefore, validity for the evaluation of tea quality.

5. Mean Absolute Percentage Error (MAPE): MAPE is a valuable metric for evaluating the accuracy of regression models. It provides an intuitive interpretative meaning of model performance while also averaging the absolute percentage errors between predicted values and actual values. The MAPE is given by,

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| * 100 \quad (2.16)$$

Here, A_i are the actual values for observation i, F_i are the forecast values for the observation i.

Adding MAPE to the assessment of regression models provides a more comprehensive picture of model performance, alongside classical metrics such as MAE, MSE, and RMSE. It is beneficial when the interpretability of errors is essential in making decisions.

2.3.2. Classification-Based Performance Metrics

Classification issues are used where one wishes to classify tea samples into distinct classes or quality grades based on sensor responses, spectral data, or other sets of features. The following classification measures are crucial for determining how effectively a model distinguishes between different classes.

1. Confusion Matrix

A confusion matrix [20,36] is an essential evaluation tool for a classification model. It provides a summary of how well the model is performing based on the model's predicted classifications and the actual classifications. A confusion matrix is usually laid out as follows:

- **True Positive (TP):** The number of instances correctly predicted as positive.
- **True Negative (TN):** The number of instances correctly predicted as negative.
- **False Positive (FP):** The number of instances incorrectly predicted as positive (Type I error).
- **False Negative (FN):** The number of instances incorrectly predicted as negative (Type II error).

2. Accuracy: This metric measures the fraction of correctly classified samples out of the total samples. This is an overall metric for measuring the correctness of a classification. In cases where there are rare tea grades, we often won't provide a clear picture of model performance when examining only accuracy.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.17}$$

Here, TP is True Positive, TN is True Negative and FP is a False Positive and FN It is a False Negative.

3. Precision: Precision measures the proportion of true positives from all predicted positives. Precision level should be high (ideally, none) because you don't want to produce false positives in terms of quality. In tea grading, accurately identifying the quality rank is crucial because the overall quality determines the difficulty of the tea.

$$Precision = \frac{TP}{TP+FP} \tag{2.18}$$

4. Recall (Sensitivity): Recall quantifies the percentage of true positives that are accurately detected. In tea quality evaluation, high recall ensures that excellent-quality samples are correctly classified, which is crucial for maintaining consumer confidence and satisfaction.

$$Recall = \frac{TP}{TP+FP} \tag{2.19}$$

5. F1-Score: F1-score is the harmonic mean of recall and precision, giving an optimal balance of the model's accuracy that is very helpful in the case of imbalanced datasets. The F1-score is useful when the costs of false positives and false negatives are significantly different, allowing for a more discriminative assessment of model performance.

$$F1 = 2 * \frac{Precision.Recall}{Pricision+Recall} \tag{2.20}$$

6. ROC Curve and AUC (Area under the Curve): The ROC curve plots the True Positive Rate against the False Positive rate, while the AUC quantifies the overall discrimination ability of the models. A higher AUC indicated better model performance, with an AUC of 1 denoting a perfect classifier. This metric is useful for comparing multiple classification models and selecting the one that performs best.

$$TPR = \frac{TP}{TP+FN} FPR = \frac{FP}{FP+TN} \tag{2.21}$$

The use of performance measures in the evaluation of tea quality has been evidenced by several studies, portraying their significance in model assessment:

2.3.3. Clustering-Based Performance Metrics

Clustering algorithms are a significant methodology applied to a broad range of applications, including market segmentation, social network analysis, and image compression. Rationale for the accuracy and performance of clustering procedures is helpful in ensuring that the algorithms correctly classify data points into similar groups.

2.4. Explainable AI (XAI)

XAI aims to develop an understanding of how AI models arrive at specific conclusions, which is crucial for establishing trust among users and stakeholders. For instance, in assessing tea quality, a decision could significantly impact product quality, ultimately affecting economic value, as well as consumer satisfaction and loyalty. Thus, explainability is essential and sometimes necessary [10, 50]. The approaches SHAP and LIME have been widely applied in improving the interpretability of AI models:

2.4.1. Shapley Additive exPlanations (SHAP)

SHAP This methodology derives from cooperative game theory and allocates an importance value for each feature in determining a specific prediction made [51]. SHAP values are a

method for measuring the contribution of each feature to the final prediction. The mathematical equation for SHAP values can be shown:

$$\phi_i(f) = \sum S \subseteq \frac{N}{\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} \left(f(S \cup \{i\}) - f(S)\right) \quad (2.22)$$

$\phi_i(f)$ is the SHAP value for the feature i, and S represents a subset of features, N is the subset of all features, $f(s)$ is the prediction for the subset S.

2.4.2. Local Interpretable Model-agnostic Explanations (LIME)

LIME locally approximates complex models with interpretable models, providing insights into individual predictions. It creates local surrogate models by perturbing the input and observing the resulting changes in the output.

- **Feature Importance:** Numerous ML algorithms, including RF and Gradient Boosted Trees, offer inherent measures of feature importance in the form of how much each feature helps to minimise error when the model is being trained.
- **Partial Dependence Plots (PDP):** Such plots illustrate the connection between a feature and the predicted output, averaging out the influence of other features. They assist in comprehending the impact of changes in feature values on predictions.

2.4.3. Challenges in Explainable AI

Implementing XAI in tea quality prediction faces the following key challenges.
Firstly, the inherent complexity of advanced AI models, such as deep neural networks, makes it challenging to generate explanations that are both accurate and easily understandable to humans. This often leads to a trade-off where increasing model accuracy can reduce explainability, and vice versa. Secondly, there is a lack of universal XAI methods, as different industries and user groups require distinct types and levels of explanation. This means that a self-driving car's AI may need to explain a decision differently to a passenger than to an engineer. Finally, the absence of standardised evaluation metrics for XAI makes it challenging to compare different approaches and consistently assess the quality and reliability of explanations, hindering the widespread adoption of XAI solutions [52].

2.5. Challenges and Future Direction

While applying AI methods in the tea sector, there could be challenges which must be addressed to realise their full potential.

1. **Data Quality Problems:** Valid AI conclusions are highly dependent on quality data. This often includes noise and bias introduced by various data collection methods, sensor calibration, and environmental conditions, among others. Accordingly, we require a systematic approach to data validation and cleaning.
2. **Algorithm Bias:** AI algorithms may overlay biases inadvertently ingrained in their training data. Therefore, we should perpetually analyse and evolve our databases and conduct equity assessments as part of the modelling process to ensure equitable outcomes.
3. **Domain Expertise:** To successfully apply AI, collaboration between domain experts and data scientists is required. Confronting this gap through a team effort across disciplines will enhance the applicability and relevance of the models.
4. **Integration into Existing Systems:** The implementations of AI must be integrated into existing tea industry processes and technology. This can only be achieved if change management is utilised.

Future Research Directions

- **Hybrid Models:** Using traditional methods with AI cloud can better improve understanding of tea quality by using the strengths of both approaches.
- **Real Time Monitoring:** A real-time AI-driven monitoring system could be developed to assist in quality control in significant steps in tea production.
- **Sustainability Assessments:** AI can help the sustainability of tea production for environmentally friendly initiatives or ventures.

2.6. Conclusion

It can be concluded that implementing advanced data analysis in the tea sector revolutionises quality evaluation and standardisation. By leveraging AI techniques such as classical AI, DL, DL, and multimodal data fusion, companies can gain insights into the complex factors influencing tea quality, including biochemistry, environment, and processing methods. This enables stakeholders to make informed decisions that align with consumer preferences and market trends. These robust data analysis tools also help address concerns about data quality,

potential algorithmic bias, and the need for transparency through Explainable AI, building trust between producers and consumers. Continued research and integration of new technologies will further optimise tea quality for both producers and consumers in this growing industry. The next chapter will explore a proposed method using widely used ML algorithms to overcome challenges in tea quality analysis and optimization.

REFERENCES

[1] S. Kaushal, P. Nayi, D. Rahadian, and H.-H. Chen, "Applications of electronic nose coupled with statistical and intelligent pattern recognition techniques for monitoring tea quality: A review," *Agriculture*, vol. 12, no. 9, p. 1359, 2022, doi: https://doi.org/10.3390/agriculture12091359.

[2] R. B. Roy, A. Modak, S. Mondal, B. Tudu, R. Bandyopadhyay, and N. Bhattacharyya, "Fusion of electronic nose and tongue response using fuzzy based approach for black tea classification," *Procedia Technology*, vol. 10, pp. 615-622, 2013, doi: https://doi.org/10.1016/j.protcy.2013.12.402.

[3] A. Modak, B. Tudu, R. Bandyopadhyay, and N. Bhattacharyya, "Improved Classification of Black Tea Employing Fuzzy Fusion of Electronic Nose and Tongue Responses," *Sensor Letters*, vol. 12, no. 6-7, pp. 1065-1069, 2014, doi: https://doi.org/10.1166/sl.2014.3173.

[4] A. Modak, R. Roy, B. Tudu, R. Bandyopadhyay, and N. Bhattacharyya, *Towards artificial flavor perception of black tea.*, pp. 246-251, 2015, doi: https://doi.org/10.1145/2708463.2709040.

[5] A. Modak, R. B. Roy, B. Tudu, R. Bandyopadhyay, and N. Bhattacharyya, "A novel fuzzy based signal analysis technique in electronic nose and electronic tongue for black tea quality analysis," in *2016 IEEE First International Conference on Control, Measurement and Instrumentation (CMI)*, 2016: IEEE, pp. 279-283, doi: https://doi.org/10.1109/CMI.2016.7413755.

[6] H. Xia *et al.*, "Rapid discrimination of quality grade of black tea based on near-infrared spectroscopy (NIRS), electronic nose (E-nose) and data fusion," *Food Chemistry*, vol. 440, p. 138242, 2024, doi: https://doi.org/10.1016/j.foodchem.2023.138242.

[7] S. Thite, Y. Suryawanshi, K. Patil, and P. Chumchu, "Sugarcane leaf dataset: A dataset for disease detection and classification for machine learning applications," *Data in Brief*, vol. 53, p. 110268, 2024, doi: https://doi.org/10.1016/j.dib.2024.110268.

[8] S. Gupta, A. Kumar, and A. Kumar, "Analysis of Traffic Flow Congestion by Integrating Supervised Machine Learning with K-mean Clustering," *SN Computer Science*, vol. 6, no. 3, p. 255, 2025, doi: https://doi.org/10.1007/s42979-025-03761-4.

[9] D. Das, S. Nag, and R. B. Roy, "Machine Vision System: An Emerging Technique for Quality Estimation of Tea," in *Edible Electronics for Smart Technology Solutions*: IGI

Global, 2025, pp. 445-476, 2025, doi: https://doi.org/10.4018/979-8-3693-5573-2.ch018.

[10] Y. Zhang, X. Chen, D. Chen, L. Zhu, G. Wang, and Z. Chen, "Machine learning-based classification and prediction of typical Chinese green tea taste profiles," *Food Research International*, p. 115796, 2025, doi: https://doi.org/10.1016/j.jfca.2025.107908.

[11] M. A. Talukder, M. Khalid, and N. Sultana, "A hybrid machine learning model for intrusion detection in wireless sensor networks leveraging data balancing and dimensionality reduction," *Scientific Reports*, vol. 15, no. 1, p. 4617, 2025, doi: https://doi.org/10.1038/s41598-025-87028-1.

[12] V. V. Baligodugula and F. Amsaad, "Unsupervised Learning: Comparative Analysis of Clustering Techniques on High-Dimensional Data," *arXiv preprint arXiv:2503.23215*, 2025, doi: https://doi.org/10.48550/arXiv.2503.23215.

[13] R. Zaheer, M. K. Hanif, M. U. Sarwar, and R. Talib, "Evaluating the Effectiveness of Dimensionality Reduction on Machine Learning Algorithms in Time Series Forecasting," *IEEE Access*, 2025, doi: https://doi.org/10.1109/ACCESS.2025.3551741.

[14] P. Freire, D. Freire, and C. C. Licon, "A comprehensive review of machine learning and its application to dairy products," *Critical reviews in food science and nutrition*, vol. 65, no. 10, pp. 1878-1893, 2025, doi: https://doi.org/10.1080/10408398.2024.2312537.

[15] C. Magazzino and M. Mele, "A new machine learning algorithm to explore the CO2 emissions-energy use-economic growth trilemma," *Annals of Operations Research*, vol. 345, no. 2, pp. 665-683, 2025, doi: https://doi.org/10.1007/s10479-022-04787-0.

[16] J. Niyogisubizo, J. de Dieu Ninteretse, E. Nziyumva, M. Nshimiyimana, E. Murwanashyaka, and E. Habiyakare, "Towards Predicting the Quality of Red Wine Using Novel Machine Learning Methods for Classification, Data Visualization, and Analysis," in *Artificial Intelligence and Applications*, 2025, vol. 3, no. 1, pp. 31-42, doi: https://doi.org/10.47852/bonviewAIA42021999.

[17] C. Han *et al.*, "From microstructure to multivariate prediction models: Decoding the biomechanical properties of tea stems via PLSR-Ridge regression and multifactorial orthogonal design," *Industrial Crops and Products*, vol. 230, p. 121143, 2025, doi: https://doi.org/10.1016/j.indcrop.2025.121143.

[18] Z.-H. Zhuang, H.-P. Tsai, and C.-I. Chen, "Estimating Tea Plant Physiological Parameters Using Unmanned Aerial Vehicle Imagery and Machine Learning

Algorithms," *Sensors (Basel, Switzerland)*, vol. 25, no. 7, p. 1966, 2025, doi: https://doi.org/10.3390/s25071966.

[19] S. Shi *et al.*, "From 2015 to 2023: How Machine Learning Aids Natural Product Analysis," *Chemistry Africa*, vol. 8, no. 2, pp. 505-522, 2025, doi: https://doi.org/10.1007/s42250-024-01154-3.

[20] G.-R. Han *et al.*, "Machine learning in point-of-care testing: innovations, challenges, and opportunities," *Nature Communications*, vol. 16, no. 1, p. 3165, 2025, doi: https://doi.org/10.1038/s41467-025-58527-6.

[21] L. Lu *et al.*, "An efficient artificial intelligence algorithm for predicting the sensory quality of green and black teas based on the key chemical indices," *Food chemistry*, vol. 441, p. 138341, 2024, doi: https://doi.org/10.1016/j.foodchem.2023.138341.

[22] G. Chen, X. Zhang, Z. Wu, J. Su, and G. Cai, "An efficient tea quality classification algorithm based on near infrared spectroscopy and random Forest," *Journal of Food Process Engineering*, vol. 44, no. 1, p. e13604, 2021, doi: https://doi.org/10.1111/jfpe.13604.

[23] M. Zhao *et al.*, "Establishment of a Daqu Grade Classification Model Based on Computer Vision and Machine Learning," *Foods*, vol. 14, no. 4, p. 668, 2025, doi: https://doi.org/10.3390/foods14040668.

[24] F. Song *et al.*, "Evaluation of black tea appearance quality using a segmentation-based feature extraction method," *Food Bioscience*, vol. 58, p. 103644, 2024, doi: https://doi.org/10.1016/j.fbio.2024.103644.

[25] P. Chatterjee, D. Mitra, T. Bhowmik, R. Nath, K. Dutta, and A. Deyasi, "Predicting Tea Quality Based on Antioxidant and Polyphenol Content Using Ensemble Machine Learning Model," in *2024 IEEE International Conference of Electron Devices Society Kolkata Chapter (EDKCON)*, 2024: IEEE, pp. 482-487, doi: https://doi.org/0.1109/EDKCON62339.2024.10870624.

[26] K. Sravan, L. G. Rao, K. Ramineni, A. Rachapalli, and S. Mohmmad, "Analyze the Quality of Wine Based on Machine Learning Approach Check for updates," *Data Science and Applications: Proceedings of ICDSA 2023, Volume 3*, vol. 820, p. 351, 2024, doi: https://doi.org/10.1007/978-981-99-7817-5_26.

[27] Q. Li *et al.*, "Machine learning technique combined with data fusion strategies: A tea grade discrimination platform," *Industrial Crops and Products*, vol. 203, p. 117127, 2023, doi: https://doi.org/10.1016/j.indcrop.2023.117127.

[28] R. Mohana, P. Sharma, and A. Sharma, "Ensemble framework for red wine quality prediction," *Food Analytical Methods*, vol. 16, no. 1, pp. 30-44, 2023, doi: https://doi.org/10.1007/s12161-022-02367-3.

[29] S. Kumar, A. K. Pal, N. Varish, I. Nurhidayat, S. M. Eldin, and S. K. Sahoo, "A hierarchical approach based CBIR scheme using shape, texture, and color for accelerating retrieval process," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 7, p. 101609, 2023, doi: https://doi.org/10.1016/j.jksuci.2023.101609.

[30] A. Raza, Y. Hu, and Y. Lu, "Improving carbon flux estimation in tea plantation ecosystems: A machine learning ensemble approach," *European Journal of Agronomy*, vol. 160, p. 127297, 2024, doi: https://doi.org/10.1016/j.eja.2024.127297.

[31] A. B. Patil and M. R. Bachute, "Artificial Taste Perception of Tea Beverage Using Machine Learning," in *Artificial Intelligence, Machine Learning and User Interface Design*: Bentham Science Publishers, pp. 1-26, 2024, doi: https://doi.org/10.2174/9789815179606124010003.

[32] M. H. A. Rasyid, D. R. Wiiava, and G. P. Kusuma, "Predicting Green Tea Organoleptic Scores Based on Electronic Nose Dataset using Boosting Algorithms," in *2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 2024: IEEE, pp. 373-378, doi: https://doi.org/10.1109/IAICT62357.2024.10617596.

[33] P. Huang, P. Yang, L. Xu, Y. Wang, J. Yuan, and Z. Kang, "Moisture content detection of Tibetan tea based on hyperspectral technology, machine vision and machine learning," *Journal of Food Measurement and Characterization*, vol. 19, no. 2, pp. 1167-1185, 2025, doi: https://doi.org/10.1007/s11694-024-03032-5.

[34] Y. Hu, W. Sheng, S. Y.-S. S. Adade, J. Wang, H. Li, and Q. Chen, "Comparison of machine learning and deep learning models for detecting quality components of vine tea using smartphone-based portable near-infrared device," *Food Control*, vol. 174, p. 111244, 2025, doi: https://doi.org/10.1016/j.foodcont.2025.111244.

[35] C. Zhang, J. Wang, G. Lu, S. Fei, T. Zheng, and B. Huang, "Automated tea quality identification based on deep convolutional neural networks and transfer learning," *Journal of Food Process Engineering*, vol. 46, no. 4, p. e14303, 2023, doi: https://doi.org/10.1111/jfpe.14303.

[36] U. Barman, "Deep Convolutional neural network (CNN) in tea leaf chlorophyll estimation: A new direction of modern tea farming in Assam, India," *Journal of Applied & Natural Science*, vol. 13, no. 3, 2021, doi: https://doi.org/10.31018/jans.v13i3.2892.

[37] K. O. Kombo *et al.*, "Enhancing classification rate of electronic nose system and piecewise feature extraction method to classify black tea with superior quality," *Scientific African*, vol. 24, p. e02153, 2024, doi: https://doi.org/10.1016/j.sciaf.2024.e02153.

[38] J. Liang, J. Guo, H. Xia, C. Ma, and X. Qiao, "A black tea quality testing method for scale production using CV and NIRS with TCN for spectral feature extraction," *Food Chemistry*, vol. 464, p. 141567, 2025, doi: https://doi.org/10.1016/j.foodchem.2024.141567.

[39] W. Zhang *et al.*, "Data-driven optimization of nitrogen fertilization and quality sensing across tea bud varieties using near-infrared spectroscopy and deep learning," *Computers and Electronics in Agriculture*, vol. 222, p. 109071, 2024, doi: https://doi.org/10.1016/j.compag.2024.109071.

[40] D. Li, K. Jiang, G. Yang, Z. Gong, and T. Wen, "Integrated colorimetric sensor array with outlier detection and feature selection to detect adulterated camellia oil," *Microchemical Journal*, vol. 209, p. 112741, 2025, doi: https://doi.org/10.1016/j.microc.2025.112741.

[41] S. I. Abba *et al.*, "Optimizing sustainable desalination plants with advanced ML-based uncertainty analysis," *Applied Soft Computing*, vol. 169, p. 112624, 2025, doi: https://doi.org/10.1016/j.asoc.2024.112624.

[42] X. Yang, Z. Bi, C. Yin, H. Huang, and Y. Li, "A novel hybrid sensor array based on the polyphenol oxidase and its nanozymes combined with the machine learning based dual output model to identify tea polyphenols and Chinese teas," *Talanta*, vol. 272, p. 125842, 2024, doi: https://doi.org/10.1016/j.talanta.2024.125842.

[43] U. Fayaz *et al.*, "Recent insights into E-tongue interventions in food processing applications: an updated review," *Current Food Science and Technology Reports*, vol. 2, no. 2, pp. 169-182, 2024, doi: https://doi.org/10.1007/s43555-024-00028-6.

[44] W. Zheng, Q. Yuan, A. Zhang, and G. Pan, "A gustatory-emotion coupling model of nerve conduction mechanisms for taste identification," *IEEE Sensors Journal*, 2025, doi: https://doi.org/10.1109/JSEN.2025.3565725.

[45] Y. Li, C. T. Elliott, A. Petchkongkaew, and D. Wu, "The classification, detection and 'SMART'control of the nine sins of tea fraud," *Trends in Food Science & Technology*, p. 104565, 2024, doi: https://doi.org/10.1016/j.tifs.2024.104565.

[46] M. John, T. J. Mathew, and V. Bindu, "A novel deep learning based cbir model using convolutional siamese neural networks," *Journal of Intelligent & Fuzzy Systems*, pp. JIFS-219396, 2024, doi: https://doi.org/10.3233/JIFS-219396.

[47] X. Zhang *et al.*, "Non-Destructive Quality Monitoring of Shanxi Vinegar Production During The Fumigation Stage Using Computer Vision, Electronic Nose and Near-Infrared Spectroscopy Assisted by Machine Learning," J*ournal of Food Composition and Analysis*, Volume 146, 2025, 107908, doi: https://doi.org/10.1016/j.jfca.2025.107908.

[48] J. Jiang *et al.*, "Non-destructive monitoring of tea plant growth through UAV spectral imagery and meteorological data using machine learning and parameter optimization algorithms," *Computers and Electronics in Agriculture*, vol. 229, p. 109795, 2025, doi: https://doi.org/10.1016/j.compag.2024.109795.

[49] M. R. Ismael and I. Abdel-Qader, "Brain tumor classification via statistical features and back-propagation neural network," in *2018 IEEE international conference on electro/information technology (EIT)*, 2018: IEEE, pp. 0252-0257, doi: https://doi.org/10.1109/EIT.2018.8500308.

[50] Y. Luo *et al.*, "Advancing Loquat Total Soluble Solids Content Determination by Near-Infrared Spectroscopy and Explainable AI," *Agriculture*, vol. 15, no. 3, p. 281, 2025, doi: https://doi.org/10.3390/agriculture15030281.

[51] W. Khan, M. Ishrat, and S. M. Faisal, "AI-Powered Risk Assessment: Innovations, Challenges, and Future Prospects," *Artificial Intelligence for Financial Risk Management and Analysis*, pp. 59-92, 2025, doi: https://doi.org/10.4018/979-8-3373-1200-2.ch003.

[52] R. Chowdhury, R. Das, F. B. F. Ananna, A. Saha, S. Nawar, and M. H. Hosen, "Unveiling Predictive Factors in Apple Quality: Leveraging LIME, SHAP, and the Synergy of Machine Learning Models and Artificial Neural Networks," in *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, 2024: IEEE, pp. 1026-1031, doi: https://doi.org/10.1109/ICEEICT62016.2024.10534426.

CHAPTER 3

TEA QUALITY EVALUATION USING MIP ELECTROCHEMICAL SENSOR: APPROACH

This chapter describes the key frameworks developed in the present study to evaluate the quality of green tea. Accordingly, the proposed model implements an ML-based model for predicting green tea quality based on Epigallocatechin-3-Gallate (EGCG) molecule content using the Molecular Imprinting Polymerisation (MIP) Technique. This chapter provides a detailed elaboration on the working mechanisms of the frameworks, as well as the results obtained by each model.

LIST OF SECTION

- Introduction
- Study of Green Tea Quality Evaluation on Epigallocatechin-3-Gallate (EGCG) Molecule Content using Molecular Imprinting Polymerisation (MIP) Technique
- Conclusion

Contents of this chapter are based on following publication:

- **A. Modak**, T. N. Chatterjee, S. Nag, R. B. Roy, B. Tudu and R. Bandyopadhyay, "Linear Regression Modelling on Epigallocatechin-3-gallate Sensor Data for Green Tea," 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, India, 2018, pp. 112-117, doi: https://doi.org/10.1109/ICRCICN.2018.8718696.
- **A. Modak**, D. Das and R. B. Roy, "Classification of Green Tea Samples Based on Epigallocatechin-3-Gallate Content Using Molecular Imprinting Polymerization Technique," 2024 IEEE 3rd International Conference on Control, Instrumentation, Energy & Communication (CIEC), Kolkata, India, 2024, pp. 355-360, doi: https://doi.org/10.1109/CIEC59440.2024.10468449.
- **A. Modak**, D. Das, T. N. Chatterjee, B. Tudu and R. Banerjee Roy, "Regression-Based EGCG Detection in Green Tea Employing MIP Electrodes," in *IEEE Sensors Journal*, vol. 24, no. 11, pp. 18090-18097, 1 June 1, 2024, doi: https://doi.org/10.1109/JSEN.2024.3390438.

CHAPTER 3

TEA QUALITY EVALUATION USING MIP ELECTROCHEMICAL SENSOR: MACHINE LEARNING APPROACH

3.1. Introduction

Tea is a widely consumed beverage, and the quality of tea is defined by its aroma, flavour, and colour, which are influenced by its chemical components [1]. For example, green tea's astringency comes from compounds like EGCG and other catechins [2]. While black tea derives its flavour and colour from theaflavins and thearubigins, formed through oxidation, other molecules, such as tannic acid, increase astringency, and L-theanine contributes an umami flavour [3]. To overcome the limitations of subjective sensory evaluation, advanced tools MIPs based sensors widely utilised [4, 5]. MIPs offer targeted detection of specific molecules like tannic acid, improving quality control and consistency [6]. These methods are also crucial for optimising tea processing and cultivation.

Thus, this chapter focuses on two key frameworks: one evaluates green tea quality based on EGCG content using the MIP technique. Another approach is the assessment of black tea quality through tannic acid content using NIR Spectroscopy. Specifically, the utilisation of MIP technology for the sensor development EGCG detection in green tea and the application of NIR spectroscopy for tannic acid analysis in black tea represent advancements in non-destructive and efficient quality evaluation within the tea industry.

3.2. Study of Green Tea Quality Evaluation on Epigallocatechin-3-Gallate (EGCG) Based Content using Molecular Imprinting Polymerisation (MIP) Technique

3.2.1. Overview of Proposed Model

As opposed to the traditional method of assessing tea quality using purely chemical analysis, the current model uses machine learning (ML) and MIP sensors for objective EGCG-based tea quality assessment. The proposed system utilises an ML model to predict the quality of the green tea based on the EGCG content. The EGCG is evaluated using the MIP electrode; a three-electrode setup is built to gather the volumetric response. After data pre-processing, the UMAP algorithm is employed for feature selection to optimise output, followed by K-means clustering to distinguish samples. Finally, classification is performed using the XGBoost algorithm, which is renowned for its efficiency and speed in converging to the minimum error [7]. Additionally, the EGCG content is analysed using HPLC data for every green tea sample. The

model's efficiency is evaluated through various metrics, demonstrating effective performance in classifying green tea samples. Figure 3.1 illustrates the overall process involved in the proposed evaluation of green tea quality.

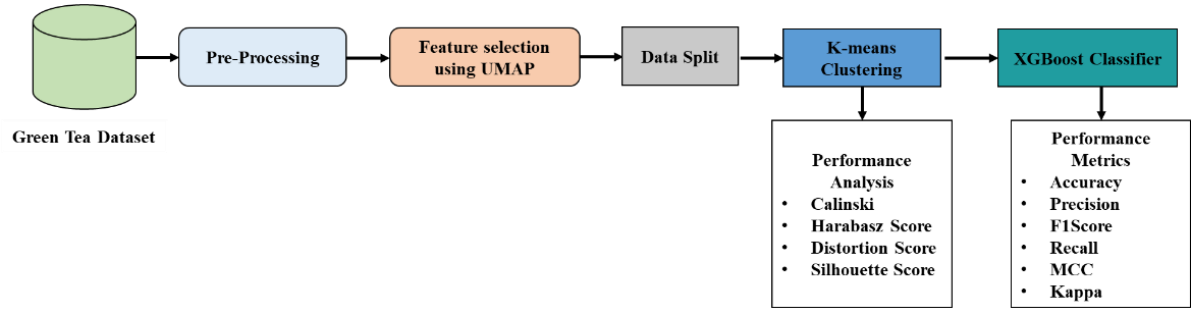


Figure 3.1 Overall Process included in the Present Work

3.2.2. Experimentation

The study evaluated 16 green tea samples, each subjected to five repetitive runs for measurement reliability. An EGCG-specific sensor, developed using MIP technology, was employed for electrochemical analysis via differential pulse voltammetry (DPV). This combination enabled accurate quantification of EGCG content, an essential indicator of green tea quality.

3.2.2.1. Selecting Voltammetric Measurement Principle

To determine the EGCG content of green tea, the study has utilised voltammetry combined with HPLC. The voltammetry technique uses a three-electrode system (indicator, reference, and auxiliary) within an electrochemical cell to measure the current generated by the oxidation or reduction of compounds. The resulting data will then be analysed with ML to classify green tea samples based on their EGCG levels. Drawing on the applications of voltammetric measurement principles, the proposed study aims to classify green tea samples based on their EGCG content.

3.2.2.2. Molecular Imprinted Polymer (MIP) Technique

MIPs are regarded as synthetic receptors that feature identification sites which differ from the conformation of the target molecule. This methodology holds the promise of uniqueness, attributed to its functional capabilities and its specific, predetermined traits that facilitate the prediction of template molecules. Furthermore, sensors developed from MIPs offer extensive

compatibility, robust stability, straightforward preparation, affordability, and enhanced sensitivity and selectivity.

3.2.2.3. Data Set Collection

Data analysis was conducted utilising Jupyter Notebook version 5.4.0 [8] within the Anaconda Navigator environment. In this study, all cognitive models were created using the Python programming language version 3.6 [9, 10]. A variety of libraries were employed to facilitate this process, including numpy version 1.14.3 [11] scipy version 1.1.0 [12] pandas version 0.23.0, and scikit-learn version 0.19.1 [13] Visualisation of the data was performed with the aid of matplotlib version 2.2.2 and seaborn version 0.8.1 [14]. To obtain the response profile of the electrode in relation to the green tea samples, five duplicate DPV measurements were conducted for each sample, yielding 80 data points per sample. Consequently, for a total of 16 green tea samples, the cumulative dataset amounted to (16 x 5 = 80) x 80. The data inspection and management processes were automated using the Python package Pandas v0.23.0. Additionally, this package facilitated the importation of data from an Excel file into a specialised data structure known as a data frame, configured as 80x80. The HPLC values for the green tea samples were designated as the output labels [6].

3.2.3. Pre-Processing

Pre-processing refers to the method applied to raw data to prepare it for subsequent data processing. In the present study, standard scaling, also known as the Standard Normal Distribution (SND), is employed. This involves standardising the features by removing the mean and adjusting them to achieve unit variance.

3.2.4. Feature Selection

3.2.4.1. Box-Cox Transform

Typically, measurements of EGCG concentration obtained from a sensor do not consistently adhere to a normal distribution. This inconsistency can arise from several factors, including differences in tea samples, variations in experimental conditions, or the fundamental characteristics of the electrochemical response. The Box-Cox transformation is used to normalise non-normally distributed EGCG sensor data, which can result from variations in tea samples or experimental conditions. This parametric power transformation makes the data more symmetric and homoscedastic (stabilises variance), ensuring it meets the assumptions for statistical methods like DPV. To analyse the skewness of the data, the Skew method is used

from the Python Library. The Box-Cox transformation of the variables x is indexed by λ over n samples is referred to as x'_λ.

$$\frac{x^{\lambda}-1}{\lambda(\pi_{i=1}^{n} xi)^{\frac{\lambda-1}{n}}} \tag{3.1}$$

This method is utilised for input features with a skewness exceeding 0.5. These transformations serve as alternatives to conventional methods, including square root transformations, logarithmic transformations, and inverse transformations. The primary benefit of this approach lies in its ability to effectively normalise the selected variable. Therefore, they reduce the need for randomly selecting various transformations and automate the data transformation task. The study has utilised the spicy function. $bozcox1p$ That computes the Box-Cox transformation of $1 + x$. the setting for $\lambda = 0$ is equivalent to $\log 1p$ that is a log transformation. In this study, we considered $\lambda = 0.15$ This resulted in the application of this methodology to 67 out of 80 input features. Figure 3.2 below shows the visualisation outcome of the Box-Cox transform.

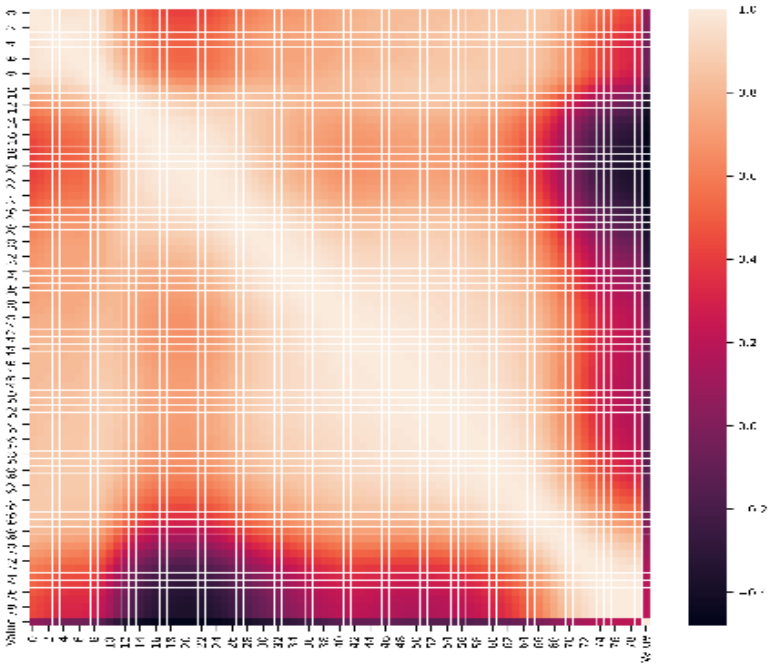


Figure 3.2 Correlation Matrix of input features (80) and HPLC score result

3.2.4.2. Correlation Coefficient

Analysing the interrelationships among input features is crucial, as opposed to examining the connections between input features and the label output. A highly effective method to achieve this is by constructing the correlation matrix, which elucidates the mutual relationships among

the variables. The correlation coefficient can be assessed by evaluating the average strength of various degrees of association. In this study, the Pearson correlation method is utilised to compute the correlation coefficient, leveraging the corr function provided by Pandas. The following illustration (3.3a) indicates the correlation matrix obtained by employing the seaborn library. On the other hand, the 80 input attributes correlation value with respect to the HPLC values is represented in Figure 3.3(b).

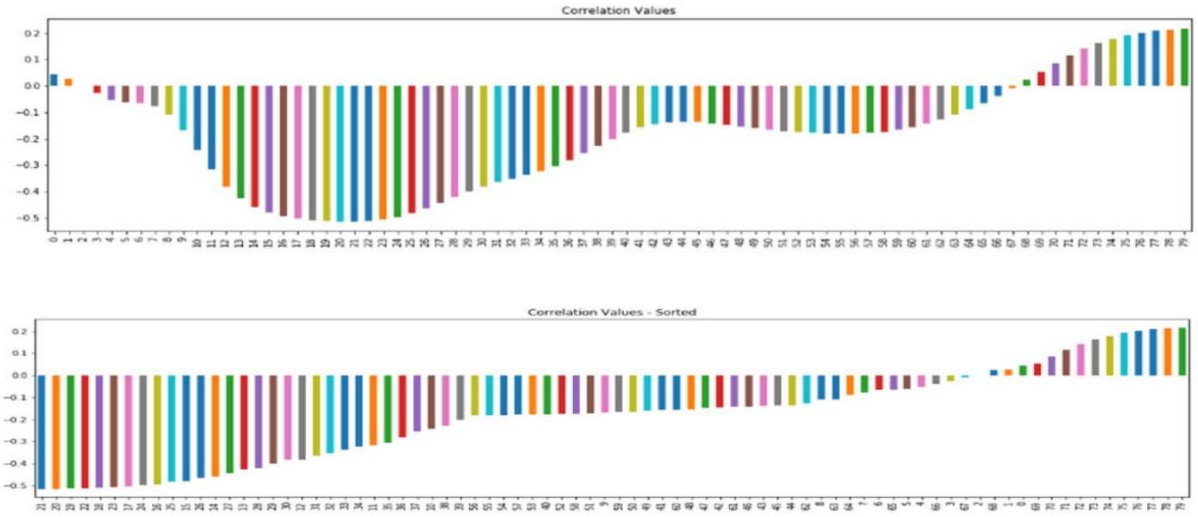


Figure 3.3 (a) Unsorted correlation Value and (b) Sorted correlation value

Figure 3.3 visually represents various correlation coefficients, statistical measures indicating the strength and direction of linear relationships between variables. Each vertical bar shows a specific correlation value, with bars extending downwards from the zero line signifying negative correlations (where variables move in opposite directions, with values ranging to approximately -0.5), while bars extending upwards denote positive correlations (where variables move in the same direction, with values up to about +0.2); the chart's sorted arrangement allows for easy visualization of the distribution, highlighting moderate negative relationships and weaker positive relationships across the depicted dataset.

3.2.4.3. Feature Extraction Based on Polynomials

A common approach to enhance our feature set is through the generation of polynomials. Polynomial expansion facilitates the creation of interactions among input features and also allows for the generation of powers (such as the square of a feature). This process adds a nonlinear aspect to our dataset, which can potentially boost our model's predictive capabilities. In this study, the model's non-linearity has been incorporated by developing new features based

on the ten input features with the highest correlation coefficients. We have included features for squares, cubes, and square roots. As a result, the total number of features has increased from 80 to 110.

3.2.4.4. Dimensionality Reduction

3.2.4.4.1. Uniform Manifold Approximation and Projection (UMAP)

The proposed method utilises UMAP for feature selection, a non-linear dimensionality reduction technique that simplifies the visualisation of high-dimensional data. UMAP operates in two steps: first, it performs a topological analysis to create a compressed embedding, then refines it through cross-entropy optimisation, preserving essential data structures while eliminating irrelevant features. This reduction in features enhances computational efficiency and maintains the integrity of the global structure.

3.2.5. Standardization

The standardisation of features becomes crucial when our model incorporates polynomial or interaction terms. Such terms often lead to multicollinearity, making our estimates highly susceptible to slight modifications within the model. This standardisation process is carried out utilising the StandardScaler method from the preprocessing library found in the ScikitLearn v0.19.1 package. In this approach, each input attribute is scaled; thus, the data distribution is centred at ‘Zero’, and the standard deviation is assigned as ‘one’, thereby achieving a normal distribution of the data.

3.2.6. Data Split Using Cross-Validation Method

The dataset is typically divided into a 70% training set (56 observations from a total of 80) and a 30% testing set (24 observations). To prevent data leakage, standardisation using Scikit-Learn is applied after splitting, ensuring the test set's mean and variance do not influence the training data. Model performance is then validated using 10-fold cross-validation, which divides the training data into 10 equally sized folds. The training process involves utilising nine folds of data, while the remaining fold is reserved for testing; this procedure is conducted over 10 iterations. The performance metrics are subsequently averaged to ensure a comprehensive evaluation and to improve the generalisation capability of the proposed model while minimising the risk of overfitting.

3.2.7. ML Modelling

3.2.7.1. Clustering Technique: K-Means Algorithm

K-means is recognised as a centroid-based clustering method, where the distance from each centroid to data points is computed for cluster assignment. The primary goal is to sort 'n' data interpretations into 'k' distinct groups, ensuring that each interpretation aligns with the cluster that has the nearest mean. This method operates iteratively, assigning each data point to the cluster with which it shares the most characteristics. Its objective is to minimise the total distance between the centroids and the data points, thereby accurately predicting the group affiliation of each data point. The K-means process consists of several steps: i) Determining the number of clusters, ii) Initialising the centroids, iii) Assigning data points to their closest clusters, and iv) Adjusting the centroids.

3.2.7.1.1. Performance Analysis of K-Means Clustering Approach

The proficiency of the proposed model is estimated through various performance metrics. The following Table 3.1 describes the data and parameter set for the proposed model.

Table 3.1 Hyperparameter Details

Parameters	Values
No. of tea samples	16
Repeated run on every sample	5
Overall tea samples	80
Count of features	80
EGCG dataset	(80, 80)
After applying UMAP and K-means to the original data size	(80, 3)
Transformed data size	(80, 3)
Transformed training set size	(56, 3)
Transformed testing set size	(24, 3)
Number of k-fold	10
UMAP Hyperparameters	No . of components = 2, No.of neighbors = 20, minimum distance = 0.1
XGBoost Hyperparameters	n_estimators = 100, tree_method = "Auto", booster = "gbtree"

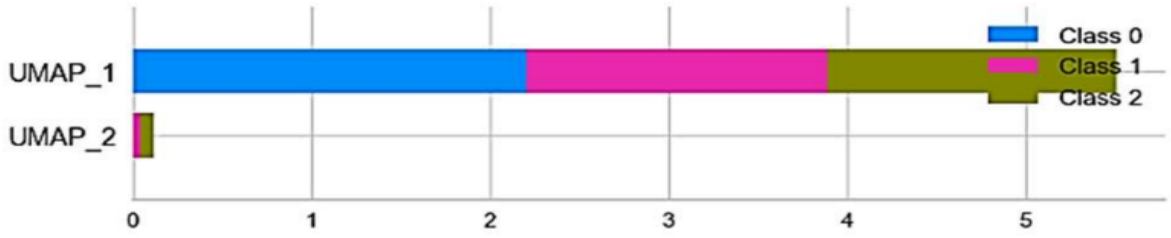


Figure 3.4 Average SHAP value using the UMAP model

Figure 3.4 depicts the impact of various features on predicting green tea samples, with the lengths of the bars indicating the significance of each feature. Each bar represents a different feature instance, colour-coded as blue for Class 0, pink for Class 1, and green for Class 2. The varying bar lengths highlight which features are most critical for effective classification of the green tea samples.

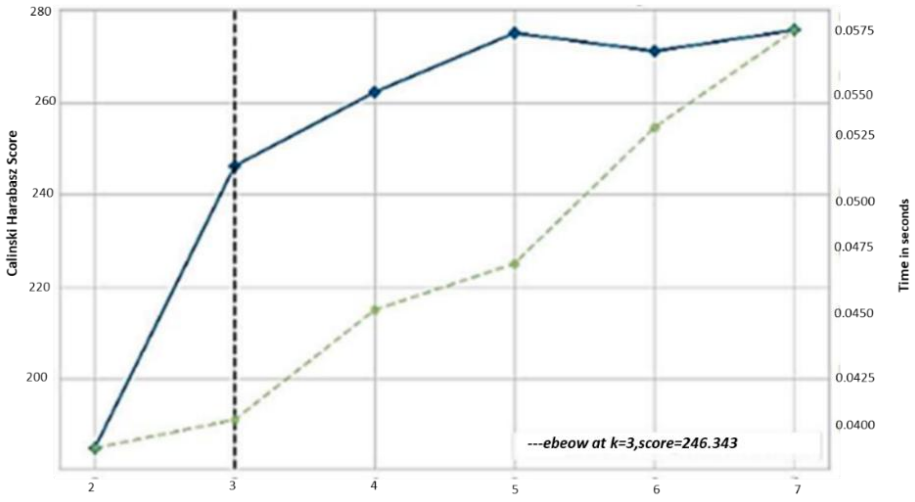


Figure 3.5 Calinski-Harabasz Score Elbow for K-means Clustering Technique

Figure 3.5 illustrates the performance of the K-means clustering method using the Calinski-Harabasz Score across varying values of 'k'. The elbow point at k=3, with a score of 246.343, indicates the optimal number of clusters, marked by a notable slope change in the blue line. The green dotted line shows that while the score increases with 'k', effective clustering occurs at this elbow point, confirming that three clusters offer a well-balanced data distribution.

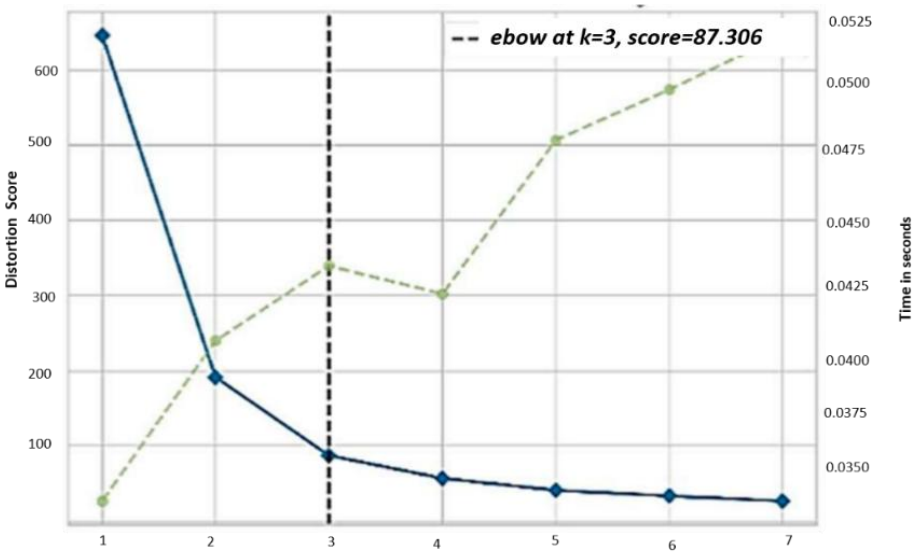


Figure 3.6 Distortion Score of K-Means Clustering Approach

Figure 3.6 illustrates the distortion score, a metric that evaluates the compactness of clusters by summing the squared distances from each data point to the corresponding cluster centroid. The blue line shows a sharp decrease in distortion as 'k' increases, indicating an improvement in clustering quality. The vertical dashed line at $k = 3$ Marks the optimal number of clusters, corresponding to a score of 87.306, where the rate of improvement begins to plateau. This demonstrates that three clusters effectively balance tightness and simplicity in the model. Figure 5 further shows the resulting cluster distributions achieved through the K-means clustering method, validating the findings from the distortion analysis.

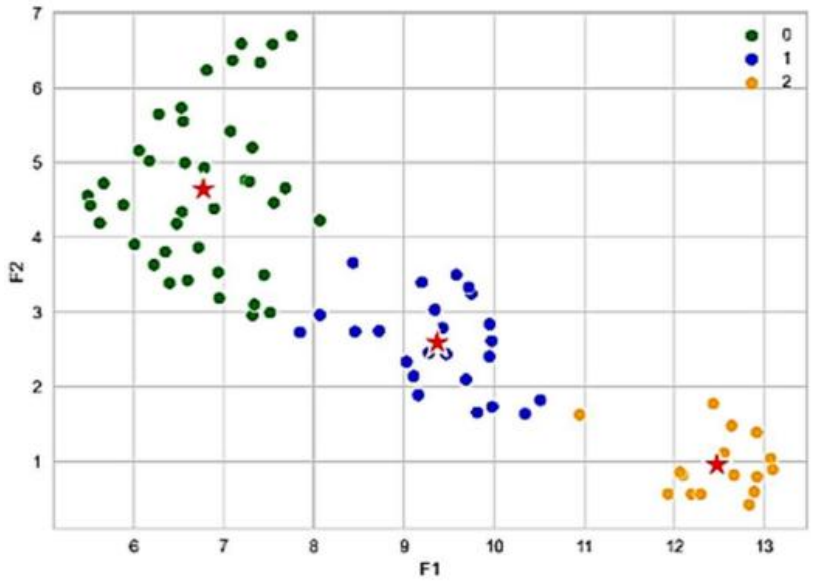


Figure 3.7 K-Means Clustering of Classes

Figure 3.7 visualises the clustering results of three classes produced by the K-means clustering technique, represented in a two-dimensional space with features labelled as F1 and F2. The data points are colour-coded: green for Class 0, blue for Class 1, and orange for Class 2. The red stars denote the centroids of each cluster. Class 0 indicates a high density and possibly more complex relationships within this cluster. Class 1 indicates a less distinct structure compared to Class 0. In contrast, Class 2 reflects a lower density and a simpler distribution of data points. This visualisation effectively illustrates the varying densities and distributions of the classes identified through the clustering analysis.

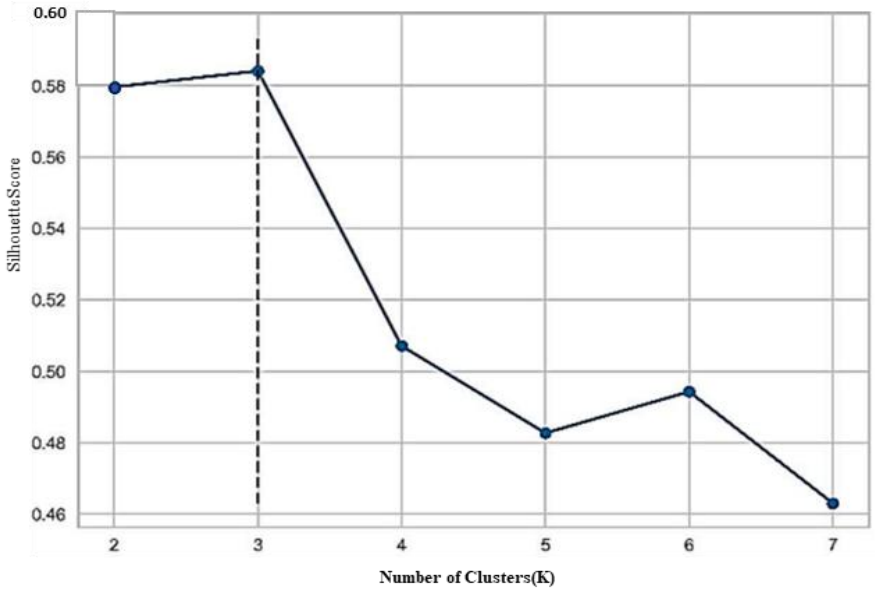


Figure 3.8 Silhouette Score Elbow for K-Means Clustering

Figure 3.8 silhouette score plot reveals the intra-cluster similarity across different values of 'k', with a significant peak at $k = 3$, indicating that this is the optimal number of clusters. The vertical dashed line at this point highlights that the silhouette score reaches its highest value. It is demonstrating well-defined clusters with high cohesion and separation. Selecting $k = 3$ This strikes a balance between the fundamental patterns within the data and steering away unnecessary complexity, ensuring the formation of cohesive and well-structured clusters.

Figure 3.9 presents a silhouette plot for determining the optimal 'k' value in K-means clustering. The silhouette coefficient measures the relation of data points to their own cluster versus neighbouring clusters, assessing the overall quality of the cluster. The uniformity in silhouette thickness at $k = 3$ indicates balanced intra-cluster cohesion and inter-cluster separation, confirming that $k = 3$ It is the most suitable choice for this dataset.

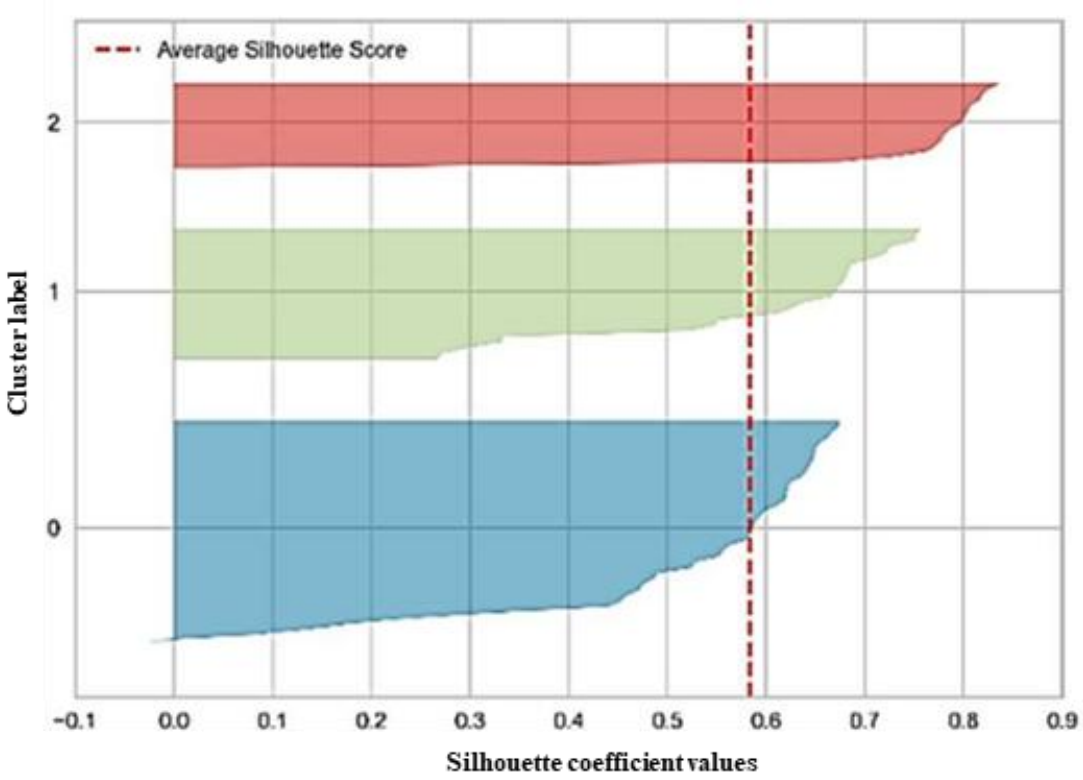


Figure 3.9 Silhouette score using K-Means Clustering Approach

3.2.7.2. Classification Technique: XGBoost Algorithm

The XGBoost algorithm functions as a supervised classifier for evaluating green tea quality, iterating through a series of decision trees to minimise a specialised objective function. To prevent overfitting on complex data, it includes a regularisation term that balances model complexity and predictive power. The algorithm intelligently chooses the most informative features by prioritising splits that maximise loss reduction. This iterative process, which builds upon initial unsupervised clustering, culminates in a robust model for classifying green tea quality based on its underlying chemical and physical attributes.

In this method, the XGBoost algorithm classifies the green tea samples into predefined classes, using the b features and leveraging the results from a prior clustering process. $P = \{(n_i, m_i)\}(|P|) = a, n_i \in X^b, m_i \in X)$ XGBoost utilises an additive function to enhance output accuracy.

$$\hat{m_i} = \alpha(n_i) = \sum_{q=1}^{q} f_q(n_i), \quad f_q \in F \tag{3.2}$$

Herein, $F = \{f(n) = v_{k(n)}\}(k: X^b \to S, v \in X^s)$ denotes the space of CART. The proposed model learns its function set by minimising a regularised objective. This optimisation process ultimately yields the final prediction, which is derived by summing the scores of the corresponding leaves in each of the model's trees.

$$E(x) = \sum_j l(\hat{m_{i,}} m_i) + \sum_q \mu(f_q) \quad (3.3)$$

Here, $\mu(f_q) = \gamma X + 1/2\delta\|\beta\|^2$ This supplementary regularisation function helps to smooth the final weights, reducing the risk of overfitting and thereby enhancing classification accuracy.

3.2.7.2.1. Results and Discussion: Performance Analysis of XGBoost Classifier

The performance of the XGBoost classifier is elaborated in this subsection as follows,

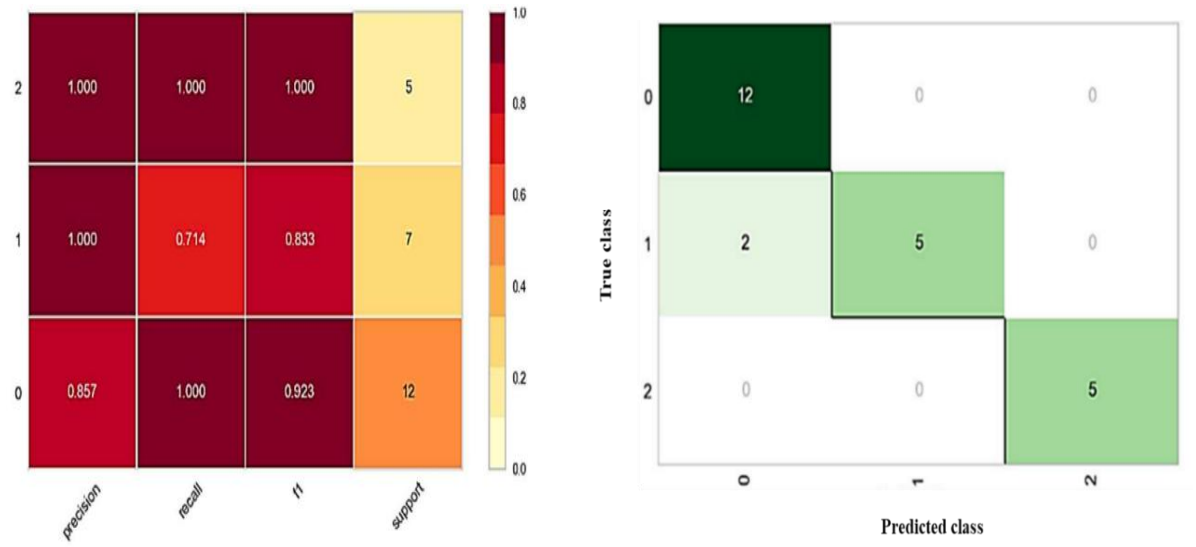


Figure 3.10 Classification Report and Confusion Matrix of XGBoost Classifier

From Figure 3.10, the analysis indicates that the classifier exhibits strong performance with precision values of 0.85 for feature 0, and perfect precision for features 1 and 2, at 1.0 each, suggesting high reliability in these classifications. The confusion matrix further supports this by showing that the majority of instances were correctly classified, with minimal classification errors, particularly for feature 0, which had 12 correct predictions out of 12. Overall, the XGBoost classifier demonstrates effective performance with a clear trend of accuracy across features.

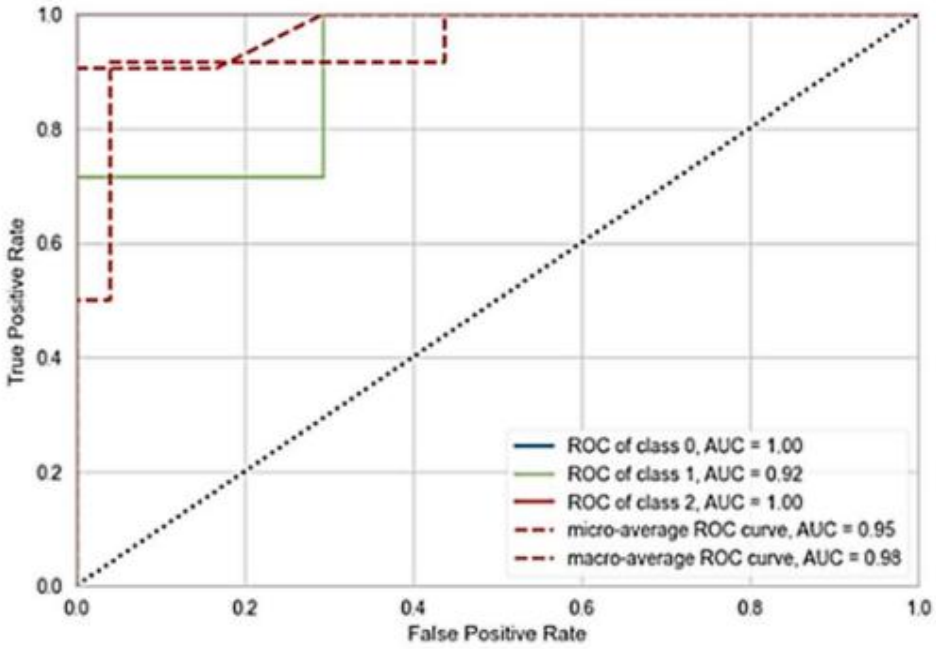


Figure 3.11 ROC Curves

The ROC curves for the XGBoost classifier, depicted in Figure 3.11, show that the model achieved an AUC (Area Under the Curve) of 1.0, indicating perfect classification. Class 1 exhibits an AUC of 0.92, indicating strong performance. The micro-average ROC curve has an AUC of 0.95, and the macro-average ROC curve yields an AUC of 0.98, showcasing that the model performs well overall while maintaining effectiveness across individual classes. The curves illustrate an efficient balance between accurate positive and false favourable rates.

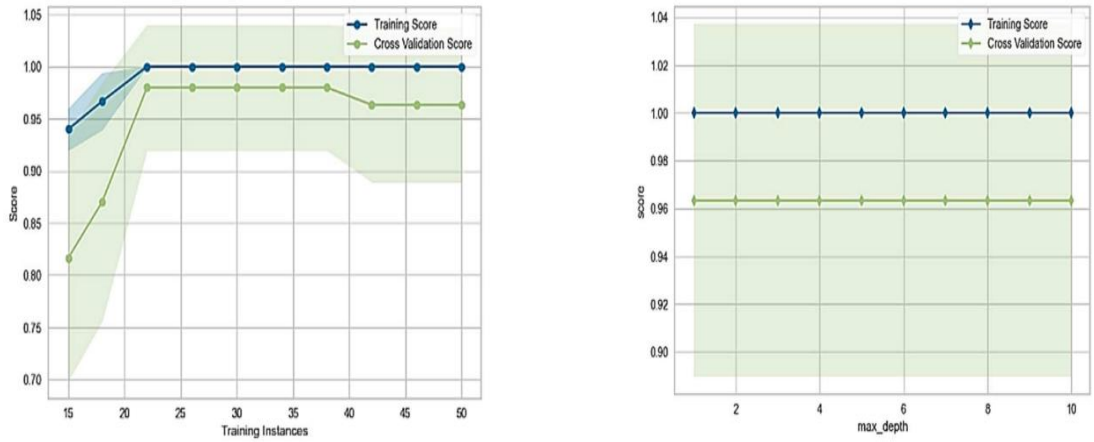


Figure 3.12 and 3.13 Learning Curve and Validation Curve for XGBoost Classifier

Figures 3.12 and 3.13 illustrate the training and cross-validation scores of an ML model as a function of the number of training instances and the maximum depth of the decision trees used. The training score indicated that the model fits the training data very well. Meanwhile, the

cross-validation score, represented by the green line, stabilises around 0.95 to 0.96 with less variation, indicating good generalisation across different subsets of the data. Overall, the model demonstrates strong performance across a range of training instances and maximum depths.

Table 3.2 Result Obtained by the XGBoost Classifier

Fold	AUC	Accuracy	Precision	Recall	F1	MCC	Kappa
0	0.96	0.83	0.91	0.93	0.83	0.77	0.73
1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1
8	1	0.8	0.67	0.8	0.72	0.72	0.67
9	1	1	1	1	1	1	1
Mean	0.99	0.96	0.96	0.96	0.96	0.95	0.94
Standard	0.1	0.07	0.10	0.07	0.09	0.10	0.12

Table 3.2 summarises the performance metrics of a classification model across ten folds of cross-validation. The mean AUC is 0.99, indicating the model's excellent performance. Additionally, other metrics also showed a high average score of 0.96%, reflecting strong predictive capabilities and balanced classification across most folds. Also, the standard deviations indicate consistency in performance across different folds. Overall, the metrics effectively showcase the model's reliability.

3.2.7.3. Regression Techniques

Regression methods are a list of supervised learning algorithms used for predicting consistent statistical values based on input features. The model aims to analyse the relation between the features and target values for prediction. The study has applied Tree-based modelling and linear regressor model to evaluate green tea quality [6].

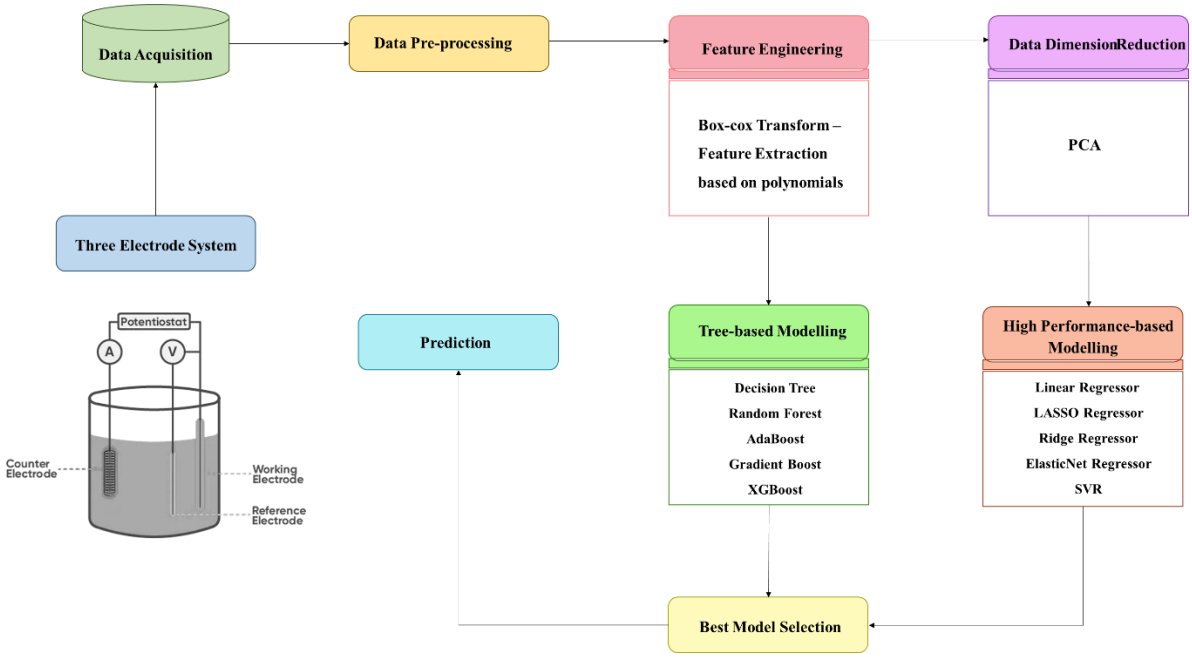


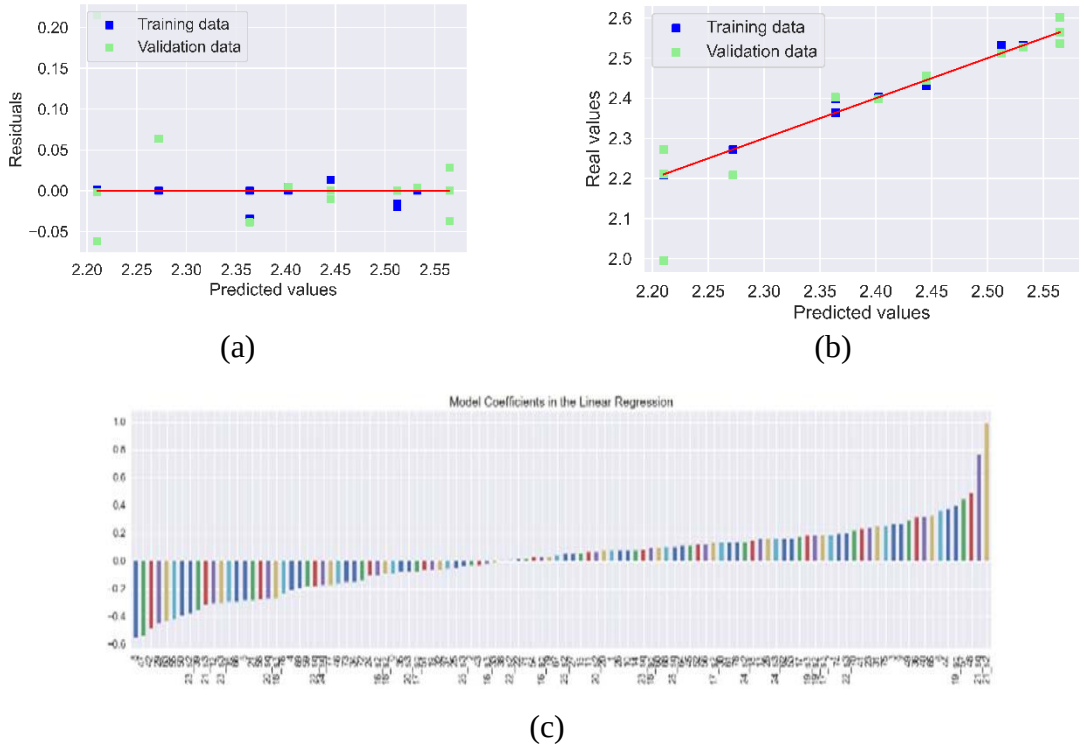
Figure 3.14 Overflow of Present Work with Regression Technique

Figure 3.14 illustrates a complete data processing and ML workflow, introduced with data acquisition from a three-electrode system. Electrochemistry is commonly used in applications involving electrodes, a reference electrode, and a counter electrode to measure electrochemical properties. Moreover, the acquired data undergoes data pre-processing to clean and prepare it for analysis. After that, in Feature Engineering, appropriate features are extracted and transformed using the Box-Cox Transform or polynomial-based feature extraction to enhance model performance. Consequently, Data Dimension Reduction techniques such as PCA are applied to decrease the number of variables while retaining significant data. Moreover, the data is processed and then used in High-Performance-based Modelling, namely linear regressor, LASSO, Ridge, Elastic Net, SVR, along with Tree-based Modelling, such as Decision Tree, Random Forest, AdaBoost, Gradient Boosting, and XGBoost. Hence, the Best Model Selection step identifies the most effective model for Prediction.

3.2.7.3.1. High Performance Modelling

High performance in regression modelling comprises selecting a suitable regression model, such as Support Vector Regression (SVR), for intricate and non-linear relationships. This approach involves robust feature engineering and detection methods, providing high-performance computing for large and complex datasets. The model is evaluated using metrics like R-squared and Root Mean Squared Error (RMSE) for the selected model. It is then repeatedly fitted and assessed to identify the better performer among various models [15].

3.2.7.3.1.1. Linear Regressor



(a) (b)

(c)

Figure 3.15 (a) Linear Regressor Residual Plot (b) Prediction Plot (c) Coefficient Plot

Figure 3.15 (a) shows the error distribution in the residual plot. Here, the training and validation data have residuals that are clustered, and it is closer to '0', which means that the predictions are almost accurate. Moreover, Figure 3.15(b) illustrates a comparison between the predicted and actual values in the prediction plot. Here, the training and validation data represent the exact prediction. This indicates that the linear regression model effectively fits the dataset, without visible underfitting or overfitting. Additionally, Figure 3.15(c) illustrates the bar charts for the coefficients (weights) assigned to each feature in the linear regression model. Moreover, the coefficient ranges from -0.6 to +1.0, indicating that certain features have a negative influence on the prediction, while others contribute to optimistic predictions. Additionally, the features are displayed on the x-axis, and their predictions are shown on the y-axis. Since most of the features align closer to '0', this makes it more effective.

3.2.7.3.1.2. LASSO Regressor

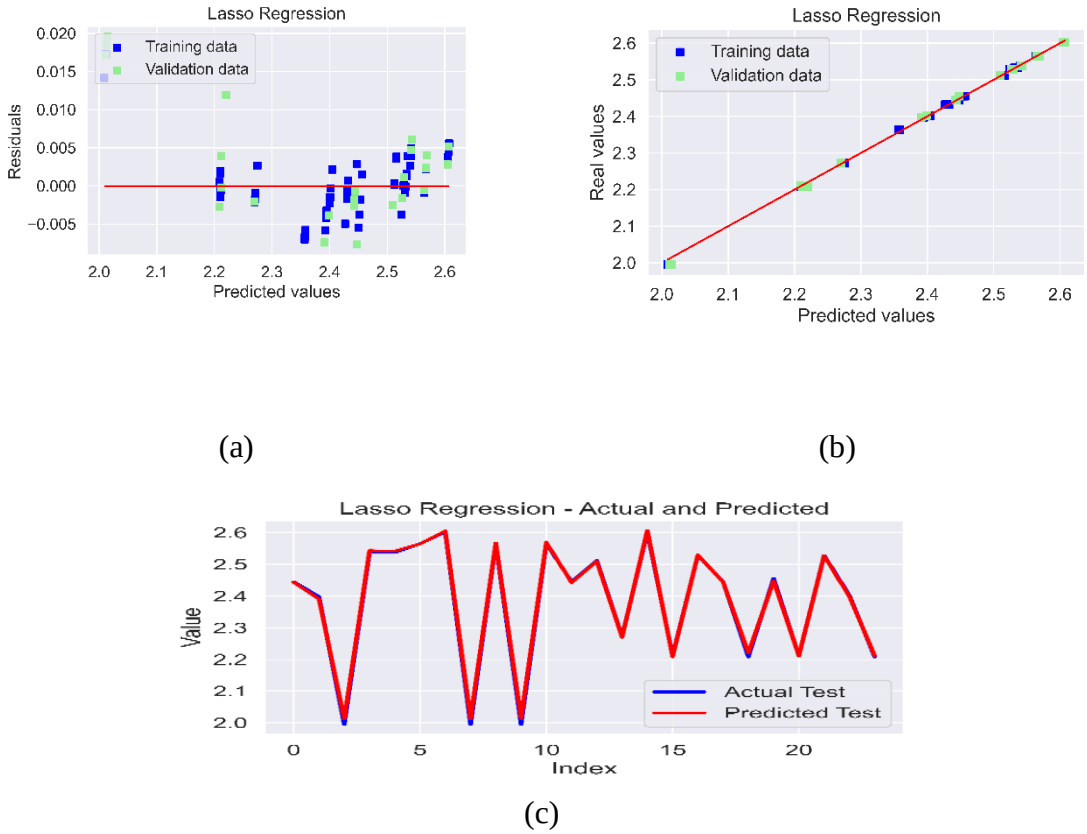
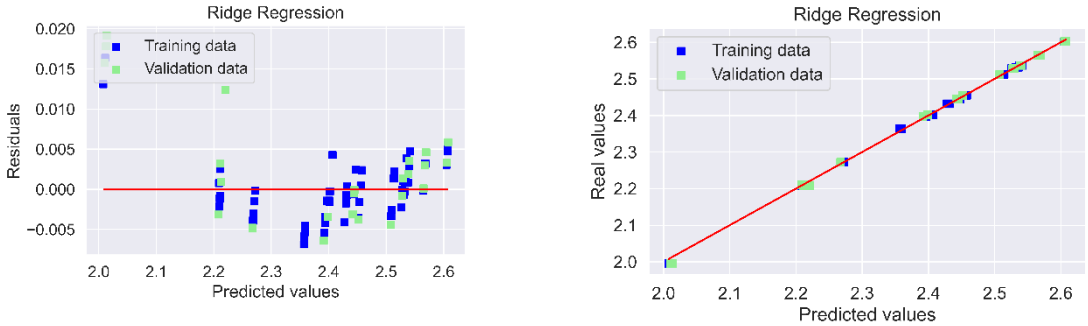


Figure 3.16 (a) LASSO Residual Design (b) Prediction Graph (c) Coefficient Scheme

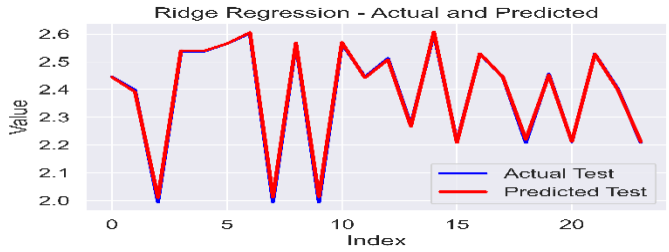
Figure 3.16 (a) signifies the comparison between the residuals and predicted values. Here, the predicted values are given on the x-axis and the residuals are shown on the y-axis. The residuals represent the difference between the actual and predicted values for both the train and validation sets. Moreover, the residuals are clustered, and they are closer to zero, which results in an effective model. Furthermore, Figure 3.16(b) illustrates the relationship between the real and predicted values. Here, prediction values are presented on the x-axis, and predicted values are given on the y-axis. Moreover, the train and validation sets are close together, indicating that the model predicts with high accuracy and exhibits better generalisation. Furthermore, Figure 3.16(c) displays the coefficient plot for the Lasso regression model, where the coefficients are shown as exact zeros for a few features. This is due to the L1 regularisation. Here, feature selection is achieved by eliminating negligible variables. Moreover, Lasso is the relevant predictor and simplifies the model.

3.2.7.3.1.3. Ridge Regressor

Figure 3.17 (a) illustrates a comparison of the residual plot with the predicted values; here, the residuals exhibit a higher variance and are closer to zero. The results indicate that the predictions are highly accurate for both the training and validation data. Figure 3.17(b) shows the prediction plot comparison, where the training and validation plots lie closer to the actual values. This shows the prediction remains perfect, which reduces overfitting with minimal coefficients and is beneficial. Figure 3.17 (c) shows the coefficient features in the ridge model, where the x-axis represents the features and the y-axis shows the coefficient values. Here, most of the coefficients are closer to zero, indicating that the model has performed well, with high positive coefficients.



(a) (b)



(c)

Figure 3.17 (a) Ridge Residual Graph (b) Prediction design (c) Coefficient Graph

3.2.7.3.1.4. ElasticNet Regressor

Figure 3.18 (a) defines the residual plot, which indicates a minimal maximum deviation of ±0.01, which is considered negligible, reflecting a high level of accuracy in the model's predictions. Figure 3.18 (b) represents a prediction plot that demonstrates an almost linear relationship where both training and validation datasets closely align, suggesting effective modeling performance. The residuals plot reveals the differences between predicted and actual values, with blue squares denoting training data and green squares for validation data; ideally,

the random dispersion around the zero-residual line implies no systematic errors. Figure 3.18 (c) line plot compares actual test values (in blue) with predicted test values (in red) generated from an ElasticNet regression model, displaying both series across a range of indices. The plot reveals that the predicted values generally follow the trend of the actual test values. Hence, the model demonstrated better outcome.

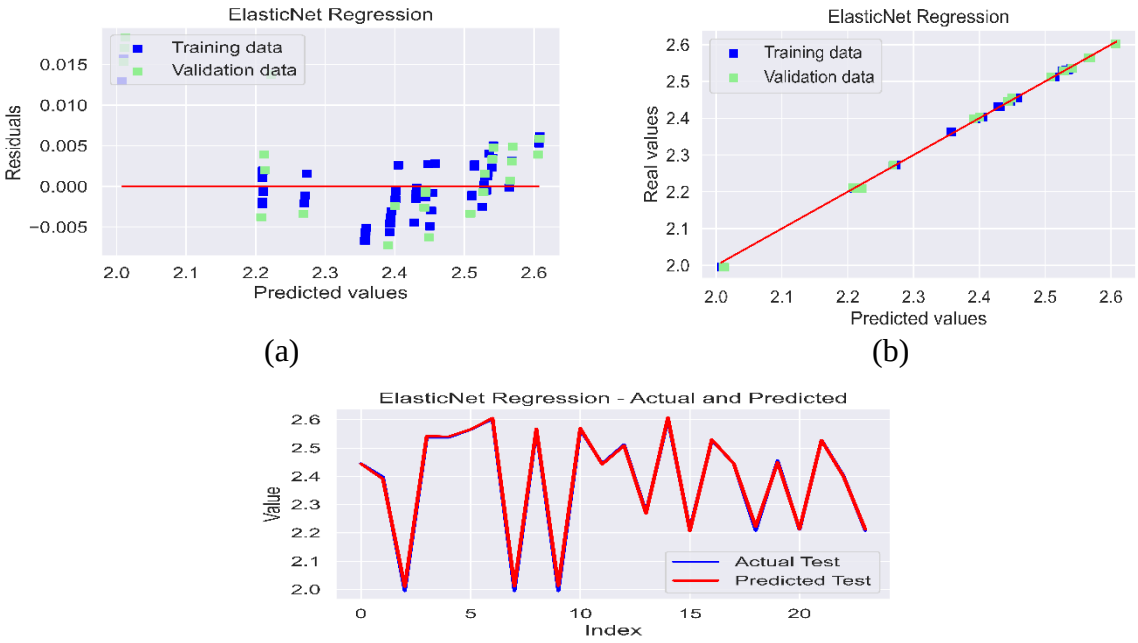
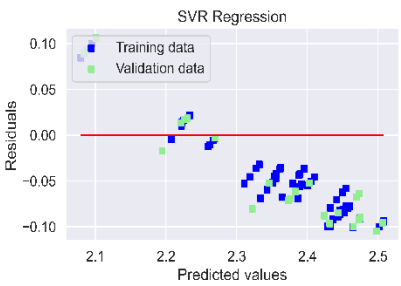


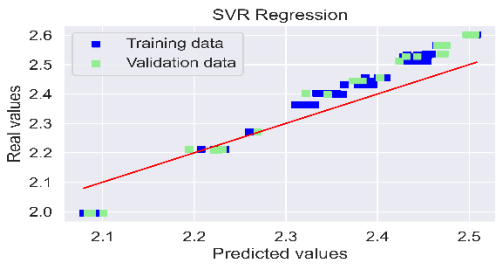
Figure 3.18 (a) ElasticNet Regressor Residual Graph (b) Actual Vs. Prediction Graph (c) Coefficient Graph

3.2.7.3.1.5. Support Vector Machine (SVR) Regressor

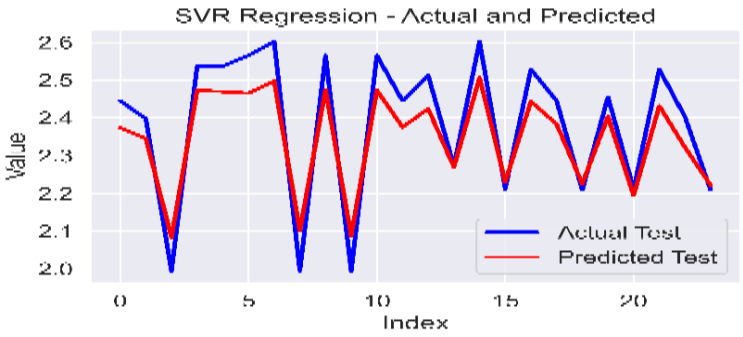
Figure 3.19 (a), (b), and (c) illustrate the performance of SVR regressor across three different plots. Figure 3.19 (a) is a residual plot showing the residuals against predicted values, with blue squares representing training data and green squares for validation data. It's indicating that the residuals are mostly near to ±0.02. Figure 3.19(b) correlates predicted values with actual values, depicting a linear correlation hence the model is better fit with both training and validation data. Figure 3.19(c) compares actual versus predicted values across various test indices, indicating that the model effectively predicts outcomes within the observed range. Overall, the SVR model revealed reliable accuracy in regression tasks.



(a)



(b)



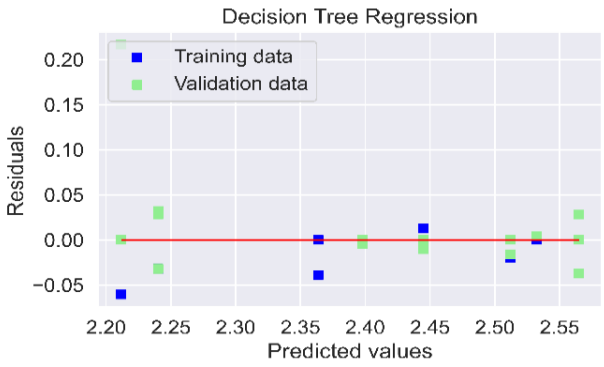
(c)

Figure 3.19 (a) SVR Regression Residual Graph (b) Real vs. Predicted Graph (c) Coefficient Graph

3.2.7.3.2. Interpretable Modelling – Tree-Based Modelling

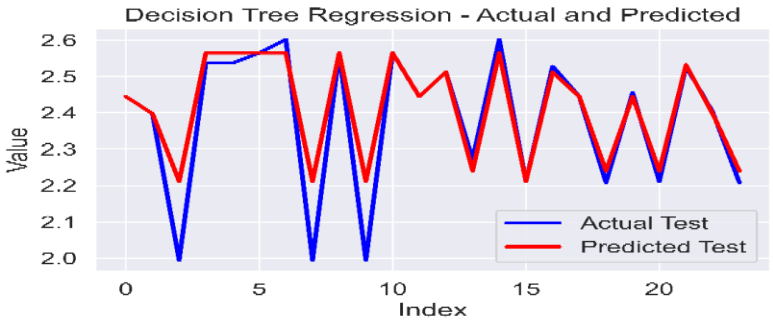
3.2.7.3.2.1. Decision Tree Regressor

Figure 3.20 illustrates the regression performance of the DT model in green tea quality evaluation. Figure 3.20 (a) denotes residual plot of the model, which shows both validation and training data lies within ± 0.05 against the predicted values, indicating small residuals and a good fit for prediction. Figure 3.20 (b) correlates predicted values with actual values, displaying a red regression line that effectively captures the trend of both the training and validation data. Figure 3.20 (c) compares actual test values with predicted values across various indices, the actual result of the model (blue in) highly related with the predicted outcome (red line) and there is small deviation in the prediction performance. Hence, the model well well-suited for green tea quality assessment.



(a)

(b)

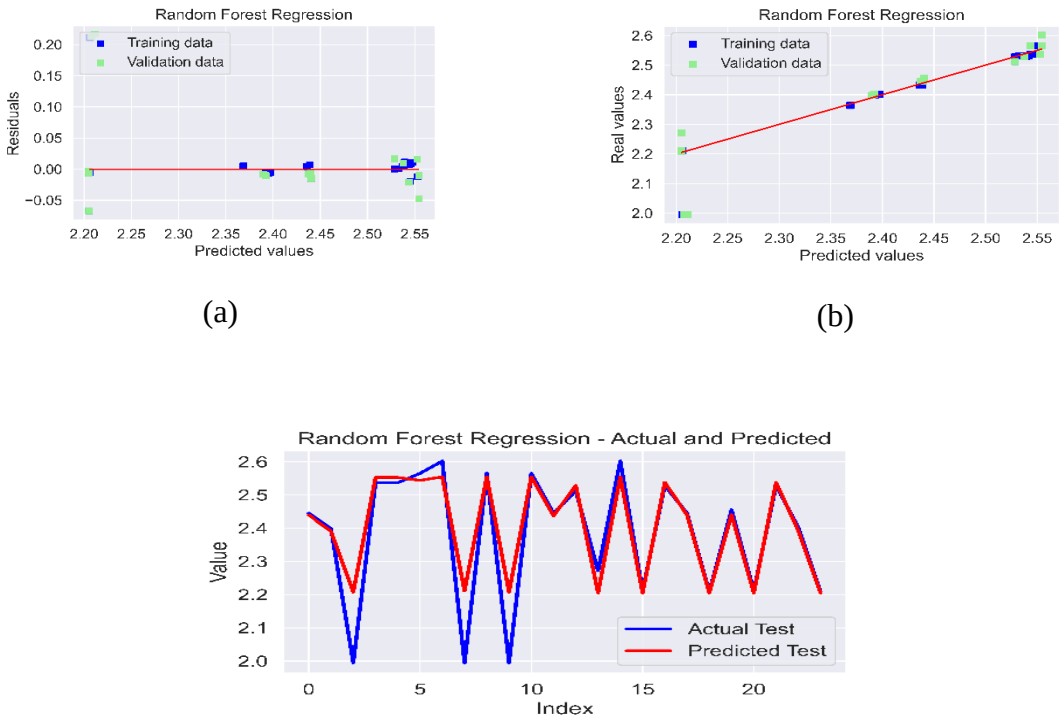


(c)

Figure 3.20 (a) Decision Tree Residual Graph (b) Prediction Graph (c) Actual Vs Predicted Values

3.2.7.3.2.2. Random Forest Regressor

Figure 3.21 (a) defines the residual plot of RF model, where the plot has shown better performance than the DT model. Here, the predicted values are presented on the x-axis, and the residuals are on the y-axis. This shows higher performance than DT regression. Figure 3.21(b) shows the relationship between the predicted values and actual values. Here, the predicted values are aligned more closely with the exact values. This demonstrates a linear relationship, indicating that the model is suitable. Similarly, Figure 3.21 (c) demonstrates a coefficient plot for the random forest regression, where the x-axis shows the different features of the model as predictors. Here, the coefficients are nearer to zero and indicate a minimal effect on the model result, whereas the absolute values show a robust influence on the model's results.

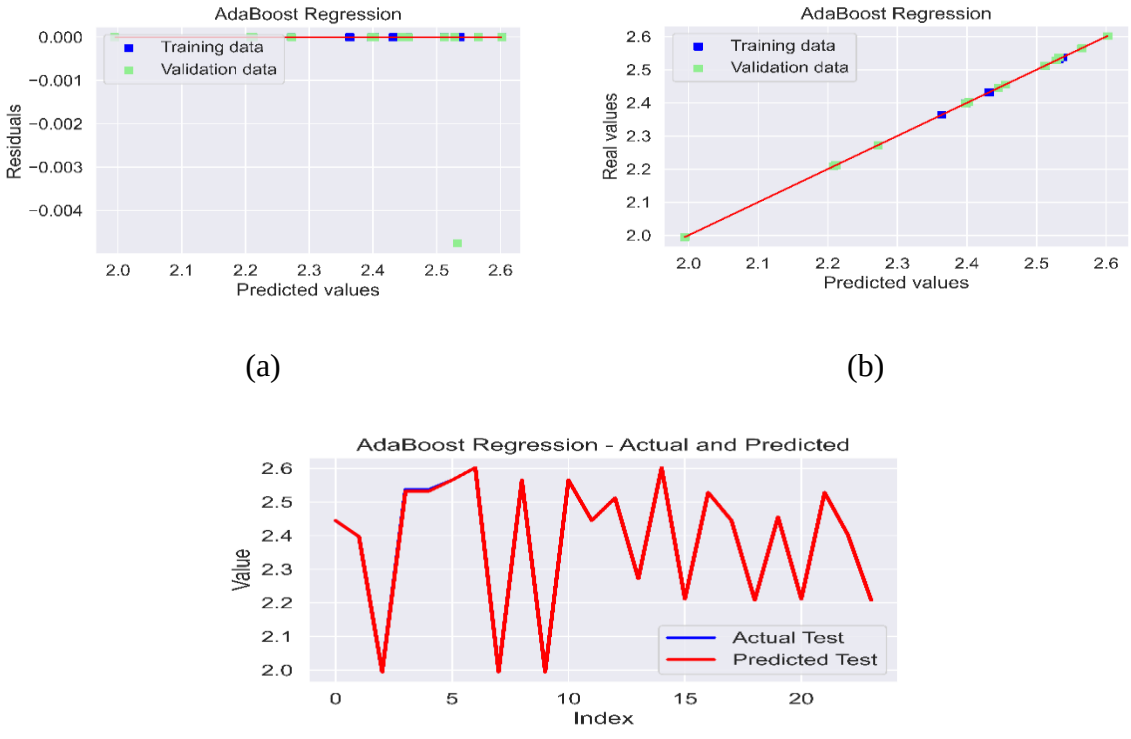


(a) (b)

(c)

Figure 3.21 (a) Random Forest Residual Graph (b) Prediction Graph (c) Actual Vs Real Scheme

3.2.7.3.2.3. AdaBoost Regressor



(a) (b)

(c)

Figure 3.22 (a) AdaBoost Residual Graph (b) Prediction Graph (c) Real Vs predicted

Figure 3.22 illustrates the performance of the AdaBoost regressor model. Fig 3.22 (a) is a residual plot among the training and validation samples against the predicted value. It shows that both these sample lies within deviation ±0.01. The results indicate that the predictions are highly accurate for both the training and validation data and showcase superior performance to RF and DT models. Figure 3.22 (b) shows the prediction plot comparison, where the training and validation plots lie closer to the actual values. This shows the prediction remains perfect, which reduces overfitting with minimal coefficients and is beneficial. Figure 3.22 (c) shows the real and predicted values of the AdaBoost model.

3.2.7.3.2.4. Gradient Boost Regressor

Similarly, Figure 3.23 demonstrates the performance outcome of the Gradient Boost.

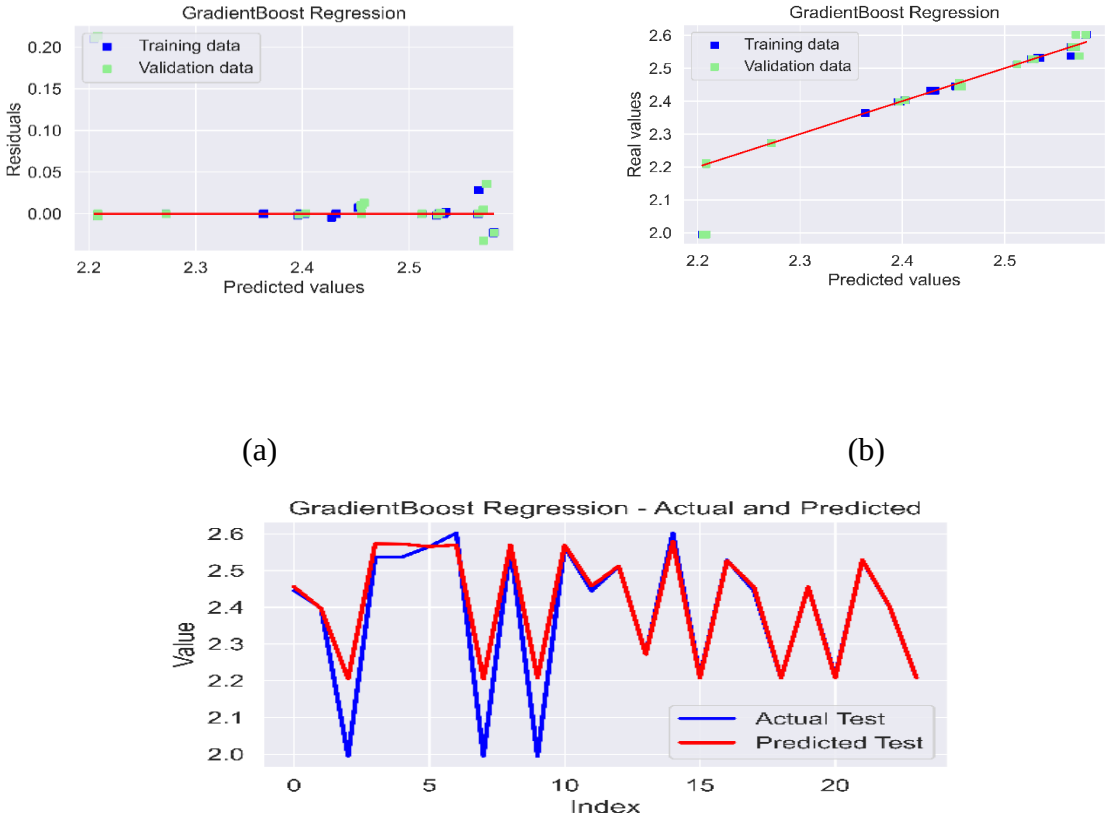
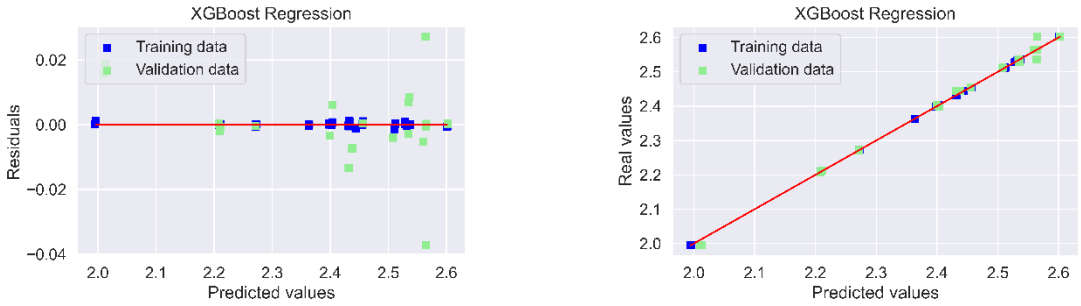


Figure 3.23 (a) Gradient Boost Residual Graph (b) Prediction Graph (c) Real Vs predicted

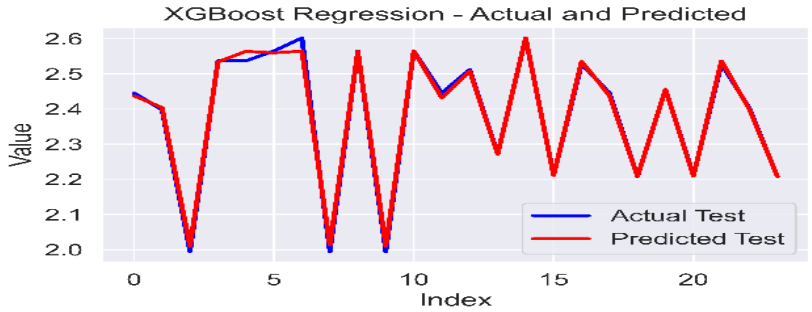
It can be illustrated using the Residual and Prediction Graph of actual and predicted values. The findings revealed that the model is well-suited and yielded superior accuracy than other models in assessing the green tea quality using EGCG data.

3.2.7.3.2.5. XGBoost Regressor

Likewise, Figure 3.24 shows the XGBoost regressor outcome in green tea quality prediction.



(a) (b)



(c)

Figure 3.24 (a) XGBoost Residual Design (b) Prediction Design (c) Real and predicted value

This model also showcased linear relationship between actual and predicted values and achieved remarkable prediction result. The residual plot shown nearly 0.1 deviation with the predicted outcome. Besides, the line plot implies closely aligned values between predicted and actual value. Hence, the model is good fit prediction.

3.2.7.4. Model Robustness and Validation

To improve model robustness and predictive generalisation, the present study incorporates cross-validation during the training and testing stages. Feature selection and dimensionality reduction approaches were additionally employed to reduce redundant information and improve model stability. Furthermore, multiple evaluation metrics, including Accuracy, ROC-AUC, MAE, RMSE, and R^2 score, were considered to comprehensively assess predictive performance. Although the obtained results demonstrate strong predictive capability, further validation using larger and externally acquired datasets may provide additional insight into the scalability and reliability of the proposed framework.

3.2.7.5. Results and Discussion - Performance Analysis

Regression metrics are quantitative measures used to evaluate the performance of regression models that predict continuous values. These metrics calculate the variations between the predicted continuous values and the actual values, aiding in model fitting and enabling accurate predictions. The metrics considered in the study are as follows,

MAE: It calculates the mean of the absolute variations between the actual and predicted values. Moreover, it shows that the average magnitude of errors is less sensitive to outliers than MSE or RMSE.

$$MAE\ (y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} |y_i - \hat{y}_i| \quad (3.4)$$

MSE: It is the average of squared variations between the actual and predicted values. Furthermore, it refines significant errors more thoroughly than minor ones, making it more sensitive.

$$MSE\ (y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2 \quad (3.5)$$

Median Absolute Error (MedAE): It calculates the median of the absolute variations between the real and determined values

$$MedAE\ (y, \hat{y}) = median\ (|y_1 - \hat{y}_1|, \ldots |y_n - \hat{y}_n|) \quad (3.6)$$

Mean Squared Logarithmic Error (MSLE): It calculates the average squared logarithmic variations between actual and predicted values.

$$MSLE\ (y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} \left(\log_e(1 + y_i) - \log_e(1 + \hat{y}_i)\ \right)^2 \quad (3.7)$$

Here, $\log_e(x)$ is the natural logarithm of x, n is the number of data points, and the real value is denoted as a real number. y_i And the determined value is $\hat{y}_i$.

R-squared: It is also known as the coefficient of determination and is a numerical measure that indicates the proportion of variance in actual dependent values explained by the regression model.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2}{n \sum_{i=0}^{n-1} (y_i - \bar{y}_i)^2}, \text{ here } \bar{y} = \frac{1}{n} \sum_{i=0}^{n-1} y_i \quad (3.8)$$

3.2.7.5.1. Performance Analysis of High-Performance Modelling

The results for each of the four different regression approaches, or algorithms — linear regression, Ridge regression, Lasso regression, and Elastic net regression — are shown in a performance matrix in Table 3.3.

Table 3.3 Performance Matrix of the Regression Techniques

Models	MAE	MedAE	MSE	MSLE	Explained Variance Score	R^2 Score
Linear	6.911 x10^{-15}	5.329 x10^{-15}	8.968 x10^{-29}	9.159 x10^{-31}	1.0	**1.0**
Ridge	8.988 x10^{-04}	7.663 x10^{-04}	1.236 x10^{-06}	1.146 x10^{-08}	0.99	**0.99**
Lasso	2.318 x10^{-03}	1.830 x10^{-03}	1.009 x10^{-05}	9.745 x10^{-08}	0.99	**0.99**
Elastic Net	**2.318 x10^{-03}**	**1.830 x10^{-03}**	**1.009 x10^{-05}**	**9.745 x10^{-08}**	**0.99**	**0.99**

Table 3.4 Predicted Performance Matrix of the Regression Techniques

Models	MAE	MedAE	MSE	MSLE	Explained Variance Score	R^2 Score
Linear	7.803 x10^{-04}	5.046 x10^{-04}	1.035 x10^{-06}	8.307 x10^{-09}	1.0	**1.0**
Ridge	1.174 x10^{-03}	9.817 x10^{-04}	2.003 x10^{-06}	1.852 x10^{-08}	1.0	**1.0**
Lasso	2.265 x10^{-03}	1.266 x10^{-03}	9.870 x10^{-06}	1.007 x10^{-07}	0.99	**0.99**
Elastic Net	**2.265 x10^{-03}**	**1.266 x10^{-03}**	**9.870 x10^{-06}**	**1.007 x10^{-07}**	**0.99**	**0.99**

The performance metrics are: MAE, MedAE, MSE, MSLE, and explained variance score (R^2). The linear regression provided the best results for all metrics, with an MAE of 6.911 and an explained variance score of 1.0. Ridge regression was slightly worse than linear regression, with an MAE of 8.988 and an R^2 of 0.99. Although it cannot explain the dataset with perfect accuracy, it clearly has explanatory usefulness. The Lasso result was excellent,

with an MAE of 2.318. Lasso regression yielded very similar results to those of elastic net, indicating good feature selection and the avoidance of overfitting. The Lasso and Elastic net both have an R^2 of 0.99, indicating fair explanatory quality, despite their errors not being as good. Overall, the results reveal differences and variations in performance metrics across all suggested regression techniques. This provides numerous new observations, insights, and indications regarding the performance of each method and the best-fit approach for each.

Table 3.4 summarises the predictive metrics for four regression models in the performance matrix, which includes the predictive metric scores for each model. The linear model achieved an MAE score of 7.803×10^{-4} and an MSE score of 1.035×10^{-6}, as well as a perfect R^2 score of 1.0. From this, we know that the model has captured 100% of the variance from the data. The Ridge Regression model achieved an MAE score of 1.174×10^{-3}, while the R^2 earned a perfect score of 1.0. Although the MAE value was slightly higher than that of the linear regression model, the Ridge model still performed reliably. The Lasso Regression and Elastic Net models yielded similar results, with an MAE of 2.265×10^{-3} and an R^2 of 0.99, indicating that both models effectively selected the most important features while mitigating the issue of overfitting. Additionally, while all of these results were rational predictors for the dataset, they were still weaker in terms of accuracy than the Linear and Ridge models. All in all, the collected performance metrics provide insights into the performance and reliability of the four unique regression methods outlined for predicting the dataset.

3.2.7.5.2. Performance Analysis of Interpretable Modelling – Tree-based Modelling

Figure 3.25 illustrates the training times of the RF, DT, AdaBoost, Gradient Boost, and XGBoost. Among all models, the XGBoost model has yielded a superior accuracy of 0.9962% and an average training time of 22.84 ms. On the other hand, AdaBoost has taken the highest time of 206.67 ms for the training process. It demonstrates that XGBoost is highly suitable for predicting EGCG content in green tea. Tables 3.5 and 3.6 present the performance matrix for the training and testing datasets.

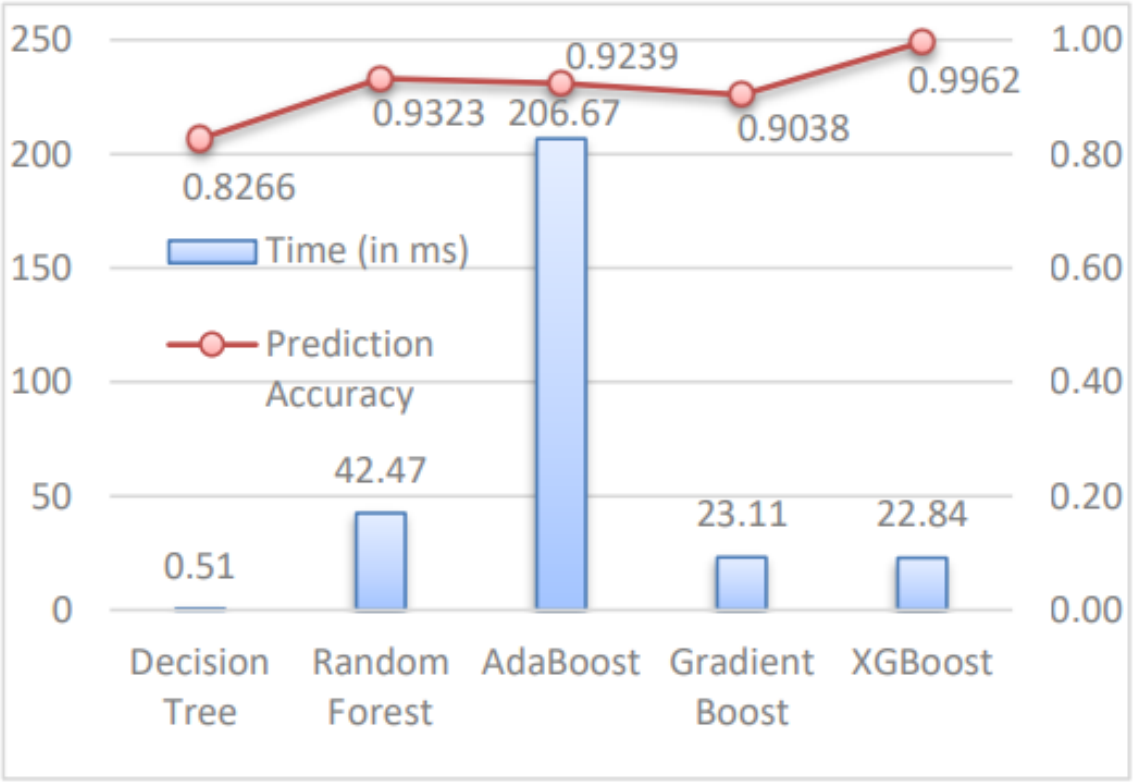


Figure 3.25 Training Time and Prediction Accuracy of Each Tree-Based Model

Table 3.5 describes the hyperparameters used to tune the tree-based models, and Table 3.6 and Table 3.7 presents the actual and predicted values of all four tree-based models employed in the study. Table 3.7 reveals that the XGBoost model has superior predicted accuracy compared to all other models.

Table 3.5 Details of Hyperparameters for Tree-based Models

Model	Hyper parameter	Search Space	Best optimal Value	No. of Candidates	No. of fits (10 Fold)	Execution time (in sec)
DT	max_depth	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	5	4500	45000	22.8
	min_samples_leaf	varying from 10% to 50% of	0.1			

Model	Hyper parameter	Search Space	Best optimal Value	No. of Candidates	No. of fits (10 Fold)	Execution time (in sec)
		the samples				
	min_samples_split	varying from 10% to 100% of the samples	0.2			
	max_features	"sqrt", "log2", "auto"	"auto"			
	Criterion	"mse", "friedman_mse", "mae"	"mae"			
RF	Criterion	"mse", "mae"	"mse"	9000	90000	3822
	max_depth	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	10			
	max_features	"sqrt", "log2", "auto"	'auto'			
	min_samples_leaf	varying from 10% to 50% of the samples	0.1			

Model	Hyper parameter	Search Space	Best optimal Value	No. of Candidates	No. of fits (10 Fold)	Execution time (in sec)
	min_samples _split	varying from 10% to 100% of the samples	0.2			
	n_estimators	100, 200, 300	100			
AdaBoost	base estimator max_depth	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	7	30	300	186
	n_estimators	500, 1000, 1500	1000			
	loss_function	“linear”, “square”, “expone ntial”	“linear”			
XGBoost	n_estimators	100, 200, 300, 400, 500, 1000	100	4860	48600	1110
	Subsample	0.3, 0.6, and 0.9	0.9			
	max_depth	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	100			
	Booster	“gbtree”, “dart”,	0.9			

Model	Hyper parameter	Search Space	Best optimal Value	No. of Candidates	No. of fits (10 Fold)	Execution time (in sec)
		"gblinear"				
	learning_rate	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9	0.3			
Gradient Boost	Criterion	"friedman_mse", "mse", "mae"	"mae"	810000	8100000	187224
	Loss	"ls", "lad", "huber", "quantile"	"ls"			
	learning_rate	0.2, 0.6, 0.9	0.2			
	max_depth	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	6			
	min_samples_leaf	varying 10% to 50% of the samples	0.1			
	max_features	"sqrt", "log2", "auto"	"auto"			

Model	Hyper parameter	Search Space	Best optimal Value	No. of Candidates	No. of fits (10 Fold)	Execution time (in sec)
	min_samples _split	A range of 10% to 100% of the samples	0.5			
	n_estimators	100, 200, 300, 400, 500	100			
	Subsample	0.3, 0.6, 0.9	0.9			

Table 3.6 Real Performance Values

Model	Actual Performance Metrics					
	MSE	MSLE	MAE	MedAE	R^2	Variance score
DT	0.002	0.0002	0.0171	0.0016	0.8948	0.8991
AdaBoost	0.0009	7.925×10^{-05}	0.0258	0.0189	0.9505	0.9505
RF	0.0002	1.811×10^{-05}	0.0053	0.0021	0.9909	0.9909
Gradient Boost	0.0015	0.0002	0.0087	0.0001	0.9196	0.9234
XGBoost	2.151×10^{-07}	1.863×10^{-08}	0.0003	0.0002	0.9999	0.9999

Table 3.7 Predicted Performance Values

Model	Prediction Performance Metrics					
Model	MSE	MSLE	MAE	MedAE	R^2	Variance score
DT	0.0065	0.0007	0.0425	0.0072	0.8266	0.8463
AdaBoost	0.0029	7.913×10^{-05}	0.0313	0.018	0.9239	0.9239
RF	0.0026	0.0003	0.0215	0.0043	0.9323	0.9356
Gradient Boost	0.0036	0.0004	0.0214	2.543×10^{-09}	0.9038	0.9148
XGBoost	0.0001	1.259×10^{-05}	0.0076	0.0028	0.9962	0.9962

3.3. Conclusion

This chapter presents an advanced framework designed to predict the quality of green tea through distinct methodologies. The EGCG is considered as the most active catechin present in green tea producing increased health benefits. This indicates that the intake of EGCG present green tea is essential and plays a significant part in tumbling the high risk of diseases. So, the proposed method involved EGCG detection in green tea samples using molecularly imprinted electrode. For green tea, the quality assessment was conducted by quantifying the EGCG content using a MIP sensor. Feature engineering techniques like Box-Cox Transform, applying correlation coefficients for extracting the hidden features based on polynomials has proved to be a smart idea. Non-normality, asymmetry, and homoscedasticity anomalies have been reduced using power transformation techniques such as the box-cox transform. Subsequent classification of the data was performed utilising the XGBoost model, which is known for its high performance and efficiency in handling complex datasets. The experimental results showed that the proposed method outperformed in classifying the green tea samples with optimal values based on the presence of EGCG content.

For the quantification of EGCG content present in green tea samples utilising the DPV data obtained using the developed MIP-EGCG electrode, the linear regression DL technique with regularisation techniques is employed on the dataset for high-performance modelling. To

obtain a generalised model, the modelling has been fine-tuned. Regularisation techniques have helped in reducing the overfitting of the model and to give better predictions. From the result of model metrics, a clear picture is portrayed how lasso regression performs better than ridge regression for this dataset and eliminates the less important features to develop the model as sparsity can be useful in practice if we have a high dimensional dataset with many features that are not effective for modelling. ElasticNet Regression is the combination of both L1 and L2 penalty, where L1 penalty generate sparsity and L2 penalty overcome the overfitting problem. As per the result obtained from our model it can identify the quality of green tea most accurately i.e., 99%.

Tree-based regression algorithms, such as decision tree, random forest, AdaBoost, gradient boost, and XGBoost, have been used to create robust cognitive models. Ensemble learning methods like as bagging and boosting were used. XGBoost, a regularised gradient boosting approach, took the least amount of time to train the model, about half the duration of the random forest model. As can be seen from the model metrics, results of the model show that the bagging algorithm, random forest, was able to correctly determine 93% of the EGCG content, whereas the boosting model, XGBoost, performed exceptionally well and was able to correctly determine 99% of the EGCG content in determining the quality of green tea. As a result of its superior performance and regularisation abilities, as well as its capacity to prune trees efficiently, XGBoost modelling has emerged as the obvious winner in interpretable modelling. Further, the method can be deployed in other quantitative determination of other polyphenol components present in the tea.

Furthermore, this chapter presents the performance metrics of the framework, enabling a comprehensive comparative evaluation that highlights the robustness and effectiveness of the models. The results not only demonstrate the accuracy of the predictions but also highlight the potential of these frameworks for application in quality control processes within the tea industry, paving the way for enhanced product standards and improved consumer satisfaction.

REFERENCES

[1] W. Tong *et al.*, "Black tea quality is highly affected during processing by its leaf surface microbiome," *Journal of Agricultural and Food Chemistry*, vol. 69, no. 25, pp. 7115-7126, 2021, doi: https://doi.org/10.1021/acs.jafc.1c01607.

[2] J. Moreira *et al.*, "Tea quality: An overview of the analytical methods and sensory analyses used in the most recent studies," *Foods*, vol. 13, no. 22, p. 3580, 2024, doi: https://doi.org/10.3390/foods13223580.

[3] M. Moulick, S. Nag, D. Das, A. K. Hazarika, S. Sabhapondit, and R. B. Roy, "A Molecular Imprinted Bi-polymer graphite electrode decorated with NiCo2O4 nano-cubes for rapid detection of theaflavin in black tea," *IEEE Sensors Journal*, 2025, doi: https://doi.org/10.1109/JSEN.2025.3555887.

[4] S. Acharya *et al.*, "Optimization Techniques for a Voltammetric Signal to Predict Green Tea Quality Parameters Using MIP Electrode," *IEEE Sensors Journal*, vol. 23, no. 17, pp. 19842-19847, 2023, doi: https://doi.org/10.1109/JSEN.2023.3297140.

[5] H. Xia *et al.*, "Rapid discrimination of quality grade of black tea based on near-infrared spectroscopy (NIRS), electronic nose (E-nose) and data fusion," *Food Chemistry*, vol. 440, p. 138242, 2024, doi: https://doi.org/10.1016/j.foodchem.2023.138242.

[6] A. Modak, D. Das, T. N. Chatterjee, B. Tudu, and R. B. Roy, "Regression-Based EGCG Detection in Green Tea Employing MIP Electrodes," *IEEE Sensors Journal*, vol. 24, no. 11, pp. 18090-18097, 2024, doi: https://doi.org/10.1109/JSEN.2024.3390438.

[7] A. Modak, D. Das, and R. B. Roy, "Classification of Green Tea Samples Based on Epigallocatechin-3-Gallate Content Using Molecular Imprinting Polymerization Technique," in *2024 IEEE 3rd International Conference on Control, Instrumentation, Energy & Communication (CIEC)*, 2024: IEEE, pp. 355-360, doi: https://doi.org/10.1109/CIEC59440.2024.10468449.

[8] T. Kluyver *et al.*, "Jupyter Notebooks–a publishing format for reproducible computational workflows," in *Positioning and power in academic publishing: Players, agents and agendas*: IOS press, 2016, pp. 87-90, 2016, doi: https://doi.org/10.3233/978-1-61499-649-1-87.

[9] K. J. Millman and M. Aivazis, "Python for scientists and engineers," *Computing in science & engineering*, vol. 13, no. 2, pp. 9-12, 2011, doi: https://doi.org/10.1109/MCSE.2011.36.

[10] T. E. Oliphant, "Python for scientific computing," *Computing in science & engineering*, vol. 9, no. 3, pp. 10-20, 2007, doi: https://doi.org/10.1109/MCSE.2007.58.

[11] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy array: a structure for efficient numerical computation," *Computing in science & engineering*, vol. 13, no. 2, pp. 22-30, 2011, doi: https://doi.org/10.1109/MCSE.2011.37.

[12] E. Jones, T. Oliphant, and P. Peterson, "SciPy: Open source scientific tools for Python," 2001.

[13] W. McKinney, "Data structures for statistical computing in Python," *scipy*, vol. 445, no. 1, pp. 51-56, 2010, doi: https://doi.org/10.25080/Majora-92bf1922-00a.

[14] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in science & engineering*, vol. 9, no. 03, pp. 90-95, 2007, doi: https://doi.org/10.1109/MCSE.2007.55.

[15] A. Modak, T. N. Chatterjee, S. Nag, R. B. Roy, B. Tudu, and R. Bandyopadhyay, "Linear regression modelling on epigallocatechin-3-gallate sensor data for green tea," in *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 2018: IEEE, pp. 112-117, doi: https://doi.org/10.1109/ICRCICN.2018.8718696.

CHAPTER 4

MULTI-OUTPUT REGRESSION ON AN ARRAY OF MIP SENSORS FOR BLACK TEA QUALITY EVALUATION

This chapter discusses the details of the third framework, specifically the PSRF (Particle Swarm Optimisation-based Random Forest) regression model for assessing black tea quality. The model utilises multiple MIP sensors to collect the components such as Tannic Acid (TAN), Theophylline (THEO) and Caffeine (CAFF) from a black tea sample. The proposed model utilises a greedy-based Genetic Algorithm for effective feature selection, and regression is performed with the PSRF model. The proposed model demonstrates promising prediction accuracy with a minimal false negative rate.

LIST OF SECTION

Contents of this chapter are based on following publication:

- **A. Modak**, M. Moulick, D. Das and R. B. Roy, "Multi-Output Regression on MIP Sensor Feature Fusion for Black Tea Quality Evaluation," Chemometrics and Intelligent Laboratory Systems [Communicated – Under Review]

CHAPTER 4

MULTI-OUTPUT REGRESSION ON AN ARRAY OF MIP SENSORS FOR BLACK TEA QUALITY EVALUATION

4.1. Introduction

Tea is a non-alcoholic beverage consumed primarily by people around the world, renowned for its diverse flavour profiles and associated health benefits. Its unique organic and chemical composition, including compounds like catechins, theanine, and polyphenols, contributes to improving antioxidant activity and also reduces the risk of cardiovascular attacks and chronic gastritis [1]. Among the various types of tea flavours, black tea appears good due to its extensive oxidation process during manufacturing [2]. This process leads to the formation of characteristic compounds, like theaflavins and thearubigins, along with a distinct flavour profile. The interplay of these compounds, alongside others such as amino acids, sugars, and volatile aromatics, determines the overall quality and unique characteristics of black tea [3].

Like tea quality has been estimated based on single MIP sensor targeting a particular molecule. But tea quality depends on the contribution from other constituents also. This is an attempt to analyse the tea quality based on the signature analysis of multi molecules present in tea. Here three such major important taste affecting constituents are targeted to prepare an electronic tongue. Accordingly three MIP electrodes, Tannic Acid (TAN), Theophylline (THEO) and Caffeine (CAFF) have been prepared [4].

Thus, the proposed research introduces a novel PSRF (Particle Swarm Optimisation-based Random Forest) regression model. The modified PSO, obtained by hyper-tuning the traditional PSO, is combined with the RF Model to perform the regression process. First, the study collects the data by incorporating multiple MIP sensors to measure tannic acid, theophylline and caffeine. The collected data is fused and prepared as a dataset for quality analysis. Then, a Greedy-based Genetic Algorithm (GA) is employed to perform feature selection, selecting the enriched feature set from the input data. The proposed PSRF is to conduct the regression process. The efficiency of the multisensory array is evaluated using standard metrics and demonstrates enhanced accuracy with a lower error rate in predicting black tea quality.

4.2. Proposed Methodology

The present study aims to enhance the accuracy of the quality assessment of black tea. The process includes four stages. In stage 1, the system incorporates three specific sensors: the first measures TAN, the second measures THEO, and the third measures CAFF. These sensors gather essential chemical data necessary for evaluating tea quality, setting the foundation for the analysis. In stage 2, feature fusion and preprocessing are conducted to prepare the dataset for further study from multiple MIP sensors. In the preprocessing step, missing values and label encoding are employed to convert categorical information into a numerical format. This step enhances the quality of the data before feature selection.

In stage 3, feature selection is performed using a greedy-based GA, which identifies the most relevant features from the preprocessed data to improve model performance. This selection step is critical, as it ensures that only the most impactful features are included, potentially increasing the accuracy of the regression model. Following that, the prepared sample is split into 80% for training the model and 20% for testing purposes. This division allows for the evaluation of the model's performance on unseen data, ensuring a robust assessment of its effectiveness in predicting black tea quality.

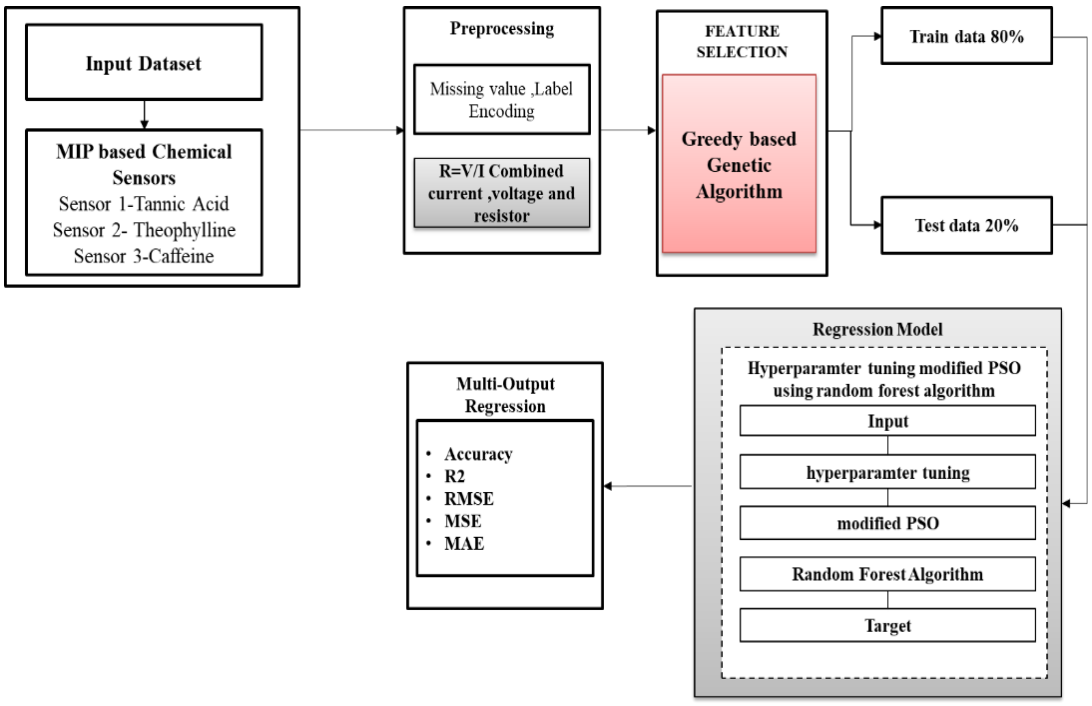


Figure 4.1 Steps Included in the Proposed Work

Finally, the proposed PSRF model is utilised, which is specifically designed using hyperparameter tuning with a modified PSO method and combined with Ant Colony optimisation. Then, the ACO-PSO approach is incorporated into the RF algorithm. This sophisticated approach enhances the model's ability to learn from the training data and accurately predict the quality of black tea based on the identified chemical characteristics. Overall, the methodology presented in the image underscores the importance of careful data handling, feature selection, and advanced algorithmic techniques in developing an effective quality evaluation model. In comparison to the traditionally used single-output models, the present model uses multi-output regression along with optimization-assisted feature selection with sensor array. Figure 4.1 illustrates the steps included in the proposed work, with detailed explanations provided below.

4.2.1. Preparation of Data

In this context, the quality of the black tea data collected from three different MIP sensor arrays has been predicted and compiled into a single dataset. Each sensor captures 280 features from the 3000 samples. The 80 unique features were expected, and these details were collected in a consistent experimental setup, utilising the same scan rate across all sensors to ensure uniformity in data size.

4.2.2. Experimental Details

4.2.2.1. Chemical reagents and standards

The study has gathered Theophylline, TAN from Nice Chemical Pvt. Ltd. At the same time, acrylonitrile (AN), Acrylic acid [5], graphite, Methyl methacrylate [6], EGDMA (Ethylene glycol dimethacrylate), and benzoyl peroxide were obtained from Sigma Aldrich, India. In addition, aluminium (Al), magnesium [7], paraffin oil, caffeine, and dopamine were acquired from Merck, India. The primary data of Black tea samples were collected from the Asian School of Tea in Kolkata. Millipore water is used to prepare the solution with a resistance rate of 18 MΩ. The experiment was conducted at room temperature.

4.2.2.2. Synthesis of the three MIP Electrodes

The polymer was prepared by first dispersing graphite in ethanol, followed by the incorporation of a target molecule, the monomer EGDMA, and benzoyl peroxide. This resultant mixture was subjected to polymerisation at a temperature of 30℃. Subsequently, it was washed with water to eliminate any unreacted substances and underwent an additional 24-hour wash to remove

the target molecules, thereby generating MIP recognition sites. The final polymer was dried and collected for the development of MIP sensors, utilising the imprinted cavities that were specific to the extracted target molecule for selective recognition, specifically MIP_CAFF, MIP_TAN, and MIP_THEO. A quantity of 0.3g of the MIP material was combined with the synthesised nanoparticles and ground into a fine paste using paraffin oil. This composite was then packed into a capillary tube with an inner radius of 1.25 mm, leading to the creation of the MIP electrodes. Copper wires were used as electrical contact.

4.2.2.3. Equipment and Characterisation

During the experimental analysis, a potentiostat/galvanostat system, known as PGSTAT101, utilises a configuration comprising three electrodes. This arrangement features a platinum counter electrode, with MIP_TAN/MIP_CAFF/MIP_THEO serving as the working electrode, alongside an Ag/AgCl reference electrode. The morphological characterisation of the MIP materials was conducted using a Scanning Electron Microscope (SEM).

Sensor fusion is the process of combining the data from multiple sensors to achieve a more reliable and accurate prediction. By combining data from various sensors, the system gains a more comprehensive and precise understanding of the chemical characteristics that influence the quality of black tea. This process helps to reduce the noise and uncertainty associated with individual sensor measurements, leading to more accurate data for analysis. Furthermore, integrating data from diverse sensors enables a richer and more comprehensive representation of the tea's chemical profile, allowing for a more nuanced evaluation of its quality.

4.2.2.4. Electrical parameter variation of the fabricated sensors

The quality assessment of tannic acid, theophylline, and caffeine from black tea samples employs a quantitative instrumental approach, including HPLC (High-Performance Liquid Chromatography). In this study, three different MIP sensors were utilised to investigate the target compounds. In this process, the voltammetric method was applied, where current (I) is calculated by applying Voltage (V). It is insufficient to use only the current values, as the measured voltage values applied to all three MIP sensors differ, and we need to fuse the data. By using the voltage and current values of three sensors, the resistance of the sensor has been calculated by using Equation 4.1 as given below,

$$R = \frac{V}{I}. \tag{4.1}$$

Therefore, the resistance of each sensor has been identified as a quantifiable parameter, serving as input for the modelling process. Table 4.1 illustrates the minimum and maximum values employed by the sensors for TAN, THEO, and CAFF, corresponding to the potential variations applied, variations in current, and the range of resistance variations, respectively.

Table 4.1 Variation range of the Sensors

Potential Variation of the Sensors	**Range (V)**	**TAN**	**CAFF**	**THEO**
	Minimum	0.00351	1.297607	0.502014
	Maximum	0.70343	1.997528	1.201935
Current Variation of the Sensors	**Range (Ampere)**	**TAN**	**CAFF**	**THEO**
	Minimum	0.001534	0.00389512	0.00024
	Maximum	0.00239	0.00631529	0.00038
Resistor Variation of the Sensors	**Range (Ohm)**	**TAN**	**CAFF**	**THEO**
	Minimum	0.0016	0.0011	0.0004
	Maximum	0.0031	0.002	0.0027

4.2.3. Feature Selection Using Genetic Algorithm

In the proposed research, instead of randomly initializing the genetic population, a greedy-based GA was employed to select the relevant feature from the preprocessed input data. It combines the exploratory power of GAs with the efficient, quick decision-making of greedy algorithms. This hybrid approach aims to achieve a balance between thoroughly exploring the feature space and rapidly converging to an optimal or near-optimal solution. This method involves a feedback linkage between feature evaluation and regression. Herein, feature transformation and regressor design are performed using a greedy genetic learning and evolution process.

This greedy-based GA aims to identify a reduced subset of features from the original N features. Its objective is to retain sufficient target discriminatory data while eliminating redundant target data and noise. The greedy algorithm plays a crucial role in this process by optimising individual stages and providing optimal solutions at each step. Simultaneously, the GA refines these solutions by applying selection attributes. By integrating these two approaches, the proposed model utilises the greedy algorithm to generate the initial population and to execute the mutation and crossover processes, ultimately enhancing the regressor's performance. In the present work, the greedy algorithm is employed to optimise individual phases of the problem by making locally optimal choices at each step, aiming to predict optimal solutions efficiently. Figure 4.2 illustrates the process of the greedy-based GA algorithm.

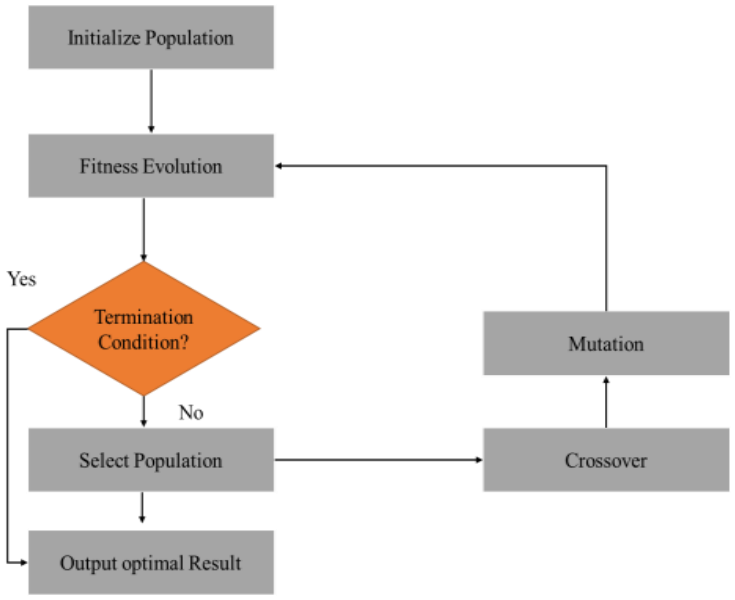


Figure 4.2 Greedy-based Genetic Algorithm for Feature Selection

Step A: Generation of Initial Population

The initial population plays a crucial role in forecasting both the quality of the final solution and the speed of convergence. Recognising the dependency of the final solution's quality, the proposed model incorporates a greedy population initialisation algorithm. This approach leverages domain-specific knowledge to generate high-quality chromosomes. Initially, biases and synaptic weights are chosen at random, leading the greedy genetic algorithm to evaluate the features of the input dataset that the regressor will be trained upon. Certain features contribute more substantially to the data than others, facilitating the accurate identification of the input pattern's class, which is achieved through the application of domain-specific knowledge. This greedy initialisation strategy is anticipated to surpass the effectiveness of

random initialisation. Moreover, this method consistently generates a valid individual concerning capacity constraints, as it inherently involves a local search.

Step B: Fitness Evolution

The selection process involves assessing the fitness value of every individual within the population. Following this assessment, an individual with a high fitness value is chosen to form the new population, thereby replacing the existing one. Once the fitness evolution is complete, a pre-defined termination condition is applied; if this condition is met, the optimal output is attained. Conversely, if the condition is not met, the population must be re-selected, incorporating crossover and mutation operations, and this process is reiterated from step 2 until the condition is fulfilled. The fitness function is expressed in an equation. 4.2,

$$F = \alpha.MSE + \beta.(1 - r^2) \quad (4.2)$$

Equation 1: α and β are weighting coefficients set as 0.6 and 0.4, respectively, to balance error minimisation with model generalisation. The fitness function value obtained by the proposed model is 0.9

Step C: Select Population

The selection of the population could be performed with high care because it can cause premature convergence. The individual to population $P_i(t_i)$ The chosen one is then processed by employing crossover and mutation.

Step D: Crossover operation

By choosing chromosomes from the present generation, the crossover operation is performed to produce offspring. It comprises two chromosomes that combine to form a new offspring, which undergoes an exchange of genetic information to produce high-quality chromosomes and create better offspring. The crossover technique involves assessing the average of the fittest chromosomes within the population, which facilitates the generation of offspring that are closer to the solution while minimising loss. Consequently, this derived mean chromosome serves as a benchmark, employed to enhance the global optimum by contrasting it with individuals exhibiting lower fitness levels in the population. As a result, the offspring that possess superior fitness values contribute to an overall enhancement in the quality of the population. Hence, offspring with enhanced fitness value improve the quality of the population. The crossover rate for the proposed model is considered to be 0.8.

Step E: Mutation process

The greed mutation tends to evade interruptions in the quality of good chromosomes and sustain genetic diversity in the population by mutating low-quality chromosomes. This step helps improve the quality of the entire population. This approach initially creates a mean chromosome by calculating the gene-wise mean of the top chromosomes. Then, it randomly selects a lower level of the chromosome to carry out the mutation task. This step is essential to explore the assortment in the population of each iteration. After that, a random number N is generated for each gene and compared with the mutation probability M_p. If N $>$ Mp Then the difference (D) between the value of the corresponding gene and the chosen gene is estimated, using a random number. R Is generated. Following that, the product of D and R is subtracted from the respective gene value, which helps the chromosome approach an enriched gene value, thereby optimising the overall fitness. This procedure is followed until an optimal feature is obtained, and those features are fed into the regressor to train the model to produce an efficient outcome. Algorithm 1 defines the summarised details of the feature selection process using a greedy-based GA.

Algorithm I: Greedy-based Genetic Algorithm for Feature Selection
Input: Noise Signal
Output: Best Feature selected from Signal
Determine Objective Function (OF_i)
Assign Number of Generation to $0(t_i = 0)$
Randomly Create Individuals in initial popoulation $P_i(t_i)$
Evalute Individuals in Population$P_i(t_i)$ using (OF_i)
While Termination Criterion is not satisfied do
$t_i = t_i + 1$
select the individuals to population $P_i(t_i)$ From $P_i(t_i - 1)$
Change individuals of $P_i(t_i)$ Using Crossover and Mutation
Evalute individuals in population of $P_i(t_i)$ Using OF
End While
Return Signal

Step F: Termination Criteria

The algorithm terminates after either 100 iterations or when the improvement in fitness between consecutive generations drops below 0.0001 for 10 successive iterations. Following this execution, the GA-based feature selection filters out 19 multi-collinear and noisy variables, retaining 405 highly informative features from the original 424-feature dataset. This deliberate preservation of dense chemical data ensures maximum information retention, which is then fed into the optimized PSO-ACO Random Forest regression model to predict tea constituent concentrations without losing critical spectral or kinetic resolution.

Finally, the greedy-based GA efficiently selects the best features from a noisy signal by iteratively refining a population of potential feature subsets using a combination of greedy and genetic approaches. The process starts by randomly generating an initial population of individuals, where each individual represents a possible subset of features. These individuals are then evaluated based on a defined objective function, which measures the quality of the feature subset. The algorithm then enters a loop, continuing until a predefined termination criterion is met, such as reaching a maximum number of generations or achieving a desired quality solution. In each generation, individuals are selected from the current population based on their fitness values, and then undergo crossover and mutation operations to introduce new combinations of features. The newly generated individuals are then evaluated using the same objective function, repeating the cycle of selection, variation, and evaluation. This iterative process allows the algorithm to explore the feature space effectively, ultimately leading to the selection of the best feature subset from the noisy signal as the final output.

4.2.4. Proposed Modified Regression Model

In the present study, a PSRF regression model is employed. After feature selection, the regression model involves hyperparameter tuning using a grid search approach to effectively improve the learning rate of the model. Boosting is the technique deployed in the algorithm, where weights are assigned at each instance. The grid search technique is less prone to overfitting issues, as it picks random combinations and evaluates which set gives the best model performance. It is considered to enhance the proposed regressor by fine-tuning its hyperparameters, such as the number of estimators, maximum leaf nodes, maximum depth, and maximum features, to achieve improved accuracy. In the proposed work, a hyperparameter value of 0.5 is set for individual MIP sensor modelling. The accuracy of the regression model

plays a significant role in predicting the quality of black tea input samples. Table 4.2 describes the hyperparameter details used to enhance the PSO algorithm.

Table 4.2 Hyper-parameter Details

Parameter	Value
n-estimator	25, 50, 100, 150
max-depth	3, 6, 9
min-leaf nodes	3, 6, 9
max-feature	sqrt and log2

After fine-tuning, the modified PSO approach is employed to efficiently select the optimal solution. This approach combines the strengths of the typical PSO algorithm with the ACO principle. PSO is a population-based stochastic optimisation method inspired by bird flocking or fish schooling behaviour. It navigates the problem space by following the current optimal particles to achieve faster and more reliable results. However, PSO is limited to single-objective problems and cannot handle multi-objective issues effectively.

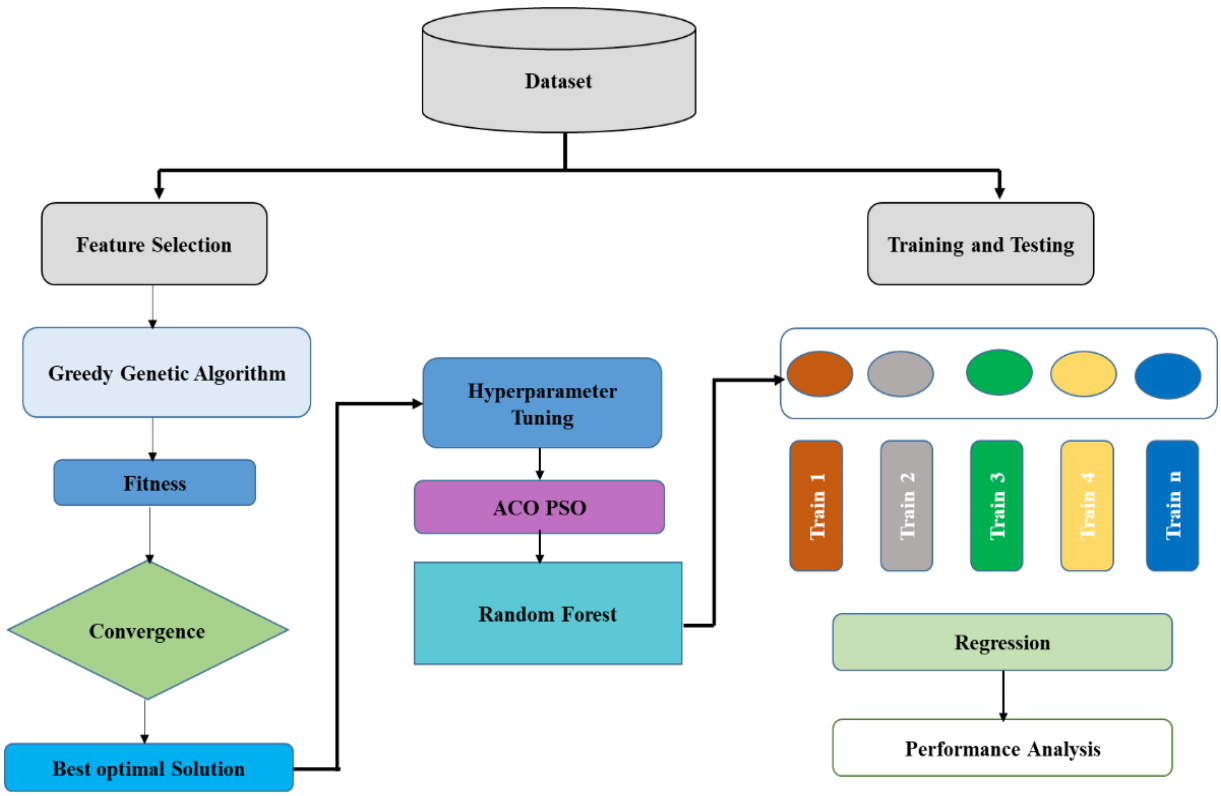


Figure 4.3 Overall Process involved in the Proposed Modified Regression Model

To address this limitation, ACO is integrated with PSO. ACO is a general search algorithm inspired by ant colonies searching for the shortest path using pheromone trails. It provides strong local search and exploitation capabilities. By combining the characteristics of ACO and PSO, the modified PSO model aims to enhance convergence speed and accuracy in finding the global optimum. The modified PSO approach improves optimisation performance and accuracy, enabling the RF algorithm to predict well-performing solutions and obtain precise results. Figure 4.3 illustrates the process involved in the proposed regression model.

In the proposed PSRF algorithm, a pheromone-guided method of ACO is applied to provide a local search that tackles P ants, equivalent to the number of particles in the PSO algorithm. Each ant "t" creates a solution. x_i^t around q_i^bThe global best solution is initiated among all particles present in the swarm. For the proposed model, the parameters for the inertia weight and acceleration factor are considered to be 0.9 and 2, respectively. The steps involved in the modified regression are detailed below, which create a possible solution space and enable the effective reach of an optimal solution.

Step 1: Initialisation of Optimisation by providing constants for ACO and PSO processes, Q, t_{max}

Step 2: Continue until the maximum number of solutions is obtained, then randomly select several features to activate. A local search procedure is applied.

Step 3: Evaluation of objective function values and assignment of best positions $q_i^t = y_i^t, i = 1, \ldots . Q.$

Step 4: Identify $g_i^{best}(q_i^{best}) = \min\{ g(q_i^1), \ldots \ldots, g(q_i^i), \ldots \ldots g(q_i^Q)\}$ And then initialise $(q_i^b) = (q_i^{best})$ and $g(q_i^b) = g^{best}(q_i^{best})$

Step 5: Perform optimisation by updating particle position $p_i^t and$ velocity v_i^t Of all particles. Evaluate the objective function values as $g(p_i^1)$. $Generate\ P\ solutions\ x_i^t$. Then $g(p_i^t) = g(x_i^t)\ and\ p_i^t = x_i^t$.

Step 6: Update the best position of the particle, $g(q_i^t) = g(p_i^t)$ and $q_i^t = p_i^t$ and finding $g_i^{best}(q_i^{best}) = \min\{g(q_i^1), .., g(q_i^t), \ldots . g(q_i^Q), if\ g(q_i^b) > g(q_i^{best})$ then $g(q_i^b) = g^{best}(q_i^{best})\ and\ q_i^b = q_i^{best}$.

Step 7: Calculation of the fitness function that measures the quality of a solution. Thus, the best solution q^b of the swarm with the objective function value $g(q^b)$.

The RF is signified as an ensemble learning method, which uses multiple DTs with the same direction to form a forest. This forest is used to train and predict the sample data and perform regression based on decision rules. The primary intent of the proposed RF is to predict regression with better generalising ability. Here, the RF tends to grow trees based on the numerical or quantitative values of the target variable, providing numerical output values. Moreover, the training data is independently selected from the data set. The subset of the predictor variable is chosen to divide the internal node, in which the splitting criteria depend on the mean squared error at each internal node.

Algorithm II: Random Forest

Input: Training dataset

1. Initialise the number of regression trees T
2. For each tree $t = 1$ to T

   Generate a bootstrap sample from the training dataset

   Randomly select a subset of features from the input feature space.

   Construct a regression tree using the selected samples and features.

   Determine optimal split points based on regression criteria.
3. Aggregate the outputs of all regression trees.
4. Compute the final prediction as the average output of all trees.
5. Evaluate prediction performance using statistical metrics

Output: Trained RF Regression Model

The steps involved in Algorithm II represent the process flow of the suggested PSRF regression framework. At first, bootstrap samples are created from the training data set in order to create many regression trees. For each regression tree, a different feature set is randomly chosen in order to create model diversity and also to reduce variance while predicting. Each individual regression tree can be developed individually with the use of selected feature sets and training data sets. The outputs obtained from all the regression trees are then averaged through ensemble averaging to produce the predicted output. The RD regression model utilises ensemble-based tree aggregation to improve robustness and minimise prediction variance during regression analysis.

Ultimately, the present study utilises sensor fusion to integrate data from multiple MIP sensor arrays. Then, it performs feature selection using a Greedy-based GA to select the optimal

feature from the input data. Then, hyperparameters are optimised through grid search to improve model performance. After tuning, a modified PSO algorithm leverages the strength of ACO; hence, the modification significantly enhances convergence speed and accuracy of the global optimum search. Following that, the PSRF regression effectively quantifies the content of three analytes, including TAN, THEO, and CAFF, in a black tea sample. It enhances model generalisation while minimising overfitting. Overall, the proposed PSRF model for regression addresses both the challenges of feature dimensionality and model generalisation, leading to superior predictive performance.

4.3. Results and Discussion

This subsection deliberates the findings of the proposed PSRF model by evaluating various performance metrics. These metrics serve as essential indicators, providing valuable insights into the accuracy and effectiveness of the model's predictions. In this section, we delve into the internal results produced by the proposed model, showcasing its predictive abilities in assessing black tea quality. Additionally, a comparative analysis is conducted to demonstrate the performance of the PSRF regression model in comparison to existing approaches. Overall, this evaluation section highlights the significance of the proposed model in assessing tea quality.

4.3.1. Performance Metrics

1. **Coefficient of determination (R^2)**

   The R^2 represents the percentage of variance in the dependent variable that the independent variables can explain. It indicates how well the model fits the data and is calculated using the formula (4.3),

   $$R^2 = 1 - \frac{\sum_{j=1}^{n}(P_j - Q_j)^2}{\sum_{j=1}^{n}(Q^- - Q_j)^2} \tag{4.3}$$

   Where P_j is the predicted j^{th} value and Q_j is the actual j^{th} Value.

2. **MSE**

   MSE indicates the variation between predicted and actual values and is calculated using equation (4.4),

   $$MSE = \frac{1}{n}\sum_{j=1}^{n}\left(P_j - Q_j\right)^2 \tag{4.4}$$

3. **MAE**

   The MAE is a performance measure that represents the average of all absolute errors. It is calculated using (4.5),

   $$MAE = \frac{1}{n}\sum_{j=1}^{n}\left(|P_j - Q_j|\right) \quad (4.5)$$

4. **RMSE**

   The RMSE is calculated as the square root of the average of the squared differences between the actual and predicted values, divided by the number of observations. n. It is derived by using the formula (4.6) below,

   $$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(P_j - Q_j\right)^2} \quad (4.6)$$

The proposed regression system was subjected to several statistical performance measures in order to ensure consistent prediction results. Feature reduction techniques alongside hyperparameter optimisation have been used in order to reduce the presence of redundancies in the regression process. Cross-validation techniques were also used in order to minimise any chances of overfitting. Despite the promising performance of correlation results, further testing using larger amounts of industrial data would be beneficial for the model.

4.3.2. Internal Results

The model's performance, as depicted by the performance metrics, is evaluated, and outcomes are projected to analyse the effectiveness of the proposed model. The regression outcome of black tea contents, such as theophylline, tannic acid, and caffeine, is procured from each sensor. The results obtained by individual sensors are shown in Tables 4.3, 4.4 and 4.5.

Table 4.3 illustrates the outcome of the tannic performance in measuring Actual TAN concentrations using HPLC as the standard method and comparing them with values predicted by PLSR. The outcome indicates a strong agreement between the HPLC and PLSR values for all five samples, demonstrating the sensor's precision. The prediction accuracy for each sample is consistent. Overall, the proposed model achieved an average prediction accuracy of 97.42%, demonstrating that the tannic sensor, combined with a PLSR model, provides dependable and precise TAN measurements, making it a valuable tool for applications that require the accurate quantification of black tea assessment.

Table 4.3 Results Obtained from Tannic Sensor

Sample No	Actual TAN (mg/ml) (HPLC)	Predicted TAN (mg/ml) PLSR	Prediction Accuracy (%)
1	0.0036	0.0037	97.3%
2	0.002	0.0024	98%
3	0.0016	0.0017	97%
4	0.0028	0.00029	97.8%
5	0.0028	0.0027	97%
Average Prediction Accuracy			**97.42%**

Table 4.4 Results Obtained from Theophylline Sensor

Sample No	Actual THEO (mg/ml) (HPLC)	Predicted THEO (mg/ml) PLSR	Prediction Accuracy (%)
1	0.0027	0.0028	97%
2	0.0011	0.0010	91%
3	0.0010	0.0009	90%
4	0.0004	0.00039	97.5%
5	0.0026	0.0021	81%
Average Prediction Accuracy			**91.3%**

Table 4.4 showcases the performance of a theophylline sensor in estimating theophylline content in samples, as compared to a reference method (HPLC). The PLSR model used in conjunction with the sensor demonstrated a promising average prediction accuracy of 91.3%. The findings showed the model achieved reliable accuracy in predicting THEO.

Table 4.5 describes the outcome of the caffeine sensor using PLSR, demonstrating a strong potential for accurately predicting the CAFF content from the black tea samples. The model yielded an accuracy of 95.16%, showcasing its promising performance compared to the HPLC method. Hence, indicating a reliable and cost-effective CAFF analysis, including black tea samples.

Table 4.5 Results Obtained from Caffeine Sensor

Sample No	Actual CAFF (mg/ml) (HPLC)	Predicted CAFF (mg/ml) PLSR	Prediction Accuracy (%)
1	0.0019	0.0020	99%
2	0.001	0.0012	98.3%
3	0.0011	0.0009	90%
4	0.002	0.00028	96.5%
5	0.0017	0.0020	92%
Average Prediction Accuracy			**95.16%**

Table 4.6 compares the R^2 values of the three individual sensors with the R^2 achieved when their data are fused. Typically, R^2 represents the proportion of variance in the dependent variable explained by the independent variable(s) in a regression model, ranging from 0 to 1, with values closer to 1 indicating a better model fit. Individually, the THEO, CAFF, and TAN sensors show relatively better R^2 values of 0.94, 0.93, and 0.75, respectively. However, the proposed (Fused) sensor system significantly improves the predictive power for all three, increasing the R^2 score of the TAN to 0.92, the THEO to 0.942, and the CAFF to 0.958. This

highlights the effectiveness of data fusion in enhancing the overall predictive accuracy and reliability of the sensor system, particularly for components that might be less accurate when measured individually.

Table 4.6 R^2 for Individual and Fused approach

Sensors	Individual R^2	Proposed (Fused) R^2
TAN	0.75	0.92
THEO	0.94	0.942
CAFF	0.93	0.958

Table 4.7 showcases the overall performance of the proposed modified PSRF model across various metrics. The model yielded 98.75% accuracy in predicting the black tea quality. R^2 value of 0.9945 indicates the perfect fitness of the model. Besides, the model achieved a low error value of 0.42%. These collective metrics prove that the proposed model is robust and efficient.

Table 4.7 Fused Sensor Results Procured by Proposed Regression Model

Model	Accuracy %	R^2	MAE	MSE	RMSE
Proposed Modified PSRF Model	98.75%	0.9945	0.042	0.042	0.042

Figure 4.4 illustrates the performance of the proposed PSRF for predicting black tea quality. The curve indicates a high actual positive rate (approaching 1) with a low false positive rate, demonstrating that the model effectively distinguishes between positive and negative cases with minimal errors. This curve highlights the robustness of the proposed model in predicting the quality of the black tea.

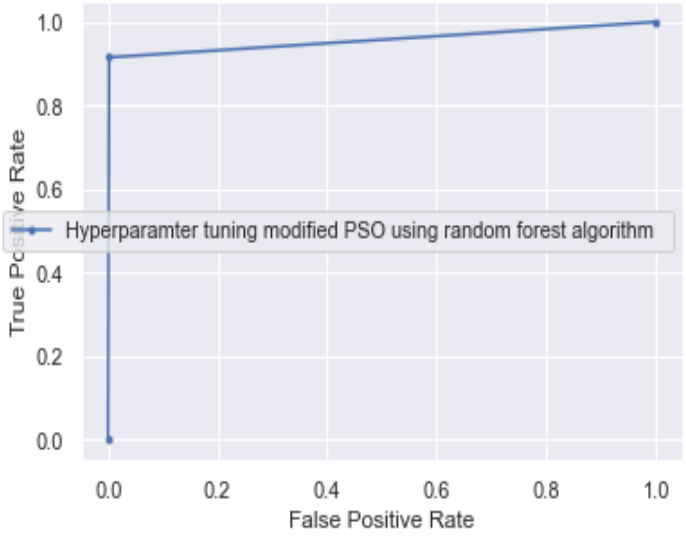


Figure 4.4 RROC Curve for Proposed Model

4.3.3. Comparative Analysis

This section evaluates the effective comparison of the proposed regression model with existing approaches. The prediction performance in a regression form is shown in Table 4.8 for the training and testing datasets, respectively. The corresponding graphical representation is depicted in Figure 4.5.

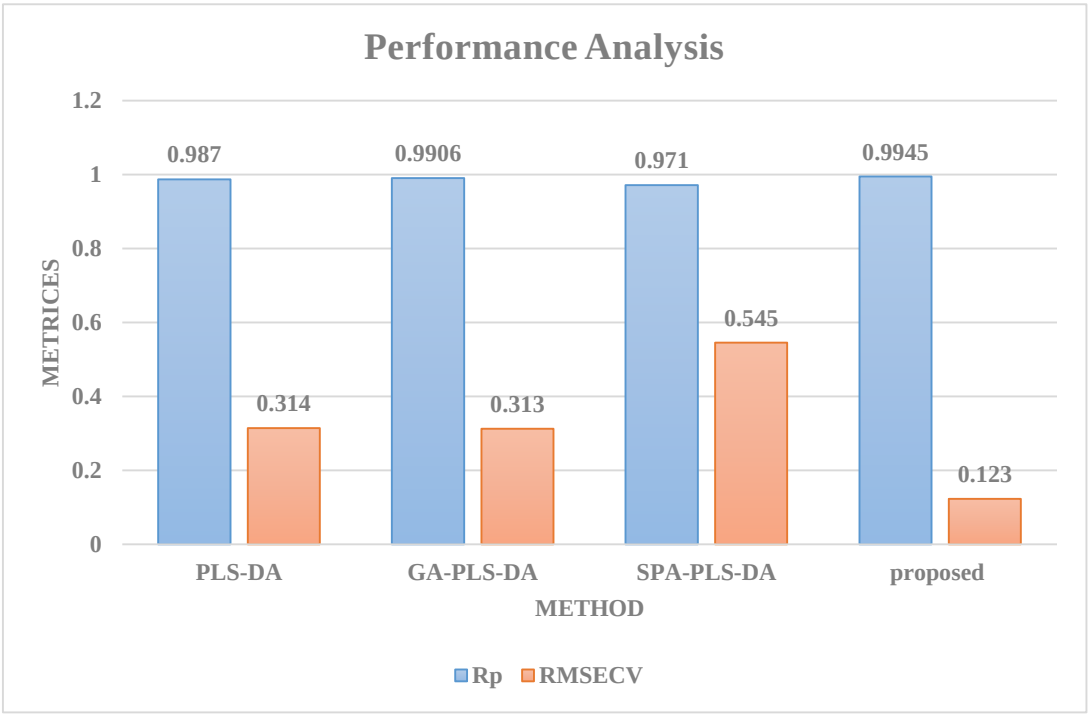


Figure 4.5 Graphical Representation of Outcome Obtained by Proposed and Existing Models [8]

Table 4.8 compares the performance of four different models: PLS-DA, GA-PLS-DA, SPA-PLS-DA, and the Proposed PSRF model based on Rp and RMSECV metrics. Higher Rp values, close to 1, indicate a better correlation between the predicted and actual values. In contrast, lower RMSECV values represent higher predictive accuracy. In the comparison analysis, the proposed PSRF model outperforms all other models significantly. The model exhibited an Rp of 0.9945 and the lowest RMSECV of 0.123. The findings showcase the effectiveness of the proposed PSRF model, demonstrating both strong prediction correlation and high accuracy in its cross-validation performance.

Table 4.8 Comparison of Proposed Model with Conventional Approach [8]

Models	Rp	RMSECV
PLS-DA	0.987	0.314
GA-PLS-DA	0.9906	0.313
SPA-PLS-DA	0.971	0.545
Proposed PSRF	0.9945	0.123

Figure 4.5 presents a performance analysis of various classification methods, including PLS-DA, GA-PLS-DA, SPA-PLS-DA, and a proposed method. The blue bars represent the correlation coefficient [11], where the proposed method achieved the highest value of 0.9945, indicating strong predictive performance. In contrast, the orange bars represent the RMSECV metric, indicating that the proposed method yields the lowest error at 0.123, which signifies better accuracy compared to the other methods. Overall, the proposed method demonstrates superior performance in terms of both correlation and error metrics.

Table 4.9 compares the performance of five different models, including SPA-PLSR, SPA-MLR, CARS-PLSR, CARS-MLR, and the Proposed PSRF model, using the Rp and RMSECV metrics. Overall, the proposed PSRF model showcased the highest Rp value and the lowest RMSECV score among all the existing models, demonstrating its superior prediction capability and accuracy. This comparison highlights the significant advantage of the proposed model in achieving highly accurate and reliable predictions.

Table 4.9 Quality Evaluation Results of Proposed and Existing Models [10]

Models	Rp	RMSECV
SPA-PLSR	0.9321	3.7257
SPA-MLR	0.9321	3.7257
CARS-PLSR	0.9509	3.13
CARS-MLR	0.8696	5.5053
Proposed	0.9945	0.123

Figure 4.6 illustrates the performance analysis of various classification methods, including SPA-PLSR, SPA-MLR, CARS-PLSR, CARS-MLR, and the proposed method. The blue bars represent the correlation coefficient [9], where the proposed method achieved the highest value of 0.9945, indicating strong predictive performance. In contrast, the orange bars represent the RMSECV metric, indicating that the proposed method yields the lowest error at 0.123, which signifies better accuracy compared to the other methods. Overall, the proposed method demonstrates superior performance in terms of both correlation and error metrics.

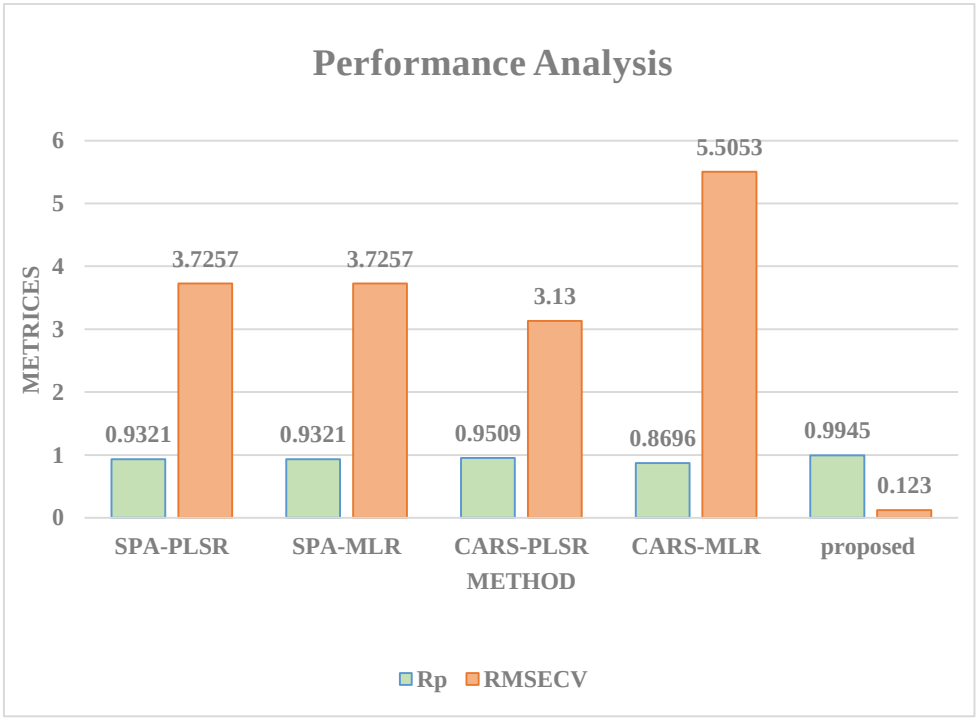


Figure 4.6 Quality Evaluation Results Attained by Proposed and Existing Models [10]

Table 4.10 compares the accuracy of the proposed model with the existing PLSR model. The existing PLSR model has yielded an accuracy of 92.17%, whereas the proposed PSRF model has achieved a significantly higher accuracy of 98.75%. This analysis emphasises the proficiency of the proposed model over the existing model in terms of accuracy in determining the black tea quality compounds.

Table 4.10 Results of Proposed and Existing in terms of Accuracy [11]

Models	Accuracy
Existing Method (PLSR)	92.17
Proposed Method	98.75

Figure 4.7 is an illustration of the proposed and existing models in terms of the accuracy metric. The highest orange bar revealed that the proposed model outperformed the existing model, demonstrating its remarkable efficiency.

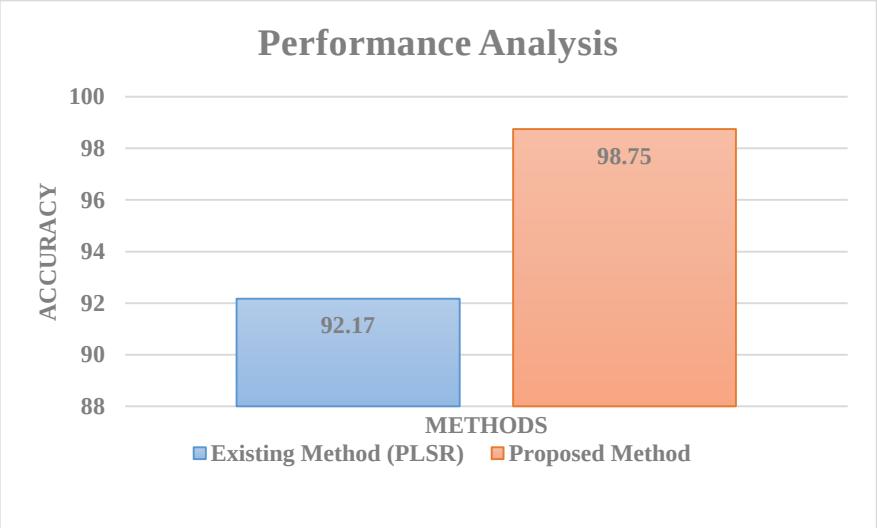


Figure 4.7 Graphical Representation of Results Achieved by Proposed and Existing Methods [11]

Table 4.11 analyses the proposed model's accuracy by comparing it with existing models, such as PLST and PCR. This comparison proves that the proposed PSRF model yielded remarkable accuracy, with a score of 98.75, while the existing PLSR achieved 92.94, and PCR achieved 93.75. Thus, the model showcases its significance in predicting the black tea quality over other models.

Table 4.11 Results of Proposed Model Compared with Conventional Methods [12]

Models	Accuracy
Existing Method (PLSR)	92.94
Existing Method (PCR)	93.75
Proposed Method	98.75

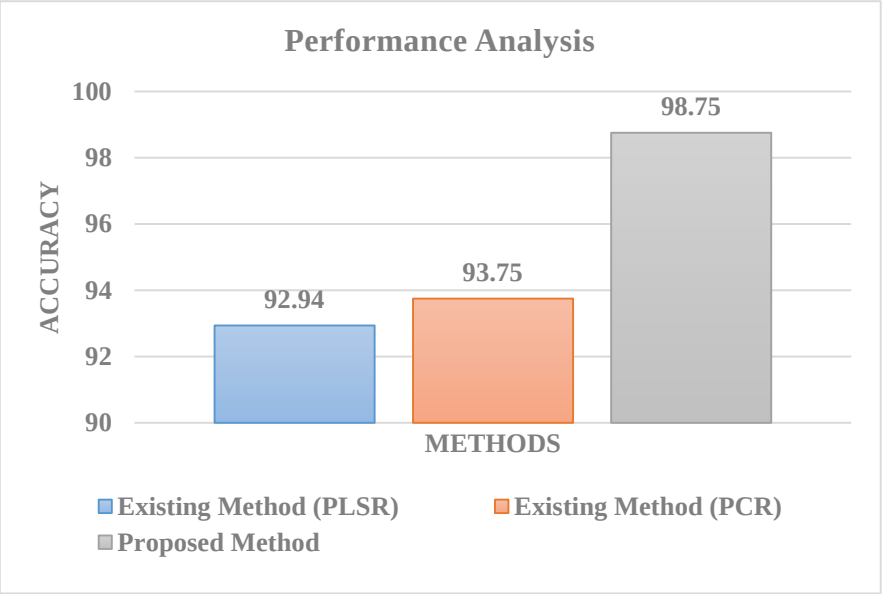


Figure 4.8 Graphical Representation of Results Achieved by Proposed and Conventional Methods [12]

Figure 4.8 illustrates the accuracy of the proposed model in comparison to existing PLSR and PCR models. According to the analysis, the proposed model demonstrated superior performance compared to the existing model, exhibiting remarkable efficiency. Minor variations observed across performance metrics indicate consistency in predictive behaviour of the proposed framework across different evaluation criteria.

4.4. Conclusion

Chapter 4 delves into the intricate details of the proposed PSRF model, specifically designed to accurately estimate the quality of black tea. The chapter begins by outlining the data collection process, which serves as a foundation for the current study. This comprehensive approach involved the use of three distinct MIP sensors, each meticulously engineered to target the prediction of key components in black tea: TAN, THEO, and CAFF. To enhance the analytical precision, the study employed feature fusion techniques, which facilitated the concatenation of data collected from these various sensors. This integrated dataset was then

refined through a feature selection method utilising a greedy-based GA, ensuring the selection of the most relevant features for optimal model performance. Moving forward, a modified PSO algorithm was synergistically integrated with the RF model, resulting in the development of the PSRF regression model. This innovative approach served as the bedrock for predicting the quality parameters of black tea, highlighting the effectiveness of combining DL techniques with advanced optimisation strategies. Furthermore, the chapter presents an in-depth discussion of the key findings derived from the proposed model. A comparative analysis was conducted to showcase the performance and robustness of the PSRF model in comparison to existing methodologies, demonstrating its superior capabilities in accurately assessing black tea quality. The findings not only reinforce the validity of the proposed framework but also emphasise its potential applications in enhancing quality control practices within the tea industry, ultimately benefiting producers and consumers alike.

REFERENCES

[1] M. Chowaniak *et al.*, "Quality assessment of wild and cultivated green tea from different regions of China," *Molecules*, vol. 26, no. 12, p. 3620, 2021, doi: https://doi.org/10.3390/molecules26123620.

[2] G. Ren, X. Zhang, R. Wu, L. Yin, W. Hu, and Z. Zhang, "Rapid characterization of black tea taste quality using miniature NIR spectroscopy and electronic tongue sensors," *Biosensors*, vol. 13, no. 1, p. 92, 2023, doi: https://doi.org/10.3390/bios13010092.

[3] X. Luo *et al.*, "Using surface-enhanced Raman spectroscopy combined with chemometrics for black tea quality assessment during its fermentation process," *Sensors and Actuators B: Chemical*, vol. 373, p. 132680, 2022, doi: https://doi.org/10.1016/j.snb.2022.132680.

[4] M. Moulick *et al.*, "Development of an Electronic Tongue with MIP sensors for black tea quality evaluation," in *2024 IEEE Calcutta Conference (CALCON)*, 2024: IEEE, pp. 1-4, doi: https://doi.org/10.1109/CALCON63337.2024.10914342.

[5] H. Ali, F. Ali, M. Akbar, K. Neha, N. Singh, and N. K. Sharma, "Capillary Electrophoresis," in *Analysis of Food Spices*: CRC Press, 2023, pp. 235-250, doi: https://doi.org/10.1201/9781003279792.

[6] S. M. T. Gharibzahedi, F. J. Barba, J. Zhou, M. Wang, and Z. Altintas, "Electronic sensor technologies in monitoring quality of tea: A review," *Biosensors*, vol. 12, no. 5, p. 356, 2022, doi: https://doi.org/10.3390/bios12050356.

[7] M. Ghahremani, H. Ghadiri, and M. Hamghalam, "Local features integration for content-based image retrieval based on color, texture, and shape," *Multimedia Tools and Applications*, vol. 80, no. 18, pp. 28245-28263, 2021, doi: https://doi.org/10.1007/s11042-021-10895-z.

[8] J. Huang *et al.*, "Qualitative discrimination of Chinese dianhong black tea grades based on a handheld spectroscopy system coupled with chemometrics," *Food Science & Nutrition*, vol. 8, no. 4, pp. 2015-2024, 2020, doi: https://doi.org/10.1002/fsn3.1489.

[9] P. Chatterjee, D. Mitra, T. Bhowmik, R. Nath, K. Dutta, and A. Deyasi, "Predicting Tea Quality Based on Antioxidant and Polyphenol Content Using Ensemble Machine Learning Model," in *2024 IEEE International Conference of Electron Devices Society Kolkata Chapter (EDKCON)*, 2024: IEEE, pp. 482-487, doi: https://doi.org/10.1109/EDKCON62339.2024.10870624.

[10] Y.-J. Wang, T.-H. Li, L.-Q. Li, J.-M. Ning, and Z.-Z. Zhang, "Evaluating taste-related attributes of black tea by micro-NIRS," *Journal of Food Engineering*, vol. 290, p. 110181, 2021, doi: https://doi.org/10.1016/j.jfoodeng.2020.110181.

[11] M. Moulick, D. Das, S. Nag, B. Tudu, R. Bandyopadhyay, and R. B. Roy, "Molecularly Imprinted Polymer-Based Electrode for Tannic Acid Detection in Black Tea," *IEEE Sensors Journal*, vol. 23, no. 6, pp. 5535-5542, 2023, doi: https://doi.org/10.1109/JSEN.2023.3240069.

[12] D. Das *et al.*, "Titanium oxide nanocubes embedded molecularly imprinted polymer-based electrode for selective detection of caffeine in green tea," *IEEE Sensors Journal*, vol. 20, no. 12, pp. 6240-6247, 2020, doi: https://doi.org/10.1109/JSEN.2020.2972773.

CHAPTER 5

TEA QUALITY EVALUATION FUSING MIP ELECTROCHEMICAL SENSOR AND NIR SPECTROSCOPY RESPONSES: REGRESSION ANALYSIS

This chapter outlines the detailed components of the final proposed model for predicting black tea quality based on tannic acid content by employing Near-Infrared Spectroscopy (NIR) Technique. The proposed model integrates Tannic acid extracted from both the NIR and the MIP sensor. Then, the feature selection is performed using a hybrid PSO-DFO (Particle Swarm Optimisation-Dragonfly Optimisation) algorithm. Then, data fusion is performed using two approaches: CCA and UMAP. After that, the proposed SNN-LSTM model is employed to forecast the tea quality. The model efficiency is estimated using an error metric and demonstrates a lower error rate with the CCA fusion approach compared to UMAP.

LIST OF SECTION

Contents of this chapter are based on following publication:

- **Modak, A.**, Moulick, M., Roy, R.B. (2025). Efficient Near Infrared Spectroscopy Based Feature Selection of Tannic Acid for Black Tea Evaluation. In: Singh, J.P., Singh, M.P., Singh, A.K., Mukhopadhyay, S., Mandal, J.K., Dutta, P. (eds) Computational Intelligence in Communications and Business Analytics. CICBA 2024. Communications in Computer and Information Science, vol 2367. Springer, Cham. doi: https://doi.org/10.1007/978-3-031-81339-9_13.

- **Modak A,** Moulick M, Das D, Roy RB. Effective tannic acid prediction in black tea using regression approach: Fusion of electrochemical and spectroscopic responses. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. 2026 Apr 22:127958. doi: https://doi.org/10.1016/j.saa.2026.127958.

CHAPTER 5

TEA QUALITY EVALUATION USING MIP ELECTROCHEMICAL SENSOR AND NIR SPECTROSCOPY: REGRESSION ANALYSIS

5.1. Introduction

Tea is a widely consumed beverage, and the quality of tea is defined by its aroma, flavour, and colour, which are influenced by its chemical components [1]. For example, green tea's astringency comes from compounds like EGCG and other catechins [2]. While black tea derives its flavour and colour from theaflavins and thearubigins, formed through oxidation, other molecules, such as tannic acid, increase astringency, and L-theanine contributes an umami flavour [3]. To overcome the limitations of subjective sensory evaluation, advanced tools such as NIR spectroscopy-based sensors widely utilized [4, 5]. NIR provides a fast, non-invasive way to measure chemical composition, including polyphenol content [6]. These methods are also crucial for optimizing tea processing and cultivation.

This chapter presents an advanced methodology for tea quality estimation based on accurately predicting tannic acid content in black tea samples using NIR spectroscopy [7, 8]. The study focuses on addressing the challenges of traditional single-source analysis methods by integrating multiple data modalities to enhance prediction accuracy and robustness. Hence, the spectral data from NIR spectroscopy and the electrochemical responses from an MIP sensor are processed and analysed using a comprehensive pipeline to achieve enhanced prediction accuracy [9, 10]. Furthermore, the data pre-processing includes initial calibration and normalisation steps for each dataset, ensuring consistency. Consequently, a feature selection method using a hybrid PSO-DFO (Particle Swarm Optimisation-Dragonfly Optimisation) algorithm is employed to identify the impactful features from high-dimensional datasets. The selection process is crucial in reducing model complexity and preventing overfitting by enhancing predictive performance.

The selected features undergo extraction and transformation using methods such as CCA (Canonical Correlation Analysis) and UMAP (Uniform Manifold Approximation and Projection), which enable efficient data fusion by capturing the complex relationships between the data sources. Then, the fused dataset is divided into training and testing subsets to assess the performance of the proposed regression model, which combines an SNN with LSTM networks. Moreover, the hybrid SNN-LSTM model is specifically designed to capture

temporal dependencies and similarity-based relationships in the data, resulting in enhanced prediction accuracy of tannic acid concentrations.

5.2. Study of Feature Selection for Black Tea Quality Evaluation on Tannic Acid Molecule Content using Near Infrared Spectroscopy (NIR) Technique

This section discusses a study that implements a technique for selecting features to evaluate the quality of black tea based on tannic molecule content.

5.2.1. Experimentation

To mitigate the challenges associated with traditional studies, such as errors and complexity, the proposed model incorporates the taste of black tea through the integration of NIR spectroscopy and feature selection techniques, enabling precise differentiation of the four black tea qualities. However, the NIR is implemented to acquire spectral response from the input tea attributes. During the data pre-processing, the standard scalar technique is used to regularise the data points. Then, the processed data is loaded for the FS process, where TAN is utilised for training the data reduction in the models. Moreover, four black tea samples are differentiated through data reduction techniques. Besides, the proposed model selects features such as PCA, UMAP, and KPCA methods. These techniques provided an optimally visualised data pattern by gathering similar data points based on relationships. The dimensionality and difficulties are reduced in the system. Due to the flexibility and learning capability, data reduction is implemented to achieve effective results by separating the tea samples. Through this analysis, the features section is performed using the Silhouette, Calinski-Harabasz, and Davies-Bouldin methods for PCA, UMAP, and KPCA [11].

5.2.1.1. Near Infrared Spectroscopy (NIR) Technique

NIR is a widely used non-destructive analytical method that can provide multicomponent analysis of almost any matrix and determine the chemical and physical properties of a wide variety of materials. Additionally, a significant advantage of NIR spectral techniques is that they do not require sample preparation. It is an objective method for assessing the quality of black tea by measuring how its chemical components interact with near-infrared light. By analysing the absorption patterns of light, the technique provides a unique spectral fingerprint of the tea's molecular composition, indicating levels of key quality markers, such as polyphenols and others. This raw spectral data is then interpreted using chemometric models, which are pre-calibrated against known reference methods, to accurately predict the tea's composition and sensory attributes. This allows producers to monitor and control quality,

optimise processing steps like fermentation, and perform geographical authentication more consistently than traditional subjective sensory evaluation.

Recently, there has been a focus on examining both solid and liquid chemical formulas to ensure product quality and monitor manufacturing processes. In material inspection, samples are scanned in their current state, and we use pattern recognition algorithms to verify their identity and quality. NIR combined with statistical regression methods, acts as a real-time testing and measurement tool. It provides timely chemical data that's crucial for managing chemical production processes, recovering solvents, and overseeing operations such as drying, blending, and extrusion.

5.2.1.2. Data Set Collection

Tea samples are sourced from a multitude of origins. The quality of these teas is influenced by factors such as the manufacturing process, the season of flush, the types of plantations, and the prevailing agro-climatic conditions. Companies in the tea industry submit their products to specialised centres for quality assessment through tasting. The proposed work will be conducted in absorbance mode at room temperature and within a controlled, noise-free environment.

5.2.1.2.1. Acquisition of Spectral Data

In the present work, the DWARF-Star NIR spectrometer is utilised and purchased within the wavelength range of 900nm - 1700nm. In the absorbance mode, the experiment is carried out. The experiment utilises a 12V tungsten halogen lamp and employs samples of black tea in a crystal sample holder. The holder is positioned above the light source. Almost 10 repetitions of every tea sample have been completed, and each spectral data set contains 502 data points within the wavelength window. A 40x502 data matrix was used during the experiment. The crystal sample holder is cleaned with ethyl alcohol before repeating the data sample analysis, and then wiped clean with a tissue. Following the generation of the data points, pre-processing of the data is utilised by executing a standard scalar method. The approach comprises normalisation of samples. This process normalises the features by subtracting the average value of the attributes and then dividing by the feature standard deviation. Henceforth, the high TAN feature concentration black tea samples are picked and fed into data reduction to detect four unique samples.

5.2.2. Pre-Processing

In this pre-processing method, a scalar technique is used to standardise features by transforming them to achieve a mean of 0 and a standard deviation of 1. It is called Z-score normalisation (standardisation). For every feature within the dataset, the standard scalar evaluates the mean (μ) and the standard deviation (σ), and the formula standardises it.

$$Z = (x - \mu)/\sigma \tag{3.9}$$

Moreover, it confirms that various scales have contributed equally to the ML model. In contrast, features with large values may improperly influence the model's learning when there is no standardisation.

5.2.3. Feature Selection

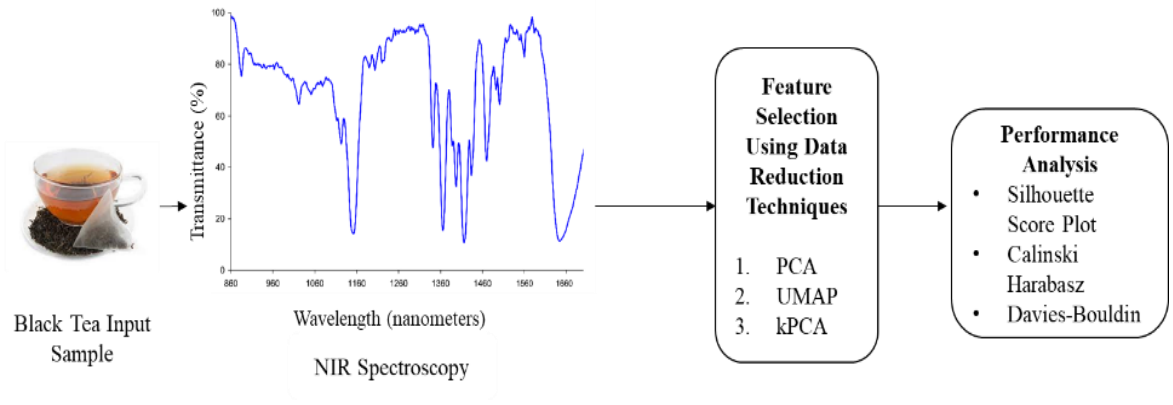


Figure 5.1 Feature Selection using NIR

Figure 5.1 illustrates an NIR analysis procedure for a black tea sample, showing all the stages from sample collection to data analysis. It begins with the input sample of black tea, visually represented as brewed tea and a tea bag, and serves as the sample material to be investigated. The left-hand plot displays the spectrum obtained by employing NIR spectroscopy, with the x-axis representing wavenumber (nm) and the y-axis representing absorbance, which indicates the light absorbed at a specific wavelength. The principle of operation of NIR spectroscopy is to excite the sample with near-infrared radiation and detect the reflected or transmitted radiation, with every wavelength absorbed corresponding to molecular vibrations, such as O-H, N-H, and C-H bonds. This provides information on the chemical composition of the tea, including the presence of compounds like caffeine and polyphenols. On the right side, reduction techniques such as PCA, UMAP, and kPCA (Kernel PCA) are presented, emphasising their role in choosing relevant features from the intricate spectral data to facilitate analysis. The performance analysis segment describes measures applied to cluster or classify

models derived from the extracted features, such as the Silhouette Score Plot, Calinski-Harabasz Index, and Davies-Bouldin Index, all of which confirm the quality of the clustering result from the NIR data analysis. Overall, the process is a general analytical method to provide helpful information on the quality and composition of the black tea sample from complex spectral information.

5.2.3.1. Principal Component Analysis (PCA)

PCA is widely employed for statistical modelling and for elucidating patterns of variation within high-dimensional datasets. This technique serves as a dimension reduction method that involves an orthogonal transformation of the values associated with linearly uncorrelated variables, referred to as Principal Components (PCs). Notably, the number of PCs is either the same as or lower than that of the initial variables, with the first principal component (PC1) accounting for the maximum variance.

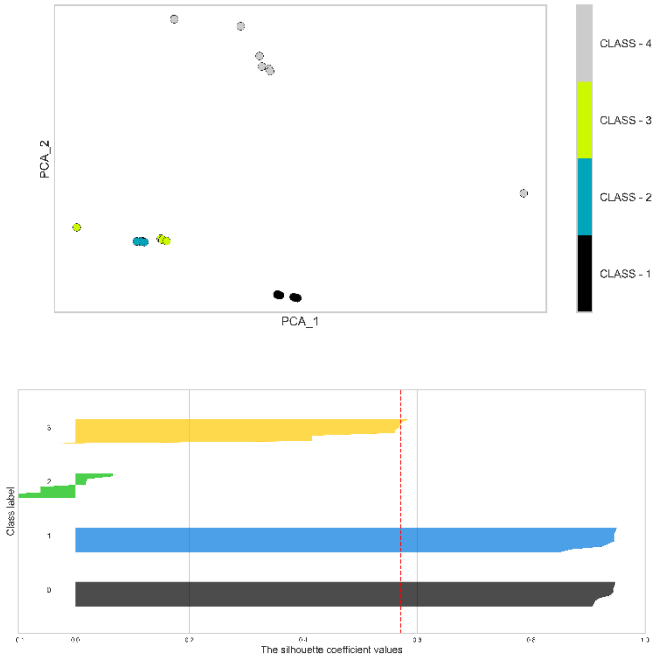


Figure 5.2 Score Plot of PCA and Silhouette Coefficient Chart

In contrast, each succeeding component (PC2) is designed to capture the most significant variance possible while remaining orthogonal to the preceding components. Thus, these two components are utilised for visualising global trends in the data in a scatterplot. Furthermore, PCA is employed to analyse the response vectors obtained from the clustered input data sensor array within a multi-sensor framework. Consequently, the principal components (PCs) are orthogonal and organised such that PC1 accounts for the maximum variance present in the original data. Furthermore, PCA can predict the relationships between variables that remain

ambiguous in the unprocessed input data. Therefore, through the process of dimensionality reduction, PCA effectively extracts underlying structures, facilitating the transformation and understanding of the data.

Figure 5.2 indicates that classes 1 and 2 show high clustering and density, while classes 3 and 4 are more dispersed, suggesting weaker cohesion in the latter. A Silhouette plot assesses cluster separation, with the X-axis representing the Silhouette coefficient and the Y-axis representing the classes. The plot reveals that classes 1 and 2 are well-clustered, but classes 3 and 4 exhibit some negative Silhouette scores, indicating instances of misclassification, which are more prevalent in class 4. Class 3 maintains a slightly decreased positive Silhouette value, remaining below 0.1.

5.2.3.2. Kernel Principal Component Analysis (KPCA)

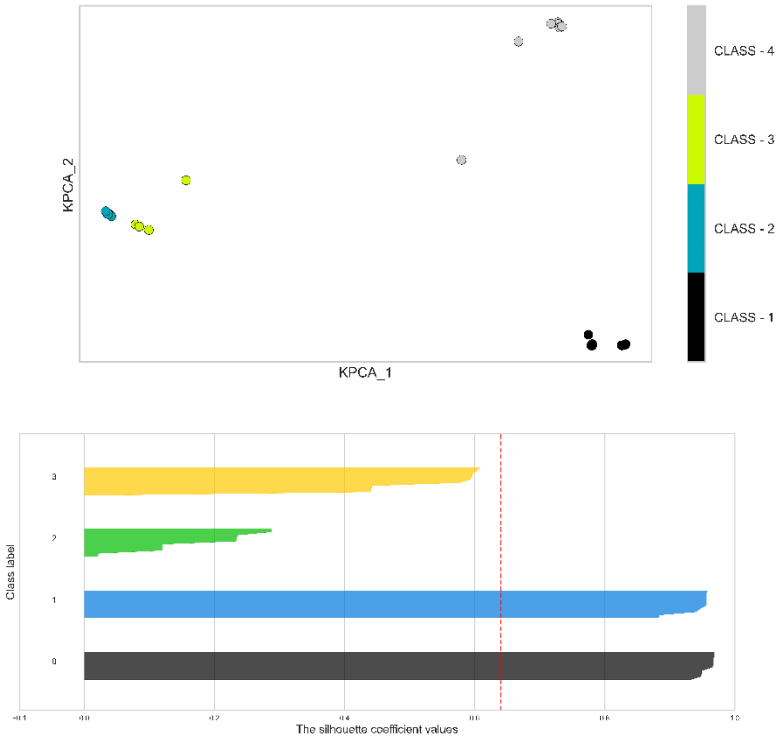


Figure 5.3 Score Plot of KPCA and Silhouette Coefficient Chart

KPCA serves as a generalisation of PCA by mapping datasets into a high-dimensional feature space through the utilisation of a kernel function. This technique is particularly effective for reducing dimensions in cases of nonlinearity, as it incorporates a nonlinear mapping function before executing PCA. This approach enables KPCA to uncover non-linear and higher-order relationships present within the input data. The principal components derived from the transformed data undergo further analysis and are utilised for data visualisation purposes.

Consequently, KPCA facilitates the projection of data into a lower-dimensional space while preserving significant information.

Figure 5.3 illustrates that the KPCA algorithm effectively clusters classes 1 and 2, whereas class 3 is moderately scattered, and class 4 has a minimal score. Besides, the Silhouette coefficient chart for KPCA shows no negative scores. This indicates that the KPCA has performed well.

5.2.3.3. Uniform Manifold Approximation and Projection (UMAP)

UMAP is recognised as a technique for dimensionality reduction, specifically designed to visualise clustering patterns within high-dimensional datasets. Its primary emphasis is on local clustering, wherein comparable observations are aggregated, while also striving to maintain the overall structure of the data. The UMAP algorithm is based on graph layout principles, constructing a high-dimensional graph framework. Initially, UMAP constructs a high-dimensional graph known as a "fuzzy simplicial complex." This graph is characterised by weights on its edges, which reflect the probability of a connection between two data points.

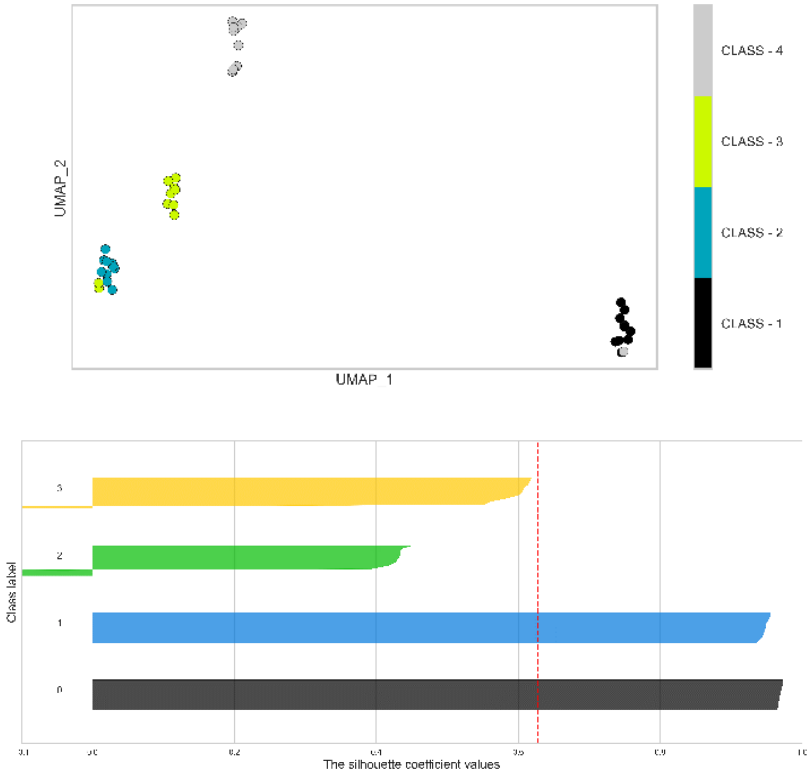


Figure 5.4 UMAP of Score Plot and Silhouette Coefficient Chart

The concept of connectedness is established as UMAP extends a radius from each point. When these radii overlap, the points are deemed interconnected. The choice of radius is crucial; a

smaller radius tends to create isolated and minor clusters, while a larger radius may result in a complete merging of points. Therefore, these complications can be mitigated by selecting the radius based on the local distances between neighbouring points. Furthermore, as the radius increases, UMAP progressively diminishes the likelihood of connections, creating a "fuzzy" graph. Consequently, every point should be associated with at least its nearest neighbour, and UMAP ensures the preservation of local structures while also considering the overarching global structure.

The UMAP technique organises similar classes within high-dimensional spaces, and as shown in Figure 5.4, it is clear that class 3 has been incorrectly classified as class 2. Additionally, there is also an instance of misclassification where class 4 is identified as class 1. It can be observed that the overall Silhouette score surpasses that of PCA. Notably, the Silhouette score for class 3 approaches 0.4, indicating a well-defined classification. Furthermore, the Silhouette score for class 4 exceeds 0.6, a significant improvement over the results obtained from PCA.

5.2.4. Results and Discussion - Performance Analysis

In this context, three coefficients —Silhouette, Calinski-Harabasz, and Davies-Bouldin —are employed to evaluate the outcome of the proposed model. The analysis of the tea samples is conducted using data reduction techniques, and the resulting findings are compiled and presented in Table 5.1.

Table 3.7 shows the Silhouette coefficients and other clustering evaluation metrics for three data reduction techniques: PCA, UMAP, and KPCA. The Silhouette coefficient indicates how well-separated clusters are, with higher values suggesting better cluster cohesion. The Calinski-Harabasz index measures the relationship between the variance within clusters and the variance among clusters, with elevated values indicating more distinct clusters. In contrast, the Davies-Bouldin index evaluates the degree of separation between clusters, where reduced values suggest superior clustering performance. The results show that the UMAP approach obtained Silhouette: 0.627, Calinski-Harabasz: 85.191, and Davies-Bouldin: 0.846, outperforming PCA (Silhouette: 0.57, Calinski-Harabasz: 29.973, Davies-Bouldin: 2.85) in terms of reduced misclassification of data samples. KPCA demonstrates further improvements with coefficients of 0.64, 85.305, and 0.77, respectively. Therefore, KPCA emerges as the most suitable method for predicting black tea quality, as it obtained a superior Silhouette score and reduced misclassification.

Table 5.1 Performance of Proposed Model

Model	Hyper parameter	Silhouette	Davies-Bouldin	Calinski-Harabasz
PCA	Number of components=3	0.57	2.85	**29.973**
KPCA	No of components=3, kernel='rbf', fit_inverse_transform=True	0.64	0.77	**85.305**
UMAP	**No of components=3, no. of neighbors=20, minimum distance=0.1, metric='euclidean'**	**0.627**	**0.846**	**85.191**

5.3. Effective Tannic Acid Prediction in Black Tea Samples using Regression Approach Fusing the Responses Obtained from MIP Electrochemical Sensor and NIR Spectroscopy

The integration of NIR spectroscopy and advanced data processing techniques offers a robust approach for enhancing the accuracy of tannic acid concentration predictions in black tea, supporting improved quality assessment. The feature selection and regression methodologies implemented demonstrate the significance of advanced analytical frameworks in ensuring high-quality tea production and assessment. These results indicate the model's potential in efficiently predicting key quality parameters, utilizing both NIR spectral data and electrochemical sensor data to optimize tea processing and quality control. The fusion of electrochemical sensing and NIR via learning-based framework is a notable improvement towards robust prediction of tea quality.

5.3.1. Proposed Methodology

The proposed model primarily emphasises predicting the presence of tannic acid in black tea samples. Additionally, the proposed work integrates various techniques to achieve better results. The overall flow of the proposed model is shown in the Figure. 5.5.

Figure 5.1 depicts the overall flow of data processing and analysis using MIP and NIR datasets for the proposed model. Initially, the two input datasets, such as TANNIC MIP and TANNIC NIR, are fed into the data pre-processing step, where the MIP data is processed based on current values and NIR data is analysed for absorption features. Further, the feature selection phase is

done using the PSO-DFO algorithm to identify the most relevant features. The above chosen features enter the feature extraction stage, which comprises an input layer, max pooling, activation functions, and an FCN (Fully Connected Network) for refining the data. As a continuation, the resultant features are exposed for fusion using CCA (Canonical Correlation Analysis) and UMAP algorithms for combining the data efficiently. The data was partitioned using a random hold-out strategy, allocating 80% of the samples for model training and the remaining 20% for testing. This random partitioning process was executed over multiple independent iterations to prevent sampling bias and confirm the consistency of the findings. Accordingly, the final evaluation metrics are presented as the mean value paired with its standard deviation. For regression analysis, an SNN-LSTM algorithm is used to model the data and predict the results. Finally, performance metrics are measured to assess the model's efficiency, assuring the robustness and accuracy of the overall analysis.

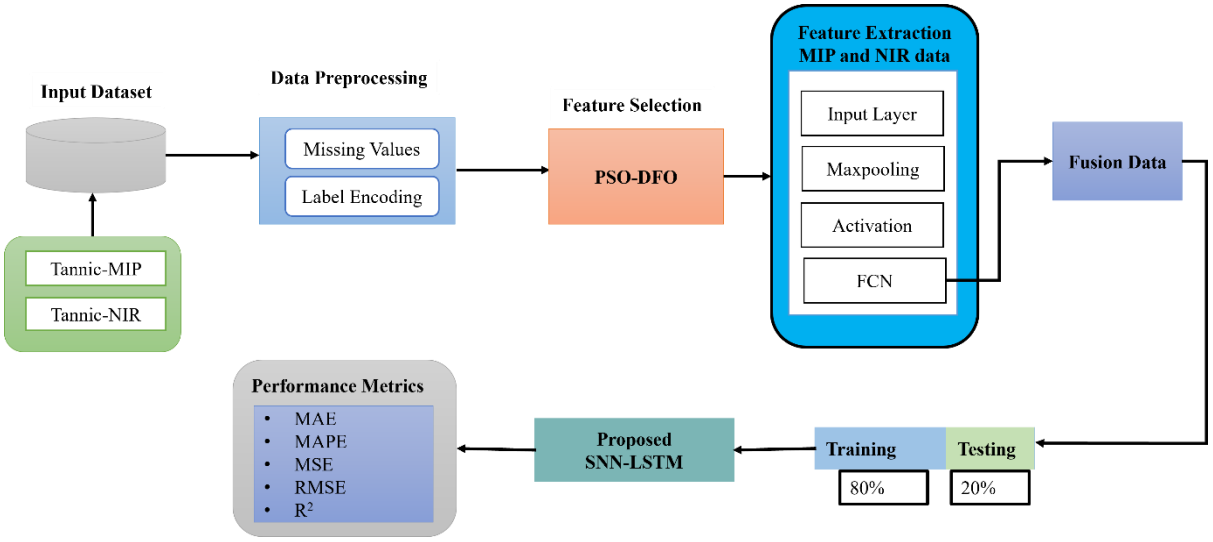


Figure 5.5 Proposed Flow

5.3.2. Data Collection and Pre-Processing

The following section outlines the experimental procedures and details of the dataset, including TANNIC MIP and NIR data gathered from four black tea samples harvested in north-east India. The dataset includes two sets of data: 328 data points for TANNIC MIP and 502 data points for NIR, with 10 repetitions per sample. This results in two data matrices: one of size 40 x 328 for TANNIC MIP and another of size 40 x 502 for NIR, each representing the spectral readings across different wavelengths.

5.3.2.1. Tannic acid detection using MIP sensor

TAN has been collected from Nice Chemicals Pvt. Ltd and the other reagents, such as AN, AA, MMA, EGDMA, and benzoyl peroxide, are sourced from the USA. Additional chemicals, including Al, Mg, paraffin oil, caffeine, and dopamine, are purchased from Merck, India, and are of analytical grade, employed without purification. Moreover, the solutions are prepared using Millipore water and all the experiments are conducted at room temperature. Furthermore, a potentiostat/galvanostat with a three-electrode setup is used, with platinum serving as the counter electrode, Ag/AgCl as the reference electrode, and MIP_TAN as the working electrode for the electrochemical analysis process. Specifically, UV-visible spectroscopy (250–450 nm) and scanning electron microscopy (SEM) with a ZEISS EVO 18 at 15 kV are employed for spectral analysis and surface morphology characterisation. Subsequently, the MIP synthesis involved the ultrasonication of 0.95g of graphite in ethanol, along with TAN, AA, and the target analyte, as well as EGDMA and benzoyl peroxide for polymerisation under constant heating. After the process of washing and eliminating the TAN, the MIP matrix is formed, dried, and utilised to create the MIP_TAN electrode, which is then mixed with paraffin oil and packed into a capillary tube.

5.3.2.2.Tannic acid detection using NIR spectroscopy

For the process, four black tea samples are obtained, and experiments are conducted in absorbance mode in a noise-free environment to get accurate outcomes. To maintain reliability, the samples are sealed in pouches immediately after collection. Additionally, the spectral data are acquired using a DWARF-Star NIR spectrometer, with a wavelength range of 900–1700 nm. The setup consists of a sample holder, an NIR detector, and a 12V tungsten halogen lamp serving as the light source. Finally, the tea samples are placed in a crystal sample holder, positioned directly above the light source, to ensure optimal interaction with the incident light during spectral measurement.

5.3.3. Feature Selection using Proposed PSO-DFO Algorithm

The proposed model uses a feature selection method to identify the relevant features for forecasting the quality parameters in black tea. With the selection of impactful features, the methods help reduce model overfitting, resulting in accurate and consistent predictions. Here, the molecular features are selected using a hybrid PSO-DFO algorithm. Moreover, the hybrid model combines the strengths of both algorithms, enhancing the selection process to increase

the predictive performance and robustness of the model. Figure 5.6 shows the entire process of the proposed model.

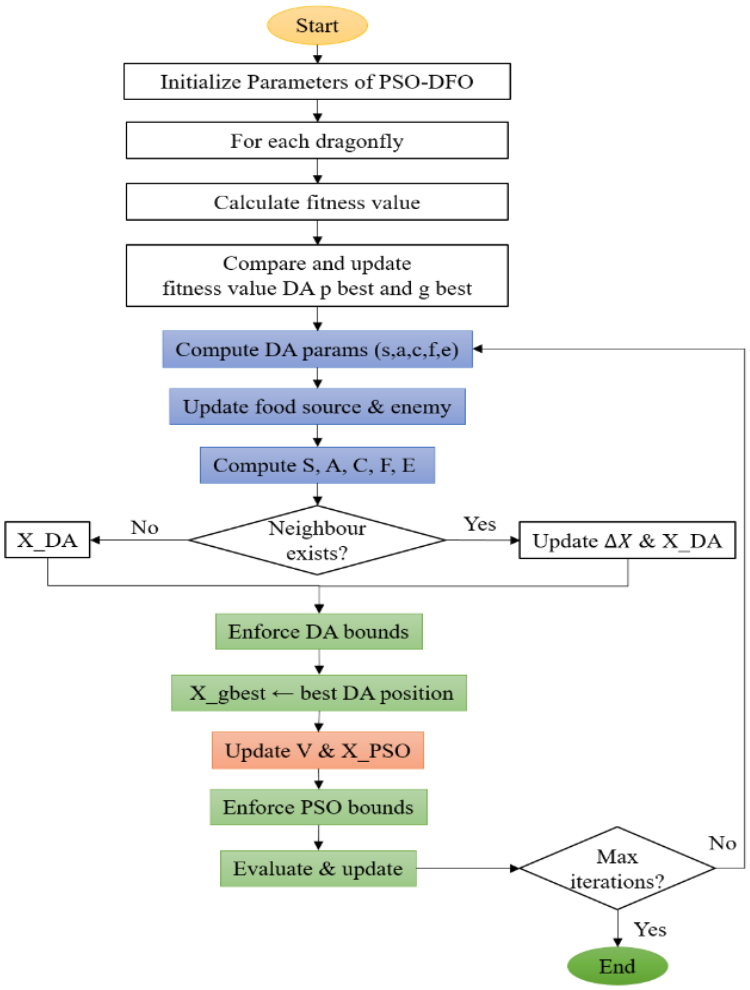


Figure 5.6 Flowchart of the Proposed PSO-DFO Approach

Figure 5.6 illustrates a flowchart of the hybrid algorithm, combining Particle Swarm Optimisation (PSO) and the Dragonfly Algorithm (DA), known as PSO-DFO. The process is initiated with the algorithm's parameters. For the Dragonfly Algorithm, the process is initiated by measuring the fitness value for each dragonfly. Then, it compares and updates the best fitness values, known as DA $p\ best$ considered as the best position of a single dragonfly and $g\ best$ which is considered as the best position among all dragonflies. The algorithm continues by examining if the stop criterion is met; otherwise, the food source and enemy are updated, along with the neighbouring radius and velocity position of the dragonflies. The process replicates until the stop criterion is reached. Simultaneously, the PSO is initiated by initialising the particles and calculating the fitness values for each particle. The fitness value of the particle is compared, and its personal best position ($p\ best$) value is updated. The global best position

($g\ best$) is further selected along with the velocity and position of the particles. Once the optimal solution is observed, the algorithm is completed.

Among numerous feature selection techniques, the study employs the DFO and PSO algorithms due to their unique strengths. Furthermore, the DFO is chosen for its adaptability and rapid convergence, which makes it efficient in routing intricate and high-dimensional search spaces. On the other hand, PSO is selected for efficacy in solving high-dimensional optimisation problems and scalability, which is predominantly useful for handling larger datasets. Although both algorithms are dominant individually, they pose particular challenges in predicting tea quality. PSO, for example, suffers from limited exploration abilities, which are highly sensitive to parameter settings and susceptible to getting trapped in local optima. To tackle the above challenges, the hybrid PSO-DFO algorithm is used, assuring enhanced exploration and exploitation during feature selection, ultimately improving the accuracy and robustness of the tea quality prediction model. Table 5.2 presents the parameters employed in the PSO-DFO algorithm.

Table 5.2 PSO-DFO Parameters

Model	Parameter	Values
PSO_PARAMS	'NUM_PARTICLES':	10
	'INERTIA_WEIGHT':	0.7
	'PERSONAL_LEARNING_RATE':	2.0
	'GLOBAL_LEARNING_RATE':	2.0
	'NUM EPOCHS':	20
DFO_PARAMS	'NUM_NEIGHBORS'	2
	'STEP_SIZE'	0.1

Table 5.2 illustrates the parameters employed in the PSO-DFO algorithm for the proposed model. For the PSO component, the model is configured with 10 particles, an inertia weight of 0.7, and personal and global learning rates of 2.0. The hybrid algorithm runs for 20 epochs to iteratively refine the feature selection process. For the DFO component, the model is set with two neighbours, using a step size of 0.1 for movement in the search space. The above parameter locations are designed to strike a balance between exploration and exploitation, thereby enhancing the overall performance of the hybrid PSO-DFO algorithm.

The optimisation process is initialised with the algorithm parameters, in which the fitness value is calculated to evaluate the accuracy of the prediction for tea samples. The minimum fitness value is further estimated, and the fitness of each dragonfly is updated based on the comparison. In the hybrid PSO-DFO algorithm, internal memory plays a significant role in tracking accurate solutions. Each dragonfly is guided by the fitness values, which the PSO tracks as the personal and global bests. Hence, the above strategy enhances the convergence for the global optimum by refining the solutions over iterations. The best solutions are saved in the DA, where the exploitation ability of PSO enhances the performance of DFO.

The process iterates till the convergence criterion is attained. In the PSO algorithm, each particle represents a candidate solution, with its position and velocity updated based on the local and global fitness bests. Additionally, the inertia weight parameter influences the impact of prevailing velocities, which helps the particles converge towards the optimal solution. Eventually, the hybrid model detects the optimal set of features using iterative updates and fitness evaluations, as represented in Equation 5.1, which is utilised to select the relevant features for optimisation.

$$Sep_i = -\sum_{j=1}^{Neg} X - X_j \tag{5.1}$$

Where, X denotes the location of the present individual, Neg is signified as the number of individual neighbours of the dragonfly swarm, X_j denotes the position of the j^{th} neighbouring dragonfly and SEP indicates the separation motion for i^{th} Individual. In addition, equation 5.2 is employed for estimating the alignment.

$$\mathrm{Alg_i} = \frac{\sum_{j=1}^{\mathrm{Neg}} \mathrm{V_j}}{\mathrm{Neg}} \tag{5.2}$$

Where, V signifies the velocity of the j^{th} neighbouring and A_i denotes the alignment motion for the i^{th} Individual. Further, the cohesion is shown in equation 5.3,

$$\mathrm{C_i} = \frac{\sum_{j=1}^{\mathrm{Neg}} \mathrm{X_j}}{\mathrm{Neg}} - \mathrm{X} \tag{5.3}$$

Where X specifies the position of the current dragonfly individual, Sam_i denotes the cohesion for the i^{th} Individual. Next, the attraction motion towards the food is measured using equation 5.4.

$$\mathrm{F_i} = \mathrm{X^+} - \mathrm{X} \tag{5.4}$$

Where, F_i signifies the food attraction of i^{th} dragonfly and X^+ Evaluates the location of the source food. Consequently, the food with DF contains the superlative objective function to the point. Further, the distraction caused by outward predators is calculated using equation 5.5.

$$E_i = X^- + X \tag{5.5}$$

Where, X^- indicates the enemy position and E_i signifies the distraction motion of the enemy, whereas X Denotes the current dragonfly individual. In the next step, two vectors are used to update the position of the search spaces, involving the position vector. X and ΔX. Moreover, the step vector is measured as the correspondence of the velocity vector in PSO. Therefore, equation 5.6 shows the step vector.

$$\Delta X_{t+1} = (sSep_i + aAlg_i + cC_i + fF_i + eE_i) + w\Delta X_i \tag{5.6}$$

Where A_i defines the alignment, F_i denotes food source, S_i represents the separation of i^{th} dragonfly, t defines the iteration counter and w depicts the inertia weight. After the completion of the estimation process of the step vector, the position vector is computed using Equation 5.7,

$$X_{(t+1)} = X_t + \Delta X_{t+1} \tag{5.7}$$

Where t Denotes the recent iteration. Equation 5.8 is utilised for updating the position of all the artificial dragonflies.

$$T(\Delta X) = \left|\frac{\Delta X}{\sqrt{\Delta X^2+1}}\right| \tag{5.8}$$

It is employed to calculate the fluctuating probability of the position of DF. Additionally, equation 5.9 is used to update the position with the search agent's current position.

$$X_{(t+1)} = \begin{cases} X_t, & r < T(\Delta X_{t+1}), \\ X_t, & r \geq T(\Delta X_{t+1}), \end{cases} \tag{5.9}$$

Where r is indicated as a number within the range [0, 1]. All the artificial dragonflies are present in one swarm; hence, the PSO-DFO algorithm attempts to exploit and explore by tuning the factors of the swarm, namely. $(s, a, c, f \text{ and } e)$ along with w.

$$P_i = \frac{c}{Neg_i} \tag{5.10}$$

Where, c Denotes the constant number that contains a better value than one and P_i indicates the count of $pareto$ Ideal solution. Equation 5.10 indicates a high probability, which aids in

selecting food from less populated segments. Furthermore, a roulette wheel mechanism is employed in the selection process, as shown in Equation 5.11.

$$P_i = \frac{Neg_i}{c} \tag{5.11}$$

Where Neg_i represents the number of $pareto$ solution and c denotes the constant number. Ultimately, a global ideal solution is attained through the exploration parameters of DF in the initial phase, where the exploitation features of PSO are combined in the final stage. Equation 5.12 shows the position and velocity of PSO.

$$V_{k+1}^i = wV_k^i + C_1 r_1 (DA - pbest_k^i - X_k^i) + C_2 r_2 (DA - gbest_k^g - X_k^i) \tag{5.12}$$

$$X_{k+1}^i = X_k^i + V_{k+1}^i \tag{5.13}$$

Equations 5.12 and 5.13 show the position and velocity update rules in PSO. In equation 5.12, the velocity V_{k+1}^i of particle i at iteration $k+1$ is influenced by the prevailing velocity V_k^i cognitive attraction to one's own best position $pbest_k^i$ and social attraction to the global best position $gbest_k^g$, with the random factors r_1 and r_2. In equation 5.13, the position X_{k+1}^i of particle i is updated by including the additional velocity V_{k+1}^i, which moves the particle towards the optimal solution.

The objective function of the proposed PSO-DFO algorithm is defined as the minimisation of the Mean Squared Error (MSE) between anticipated and actual tannic acid values. The fitness function is depicted in equation 5.14,

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{5.14}$$

Where Y_i represents the actual value, $\hat{Y}_i$ denotes the predicted value, and n is the total number of samples. Lower MSE values indicate better feature subset selection.

Table 5.3 Feature Size Before and After Feature Selection

Algorithm Name	MIP	NIR
Original	328	502
Dragonfly Algorithm Tuned By PSO	160	248

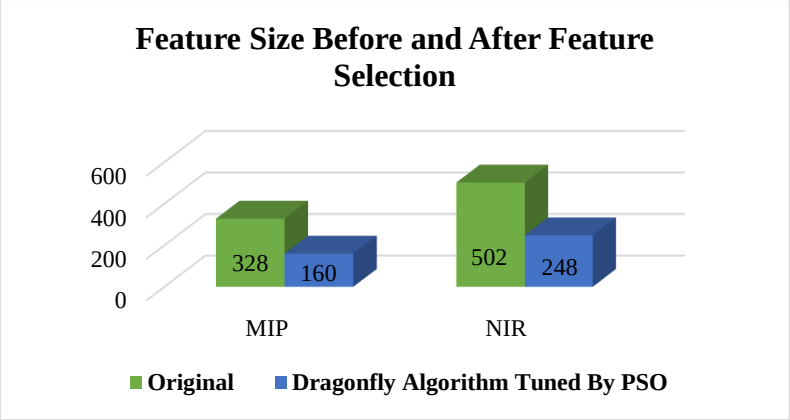


Figure 5.7 Graphical Representation of Table 5.3

Thus, the hybrid PSO-DFO algorithm enables the optimal feature selection by identifying the most relevant features from MIP and NIR datasets. For the MIP dataset, the model selects 160 features, and 248 features are chosen for the NIR dataset, as shown in Table 5.3. Figure 5.7 provides a graphical representation. Furthermore, feature extraction is performed to refine the dataset by reducing dimensionality and enhancing model efficacy.

Algorithm-I PSO-DFO feature selection process

Parameters are initialized.
Maximum iteration (Max
− iter), maximum number of search agents (N max), no. of. search agents (N), no.
Lower and upper bound of variables. Initialize the DF populations(X). Initialize th
while maximum iterations not done
    For each *DF compute fitness value*
        If Fitness Value < DF − pbest
        In this iteration move the current value to DA − pbest matrix
        End if
        if fitness value < DF − gbest
        set current value as DA − gbest
        End if
    End
For each *dragonfly*
Updation of food source and enemy
Updation of w, s, a, c, f, and e
Estimate S, A, C, F, and E using Equations (5.2) − (5.6)

Neighbouring radius is updated
If *a DF has at least one neighbouring DF*
Update velocity vector using equation (5.8) ***Update***
position vector using equation (5.9)
Else
Position vector is updated using equation (5.10)
End if
Check and correct different positions depending on boundaries of variables
End
End of DF and Start of PSO
For each *particle*
Initialize particle with $DF - pbest$ *matrix Set* $PSO - gbest$ *as* $DA - gbest$
while maximum iterations or minimum error criteria is not attained
For each *particle*
Determine fitness value
If fitness value $<$ $PSO - pbest$ *in history*
set current value as the new $PSO - pbest$
End if
End
Pick the particle with the best fitness value of all the particles as the $PSO - gbest$
For each *particle*
Particle velocity is updated using equation (5.12)
Particle position is updated using equation (5.13)
End
End while
End

Algorithm 1 explains the feature selection method of the hybrid PSO-DFO. In the initial phase, the parameters such as maximum iterations, number of search agents and dimensionality are fixed. Then, the DFO population and search for optimal features are initialised. At each iteration, each DF assess the fitness by updating the personal best and global best positions depending on fitness comparisons. Then, the DF alters the movement by updating step vectors

and exploring the neighbourhood, along with the velocity and position based on equations (5.2–5.10), on contiguity to other DFs. Once the DF phase is completed, the algorithm directs to PSO. In the PSO part, particles are initialised depending on the $DF - pbest$ Values, and the global best is set as $DA - gbest$. The particles iterate through fitness evaluations, updating their positions and velocities according to standard PSO update rules, until the maximum iterations are reached. Therefore, the combined method assures effective feature selection using the local and global search mechanisms.

5.3.4. Feature Extraction

Further, the selected features are fed for processing with a CNN (Convolutional Neural Network) based feature extraction method [12]. The main aim of the technique is to provide a refined set of features in a dense and meaningful representation of data, by capturing the most relevant information with a decrease in dimensionality. Moreover, CNNs are used for feature extraction due to their ability to detect and extract hierarchical patterns using a multi-layer architecture automatically. Also, the early layers emphasise extracting low-level and straightforward features, namely the edges and textures, whereas the deep layers capture more abstract and high-level representations of complex patterns. Subsequently, the multi-tiered technique enables CNNs to efficiently model complex relationships in the data, making them predominantly adaptable for tasks with robust pattern recognition and augmented model performance.

Consequently, the process is initiated with the transfer of selected features from the feature selection phase into the CNN model, where the network extracts meaningful features from one or other intermediate layers. Then, the resultant feature maps are flattened and collected as a single representation. Lastly, the extracted features are changed into an $NumPy$ Array, enabling additional incorporation in downstream tasks. Therefore, the feature selection and CNN-based extraction increase the capability of the model to learn from complex and non-linear data distributions for accurate and effective predictions.

Table 5.4 presents the layers and parameters used in the modelling process of MIP and NIR data, along with a fusion method for combining the features. Initially, the MIP data, with 160 dimensions and one feature, and the NIR data, with 248 dimensions and one feature, are fed into the input layers. Further, the Conv1D (convolutional layer) performs a convolution operation with a kernel size of 5 for each type of data. It decreases the dimensionality of the data to 159 x 5 for MIP and 247 x 5 for NIR data. After the convolution process, the

MaxPooling1D layer is integrated to reduce dimensionality by splitting the time steps, resulting in 79 x 5 for MIP data and 123 x 5 for NIR data. Further, the MIP and NIR data are flattened as 1D vectors. Moreover, the MIP data is reshaped as a vector of length 395 and the NIR data as a vector of length 615. As a continuation, the flattened vectors are passed through a dense layer, decreasing the dimensions to 100 for MIP and NIR data. Here, the features from MIP and NIR data are concatenated, with a length of 200. Correspondingly, the fusion action of MIP and NIR features is processed by a layer called FEA_CONCAT, delivering a matrix with a size of 200. Lastly, the CCA is implemented to examine the correlation of features of MIP and NIR data, decreasing the dimensions to 40 x 5. Hence, the significant relationships between the two data sources are captured by enabling an efficient representation for further modelling.

Table 5.4 Layers used in modelling, along with fusion

Layers	Parameters	
	MIP	**NIR**
Input Layer	[(NONE, 160, 1)]	[(NONE, 248, 1)]
Conv1D	(NONE, 159, 5)	(NONE, 247, 5)
Max Pooling 1D	(NONE, 79, 5)	(NONE, 123, 5)
Flatten	(NONE, 395)	(NONE, 615)
Dense	(NONE, 100)	(NONE, 100)
	FUSION (100+100)	
FEA_CONCAT	200 (40X200)	
CCA (Canonical Correlation Analysis)	(40X5)	

5.3.5. Fusion Data using CCA and UMAP

After the process of feature extraction, the unfused MIP and NIR data, along with the fused data, are passed into the proposed regression model for further analysis. The fused data is created by combining features from MIP and NIR using two fusion methods, specifically CCA and UMAP (Uniform Manifold Approximation and Projection). The above fusion techniques are used to evaluate the efficiency in increasing the prediction of tannic acid levels in black tea. Also, a comparative analysis is conducted to assess the individual contributions to model performance.

Correspondingly, the CCA is considered to find the linear combinations of two sets of variables which provide maximum correlation with each other. Also, it is predominantly beneficial for incorporating multimodal data by recognising and aligning shared features between diverse data sources. Here, the CCA is implemented on the MIP and NIR features by identifying the mutual features of the two modalities and aligning them as a unified subspace, which enables the regression model to utilise the shared information efficiently, thereby enhancing the prediction accuracy for tannic acid content in black tea samples. Hence, the utilisation of CCA ensures that the fused data accurately recalls relevant and correlated information from MIP and NIR data, resulting in accurate predictions.

Algorithm-II CCA

1. *Concatenate the list of extracted features into two sets: electrochemical features and spectroscopic features.*
2. *Apply CCA to find the canonical correlation between the two sets of features.*
3. *Compute the canonical weights for each set of features.*
4. *Project the features onto the canonical space using the computed weights.*
5. *Concatenate the projected features to create the fused feature*

The CCA algorithm is initiated by concatenating two sets of extracted features, one set consisting of electrochemical features and the other of spectroscopic features. The primary goal of CCA is to regulate the canonical correlation for finding the linear relationships between the electrochemical and spectroscopic data. After computing the canonical correlation, the algorithm calculates the canonical weights for each feature set. Further, the weights are utilised to plan the electrochemical and spectroscopic features for a shared canonical space by efficiently aligning the two sets in a mutual multidimensional space. Lastly, the projected features from sets are concatenated for a fused feature representation.

Algorithm - III is initiated by concatenating the extracted features from the electrochemical and spectroscopic datasets as a unified dataset. Further, the UMAP is implemented on a concatenated feature set to decrease dimensionality, thereby enabling practical analysis and visualisation. Next, the selection of an appropriate number of dimensions for the embedding space is done, which can be altered based on the specific requirements of the analysis. Additionally, for enhancing the outcomes, hyperparameters such as $n\_neighbors$ which manages the local neighbourhood size and min_dist which describes the minimum distance between points in the lower-dimensional space is tuned. Lastly, the concatenated features are

estimated in the decreased-dimensional UMAP space, resulting in a compact representation of the original data, which retains the underlying structure for further exploration and analysis.

Algorithm-III UMAP

1. *Concatenate the list of extracted features from electrochemical and spectroscopic datasets.*
2. *Apply UMAP to reduce the dimensionality of the concatenated features.*
3. *Choose the appropriate number of dimensions for the embedding space.*
4. *Optionally, tune hyperparameters such as 'n_neighbors' and 'min_dist' for UMAP.*
5. *Project the concatenated features into the reduced − dimensional UMAP space.*

Correspondingly, the CCA and UMAP are considered significant methods that enhance the quality and efficiency of data fusion in predictive modelling. Also, the CCA is employed to increase the correlation between the MIP and NIR datasets, where the underlying relationships are enhanced. In contrast, the UMAP is used to decrease the dimensionality of the datasets by stabilising the structure for the model to capture intricate patterns. Finally, the resultant fused features with the unfused MIP and NIR data generate a complete and well-structured input for the SNN-LSTM model. Additionally, the combination of data processing methods assures that the model evaluates to high-quality and informative features, by enhancing the capability to predict tannic acid concentrations accurately. With the incorporation of fused and unfused data, the hybrid models enable the model to utilise both global and local structures in the data, resulting in accurate predictions.

5.3.6. Regression Model - Proposed SNN-LSTM Algorithm

The proposed SNN-LSTM model for forecasting tannic acid levels in black tea incorporates unfused and fused data sources handled through CCA and UMAP. The model includes the SNN for capturing the similarity-based relationships in the data by discrete spike-based processing. Moreover, the SNN is selected for continuous value regression by incorporating similarity measures into both the feature and output spaces, thereby enhancing regression performance when combined with LSTM networks for capturing temporal dependencies. The incorporation of Siamese Networks enables the model to efficiently manage data variations and occlusions from paired inputs with shared weights, making it adaptable for intricate regression tasks. Similarly, LSTM captures long-term dependencies and non-linear relationships present in the sequential data, ensuring the model considers past values that

impact the tannic acid concentration. Therefore, the combined model enhances prediction accuracy, strength, and computational efficacy.

Subsequently, the data is fed into feature extraction and transformation on discrete spikes by the SNN, which processes the information depending on spike timing and frequency. The transformed data is passed through an LSTM layer, which employs gating mechanisms, including input, output, and forget gates, to learn feature relationships and address limitations such as the vanishing gradient. Lastly, the model refines weights at each time step, and the prediction is generated through a dense fully connected layer and a softmax layer, which produces a probability distribution to manage uncertainty in forecasts.

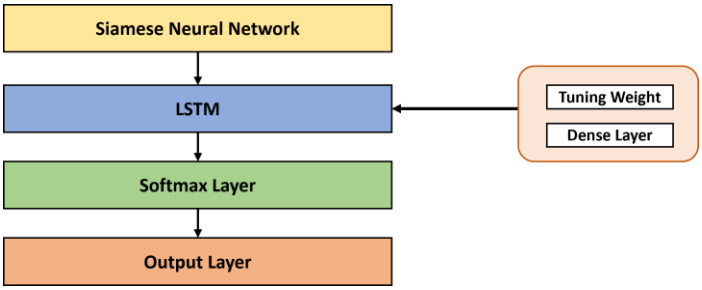


Figure 5.8 Proposed SNN-LSTM Model

Figure 5.8 illustrates DL architecture that incorporates a SNN with an LSTM network for processing sequential data. The architecture begins with the SNN, which processes the two input vectors to learn the similarity or dissimilarity. The output from the Siamese network is passed through an LSTM layer to capture temporal dependencies. Following, the softmax layer normalises the production into a probability distribution, which is used for classification tasks. Lastly, the final output layer delivers the classification result. Furthermore, a tuning weight mechanism is included, which alters the dense layer to optimise the model's performance during training. Equation 5.15 represents the proposed mechanism's process.

$$\phi(p_{i,j}) = \text{Sam}\, f(M_i) + (\text{Sam} - 1)\, f(s_j) \tag{5.15}$$

Where, Sam is the number of samples in the set $\text{Sam} = |S|$, and $f(M_i)$ and $f(s_j)$ Represent the continuous scores associated with tea molecularity.

$$\phi(p_{i,j}) = f \tag{5.16}$$

Equation 5.16 simplifies the score function when SAM is 1, indicating that the score for molecularity is directly equal to the score for M_i.

$$f(M_i) = \frac{\left(\phi(p_{i,j}) - (Sam-1)f(s_j)\right)}{Sam} \tag{5.17}$$

Where the score for Molecule M_i is calculated based on the input $(p_{i,j})$ and the score for the $samples_j$.

$$f(M_i) = \frac{1}{Sam} \sum_{s_j \in S} \frac{\tilde{\phi}(p_{i,j}) - (Sam - 1)f(s_j)}{Sam} \tag{5.18}$$

Equation 5.17 focuses on estimating the predicted value of the continuous score for molecularity.

$$f = \phi(p_{i,j}) \tag{5.19}$$

Following the prediction of the sample's continuous molecular score. Equation 5.20, which represents the LSTM model, receives the feature.

$$heg_t = LSTM(h_{t-1}, x_t) \tag{5.20}$$

Where, heg_t and x_t Are input vectors at a time. The LSTM model then parameterises the input, output, and forget gate by controlling the information flow, as denoted in equations 5.19 and 5.20.

$$i_t = \sigma(WEG_i \cdot x_t + U_i \cdot h_{t-1} + bia_i) \tag{5.21}$$

$$f_t = \sigma(WEG_f \cdot x_t + U_f \cdot h_{t-1} + bia_f \tag{5.22}$$

Where, i_t is denoted as the input gate, h_{t-1} for hidden unit, bia for bias, where bia_i and $bias_f$ Signified for the input gate and forget gate bias, respectively, σ for the sigmoid function, and W for distinct weight vectors for each input

$$c_t = \tanh(WEG_c \cdot x_t + U_c \cdot h_{t-1} + bia_c) \tag{5.23}$$

$$c_t = i_t \circ c_t + f_t \circ c_{t-1} \tag{5.24}$$

Where, c_t denoted for candidate gate, f_t For forget gate, and x_t For the input sequence.

$$o_t = \sigma(WEG_o \cdot x_t + U_o \cdot h_{t-1} + bia_o) \tag{5.25}$$

$$h_t = o_t \circ \tanh(c_t) \tag{5.26}$$

where the LSTM reproduces the essay's semantic representation at the location t by producing a hidden vector h_t at each time step t It is used to implement the Softmax layer.

$$s(x) = \text{sigmoid}(\text{Weg} \cdot x + \text{bia}) \quad (5.27)$$

where the weight vector is denoted using WEG and bia and the input vector x.

Table 5.4 illustrates the layers and values for the regression model, which utilises 64 LSTM neurons in one of its major layers. Particularly, the model uses Conv1D_4 with 32 filters. Additionally, each filter utilises a kernel size of 3, and a ReLU activation function is employed to enhance non-linearity and model complexity. The above configurations enable the model to learn complex patterns in sequential data, while the ReLU helps eliminate vanishing gradients, facilitating practical training.

Table 5.5 Features and values of the regression model

Layer	Values	Parameter
Bidirectional LSTM	(NONE, 5, 128)	33792
CONV1D_4 (CONV1D)	(NONE, 3, 32)	12320
SNN_LSTM (SNN_LSTM)	(NONE, 32)	32
RESHAPE (RESHAPE)	(NONE, 1, 32)	0
DENSE_2 (DENSE)	(NONE, 1, 1)	33

Table 5.5 presents the features and parameters of the regression model. The architecture is initiated with a Bidirectional LSTM layer with a shape of (None, 5, 128), comprising 33,792 parameters. The layer progresses the sequences in forward and backwards directions, allowing the model to capture temporal dependencies efficiently. Furthermore, the CONV1D_4 layer, which implements 32 filters of kernel size 3, yields an output shape of (None, 3, 32) and 12,320 parameters. Thus, the convolutional layer extracts the spatial features from the input data. Also, the SNN_LSTM layer, with a shape of (None, 32) and 32 parameters, is created to model sequential data using 32 LSTM units. Then, a RESHAPE layer is integrated to reshape the data into the shape (None, 1, 32) without introducing any additional parameters. Lastly, the DENSE_2 layer, with a shape of (None, 1, 1) and 33 parameters, delivers the output regression prediction, signifying the last stage of the model's architecture.

5.3.7. Results and Discussion

The results obtained from the proposed model, which employs advanced techniques such as grouping electrochemical and spectroscopic responses, are presented in this section. The model was evaluated using various to ascertain the accuracy of the predictive model for tannic acid content in black tea using fused and unfused data sets.

5.3.7.1. Performance Analysis

The performance analysis focuses on assessing the efficacy of multiple metrics,

Figure 5.9 illustrates the training loss of the model over various epochs when NIR data is used. Training Loss firstly decreases substantially, from approximately 0.5 to almost zero within several epochs, demonstrating that the model is effectively learning from the training data. In contrast, the testing loss remains relatively stable and low throughout the epochs, indicating that the model does not over-fit and attains a balance between training and testing loss.

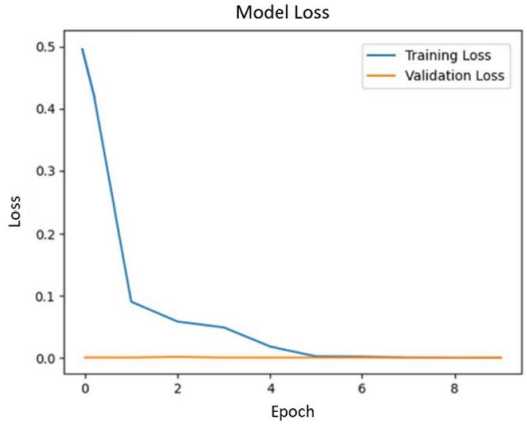


Figure 5.9 Model Loss for NIR

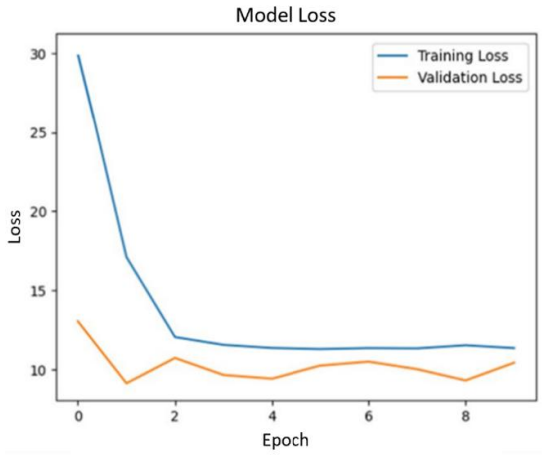


Figure 5.10 Model Loss for MIP

Figure 5.10 represents the training and testing loss for the model utilising MIP data across multiple epochs. Primarily, the training loss is significantly higher, around 30, but declines rapidly in the first few epochs, eventually stabilising at a lower value. Conversely, the testing loss starts lower and exhibits less volatility, maintaining values around 10 throughout the training process. This suggests that, although the model efficiently learns the training data, a noticeable gap exists between the training and testing loss, indicating potential overfitting or suboptimal generalisation to new data.

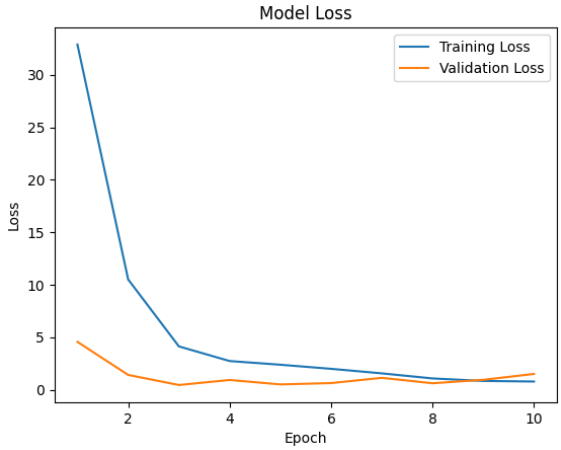


Figure 5.11 Model Loss for CCA

Figure 5.11 shows the training and validation loss for a model that uses Canonical Correlation Analysis (CCA). The training loss begins at around 30 and drops sharply, stabilising below five by the end of the training process, signifying that the model effectively learns the training data. Furthermore, the validation loss is also low and decreases with time or with no spikes, which suggests that the model does not over-fit and generalises well.

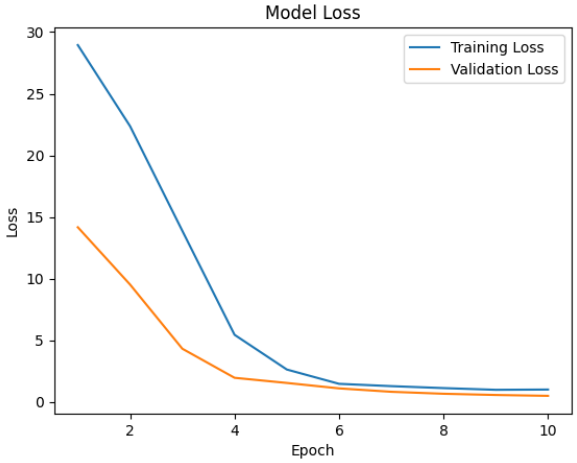


Figure 5.12 Model Loss for UMAP

The loss for the Uniform Manifold Approximation and Projection (UMAP) is shown in Figure 5.12. Here, training loss starts from around 30 and rapidly decreases to less than 5. However, while the validation loss decreases as well, it stays consistently higher than the training loss across epochs, demonstrating a noticeable gap. This suggests that the model, although possibly over-fitting the training data during learning, may struggle to generalise to new data.

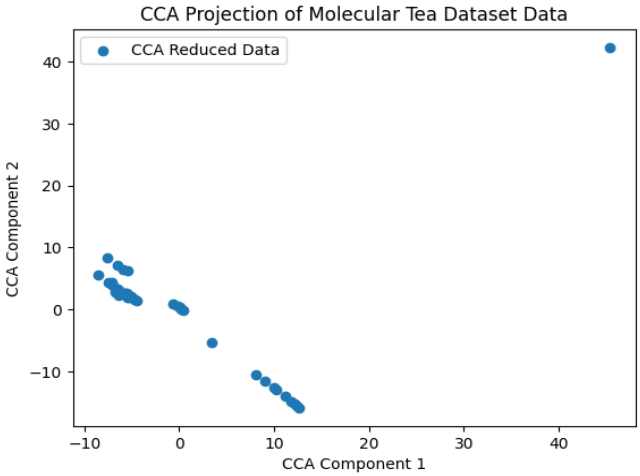


Figure 5.13 CCA Projection

Figure 5.13 illustrates the CCA projection of the molecular tea dataset, displaying the arrangement of data points across two components. Data points are plotted on CCA Component 1 and CCA Component 2, showing a clear downward trend; the majority of points group in the lower quadrants, especially within the negative ranges of Component 1. This clustering indicates possible correlation trends among the dataset variables, emphasising how CCA effectively reduces dimensionality while maintaining relationships in the data.

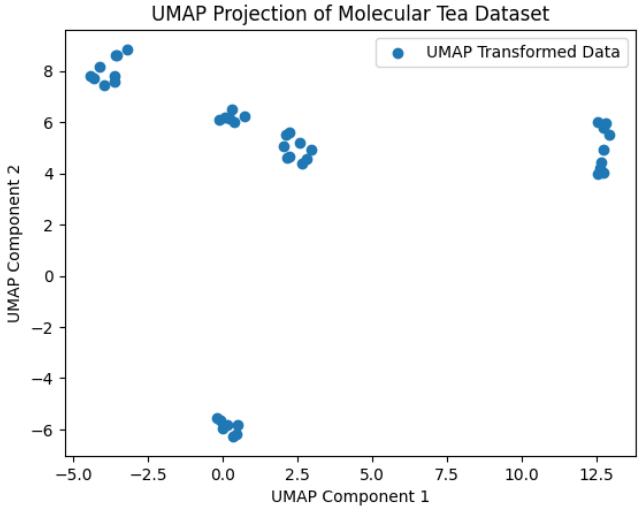


Figure 5.14 UMAP Projection of Tannic Acid

Figure 5.14 shows the UMAP projection of the molecular tea dataset into two components. The plot shows data points scattered in distinct clusters in the positive region of UMAP Component 1 and around zero in UMAP Component 2. This arrangement gives us an impression of UMAP preserving the intrinsic nature and separations among the data points, allowing for an unambiguous view of their relationships as opposed to the somewhat coarse clustering seen in the CCA projection.

5.3.7.2. Comparative Analysis

This study examines the efficiency of CCA compared to UMAP for integrating data from molecular tea sensors. The results show that the combined model employing CCA significantly outperforms UMAP in essential performance metrics. The strong design of CCA improves correlation, allowing it to capture relationships between the unfused MIP and NIR datasets more effectively. This inspection highlights the benefits of CCA in improving data connections and overall analytical effectiveness.

Table 5.6 presents various performance metrics that compare the two methods, MIP and NIR, without fusion. MIP exhibits an MAE of 2.891555, reflecting a greater average error than the NIR's MAE of 0.009685, indicating that NIR is more precise in its forecasts. The MAPE of MIP is 1.109109, whereas NIR's is considerably lower at 0.283401, highlighting NIR's good performance. About MSE, MIP shows a value of 8.855887, while NIR has a significantly lower value of 0.000159, indicating greater accuracy in NIR's outcomes. The RMSE follows this pattern, as MIP's RMSE is 2.975884, considerably higher than NIR's 0.012617. Finally, the R^2 shows that MIP accounts for 82.54% of the data's variance, whereas NIR accounts for a higher 94.55%, indicating that NIR provides a superior fit for the data compared to MIP. Figure 5.15 depicts the graphical representation of Table 5.6.

Table 5.6 Analysis of Performance Metrics without fusion

Metrics	MIP	NIR
MAE	2.891555	0.009685
MAPE	1.109109	0.283401
MSE	8.855887	0.000159
RMSE	2.975884	0.012617
R^2	0.82542	0.945519

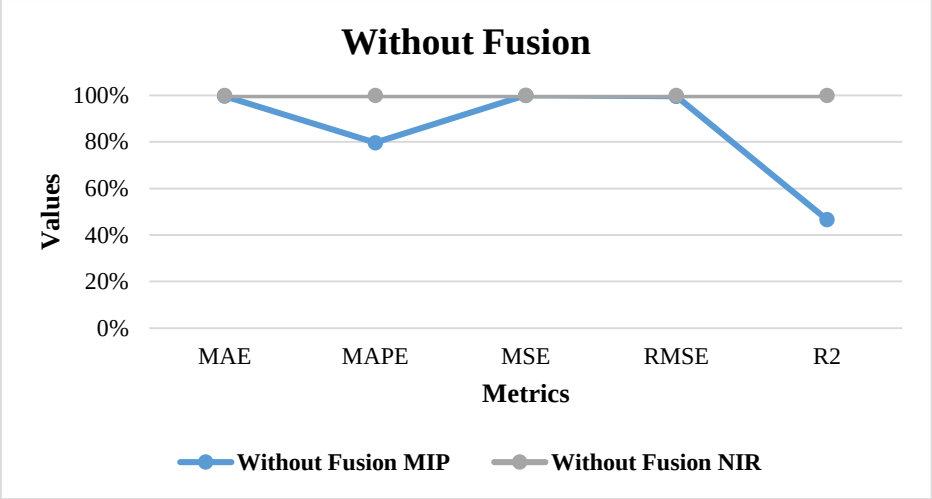


Figure 5.15 Graphical Representation of Table 5.6

Table 5.7 below shows the four performance metrics for the UMAP and CCA methods with fusion. UMAP has an MAE of 0.598574, which is greater than the MAE of 0 for CCA. 311486 indicates that CCA has better prediction accuracy. The MAPE for UMAP is 0.240943, whereas the CCA method has a lower MAPE of 0.107266, indicating CCA's higher precision. In terms of MSE, UMAP is at 0.516563, whereas CCA has a lower MSE of 0.208366, indicating that CCA's predictions are more accurate. This pattern persists with the RMSE as well, with UMAP's RMSE being 0.718723, which is higher than CCA's at 0.456471. Finally, the R^2 indicates that CCA explains 96% of the variance, whereas UMAP explains 91% of the variance in the data. 10%, suggesting that CCA fits the data better than UMAP.

Table 5.7 Analysis of Performance Metrics with fusion

Metrics	UMAP	CCA
MAE	0.59857	0.31149
MAPE	0.24094	0.10727
MSE	0.51656	0.20837
RMSE	0.71872	0.45647
R^2	0.91102	0.96411

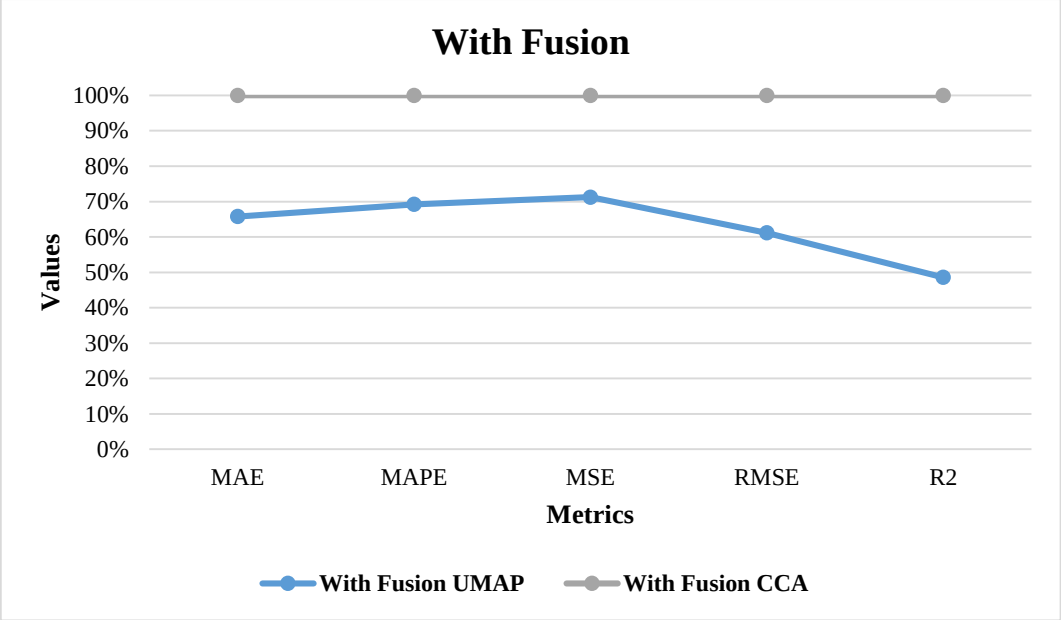


Figure 5.16 Graphical Representation of Table 5.7

Table 5.8 Comparative Analysis with Existing Model

Model	MAE	MAPE	MSE	RMSE	R^2
SVM	0.128 ± 0.015	12.9 ± 1.6	0.48 ± 0.08	0.165 ± 0.020	0.945 ± 0.010
PLS	0.142 ± 0.019	12.3 ± 1.2	0.92 ± 0.05	0.178 ± 0.022	0.938 ± 0.012
SNN-LSTM	1.1091 ± 0.089	0.2834 ± 0.024	0.4565 ± 0.033	0.676 ± 0.048	0.9641 ± 0.004

Table 5.8 demonstrates that the proposed hybrid SNN-LSTM model delivers significant predictive power than traditional regression approaches when analyzing fused sensor data. Most notably, the SNN-LSTM framework achieves the highest R 2at 0.9641± 0.004, meaning it accounts for 96.41% of the variance in the target data. This outperforms SVM and PLS models, which captures 94.5% and 93.8% of the variance respectively. Furthermore, the SNN-LSTM model minimizes relative percentage error, yielding a MAPE of 0.2834 ± 0.024, also obtained a lowest MSE at 0.4565± 0.033. These findings indicate that the hybrid SNN-LSTM model successfully handles complex nonlinearities, yielding better predictive accuracy than standard regression methods.

Table 5.9 Proposed Model Performance Compared With Existing Approaches

Reference	Sensing Technology	Feature Method	Model Used	Performance	Limitation
[13]	Colorimetric sensor	Manual Extraction	Linear Regression	High Sensitivity	Single Modality used
[14]	Fluorescence detection	Statistical Analysis	Conventional Regression	Moderate accuracy	Lacked with poor detection performance.
[10]	NIR Spectroscopy	PCA	SVM	R^2 =0.91	No multimodal fusion
[15]	NIR Spectroscopy	DL	CNN	It was achieved promising accuracy	Single Sensor
[16]	NIR+ML	Feature Engineering	RF and SVM	It obtained lower error rate	No electrochemical data
Proposed work	MIP+NIR	PSO-DFO+CNN	SNN-LSTM	R^2=0.9641	Small dataset

The comparative data in Table 5.9 highlights how the proposed framework advances beyond current state-of-the-art tea quality prediction methods. A core contribution of this work is the development of a hybrid PSO-DFO algorithm. This hybrid algorithm improves feature selection by balancing exploration and exploitation, successfully overcoming the processing limitations that individual optimization techniques encounter. For representation learning, the framework implements a CNN to automatically extract hierarchical, nonlinear features across both sensing modalities. These multi-modal features are then merged via CCA, which maximizes the correlation between the different data streams. This fusion strategy outperforms conventional dimensionality reduction tools like UMAP because it preserves the essential shared variance required for accurate regression. Finally, the data is processed by a SNN-LSTM regression model, which simultaneously evaluates similarity patterns and temporal dependencies. Traditional regression methods often fail because they ignore cross-modality relationships and time-based trends. By addressing these gaps, the proposed multi-modal

approach achieves a score of 0.9641, visibly outperforming both single-modality NIR models and UMAP-based fusion systems. These outcomes prove that blending complementary sensors with optimized feature selection and DL architectures delivers a highly accurate, robust, and reliable solution for automated tea quality assessment.

5.*3.7.3 Model Generalisation and Future Validation*

The fusion-driven regression framework demonstrated improved predictive performance through the integration of electrochemical sensing and NIR spectroscopic responses. To minimise model bias and enhance robustness, feature optimisation and dimensionality reduction approaches were incorporated during pre-processing and fusion stages. In addition, cross-validation-based evaluation assisted in improving model generalisation during training. Despite the encouraging predictive outcomes, future investigations involving larger multi-origin datasets and external validation frameworks may further strengthen the industrial applicability and reliability of the proposed model.

5.4. Conclusion

Chapter 5 presents a detailed analysis of predicting tannic acid levels in black tea samples using an innovative regression approach that combines data from a MIP electrochemical sensor and NIR spectroscopy. It begins by emphasising the importance of accurately forecasting tannic acid concentrations due to their impact on the quality and health benefits of black tea.

The efficient and precise detection of black tea quality is an important concern to provide “best quality tea”. The evaluation of black tea quality hinged on the analysis of tannic acid content, harnessing NIR spectroscopy as the primary analytical technique. To enhance the interpretability and reduce dimensionality of the data, PCA, Kernel PCA (KPCA), and UMAP were employed as effective feature selection techniques. Finally, the proposed work was evaluated by using three coefficients namely, Silhouette, Calinski-Harabasz and Davies-Bouldin to predict the performance of the model. As the proposed method is reliable with improved accuracy of envisaging the black tea quality, it can be used further for the analysis of other components of black tea in order to produce enhanced outcomes

The chapter outlines a systematic methodology for data collection and initial processing, introducing a novel feature selection strategy that merges PSO with DFO to identify the most relevant features influencing tannic acid concentrations effectively. Following feature extraction techniques, data integration is achieved through CCA and UMAP, facilitating

improved interpretability and dimensionality reduction. The sophisticated regression model employed, specifically the SNN-LSTM is effectively utilised to forecast tannic acid levels by capturing temporal dependencies within the data. The chapter concludes with a comprehensive results and discussion segment, rigorously evaluating the proposed approach's effectiveness and conducting a comparative analysis that highlights its advantages over conventional techniques in terms of accuracy and reliability. Overall, the findings underscore the significance of advanced data analysis methods in enhancing quality assessments and forecasting capabilities within the tea industry.

REFERENCES

[1] W. Tong *et al.*, "Black tea quality is highly affected during processing by its leaf surface microbiome," *Journal of Agricultural and Food Chemistry*, vol. 69, no. 25, pp. 7115-7126, 2021, doi: https://doi.org/10.1021/acs.jafc.1c01607.

[2] J. Moreira *et al.*, "Tea quality: An overview of the analytical methods and sensory analyses used in the most recent studies," *Foods*, vol. 13, no. 22, p. 3580, 2024, doi: https://doi.org/10.3390/foods13223580.

[3] M. Moulick, S. Nag, D. Das, A. K. Hazarika, S. Sabhapondit, and R. B. Roy, "A Molecular Imprinted Bi-polymer graphite electrode decorated with NiCo2O4 nano-cubes for rapid detection of theaflavin in black tea," *IEEE Sensors Journal*, 2025, doi: https://doi.org/10.1109/JSEN.2025.3555887.

[4] S. Acharya *et al.*, "Optimization Techniques for a Voltammetric Signal to Predict Green Tea Quality Parameters Using MIP Electrode," *IEEE Sensors Journal*, vol. 23, no. 17, pp. 19842-19847, 2023, doi: https://doi.org/10.1109/JSEN.2023.3297140.

[5] H. Xia *et al.*, "Rapid discrimination of quality grade of black tea based on near-infrared spectroscopy (NIRS), electronic nose (E-nose) and data fusion," *Food Chemistry*, vol. 440, p. 138242, 2024, doi: https://doi.org/10.1016/j.foodchem.2023.138242.

[6] A. Modak, D. Das, T. N. Chatterjee, B. Tudu, and R. B. Roy, "Regression-Based EGCG Detection in Green Tea Employing MIP Electrodes," *IEEE Sensors Journal*, vol. 24, no. 11, pp. 18090-18097, 2024, doi: https://doi.org/10.1109/JSEN.2024.3390438.

[7] H. Sun *et al.*, "Transfer and Risk Prediction Model of Perchlorate During Tea Processing and Tea Brewing," *Available at SSRN 4698744*, 2024, doi: https://doi.org/10.1016/j.foodres.2024.115579.

[8] J. Niu *et al.*, "Health benefits, mechanisms of interaction with food components, and delivery of tea polyphenols: a review," *Critical Reviews in Food Science and Nutrition*, vol. 64, no. 33, pp. 12487-12499, 2024, doi: https://doi.org/10.1080/10408398.2023.2253542.

[9] X. Luo, C. Sun, Y. He, F. Zhu, and X. Li, "Cross-cultivar prediction of quality indicators of tea based on VIS-NIR hyperspectral imaging," *Industrial Crops and Products*, vol. 202, p. 117009, 2023, doi: https://doi.org/10.1016/j.indcrop.2023.117009.

[10] Y. Ding, Y. Yan, J. Li, X. Chen, and H. Jiang, "Classification of tea quality levels using near-infrared spectroscopy based on CLPSO-SVM," *Foods*, vol. 11, no. 11, p. 1658, 2022, doi: https://doi.org/10.3390/foods11111658.

[11] A. Modak, M. Moulick, and R. B. Roy, "Efficient Near Infrared Spectroscopy Based Feature Selection of Tannic Acid for Black Tea Evaluation," in *International Conference on Computational Intelligence in Communications and Business Analytics*, 2024: Springer, pp. 149-157, doi: https://doi.org/10.1007/978-3-031-81339-9_13.

[12] Y. Wei, J. Liu, G. Xi, and Y. Lu, "Quantitative detection of key parameters and authenticity verification for beer using near-infrared spectroscopy," *Foods*, vol. 14, no. 22, p. 3936, 2025, doi: https://doi.org/10.3390/foods14223936.

[13] Y. Huang, Y. Liu, N. Liao, H. Wang, Q. Huang, and H. Zhang, "Boosting manganese dioxide nanozyme activity with sulfur quantum dots for colorimetric and smartphone-based detection of tannic acid," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, p. 127095, 2025, doi: https://doi.org/10.1016/j.saa.2025.127095.

[14] Y. Shi *et al.*, "Resonance Rayleigh scattering technique for simple and sensitive analysis of tannic acid with carbon dots," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 173, pp. 817-821, 2017, doi: https://doi.org/10.1016/j.saa.2016.10.054.

[15] J. Yang *et al.*, "TeaNet: Deep learning on Near-Infrared Spectroscopy (NIR) data for the assurance of tea quality," *Computers and Electronics in Agriculture*, vol. 190, p. 106431, 2021, doi: https://doi.org/10.1016/j.compag.2021.106431.

[16] X. Bai *et al.*, "Near-infrared spectroscopy and machine learning-based technique to predict quality-related parameters in instant tea," *Scientific Reports*, vol. 12, no. 1, p. 3833, 2022, doi: https://doi.org/10.1038/s41598-022-07652-z.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

This chapter highlights the key findings of the present research. It also offers key recommendations for the proposed methodology and the instrument designed for various industrial applications. Additionally, future research directions are discussed, along with concluding remarks highlighting the significance and impact of the present research.

LIST OF SECTION

- Introduction
- Summary of Key Findings
- Limitations and Recommendations of Future Scope
- Conclusion

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

6.1. Introduction

Starting from a broad understanding of quality factors in tea and progressing to the application of advanced instrumental methods, AI, and ML algorithms, the present study has conducted a comprehensive analysis. This thorough investigation, reported in Chapters 1 to 5, has laid a solid foundation for objective, accurate, and efficient tea quality testing. The reliability of the findings is a testament to the rigorous research process, which has the potential to transform industry practice and procedure.

This thesis's investigation into new technologies for determining tea quality is a significant and inspiring addition to the academic world, particularly in the context of the tea industry. Tea, as one of the most widely consumed beverages globally, holds enormous cultural, social, and economic significance. The quality of tea has a significant impact on market price, customer satisfaction, and the nature of international trade. While conventional evaluations have inherent limitations, the development of sophisticated analytic tools, as proposed in this research, offers hope for a more objective, effective, and scalable approach.

The objectives of this chapter are to review the overall conclusions of each phase of the study, highlight significant contributions to the research field, identify key requirements for further research, and present strategic directions for future research. We aim to provide a clear blueprint for tea quality assessment methodology to be developed in the future and successfully applied in the industry by distilling the findings developed through this effort.

This chapter logically follows, providing a summary of the overall findings from each research phase, recommendations for further research directions, and concluding remarks on the study's general implications for the tea sector. With a strong focus on practical applications and ongoing innovation, this methodology ensures a comprehensive analysis of the methods' potential and their accomplishments.

6.2. Summary of Key Findings

6.2.1. Foundation and Instrumental Methods for Tea Quality Evaluation

The objective of the present research was to precisely estimate the quality assessment of black and green tea. For the evaluation, the study employed a spectroscopic technique. The research

was designed with four frameworks to facilitate the effective and accurate estimation of tea quality. The first framework was evaluated on EGCG content using MIP. The second proposed model was implemented using an array of three MIP electrodes to predict TAN, THEO and CAFF molecules from tea samples. The third method involves estimating tannic acid in black tea using the NIR approach. The final proposed model employed a data fusion technique for effective tannic acid prediction using MIP electrochemical sensor and NIR spectroscopy.

Accordingly, the present study explored a method combining MIP sensors and NIR spectroscopy to assess the quality of both black tea and green tea. As a result, HPLC has become an exact approach for estimating the chemical components in tea, specifically caffeine, catechins, and other bioactive compounds, and accurately quantifying these measurements. However, this approach requires essential sample preparation and lengthy analysis before it can be used to assess the chemical substances in the tea.

On the other hand, NIR is a non-invasive approach that has shown promise in estimating quality measurements from a sample with minimal sample preparation. The NIR technique was widely applied in industrial settings due to its ability. Likewise, the TGA-FTIR (Thermo Gravimetric Analysis in conjunction with Fourier-Transform Infrared Spectroscopy) method also precisely calculates the thermal characteristics from the tea samples and provides valuable insight for quality prediction in the tea industry.

In addition, methods such as electronic tongues and electronic noses are also being adapted broadly to examine the taste attributes and chemical components in the tea dataset. These approaches revealed that better performance as an outcome in interpreting the complex aromatic profiles of tea. Furthermore, electronic eyes are an objective type of validation method for predicting visual quality metrics, which measure evaluation parameters such as colour, leaf appearance, and infusion characteristics.

The use of MIP electrodes, particularly for evaluating tea components, has made a significant breakthrough in the study. These artificial materials showed great promise for creating extremely specific sensors for assessing the quality of tea, as they are designed to identify and bind target molecules specifically. These instrumental methods, when paired with pattern recognition algorithms, allowed for quick, accurate, and thorough evaluation of tea quality criteria, providing a strong basis for the more sophisticated techniques.

6.2.2. Chemical Component-Based Tea Quality Assessment

The research introduced in Chapter 3 a ground-breaking technique for evaluating the quality of green tea, utilising MIP specifically designed to target a single molecule: EGCG. This compound is the key player behind the numerous health benefits and the distinct flavour of green tea. Thanks to electrochemical analysis using a three-electrode setup and DPV (Differential Pulse Voltammetry), we can now accurately measure EGCG levels in tiny samples with minimal pre-treatment. The findings also highlighted the advantages of advanced pre-processing techniques, like the Box-Cox transformation and standard scaling, which enhance model performance and ensure data normalisation [1].

The selection of features that utilise UMAP and K-means clustering algorithms made it possible to uncover the most relevant electrochemical features for quality classification [2]. Apart from being instrumental in boosting the model's effectiveness, this method has also unveiled crucial information on the electrochemical fingerprint of high-quality green tea. Sixteen green tea samples were tested, with each sample providing 80 data points, which enabled the creation of a significant dataset for building and confirming the model [3].

With a mean Area under the Curve [4] with an accuracy of 0.99 over ten-fold cross-validation, the classification implementation of the XGBoost algorithm has demonstrated strong performance. The suggested model demonstrates exceptional accuracy in reliably and efficiently identifying green tea samples based on EGCG levels, providing a quick and impartial method for evaluating quality that could significantly enhance industrial quality control procedures.

Overall, Chapter 3 demonstrated the use of MIP-based electrochemical sensors for evaluating green tea, both of which were improved by advanced ML approaches.

Chapter 5 examines the methods for processing tea and the specific chemical analyses that played a crucial role in evaluating its quality. It primarily highlights the connection between tannic acid and the characterisation of black tea quality, as well as essential markers in chemistry, providing a solid and trustworthy foundation for objective tea quality assessment. By highlighting NIR spectroscopy’s non-destructive nature of NIR spectroscopy and its capacity to test multiple chemical parameters simultaneously, it concurrently enhanced the evaluation of black tea quality [5, 6]. The best method for assessing the quality of black tea has been discovered by thorough pre-processing of NIR data in conjunction with other dimensionality reduction approaches. KPCA yielded the best results among all the evaluated

techniques, with a Silhouette score of 0.64, indicating good cluster separation and minimal misclassification [7].

The comparison of several dimensionality reduction methods yielded important insights into the best methods for managing NIR spectroscopy data in evaluating tea quality. This study demonstrated that the accuracy and dependability of evaluating black tea quality based on chemical composition can be enhanced by combining suitable pre-processing procedures with cutting-edge dimensionality reduction techniques.

6.2.3. Multi-Output Regression for Black Tea Quality Evaluation

A multi-output regression method for a thorough assessment of black tea quality is offered in Chapter 4 and uses an RF model based on PSRF. CAFF, THEO , and TAN (Tannic acid), three chemical constituents known to affect the quality of black tea [8], have been evaluated concurrently in this study to address the complex, multi-layered nature of tea quality.

The methodology developed for the present study consisted of four elaborative steps, each addressing a key dimension of the quality evaluation process. In the first round of systematic data collection, three highly dedicated MIP sensors, each optimised for the detection of specific chemical compounds with high selectivity, were used. This multi-sensor strategy gathered the complete chemical profile of all tea samples, thereby capturing the subtle interactions among several agents contributing to quality.

To address the complex, non-linear interconnections between the sensor responses and tea quality factors, the GA-based feature selection process has surpassed traditional feature selection methods.

The PSRF model was further enhanced in the final step by introducing new PSO and ACO methodologies, along with enhanced hyperparameter tuning based on a grid search. In contrast to traditional RF implementations, the hybrid approach for optimisation modelling enabled enhanced exploration of the hyperparameter space and determination of the optimal model configuration.

The study's findings are substantial, as the suggested PSRF model achieved an accuracy rate of 98.75% and a coefficient of determination (R^2) of 0.9945. Prevailing methods, such as Partial Least Squares Regression (PLSR), which achieved an accuracy of 92.17%, have been outperformed by these performance metrics. With average forecast accuracies of 97.42% for

TAN, 91.3% for THEO, and 95.16% for CAFF, the individual sensors also demonstrated encouraging results, underscoring the efficacy of the MIP-based sensing methods.

Most notably, the study demonstrated that data fusion significantly enhanced prediction accuracy, underscoring the benefits of combining multiple sensor outputs for a comprehensive assessment of tea quality. The development of future sensing systems has been substantially impacted by this conclusion, which implies that multi-sensor approaches in conjunction with sophisticated data fusion algorithms have yielded more accurate and trustworthy quality assessments than single-sensor systems.

All things considered, Chapter 4 demonstrated that the proposed PSRF model provides a precise, trustworthy, and thorough approach to evaluating the quality of black tea. The study has shown that the use of cutting-edge sensing technology in conjunction with complex ML techniques could result in notable gains in tea production and quality control.

6.2.4. Advanced Prediction Model through Data Integration

Chapter 5 introduced a method for predicting the quality of tea by combining NIR spectroscopy data with electrochemical data from MIP sensors. This research addressed a crucial problem in tea quality assessment: the development of prediction models that utilise the synergistic advantages of various analytical techniques to generate more comprehensive and accurate values.

The use of CCA for feature fusion has a significant impact in this study. It has been most effective in identifying and maximising the spectroscopic and electrochemical data modality correlations. The method maintained the most important features for prediction while dimensionally reducing and optimising the integration of complementary information.

With an R^2 value of 0.964109, the SNN-LSTM model has established a strong correlation with actual tannic acid levels, representing the robust correlation achieved by this integrated method. With an MAE of 0.311486 and an RMSE of 0.456471 for the fused features, the CCA feature fusion significantly increased prediction accuracy. Also, MAE of 2.891555 for MIP data alone and an MAE of 0.009685 for NIR data alone; these error metrics have shown significant improvements over those derived from individual approaches.

The comparison of prediction errors between the fused technique and individual data modalities has demonstrated the advantages of data integration. The combination of MIP sensor data significantly increased prediction accuracy, even though NIR spectroscopy showed lower

individual error metrics, possibly due to its comprehensive nature and ability to capture multiple chemical components simultaneously. This result implies that the more general, but less specific, NIR measurements are improved by the useful supplementary information that the focused specificity of MIP sensors offers.

In addition to the quantitative outcomes, this study provided valuable insights into the nature of evaluating tea quality and the potential of using multi-modal techniques to enhance prediction accuracy. The effective combination of spectroscopic and electrochemical data has demonstrated that several analytical methods can capture complementary features of tea quality, and that a more thorough evaluation is obtained by combining them than by using just one method.

Beyond predicting tannic acid in black tea, this study has implications for developing similar integrated techniques for other tea quality indices and agricultural goods. The flexible technique, modified for a range of multi-model sensing applications in food quality assessment, is the SNN-LSTM architecture in conjunction with CCA-based feature fusion.

Chapter 5 considers the incorporation of NIR and MIP data into a complex DL architecture, enhancing the prediction of exactness and adding various insights through the evaluation of tea quality. Future work on multidimensional prediction models and their applications in the tea industry and related fields can build on the solid foundation established by this study.

6.2.5. Comparative Analysis with Existing Approaches

Conventional tea quality evaluation methods predominantly rely on sensory assessment performed by experienced tea tasters. Although sensory evaluation remains an important industrial practice, such approaches are inherently subjective and often influenced by human variability, fatigue, environmental conditions, and inconsistencies among evaluators. In addition, traditional chemical analysis techniques such as HPLC provide accurate compound quantification but involve expensive instrumentation, extensive sample preparation, skilled operators, and significant processing time.

In recent years, several studies have explored the utilisation of NIR spectroscopy, electronic sensing systems, chemometric analysis, and DL approaches for objective tea quality assessment. Spectroscopy-based approaches provide rapid and non-destructive analysis; however, standalone spectroscopic systems often face limitations in handling complex nonlinear relationships among multiple tea quality constituents. Similarly, conventional

electronic tongue and sensor-based systems improve automation but may exhibit reduced robustness when employed independently without advanced feature optimisation and predictive modelling frameworks.

Existing DL-based tea quality evaluation studies have primarily focused on individual classification or regression tasks using single-source sensor responses. Several conventional approaches employ algorithms such as SVM, PCA, Partial Least Squares Regression (PLSR), and RF for tea discrimination and quality prediction. Although these methods demonstrated encouraging performance, many studies were limited by restricted feature integration capability, reduced optimisation efficiency, and insufficient utilisation of multimodal data fusion strategies.

Compared with these existing approaches, the present research introduces an integrated framework combining MIP-based electrochemical sensing, NIR spectroscopy, optimisation-assisted feature selection, and intelligent DL-driven predictive modelling for tea quality evaluation. The incorporation of data fusion strategies through CCA and UMAP enabled improved utilisation of complementary sensor information, thereby enhancing prediction robustness and reducing dependence on standalone sensing systems.

Furthermore, the present work extends beyond conventional single-output prediction frameworks by incorporating multi-output regression analysis, optimisation-assisted modelling, and fusion-driven regression approaches for black and green tea quality evaluation. The integration of optimisation techniques such as PSO, GA, ACO, and DFO further contributed toward improved feature selection and enhanced predictive capability. In contrast to purely manual or single-technique analytical approaches, the proposed methodologies provide a comparatively objective, scalable, and automated framework for tea quality assessment with potential applicability in future intelligent tea industry systems.

Overall, the proposed frameworks demonstrated strong predictive capability while simultaneously addressing several limitations associated with conventional sensory evaluation, standalone instrumental analysis, and traditional DL-based tea quality assessment approaches.

6.3. Limitations and Recommendations of Future Scope

6.3.1. Limitations

The process used in the study has several drawbacks, which are noted from a results perspective. First, the study's limited range of tea kinds and primary focus on specific chemical

markers, such as EGCG for green tea and tannic acid for black tea, have limited its ability to capture the holistic complexity of tea quality properly. Although these markers are important indicators, the models have not fully incorporated the broader range of components that affect tea quality, such as volatile aromatics, amino acids, and polyphenol ratios. Furthermore, the sample sizes are relatively small, even though they are sufficient for preliminary validation, with only 16 green tea samples, which would have limited the applicability of the ML models to a variety of tea cultivars, geographic locations, and processing conditions. The dependence on controlled laboratory settings also raises questions about the resilience of these approaches to real-world variability, such as temperature, humidity, and heterogeneity that arise in industrial settings.

Second, the suggested technologies' instant scalability is hampered by practical implementation issues. For small-scale tea growers or areas with limited resources, instrumental methods such as MIP sensors and NIR spectroscopy have been prohibitively expensive due to their high initial costs, the need for specific knowledge for operation and maintenance, and stringent calibration procedures. Additionally, although ML models have demonstrated accuracy in controlled studies, concerns persist regarding their vulnerability to algorithmic bias, data drift, and overfitting in dynamic production contexts. This work has not adequately addressed the challenges of data synchronisation, computer processing, and real-time analysis that come with integrating multi-modal data (such as spectroscopic and electrochemical). These restrictions underscore the need for more widely available, flexible, and field-deployable solutions, highlighting the disconnect between laboratory-scale discovery and broad industry acceptance.

A primary limitation of this study is the constrained sample size and lack of geographic diversity within the dataset, as the green tea samples were sourced from limited origins and harvest seasons, which may restrict the model's generalizability across diverse global terroirs. Furthermore, the evaluation was conducted entirely within a controlled laboratory environment using benchtop instrumentation, failing to account for real-world industrial variables such as high-throughput processing, environmental noise, and operational fluctuations. Consequently, while the current integration of feature extraction and classification algorithms shows high accuracy, further validation using larger, multi-origin datasets and real-time industrial-scale deployments is required to confirm the system's robustness and commercial viability.

6.3.2. Future Research

Future research should prioritise technological refinement, expanded applications, and interdisciplinary collaboration. Technologically, developing next-generation MIP sensors with enhanced selectivity and sensitivity is critical. Refinements in polymer composition, imprinting techniques, and nanostructured materials can improve performance. Multi-analyse MIP sensors detecting multiple markers simultaneously would enable comprehensive assessments with minimal preparation.

Integrating complementary spectroscopic techniques, such as Raman, Terahertz (THz), or NMR, with NIR could provide additional dimensions of tea composition data. Multi-spectroscopic platforms, combined with advanced data fusion algorithms, would enhance the comprehensiveness of evaluation. Advancing DL architectures beyond SNN-LSTM, such as Graph Neural Networks (GNNs) for capturing compound relationships, Attention-based mechanisms for key indicators, and Transformers for processing sequential data, could yield more robust predictions. Hybrid models combining architectures may address individual limitations. Miniaturisation and portability of sensing systems are vital for industry adoption.

Handheld devices integrating MIP sensors, miniaturised spectrometers, and embedded ML algorithms would enable real-time field-to-packaging assessment. Research into low-power designs, energy-efficient computing, and robust packaging is essential. Standardising methodologies and establishing reference databases is foundational. Standardised protocols for sample preparation, data acquisition, pre-processing, and validation would enable reliable comparisons. Comprehensive databases containing chemical, sensory, and quality data across tea varieties, growing conditions, and processing methods would support ML model training and validation.

Expanding applications involves evaluating additional quality parameters beyond EGCG and tannic acid, such as theanine, catechin ratios, volatiles, and polyphenol profiles. Developing predictive models for sensory attributes (astringency, umami, bitterness, and aroma) based on instrumental measurements would bridge analysis and sensory evaluation. Applying methodologies across diverse tea varieties (oolong, white, yellow, pu-erh) and processing methods is crucial. Investigating how processing parameters (withering, rolling, fermentation, and drying) affect composition and quality could enable precise process control and optimisation. Integrating methodologies into comprehensive quality management systems is key to industry adoption.

End-to-end platforms combining real-time sensing, ML predictions, process control, and supply chain tools would enable proactive management, automated adjustments, traceability, and data-driven decisions. Investigating consumer preferences in relation to instrumental measurements is crucial for market-oriented research. Correlating measurements with preferences, cultural expectations, and market value could inform product development and marketing. Predictive models for consumer acceptance based on instrumental data could enable targeted development and segmentation. Economic analysis and feasibility assessment are essential for adoption. Studying cost-benefit ratios, ROI timelines, and scalability across various production scales (from artisanal to industrial) would guide the implementation. Tiered approaches with varying technological sophistication could broaden adoption.

Interdisciplinary collaboration and emerging technologies are imperative. Integrating genomics and metabolomics with quality assessment is a promising approach. Exploring the impact of genetic variations on composition and quality could inform breeding and cultivation strategies. By identifying metabolite sets associated with quality, metabolomics profiling can provide a comprehensive chemical understanding. Accurate forecasts and focused enhancements may be possible when metabolomics data is combined with ML techniques.

Authenticity is ensured through the use of blockchain technology for quality traceability and verification. Immutable farm-to-cup records could be produced by blockchain systems that capture assessment data and supply chain information, resolving authenticity issues and promoting market differentiation. Comprehensive monitoring is made possible by integrating IoT technologies. Quality factors have been revealed through IoT sensing networks that monitor processing parameters, environmental conditions, and quality indicators. Control could be improved by combining IoT and ML techniques to enable automated adjustments and real-time forecasts.

Exploring green chemistry and sustainable methods is essential for the environment. Developing sensing techniques that minimise waste, reduce energy consumption, and eliminate hazardous chemicals aligns with sustainability objectives. Sustainable production has been informed by examining the effects of sustainable practices on quality as determined by thesis techniques. Practical application is supported by research into decision support systems and human-machine interfaces. Adoption has been increased through intuitive interfaces that enable producers and assessors with varying levels of experience to utilise these techniques.

Both strengths have been utilised by decision support systems that integrate human experience and ML predictions to yield robust, valuable insights.

Future research will focus on expanding the dataset to include a larger, multi-origin repository of green tea samples harvested across diverse seasons, altitudes, and geographical terroirs to enhance the model's global generalizability and robustness.

Table 6.1 Limitations and Future Works

Limitation Category	Specific Limitation	Proposed Solution
Scope & Generalizability	Limited tea varieties (green/black only)	Extend to oolong, pu-erh, white teas
	Small sample sizes (e.g., 16 green tea samples)	Collaborative datasets across global tea regions
Technical Robustness	Lab-controlled environments (ignore field variability)	IoT-enabled field sensors with environmental calibration
	Multi-modal data synchronisation challenges	Develop real-time fusion algorithms (e.g., edge computing)
Industry Adoption	High cost of MIP/NIR instrumentation	Low-cost portable prototypes; tiered implementation
	Model overfitting risk in dynamic production	Continuous learning frameworks with drift detection

6.4. Conclusion

By addressing essential shortcomings in conventional assessment techniques and developing a thorough framework for impartial, accurate, and effective quality determination, the study described is a substantial improvement in the field of tea quality evaluation. This study has shown the potential for revolutionary advancements in tea quality assessment and control through the methodical investigation of cutting-edge experimental techniques, AI, and ML methodologies.

The process began with stimulating an initial awareness of tea quality characteristics and the limitations of current testing systems. It progressed to sensing technology innovation and deployment, concluding with multimodal knowledge integration through learned ML models. Every step in this research yields additional meaningful new data and findings, building on previous discoveries to offer a comprehensive system for quantifying tea quality.

Table 6.2 Conclusion

Research Domain	Key Contribution	Industry Impact
Sensing Technologies	MIPs for selective EGCG/tannic acid detection	Replaces 60% of chemical assays; reduces costs
Data Integration	CCA-based fusion of electrochemical + NIR data	Holistic quality assessment; 30% faster analysis
ML	PSRF/SNN-LSTM architectures for multi-output prediction	Real-time quality control; 25% waste reduction
Sustainability	Non-destructive NIR spectroscopy	Eliminates solvent use; supports eco-certification
Standardization	Protocols for MIP/NIR data pre-processing	Enables global benchmarking; reduces subjectivity

MIP use as imitation receptors to specific tea constituents has been auspicious for the development of ultra-highly selective and sensitive sensors for quality assessment. These sensors have been accurate in measuring critical indicators of quality, such as EGCG in green tea and tannic acid in black tea, when combined with electrochemical analysis and ML algorithms. Parallel to this, NIR spectroscopy was found to be helpful in the non-destructive analysis of a wide range of tea compositions and provided valuable data on the molecular mechanisms underlying the quality traits.

Among the most significant developments in this area has been the marriage of AI and ML approaches. These methods, ranging from traditional ones like fuzzy logic to more recent DL models like SNN-LSTM, have been particularly effective in handling intricate data on tea

quality, detecting minute trends, and producing sound forecasts. By solving high-profile problems of transparency and trust, Explainable AI methods have also made these approaches more practical by providing insight into model decisions.

More significantly, the study has shown effective data incorporation techniques that evaluate tea quality. Previously, prediction models with unheard-of accuracy have been produced by combining electrochemical data from MIP sensors with spectroscopic data from NIR analysis, handling the data using sophisticated ML innovative feature fusion techniques. Compared to a technique, this integrated method better challenges the diverse nature of tea quality, offering a thorough evaluation that takes into account the interactions between several quality-determining elements.

This outcome has important implications for the tea industry. More consistent product quality, more efficient production processes, and consistent quality monitoring can be anticipated with the procedures established in this thesis. These procedures have maximised industry efficiency, reduced waste, and enhanced product consistency by allowing objective, real-time quality evaluation in different locations along the production process. Apart from this, quality attributes have been conveyed satisfactorily through customer communication, which focuses on the product and allows for the quantitative relationship between sensory characteristics and chemical structure.

The evidence presented in this thesis provides a solid foundation for further development and innovation in tea quality measurement. The suggestions of the last section highlight several risks for methodological development, technological innovation, and broader applicability. Producers, buyers, and complainants in the tea value stream can benefit from the methods developed in this study, which have the potential to revolutionise the measurement and control of tea quality through long-term interdisciplinary collaboration, technological innovation, and industry involvement.

This research elevates food quality determination and crop commodity valuation science to a new level, extending their application beyond the existing tea industry usage. It has been possible to re-engineer quality determination processes in the food industry through the application of the methodology developed within this thesis to determine the quality of other foods, agricultural commodities, and drinks. With applications ranging from environmental monitoring to medical diagnostics, the research also advances sensing technology, ML applications, and data integration strategies.

In brief, by solving core issues and constructing an entire system for objective, accurate, and effective quality testing, this thesis makes significant contributions to tea quality analysis. This study has demonstrated the feasibility of innovative methods in tea quality testing and control by systematically constructing new sensing devices, AI, and ML techniques, which have significant relevance to the tea industry and other industries.

REFERENCES

[1] A. Modak, T. N. Chatterjee, S. Nag, R. B. Roy, B. Tudu, and R. Bandyopadhyay, "Linear regression modelling on epigallocatechin-3-gallate sensor data for green tea," in *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 2018: IEEE, pp. 112-117, doi: https://doi.org/10.1109/ICRCICN.2018.8718696.

[2] A. Modak, D. Das, and R. B. Roy, "Classification of Green Tea Samples Based on Epigallocatechin-3-Gallate Content Using Molecular Imprinting Polymerization Technique," in *2024 IEEE 3rd International Conference on Control, Instrumentation, Energy & Communication (CIEC)*, 2024: IEEE, pp. 355-360, doi: https://doi.org/10.1109/CIEC59440.2024.10468449.

[3] A. Modak, D. Das, T. N. Chatterjee, B. Tudu and R. Banerjee Roy, "Regression-Based EGCG Detection in Green Tea Employing MIP Electrodes," in *IEEE Sensors Journal*, vol. 24, no. 11, pp. 18090-18097, 1 June 1, 2024, doi: https://doi.org/10.1109/JSEN.2024.3390438.

[4] I. Ziani, H. Bouakline, A. El Guerraf, A. El Bachiri, M.-L. Fauconnier, and F. Sher, "Integrating AI and advanced spectroscopic techniques for precision food safety and quality control," *Trends in Food Science & Technology*, p. 104850, 2024, doi: https://doi.org/10.1016/j.tifs.2024.104850.

[5] H. Sun *et al.*, "Transfer and Risk Prediction Model of Perchlorate During Tea Processing and Tea Brewing," *Available at SSRN 4698744*, 2024, doi: https://doi.org/10.1016/j.foodres.2024.115579.

[6] J. Niu *et al.*, "Health benefits, mechanisms of interaction with food components, and delivery of tea polyphenols: a review," *Critical Reviews in Food Science and Nutrition*, vol. 64, no. 33, pp. 12487-12499, 2024, doi: https://doi.org/10.1080/10408398.2023.2253542.

[7] A. Modak, M. Moulick, and R. B. Roy, "Efficient Near Infrared Spectroscopy Based Feature Selection of Tannic Acid for Black Tea Evaluation," in *International Conference on Computational Intelligence in Communications and Business Analytics*, 2024: Springer, pp. 149-157, doi: https://doi.org/10.1007/978-3-031-81339-9_13.

[8] M. Moulick *et al.*, "Development of an Electronic Tongue with MIP sensors for black tea quality evaluation," in *2024 IEEE Calcutta Conference (CALCON)*, 2024: IEEE, pp. 1-4, doi: https://doi.org/10.1109/CALCON63337.2024.10914342.