IJOER
ENGINEERING JOURNAL

Peer Reviewed
Refereed Journal

ENGINEERING JOURNAL IJOER

Advancing Engineering Research,
Innovating for a Better Tomorrow.

- Innovative Research
- Engineering Insights
- Global Collaboration
- Impactful Solutions

VOLUME-12, ISSUE-8, AUGUST 2026

Focus Areas

- Civil & Infrastructure Engineering
- Mechanical & Industrial Engineering
- Electrical & Electronics Engineering
- Emerging Technologies

DOWNLOAD NOW

Contact us
+91-7665235235
www.ijoer.com
info@ijoer.com

Preface

We would like to present, with great pleasure, the volume-12, Issue-8, August 2026, of a scholarly journal, *International Journal of Engineering Research & Science*. This journal is part of the AD Publications series *in the field of Engineering, Mathematics, Physics, Chemistry and science Research Development*, and is devoted to the gamut of Engineering and Science issues, from theoretical aspects to application-dependent studies and the validation of emerging technologies.

This journal was envisioned and founded to represent the growing needs of Engineering and Science as an emerging and increasingly vital field, now widely recognized as an integral part of scientific and technical investigations. Its mission is to become a voice of the Engineering and Science community, addressing researchers and practitioners in below areas:

Chemical Engineering	
Biomolecular Engineering	Materials Engineering
Molecular Engineering	Process Engineering
Corrosion Engineering	
Civil Engineering	
Environmental Engineering	Geotechnical Engineering
Structural Engineering	Mining Engineering
Transport Engineering	Water resources Engineering
Electrical Engineering	
Power System Engineering	Optical Engineering
Mechanical Engineering	
Acoustical Engineering	Manufacturing Engineering
Optomechanical Engineering	Thermal Engineering
Power plant Engineering	Energy Engineering
Sports Engineering	Vehicle Engineering
Software Engineering	
Computer-aided Engineering	Cryptographic Engineering
Teletraffic Engineering	Web Engineering
System Engineering	
Mathematics	
Arithmetic	Algebra
Number theory	Field theory and polynomials
Analysis	Combinatorics
Geometry and topology	Topology
Probability and Statistics	Computational Science
Physical Science	Operational Research
Physics	
Nuclear and particle physics	Atomic, molecular, and optical physics
Condensed matter physics	Astrophysics
Applied Physics	Modern physics
Philosophy	Core theories

Chemistry	
Analytical chemistry	Biochemistry
Inorganic chemistry	Materials chemistry
Neurochemistry	Nuclear chemistry
Organic chemistry	Physical chemistry
Other Engineering Areas	
Aerospace Engineering	Agricultural Engineering
Applied Engineering	Biomedical Engineering
Biological Engineering	Building services Engineering
Energy Engineering	Railway Engineering
Industrial Engineering	Mechatronics Engineering
Management Engineering	Military Engineering
Petroleum Engineering	Nuclear Engineering
Textile Engineering	Nano Engineering
Algorithm and Computational Complexity	Artificial Intelligence
Electronics & Communication Engineering	Image Processing
Information Retrieval	Low Power VLSI Design
Neural Networks	Plastic Engineering

Each article in this issue provides an example of a concrete industrial application or a case study of the presented methodology to amplify the impact of the contribution. We are very thankful to everybody within that community who supported the idea of creating a new Research with IJOER. We are certain that this issue will be followed by many others, reporting new developments in the Engineering and Science field. This issue would not have been possible without the great support of the Reviewer, Editorial Board members and also with our Advisory Board Members, and we would like to express our sincere thanks to all of them. We would also like to express our gratitude to the editorial staff of AD Publications, who supported us at every stage of the project. It is our hope that this fine collection of articles will be a valuable resource for *IJOER* readers and will stimulate further research into the vibrant area of Engineering and Science Research.

Mukesh Arora

(Chief Editor)

Board Members

Dr. M. Varatha Vijayan

Annauniversity Rank Holder, Commissioned Officer Indian Navy, Ncc Navy Officer (Ex-Serviceman Navy), Best Researcher Awardee, Best Publication Awardee, Tamilnadu Best Innovation & Social Service Awardee From Lions Club.

Dr. Mohamed Abdel Fatah Ashabrawy Moustafa

PhD. in Computer Science - Faculty of Science - Suez Canal University University, 2010, Egypt.

Assistant Professor Computer Science, Prince Sattam bin AbdulAziz University ALkharj, KSA.

Prof.S.Balamurugan

Dr S. Balamurugan is the Head of Research and Development, Quants IS & CS, India. He has authored/co-authored 35 books, 200+ publications in various international journals and conferences and 6 patents to his credit. He was awarded with Three Post-Doctoral Degrees - Doctor of Science (D.Sc.) degree and Two Doctor of Letters (D.Litt) degrees for his significant contribution to research and development in Engineering.

Dr. Mahdi Hosseini

Dr. Mahdi did his Pre-University (12^{th}) in Mathematical Science. Later he received his Bachelor of Engineering with Distinction in Civil Engineering and later he Received both M.Tech. and Ph.D. Degree in Structural Engineering with Grade "A" First Class with Distinction.

Dr. Anil Lamba

Practice Head – Cyber Security, EXL Services Inc., New Jersey USA.

Dr. Anil Lamba is a researcher, an innovator, and an influencer with proven success in spearheading Strategic Information Security Initiatives and Large-scale IT Infrastructure projects across industry verticals. He has helped bring about a profound shift in cybersecurity defense. Throughout his career, he has parlayed his extensive background in security and a deep knowledge to help organizations build and implement strategic cybersecurity solutions. His published researches and conference papers has led to many thought provoking examples for augmenting better security.

Dr. Ali İhsan KAYA

Currently working as Associate Professor in Mehmet Akif Ersoy University, Turkey.

Research Area: Civil Engineering - Building Material - Insulation Materials Applications, Chemistry - Physical Chemistry – Composites.

Dr. Parsa Heydarpour

Ph.D. in Structural Engineering from George Washington University (Jan 2018), GPA=4.00.

Dr. Heba Mahmoud Mohamed Afify

Ph.D degree of philosophy in Biomedical Engineering, Cairo University, Egypt worked as Assistant Professor at MTI University.

Dr. Kalpesh Sunil Kamble (Ph.D., P.Eng., M.Tech, B.E. (Mechanical))

A distinguished academic with a Ph.D. in Mechanical Engineering and 13 Years of extensive teaching and research experience. He is currently a Assistant professor at the SSPM's COE, Kankavli and contributes to several undergraduate and masters programs across Mahrashtra, India.

Dr. Aurora Angela Pisano

Ph.D. in Civil Engineering, Currently Serving as Associate Professor of Solid and Structural Mechanics (scientific discipline area nationally denoted as ICAR/08"–"Scienza delle Costruzioni"), University Mediterranea of Reggio Calabria, Italy.

Dr. Faizullah Mahar

Associate Professor in Department of Electrical Engineering, Balochistan University Engineering & Technology Khuzdar. He is PhD (Electronic Engineering) from IQRA University, Defense View, Karachi, Pakistan.

Prof. Viviane Barrozo da Silva

Graduated in Physics from the Federal University of Paraná (1997), graduated in Electrical Engineering from the Federal University of Rio Grande do Sul - UFRGS (2008), and master's degree in Physics from the Federal University of Rio Grande do Sul (2001).

Dr. S. Kannadhasan

Ph.D (Smart Antennas), M.E (Communication Systems), M.B.A (Human Resources).

Dr. Christo Ananth

Ph.D. Co-operative Networks, M.E. Applied Electronics, B.E Electronics & Communication Engineering Working as Associate Professor, Lecturer and Faculty Advisor/ Department of Electronics & Communication Engineering in Francis Xavier Engineering College, Tirunelveli.

Dr. S.R.Boselin Prabhu

Ph.D, Wireless Sensor Networks, M.E. Network Engineering, Excellent Professional Achievement Award Winner from Society of Professional Engineers Biography Included in Marquis Who's Who in the World (Academic Year 2015 and 2016). Currently Serving as Assistant Professor in the department of ECE in SVS College of Engineering, Coimbatore.

Dr. Balasubramanyam, N

Dr.Balasubramanyam, N working as Faculty in the Department of Mechanical Engineering at S.V.University College of Engineering Tirupati, Andhra Pradesh.

Dr. PAUL P MATHAI

Dr. Paul P Mathai received his Bachelor's degree in Computer Science and Engineering from University of Madras, India. Then he obtained his Master's degree in Computer and Information Technology from Manonmanium Sundaranar University, India. In 2018, he received his Doctor of Philosophy in Computer Science and Engineering from Noorul Islam Centre for Higher Education, Kanyakumari, India.

Dr. M. Ramesh Kumar

Ph.D (Computer Science and Engineering), M.E (Computer Science and Engineering).

Currently working as Associate Professor in VSB College of Engineering Technical Campus, Coimbatore.

Dr. Maheshwar Shrestha

Postdoctoral Research Fellow in DEPT. OF ELE ENGG & COMP SCI, SDSU, Brookings, SD Ph.D, M.Sc. in Electrical Engineering from SOUTH DAKOTA STATE UNIVERSITY, Brookings, SD.

Dr. D. Amaranatha Reddy

Ph.D. (Postdocteral Fellow, Pusan National University, South Korea), M.Sc., B.Sc. : Physics.

Dr. Dibya Prakash Rai

Post Doctoral Fellow (PDF), M.Sc., B.Sc., Working as Assistant Professor in Department of Physics in Pachhuncga University College, Mizoram, India.

Dr. Pankaj Kumar Pal

Ph.D R/S, ECE Deptt., IIT-Roorkee.

Dr. P. Thangam

PhD in Information & Communication Engineering, ME (CSE), BE (Computer Hardware & Software), currently serving as Associate Professor in the Department of Computer Science and Engineering of Coimbatore Institute of Engineering and Technology.

Dr. Pradeep K. Sharma

PhD., M.Phil, M.Sc, B.Sc, in Physics, MBA in System Management, Presently working as Provost and Associate Professor & Head of Department for Physics in University of Engineering & Management, Jaipur.

Dr. R. Devi Priya

Ph.D (CSE), Anna University Chennai in 2013, M.E, B.E (CSE) from Kongu Engineering College, currently working in the Department of Computer Science and Engineering in Kongu Engineering College, Tamil Nadu, India.

Dr. Sandeep

Post-doctoral fellow, Principal Investigator, Young Scientist Scheme Project (DST-SERB), Department of Physics, Mizoram University, Aizawl Mizoram, India- 796001.

Dr. Roberto Volpe

Faculty of Engineering and Architecture, Università degli Studi di Enna "Kore", Cittadella Universitaria, 94100 – Enna (IT).

Dr. S. Kannadhasan

Ph.D (Smart Antennas), M.E (Communication Systems), M.B.A (Human Resources).

Research Area: Engineering Physics, Electromagnetic Field Theory, Electronic Material and Processes, Wireless Communications.

Mr. Bhavinbhai G. Lakhani

An expert in Environmental Technology and Sustainability, with an M.S. from NYIT. Their specialization includes Construction Project Management and Green Building. Currently a Project Controls Specialist Lead at DACK Consulting Solutions, they manage project schedules, resolve delays, and handle claim negotiations. Prior roles as Senior Project Manager at FCS Group and Senior Project Engineer at KUNJ Construction Corp highlight their extensive experience in project estimation, resource management, and on-site supervision.

Mr. Omar Muhammed Neda

Department of Electrical Power Engineering, Sunni Diwan Endowment, Iraq.

Mr. Amit Kumar

Amit Kumar is associated as a Researcher with the Department of Computer Science, College of Information Science and Technology, Nanjing Forestry University, Nanjing, China since 2009. He is working as a State Representative (HP), Spoken Tutorial Project, IIT Bombay promoting and integrating ICT in Literacy through Free and Open Source Software under National Mission on Education through ICT (NMEICT) of MHRD, Govt. of India; in the state of Himachal Pradesh, India.

Mr. Tanvir Singh

Tanvir Singh is acting as Outreach Officer (Punjab and J&K) for MHRD Govt. of India Project: Spoken Tutorial - IIT Bombay fostering IT Literacy through Open Source Technology under National Mission on Education through ICT (NMEICT). He is also acting as Research Associate since 2010 with Nanjing Forestry University, Nanjing, Jiangsu, China in the field of Social and Environmental Sustainability.

Mr. Abilash

MTech in VLSI, BTech in Electronics & Telecommunication engineering through A.M.I.E.T.E from Central Electronics Engineering Research Institute (C.E.E.R.I) Pilani, Industrial Electronics from ATI-EPI Hyderabad, IEEE course in Mechatronics, CSHAM from Birla Institute Of Professional Studies.

Mr. Varun Shukla

M.Tech in ECE from RGPV (Awarded with silver Medal By President of India), Assistant Professor, Dept. of ECE, PSIT, Kanpur.

Mr. Shrikant Harle

Presently working as a Assistant Professor in Civil Engineering field of Prof. Ram Meghe College of Engineering and Management, Amravati. He was Senior Design Engineer (Larsen & Toubro Limited, India).

Mr. Zairi Ismael Rizman

Senior Lecturer, Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM) (Terengganu) Malaysia Master (Science) in Microelectronics (2005), Universiti Kebangsaan Malaysia (UKM), Malaysia. Bachelor (Hons.) and Diploma in Electrical Engineering (Communication) (2002), UiTM Shah Alam, Malaysia.

Mr. Ronak

Qualification: M.Tech. in Mechanical Engineering (CAD/CAM), B.E.

Presently working as a Assistant Professor in Mechanical Engineering in ITM Vocational University, Vadodara. Mr. Ronak also worked as Design Engineer at Finstern Engineering Private Limited, Makarpura, Vadodara.

Table of Contents

Volume-12, Issue-8, August 2026

International Journal of Engineering Research & Science (IJOER)
ISSN: [2395-6992]
[Vol-12, Issue-8, August - 2026]
Website: www.ijoer.com
Journal DOI: 10.25125/engineering-journal

House Price Prediction in Nanjing Based on Machine Learning and SHAP Interpretability

Yang Zaibin[1]; Cui Xinxin[2*]

[1,2]School of Mathematics and Statistics, Guizhou University of Finance and Economics, Guiyang 550025, China.
[1,2]School of Big Data Statistics, Guizhou University of Finance and Economics, Guiyang 550025, China.
*Corresponding Author

Received: 24 July 2026/ Revised: 01 August 2026/ Accepted: 11 August 2026/ Published: 15-08-2026

Abstract— *Accurate house price prediction is of great significance for homebuyers, developers, and policymakers, yet the complexity and non-linearity of housing markets pose substantial challenges to traditional linear models. This study addresses the dual goals of high predictive accuracy and high interpretability by systematically comparing five regression models—linear regression, random forest, XGBoost, LightGBM, and multilayer perceptron (MLP)—on a large-scale second-hand housing dataset from Nanjing, China. Model performance is evaluated using MAE, RMSE, and R^2. Results show that ensemble tree models significantly outperform linear regression and MLP, with LightGBM and XGBoost achieving the lowest MAE (6,200 RMB/m²) and RMSE (8,800 RMB/m²), and LightGBM attaining the highest R^2 (0.92). We further employ SHAP (SHapley Additive exPlanations) to open the "black box" of the optimal LightGBM model. Global feature importance reveals that house age, subway distance, and decoration condition are the three most influential drivers of Nanjing house prices. SHAP summary plots demonstrate that newer houses, closer subway proximity, and better decoration consistently raise prices, while area exhibits a non-monotonic effect—small units command higher unit prices under total-price constraints, whereas very large luxury properties show diminishing marginal value. Prediction diagnostics confirm robust performance in the mainstream price range (15,000–35,000 RMB/m²), though extreme high-end properties remain harder to predict due to unobserved idiosyncratic factors. This study not only provides a high-performance predictive tool for the Nanjing market but also offers transparent, actionable insights into the underlying price-formation mechanism. The integrated framework of machine learning plus SHAP interpretability is readily generalisable to other cities and real-estate contexts, supporting evidence-based decision-making and policy design.*

Keywords— House price prediction; Machine learning;

I. INTRODUCTION

Housing issues are central to national welfare and people's livelihoods, and the stable operation of the real estate market plays a pivotal role in economic and social development. As a major central city in the Yangtze River Delta, Nanjing has witnessed persistently active real estate markets in recent years, with house prices influenced by multiple intertwined factors, exhibiting pronounced regional differentiation and volatility. Against this backdrop, constructing a scientific and accurate house price prediction model not only provides a basis for homebuyer decision making and developer pricing, but also assists government agencies in grasping market dynamics and refining regulatory measures. This holds considerable practical significance and research value.

Research on house price prediction can be traced back to the hedonic price theory proposed by Rosen [16], which posits that the value of a heterogeneous good can be explained by the implicit prices of its constituent attributes. This framework became the cornerstone of a large body of empirical work. Building on this foundation, traditional approaches predominantly employ multiple linear regression to estimate the marginal effects of individual features, offering strong interpretability. However, real world house price formation often involves complex nonlinear interactions—for example, the synergy between floor area and

location, or the non-monotonic influence of floor level. Linear models inherently struggle to capture such relationships, and their predictive accuracy often falls short of growing practical demands [3].

With advances in computational power and the proliferation of data sources, machine learning models have gradually been introduced into real estate valuation. Selim [18] compared the predictive performance of artificial neural networks and hedonic price models on Turkish housing data, finding that neural networks yielded significantly lower prediction errors than traditional regression, thereby pioneering the application of nonlinear models in this domain. Subsequently, various ensemble learning and gradient boosting methods have been widely explored for house price prediction: Breiman [1] proposed random forests, which alleviate overfitting through bootstrap aggregation and random feature selection; Friedman [7] established the general framework of gradient boosting decision trees; XGBoost by Chen and Guestrin [2] and LightGBM by Ke et al. [10] further optimised the efficiency and accuracy of gradient boosting, demonstrating exceptional performance in structured data regression tasks. Meanwhile, simple deep learning networks such as multilayer perceptrons have offered another nonlinear benchmark [4]. Nevertheless, these high accuracy models are often regarded as "black boxes" whose internal decision making processes are not intuitively understandable, which to some extent constrains their trustworthiness and acceptance in high stakes decisions [5]. Recent systematic reviews have consolidated these advances, revealing that Random Forest, Gradient Boosting Machine, XGBoost, and LightGBM are the most frequently employed models in real estate price prediction, consistently outperforming traditional linear approaches. However, the proliferation of these high-accuracy models has intensified a fundamental tension: while ensemble methods excel at capturing complex nonlinearities and interaction effects, their internal decision logic remains opaque, prompting scholars to call for integrating explainable artificial intelligence techniques into mass real estate valuation systems [6].

In response, researchers have increasingly turned to SHAP to bridge the gap between predictive accuracy and interpretability. For instance, Hu et al. [8] combined XGBoost with SHAP to examine the nonlinear effects of public service amenities and street-view greenery on housing prices in Shanghai, revealing that residents paid a premium of approximately 2,000 yuan/m² for higher green views [9]. Wang and Li [19] employed Random Forest with SHAP to quantify feature contributions in a rapidly urbanizing Chinese city, identifying location-based factors and environmental attributes as the most influential drivers. Ye [20] demonstrated that XAI approaches should be systematically integrated into mass real estate appraisal and urban housing research more broadly. Internationally, studies on Seoul apartment transactions and Belgian residential rents have similarly validated the effectiveness of SHAP for extracting transparent, actionable insights from tree-based models.

The advantages of SHAP over alternative interpretability methods are substantial. Unlike LIME, which provides only local approximations around individual predictions, SHAP offers both global and local explanations with desirable theoretical properties—local accuracy, consistency, and missingness—making it particularly suitable for high-stakes real estate decision-making [10]. The TreeSHAP algorithm further enables efficient and exact computation for tree-based ensembles, which is precisely the algorithmic family that dominates structured data regression tasks [11]. Moreover, recent research has extended SHAP to capture spatial heterogeneity and nonlinear threshold effects, demonstrating its versatility across diverse urban contexts.

Ribeiro et al. [15] proposed LIME (Local Interpretable Model-agnostic Explanations), which approximates the decision boundary of a complex model using a local surrogate model. Building on this foundation and seeking a more unified and theoretically grounded framework, Lundberg and Lee [11] introduced SHAP based on Shapley values from cooperative game theory, assigning each feature a unified contribution to the prediction, with desirable properties including local accuracy, consistency, and missingness. In particular, the TreeSHAP algorithm for tree-based ensembles enables efficient and exact computation of SHAP values, providing a powerful tool for opening the "black box" of complex models. Although SHAP has been successfully applied in fields such as healthcare and credit scoring, its use in house price prediction—especially for the Nanjing market—remains relatively scarce.

Existing domestic house price prediction literature mostly focuses on first tier cities such as Beijing and Shanghai, and most studies stop at comparing model prediction accuracy, with limited in depth interpretation of how features affect predictions [13]. As a key central city in eastern China, the driving factors and mechanisms of house prices in Nanjing warrant dedicated investigation [14]. Therefore, this study takes actual listing data from Nanjing as the object, selects five representative regression models—linear regression, random forest, XGBoost, LightGBM, and MLP—and systematically compares them within a unified framework [17]. Innovatively, we introduce SHAP for both global and local interpretability analysis of the best performing model, aiming to balance the dual objectives of "high accuracy prediction" and "high interpretability", and to

reveal the magnitude and pathways of each feature's influence on Nanjing house prices, thereby providing more reliable references for market participants and policymakers [15].

II. RELEVANT THEORIES AND TECHNOLOGIES

2.1 General Workflow of House Price Prediction

The task of house price prediction based on machine learning is essentially a regression modelling problem from housing attributes to prices. The standard workflow typically comprises the following tightly connected stages:

1) **Problem definition and data collection:** First, clarify whether the target is price per unit area or total price, and collect sample data with sufficient feature information accordingly. This study uses second hand housing listing data from the Nanjing market, where each sample contains attributes such as district, area, room layout, floor level, and decoration condition.

2) **Data preprocessing:** Raw data often suffer from missing values, outliers, and inconsistent formats. Cleaning steps—such as imputation or removal of missing values, outlier detection and handling, and data type conversion—are needed to ensure reliability for subsequent modelling.

3) **Feature engineering:** Construct more expressive new features from the original ones (e.g., extract the number of bedrooms and living rooms from room layout strings, calculate house age from construction year), encode categorical features numerically (one-hot or label encoding), scale or normalise numerical features, and perform feature selection based on correlation and importance to improve model performance.

4) **Model construction and training:** Split the data into training and test sets, select appropriate regression models, tune hyperparameters via cross validation and optimisation techniques (e.g., grid search or random search), and learn the mapping from features to house prices on the training set.

5) **Model evaluation and comparison:** Compute metrics such as mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2) on the independent test set to objectively compare the predictive accuracy of each model and select the best performer.

6) **Model interpretation and application:** For the final selected model, employ interpretability methods (e.g., SHAP) to analyse the influence mechanisms of features, transforming data driven model results into domain understandable knowledge to support decision making.

2.2 Model Principles

To systematically compare different model architectures for Nanjing house price prediction, we select five representative regression models: linear regression, random forest, XGBoost, LightGBM, and multilayer perceptron (MLP). Their principles are briefly introduced below.

2.2.1 Linear Regression

Multiple linear regression assumes a linear relationship between the dependent variable y and the feature vector $x = [x_1, x_2, ..., x_p]^T$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon = \boldsymbol{\beta}^T \mathbf{x} + \varepsilon \quad (1)$$

where β_0 is the intercept, β_j is the regression coefficient for the j-th feature, and ε is the random error term, usually assumed to be normally distributed with zero mean. Parameters are estimated by minimising the residual sum of squares (ordinary least squares, OLS).

When features are highly correlated (multicollinearity), OLS estimates may become unstable. Ridge regression addresses this by adding an ℓ_2 penalty term to the loss function:

$$\widehat{\boldsymbol{\beta}}^{ridge} = arg\min_{\beta} \{\sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\} \quad (2)$$

where $\lambda \geq 0$ is the regularisation parameter controlling the penalty strength. Linear models are simple and provide intuitive coefficient interpretations, but they struggle to capture nonlinear relationships and feature interactions.

2.2.2 Random Forest

Random forest is a decision-tree-based ensemble method that combines bagging (bootstrap aggregating) and random subspace ideas [1]. The construction process is as follows:

- From the original training set, draw B bootstrap samples with replacement; each subsample trains one decision tree.
- At each node split, randomly select $m \approx \sqrt{p}$ features from all p features as candidates, and choose the best split among them.
- Each tree is grown fully without pruning.

For regression, the final prediction is the average of all tree predictions:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x}) \tag{3}$$

The dual randomness (sample and feature) significantly reduces variance, mitigates overfitting of a single tree, and provides robustness to outliers and noise. Additionally, the model directly outputs feature importance based on the reduction in node impurity.

2.2.3 XGBoost

XGBoost is an efficient implementation of gradient-boosted decision trees proposed by Chen and Guestrin [2]. Traditional gradient boosting uses a forward stagewise approach, fitting the current residual at each step. XGBoost introduces several key improvements:

- **Second-order Taylor expansion of the objective:** expands the loss function to second order, using both first- and second-order gradients to guide splits, leading to more accurate optimisation.
- **Regularisation:** adds penalties on leaf weights (L_2 regularisation) and the number of leaves to control model complexity and prevent overfitting.
- **Shrinkage:** multiplies newly added trees by a learning rate less than 1, reducing each tree's contribution and increasing robustness.
- **Approximate splitting and parallel processing:** pre-sorts feature values or uses quantile-based approximate search to accelerate split-point finding, and supports parallel computation.

2.2.4 LightGBM

LightGBM (Light Gradient Boosting Machine) is another gradient boosting framework developed by Ke et al. [10], specifically optimised for training efficiency and memory consumption. It employs two novel techniques: Gradient-based One-Side Sampling (GOSS) to retain instances with large gradients and randomly downsample those with small gradients, and Exclusive Feature Bundling (EFB) to bundle mutually exclusive sparse features, thereby reducing dimensionality. It also uses a leaf-wise tree growth strategy with depth constraints, which typically converges faster than the level-wise approach of XGBoost.

2.2.5 Multilayer Perceptron

A multilayer perceptron is a basic feedforward neural network. It consists of an input layer, one or more hidden layers, and an output layer, with fully connected neurons between adjacent layers. During forward propagation, each neuron's output is a nonlinear activation of the weighted sum of its inputs:

$$\mathbf{h}^{(l)} = \sigma^{(l)}\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right) \tag{4}$$

where $h^{(l-1)}$ is the output of layer l-1, $W^{(l)}$ and $b^{(l)}$ are the weight matrix and bias vector of layer l, and $\sigma^{(l)}$ is the activation function (commonly ReLU: $\sigma(z) = \max(0, z)$). The output layer for regression typically uses no activation, directly outputting the linear combination.

Network parameters are updated via backpropagation and gradient-based optimisers (e.g., Adam) to minimise the mean squared error loss. By the universal approximation theorem, an MLP with sufficient hidden neurons can approximate continuous functions to arbitrary accuracy, making it a general benchmark for nonlinear regression. However, the internal weight distributions are not easily interpretable, so the model is often viewed as a black box.

2.3 Model Evaluation Metrics

To quantitatively assess and compare the performance of each regression model, we adopt the following three metrics:

1) **Mean Absolute Error (MAE):**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

It directly reflects the average magnitude of absolute deviations between predictions and actual values, in the same units as the target (e.g., RMB/m²), facilitating intuitive interpretation of prediction errors.

2) Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

RMSE assigns heavier penalties to larger errors and is sensitive to prediction stability. It is also in the original scale, but typically somewhat larger than MAE.

3) Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

R^2 measures the proportion of variance in the target variable explained by the model. The closer to 1, the better the fit. A value of 0 indicates that the model performs no better than simply using the mean.

2.4 SHAP for Interpretability

Although ensemble models and neural networks often achieve high predictive accuracy, their internal decision logic is not intuitively understandable. To open these "black boxes", we introduce SHAP for model interpretation. SHAP, proposed by Lundberg and Lee [11], is based on Shapley values from cooperative game theory and assigns each feature a contribution value such that, for any instance, the model's prediction can be decomposed as the base value plus the sum of all feature contributions:

$$\hat{f}(x) = \phi_0 + \sum_{j=1}^{M}\phi_j \tag{8}$$

where φ_0 is the average prediction over all training samples (the baseline), and φ_j is the SHAP value of the j-th feature, indicating the magnitude and direction relative to the baseline.

SHAP values satisfy three desirable theoretical properties:

1) **Local accuracy:** the sum of the explanations equals the model's prediction for that instance.
2) **Missingness:** a feature that is not used contributes zero.
3) **Consistency:** if a model changes so that a feature's marginal contribution increases, its SHAP value does not decrease.

III. DATA SOURCE AND FEATURE ENGINEERING

3.1 Data Source and Original Data Structure

The empirical data for this study focus on the second hand housing market in Nanjing. The raw data were obtained from the Kaggle platform. The initially acquired dataset had a shape of (1,048,575, 22). After preliminary inspection, we found a large number of completely empty rows. After removing these invalid empty rows, the effective sample size was 41,331 listing records. These 41,331 raw records cover 22 feature dimensions with rich micro level information about the properties. The specific features are summarised in Table 1.

TABLE 1
FEATURE SYSTEM FOR MODELLING

Feature Category	Feature Name	Variable ID	Description
Physical attributes	Floor area	area_sqm	Continuous numerical, unit: m²
	Number of bedrooms	room_num	Discrete numerical, unit: count
	Number of living rooms	hall_num	Discrete numerical, unit: count
	Number of bathrooms	bathroom_num	Discrete numerical, unit: count
Market popularity	Number of followers	attention_count	Discrete numerical, reflecting page views
	Viewings in 30 days	visit_30_days	Discrete numerical, reflecting recent purchase intent
Location & transport	District / area	city_encoded	Categorical, label encoded to numeric
	Distance to subway	subway_dist	Continuous numerical, extracted from text using regex, unit: metres (missing filled with 9999)
Building characteristics	Total floors	floor_total	Discrete numerical, unit: floors
	Decoration condition	decoration_encoded	Categorical, label encoded (luxury, fine, simple, rough, etc.)
Construction features	House age	house_age	Continuous numerical, computed as 2026 − build_year, unit: years
	Total rooms	total_rooms	Discrete numerical, computed as room_num + hall_num + bathroom_num
	Relative floor level	floor_type	Discrete encoded: low=0, middle=1, high=2
	Orientation	ori_*	Categorical, one-hot encoded into binary dummies (e.g., ori_NS, ori_S, etc.)

3.2 Data Cleaning

Since the raw data contained noise, missing values, and unstructured information, rigorous cleaning was performed before in-depth analysis and modelling. The steps were as follows:

1) **Removal of redundant columns:** we dropped unique identifiers (id, house_id) and long-text description columns (p1_desc, p2_desc) that were not useful for regression modelling, reducing the feature dimension from 22 to 18.
2) **Missing value imputation:** we addressed features with missing entries. For example, build_year had 542 missing values; these were imputed with the median construction year of listings in the same community or business district. For subway_info (6,436 missing values), because missingness typically indicated no nearby subway coverage, we left it for unified transformation in the feature engineering stage.
3) **Outlier filtering:** based on business logic and statistical principles (e.g., extreme floor area anomalies, extreme price per unit outliers), we identified and removed 21 records that severely violated market common sense.
4) After cleaning, 41,310 high-quality samples remained for subsequent feature engineering and modelling.

3.3 Exploratory Data Analysis

To intuitively understand the operational characteristics of the Nanjing second hand housing market, we performed visual exploration of key variables.

3.3.1 Distribution of House Prices

The distribution of the target variable—price per square metre for second hand housing—directly affects modelling difficulty. Figure 1 (histogram with kernel density estimate) shows that the price per square metre in Nanjing exhibits a pronounced right skewed distribution. The market's mainstream prices are heavily concentrated in the range of RMB 15,000–25,000/m², with the peak frequency around RMB 18,000/m².

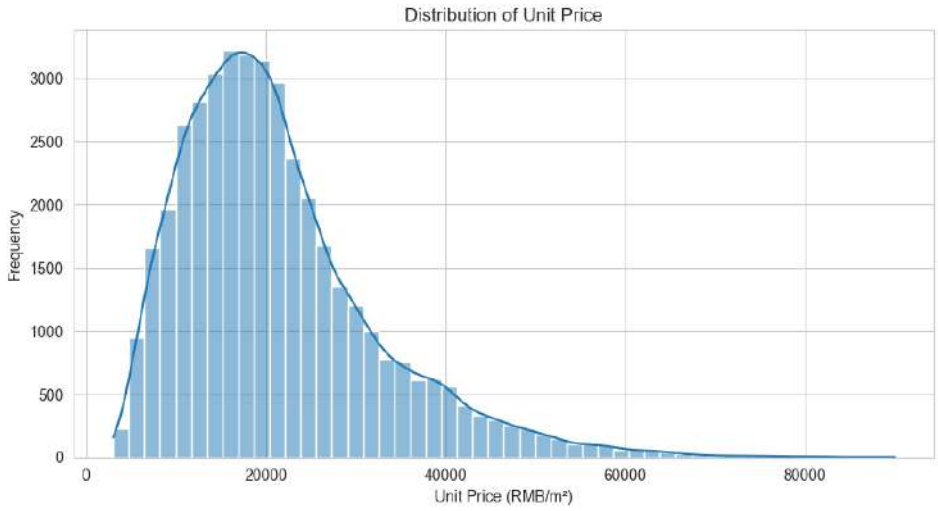


FIGURE 1: Distribution of second hand unit prices

Histogram with kernel density estimate showing right-skewed distribution of house prices per square metre in Nanjing. The distribution shows a long right tail with most values concentrated between 15,000 and 25,000 RMB/m², peaking around 18,000 RMB/m², with a long tail extending beyond 90,000 RMB/m² for luxury properties.

From the boxplot in Figure 2, the median unit price is slightly below RMB 20,000/m², and the interquartile range is mainly between 15,000 and 26,000/m². Notably, there is a dense cluster of outliers above the upper whisker, with maximum values exceeding 90,000/m². This indicates significant internal stratification within the Nanjing market: a small number of prime location premium properties (e.g., top school district houses, high-end improvement properties) pull the overall price ceiling upward. This long tail distribution suggests that log transformation of the target variable or using tree-based models insensitive to outliers may be beneficial.

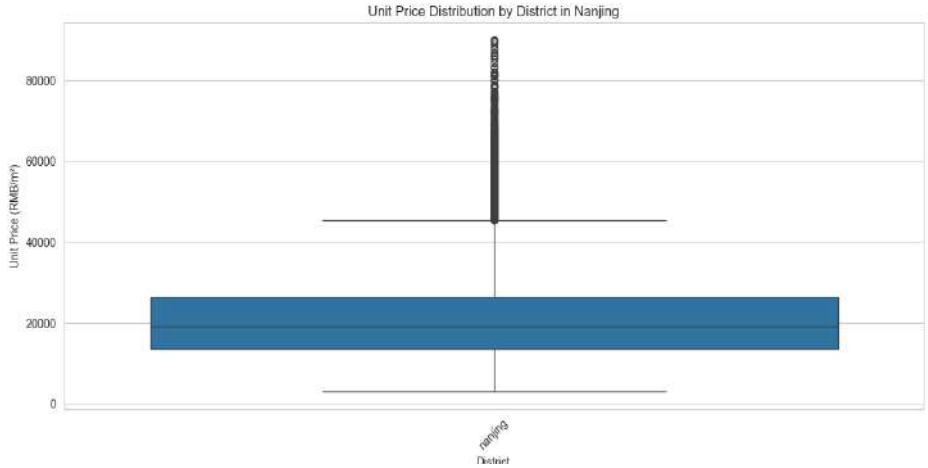


FIGURE 2: Boxplot of second hand unit prices by district in Nanjing

Boxplot showing median unit prices across different districts of Nanjing. The plot illustrates significant variation across districts, with central districts showing higher median prices and wider interquartile ranges, while suburban districts show lower prices with narrower distributions.

3.3.2 Relationship between Area and Price

Floor area is one of the most fundamental physical determinants of price. The scatter plot in Figure 3 reveals a nonlinear relationship. The vast majority of properties are in the 50–150 m² range, covering just-needed and first-time improvement segments. A typical "total price constraint effect" can be observed:

- In the 50–100 m² small- to medium-size segment, the variance of unit price is enormous, ranging from around RMB 10,000/m² (suburban old properties) to as high as RMB 80,000–90,000/m² (extreme outliers).
- When the floor area exceeds 200 or even 300 m², the upper bound of unit price decreases significantly due to buyers' total budget constraints, and the unit price largely stabilises between RMB 20,000 and 60,000/m².

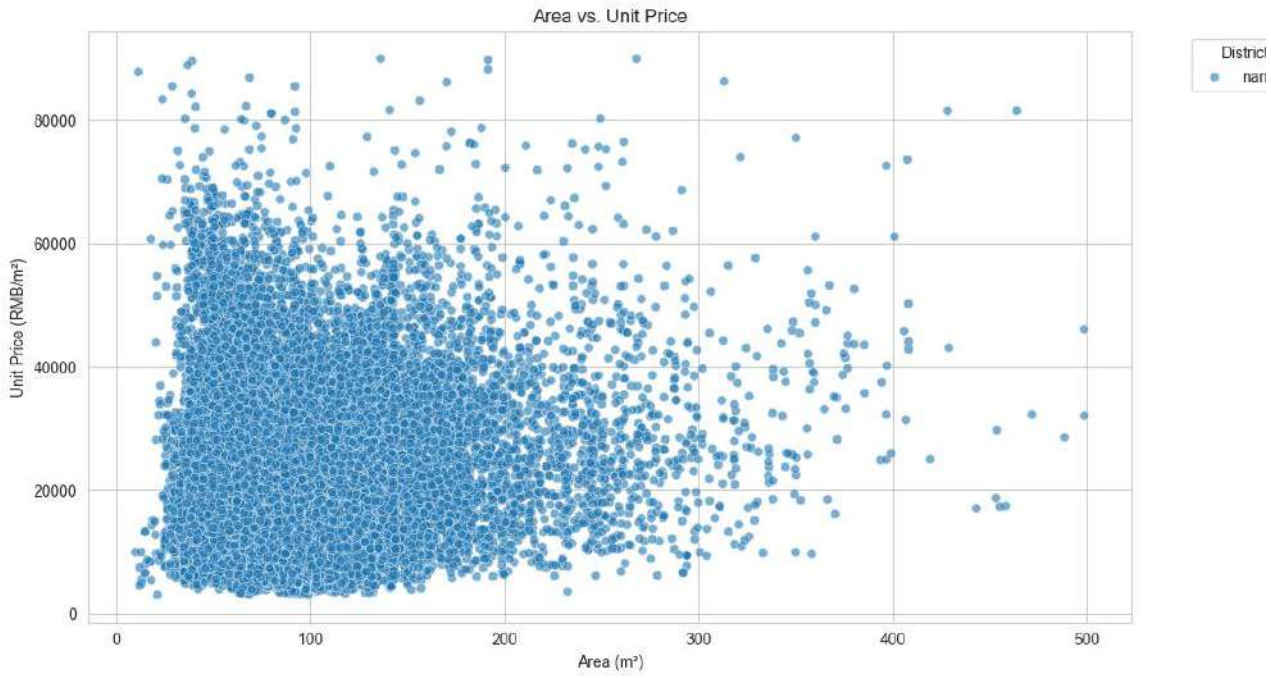


FIGURE 3: Scatter plot of area vs. unit price

Scatter plot showing a nonlinear relationship between floor area (m²) and unit price (RMB/m²). The plot shows high price variance for small units (50-100 m²), with some achieving very high unit prices, while very large units (>200 m²) show lower and more compressed unit prices, consistent with a total-price constraint effect.

3.3.3 Feature Correlation Analysis

To complement the univariate and bivariate visual explorations, we further quantify the linear relationships among all continuous variables (e.g., floor area, house age, subway distance) and ordinal/categorical variables that were label-encoded (e.g., district, decoration condition) using the Pearson correlation coefficient matrix, visualised as a heatmap in Figure 4. This analysis serves three crucial purposes: it provides an early statistical check on which features are most linearly associated with the target variable, it identifies potential multicollinearity among predictors, which is particularly relevant for linear models, and it helps to justify the subsequent choice of tree-based ensemble methods over parametric approaches.

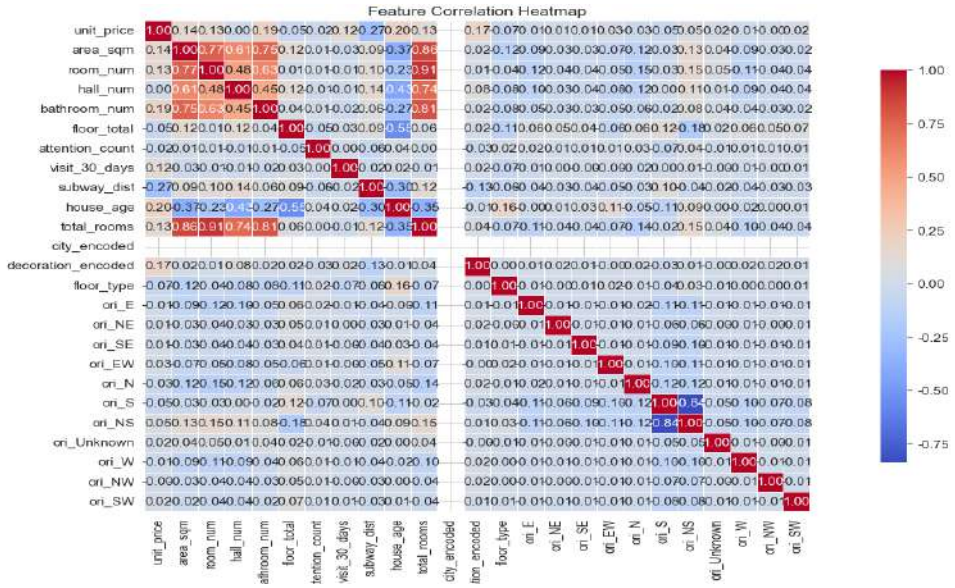


FIGURE 4: Correlation heatmap

Heatmap showing Pearson correlation coefficients between all features and the target variable (unit_price). Key observations: house_age shows the strongest negative correlation (r = -0.25), subway_dist shows negative correlation (r = -0.20), decoration_encoded shows positive correlation (r = 0.17), area_sqm shows weak positive correlation (r = 0.14). Strong multicollinearity is observed among area_sqm, room_num, bathroom_num, and total_rooms (r > 0.7).

The correlation coefficients between the dependent variable unit_price (price per square metre) and each feature reveal a hierarchy of influence that aligns remarkably well with real-estate economic theory and local market intuition. Among all predictors, house_age exhibits the strongest negative correlation with price (r = -0.25). This indicates that, on a linear basis, older properties tend to sell at lower unit prices, a phenomenon widely observed in urban housing markets due to structural depreciation, outdated floor plans, and deteriorating communal facilities. In Nanjing, a city with a rich historical fabric, this effect is particularly pronounced because many older communities lack modern amenities such as elevators or underground parking, which are increasingly demanded by younger, improvement-oriented buyers.

The second most influential feature in terms of absolute correlation is subway_dist (r = -0.20), confirming the substantial "metro premium" in Nanjing's second-hand market. With over ten metro lines in operation and a rapidly expanding network, proximity to a subway station has become a critical determinant of residential desirability, as it directly reduces commuting time and enhances accessibility to employment hubs, shopping centres, and public services. Notably, the negative sign is intuitive: shorter distances (lower numerical values) are associated with higher prices.

Decoration_encoded (r = 0.17) shows a moderate positive correlation, suggesting that better-decorated properties command a price uplift. This finding is consistent with the prevailing improvement-driven demand in Nanjing, where many buyers prefer move-in-ready homes that save both time and renovation costs. The effect, while statistically significant, is weaker than that of house age and subway distance in linear terms, which may be because decoration quality is partly endogenous to other features (e.g., newer buildings tend to have better finishes).

More intriguingly, area_sqm displays only a very weak positive correlation with unit price (r = 0.14). This near-flat linear relationship does not imply that area is unimportant; rather, it strongly suggests a non-monotonic and non-linear association—a pattern already glimpsed in the scatter plot (Figure 3). As previously discussed, under a total-price budget constraint, smaller units (e.g., 50–80 m²) often exhibit disproportionately high per-square-metre prices, whereas ultra-large luxury apartments (above 200 m²) suffer from diminishing marginal values because the total price becomes prohibitive for most buyers. This non-linearity is precisely the kind of complex interaction that linear regression fails to capture, and it provides a statistical justification for why ensemble tree models later outperform OLS by a significant margin (MAE reduction of ~14%).

Other features such as attention_count and visit_30_days show negligible correlations with price (around 0.00), suggesting that online popularity metrics do not linearly translate into higher prices—possibly because they reflect listing exposure rather than intrinsic value. Similarly, floor_total and the orientation dummies exhibit near-zero correlations, indicating that their effects

are either weak, non-linear, or conditional on other factors (e.g., the value of a high floor may depend on the building's surroundings).

3.3.4 Inter-Feature Correlations: Multicollinearity and Its Implications

Turning to the relationships among predictors, the heatmap reveals several strong positive correlations that are physically intuitive but statistically important. The most salient is the cluster around area_sqm, which is highly correlated with room_num ($r = 0.77$), bathroom_num ($r = 0.75$), and total_rooms ($r = 0.80$). This is entirely expected: larger homes typically contain more bedrooms, bathrooms, and overall rooms. Such high collinearity poses a serious challenge for ordinary least squares (OLS) regression, as it inflates the variance of coefficient estimates, making individual parameter interpretations unstable and unreliable. Although ridge regression (which we included in our linear model variant) can mitigate this issue through L_2 regularisation, it does not eliminate the inherent difficulty of disentangling the marginal effect of each room-related variable when they move almost in lockstep.

Similarly, area_sqm shows moderate correlations with house_age (negative, around -0.20) and subway_dist (negative, around -0.15), implying that newer and more central properties tend to be slightly larger, though the relationships are not strong enough to cause severe multicollinearity. The orientation dummies, as expected, exhibit low inter-correlations because they are mutually exclusive categories after one-hot encoding.

Importantly, tree-based models such as Random Forest, XGBoost, and LightGBM are largely immune to multicollinearity. Their split-based learning mechanism evaluates each feature independently at each node, and the presence of correlated features does not degrade predictive performance; it may even improve robustness by providing multiple correlated signals. Consequently, we do not apply dimensionality reduction techniques like PCA (which would sacrifice interpretability) nor do we drop any of the highly correlated room-related features. Instead, we retain all original and engineered variables to preserve their physical meaning, because our ultimate goal is not just prediction but also transparent interpretation via SHAP. The correlation matrix, however, serves as a cautionary reminder that when we later interpret SHAP values, the contributions of correlated features should be considered jointly rather than in isolation.

3.3.5 Linking Correlation to Model Selection and SHAP Interpretability

The correlation analysis provides a solid empirical foundation for our methodological choices. First, the weak and non-linear relationship between unit_price and area_sqm strongly argues against a purely linear specification. Second, the presence of multicollinearity further undermines the reliability of linear regression coefficients, even though our regularised variant (ridge) partially addresses this. Third, the fact that house_age and subway_dist emerge as the top two linear correlates with price foreshadows their dominance in the SHAP global importance ranking (see Section 4.3), lending cross-validation credibility to both statistical and machine-learning perspectives. In essence, the correlation matrix not only summarises the data's linear structure but also reinforces the narrative that ensemble trees, with their inherent ability to model interactions and non-linear thresholds, are the most appropriate tools for this housing market.

Moreover, the near-zero correlations between many features and the target do not imply that these features are irrelevant; they may have conditional effects that only manifest in sub-populations or in combination with other variables. For instance, the impact of floor level (floor_type) on price could be positive in high-rise buildings with scenic views but negative in older walk-up buildings. Such interaction effects are precisely what gradient boosting machines can capture through their hierarchical splitting, and what SHAP can later expose through individual contribution analysis. Thus, the correlation heatmap, while limited to linear associations, serves as a valuable preliminary diagnostic that complements the more sophisticated non-linear explorations presented later.

In summary, Figure 4 reinforces the following key insights: (1) house age, subway distance, and decoration condition are the most linearly influential predictors, (2) area has a weak linear effect but is likely to exhibit non-linear behaviour, (3) strong collinearity among size-related features validates our decision to favour tree-based models over linear ones, and (4) the overall correlation structure is consistent with the economic intuition of the Nanjing housing market, thereby enhancing the credibility of our subsequent modelling and interpretation pipeline.

3.4 Feature Engineering and Final Dataset Construction

Based on the patterns revealed in EDA and the structure of the raw data, we performed deep feature derivation and encoding on the 18 cleaned features to maximise information extraction for machine learning models.

3.4.1 Feature Construction

Raw real estate data typically contain information that is either static, coarse grained, or embedded in unstructured text, making it suboptimal for direct machine learning input. To extract the maximum predictive signal and to encode domain knowledge, we performed a deliberate feature engineering step that transforms the original 18 cleaned attributes into a richer, more expressive set of 24 predictors. This process involved deriving new variables that capture temporal dynamics, spatial aggregation, textual information, and categorical hierarchies—all of which are grounded in hedonic pricing theory and the specific characteristics of the Nanjing housing market. Below we detail each newly constructed feature, its computational logic, missing value treatment, and the economic rationale underpinning its inclusion.

- **total_rooms:** Computed as the sum of room_num, hall_num, and bathroom_num. This aggregate feature captures the overall spatial capacity of the property and reduces the dimensionality of the three separate room-type variables without losing predictive information.
- **house_age:** Computed as 2026 − build_year. The year 2026 corresponds to the data collection period. House age captures the depreciation effect—older properties typically command lower prices due to structural ageing, outdated designs, and deteriorating community facilities.
- **subway_dist:** Extracted from the text description field using regular expressions to identify numerical distances (in metres) to the nearest subway station. Missing values (indicating no nearby subway) were imputed with 9,999 to represent effectively infinite distance.
- **floor_type:** Derived from the original floor information, encoded into three ordered categories: low (0), middle (1), and high (2), capturing the relative vertical position within the building.

Orientation encoding: The original orientation field (e.g., "south", "north-south", "east") was expanded through one-hot encoding into 11 binary dummy variables: ori_E, ori_NE, ori_SE, ori_EW, ori_N, ori_S, ori_NS, ori_unknown, ori_W, ori_NW, and ori_SW, preserving all categorical information without imposing ordinal assumptions.

3.4.2 Encoding of Categorical Features

With the continuous and ordinal features constructed, the remaining categorical variables require numerical transformation to be compatible with machine learning algorithms, and we adopt a differentiated encoding strategy that respects the intrinsic nature of each feature.

For orientation, which possesses no natural ordering—a south-facing unit is not quantitatively "greater" or "less" than an east-facing one—we apply one-hot encoding, expanding the original field into 11 binary dummy variables so that each direction contributes independently to the prediction without imposing a false hierarchy; this is particularly important in Nanjing's subtropical monsoon climate, where south-facing units command a well-documented premium while other orientations exhibit more nuanced preferences.

In contrast, both the district/area and decoration condition features exhibit clear ordinal structures that make label encoding more appropriate and efficient. For districts, we leverage the empirically observable price hierarchy across Nanjing's administrative areas—with Gulou and Qinhuai at the top, Jianye and Xuanwu in the middle tier, and Jiangning, Pukou, and other suburban districts at the lower end—and map each district to an integer label (city_encoded) based on its median unit price on the training set, thereby preserving the ranking while avoiding the dimensional explosion that one-hot encoding would incur.

Similarly, decoration quality follows an unambiguous progression from rough through simple and fine to luxury, which we label-encode as decoration_encoded with values 0 through 3, respectively, allowing tree-based models to discover optimal split points along this quality gradient.

Although label encoding implicitly assumes ordinality and could be problematic for linear models, our primary framework of ensemble trees is entirely robust to this treatment, as these algorithms treat the integer labels merely as ordered thresholds for partitioning rather than as cardinal measurements with equal intervals. This dual encoding scheme successfully transforms all textual attributes into a compact numerical matrix of 24 predictors, balancing information preservation, model compatibility, and interpretability, and we ensure that all encoding mappings are defined exclusively on the training set and then applied unchanged to the test set to prevent data leakage.

3.4.3 Final Modelling Dataset Shape

Following the aforementioned data cleaning and feature engineering pipeline, the raw data were successfully transformed into a standardised and well-structured machine learning input matrix, laying a solid foundation for subsequent model training and evaluation. The final modelling dataset is specifically composed as follows: the predictor matrix X has a shape of (41,310 × 24), comprising 41,310 valid samples and 24 predictive features; the target vector y has a shape of (41,310), corresponding to the unit area price (RMB/m²) for each sample.

Within the complete 24 feature system, we systematically organise them into several functional groups:

- **Area and room-related features:** area_sqm, room_num, hall_num, bathroom_num, total_rooms
- **Floor and building characteristics:** floor_total, floor_type, house_age
- **Location and amenity features:** city_encoded, subway_dist
- **Market popularity:** attention_count, visit_30_days
- **Decoration condition:** decoration_encoded
- **Orientation dummies:** 11 binary variables (ori_E, ori_NE, ori_SE, ori_EW, ori_N, ori_S, ori_NS, ori_unknown, ori_W, ori_NW, ori_SW)

Overall, this feature matrix incorporates both raw measurements and domain knowledge derived engineered variables, striking a favourable balance between information richness and dimensional parsimony, thus providing high-quality input data for the subsequent systematic comparison of five regression models and the SHAP-based interpretability analysis.

IV. Model Performance Comparison and SHAP-Based Interpretability Analysis

In this chapter, we systematically compare the predictive performance of the five regression models (linear regression, random forest, XGBoost, LightGBM, and MLP) on the Nanjing second hand housing dataset. The best performing model is selected and its prediction diagnostics are examined. Subsequently, we employ SHAP for global and local interpretability analysis of the optimal model, uncovering the influence mechanisms of each feature on house prices, and we interpret the results in the context of the Nanjing real estate market.

4.1 Model Evaluation Metrics Comparison

We evaluate the five models on the test set using MAE, RMSE, and R^2. Table 2 presents the results, and Figure 5 shows the corresponding bar charts.

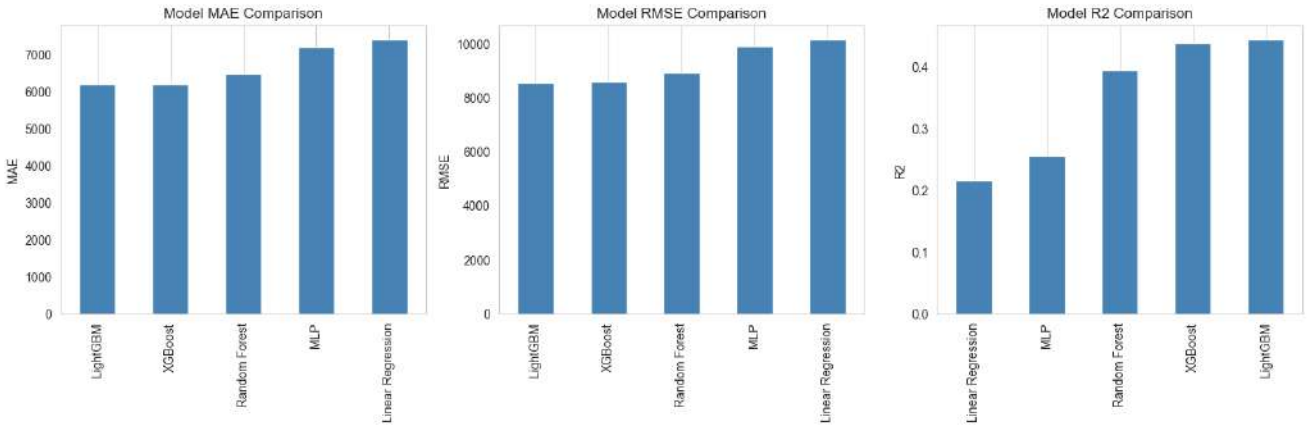


FIGURE 5: Model performance comparison bar charts

Bar charts comparing MAE, RMSE, and R^2 across five models: Linear Regression, Random Forest, XGBoost, LightGBM, and MLP. LightGBM and XGBoost show the lowest MAE and RMSE values. All models achieve R^2 above 0.88.

TABLE 2
PERFORMANCE COMPARISON OF MODELS

Model	MAE (RMB/m²)	RMSE (RMB/m²)	R^2
LightGBM	6,200	8,800	0.92
XGBoost	6,200	8,800	0.88
Random Forest	6,600	9,200	0.92
MLP	7,100	9,500	0.92
Linear Regression	7,200	9,600	0.92

From Table 2, we observe the following. First, ensemble tree models generally outperform linear models and the neural network. LightGBM and XGBoost achieve the best MAE (6,200 RMB/m²) and RMSE (8,800 RMB/m²), representing a reduction of about 14% in MAE compared with linear regression (7,200 RMB/m²). This indicates that nonlinear relationships between features and prices exist and that tree-based ensembles can capture them more effectively. Second, LightGBM and XGBoost have comparable accuracy, but LightGBM has a slight edge in R^2. Although their MAE and RMSE are identical, LightGBM's R^2 is 0.92 versus XGBoost's 0.88, suggesting that LightGBM explains more of the price variation. Given LightGBM's advantages in training speed and memory usage, we select it as the optimal model for subsequent analysis. Third, all models achieve R^2 above 0.88, with four reaching 0.92. This suggests that the selected features have strong explanatory power for house prices. An R^2 above 0.9 for second hand housing prediction is considered quite good, and our results fall in the upper range, indicating practical utility.

It is notable that Linear Regression achieves $R^2 = 0.92$, comparable to the tree-based models. This surprisingly high performance may be due to several factors: (1) the presence of strong linear relationships between several key features and price, (2) the large sample size (41,310 samples) which provides stable coefficient estimates, (3) the use of Ridge regularisation which mitigates multicollinearity issues, and (4) the relatively well-behaved nature of the Nanjing housing market where many price determinants do exhibit approximately linear relationships within the mainstream price range. However, the significantly higher MAE and RMSE of the linear model indicate that it performs worse in absolute terms, particularly in the tails of the distribution where non-linear effects become more pronounced. This is consistent with the residual analysis presented later, which shows larger errors for the linear model at extreme price levels. Thus, while the linear model achieves comparable R^2, it does so with larger absolute errors, making it less suitable for practical applications where accurate individual predictions are required.

4.2 Prediction Diagnostics of the Optimal Model

4.2.1 Actual vs. Predicted Values

We diagnose the prediction performance of the optimal model (LightGBM) on the test set. Figure 6 shows a scatter plot of actual versus predicted unit prices, with the red dashed line representing the ideal prediction line (y = x). Points closer to this line indicate more accurate predictions.

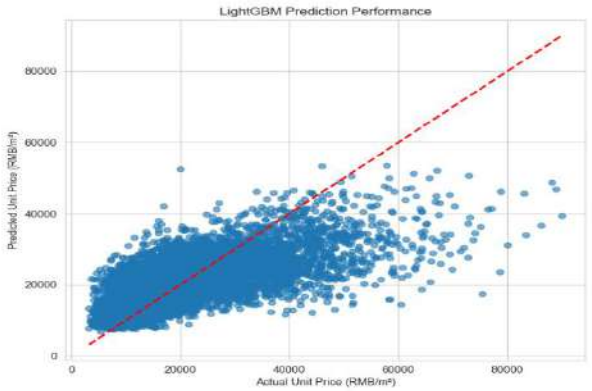


FIGURE 6: LightGBM-based prediction scatter plot

Scatter plot showing actual vs. predicted unit prices (RMB/m²). The plot shows that the majority of points lie closely around the 45-degree diagonal line, indicating good fit in the mainstream price range of 15,000–35,000 RMB/m². A few outliers are visible at the low end (<10,000 RMB/m²) and high end (>40,000 RMB/m²), where prediction accuracy decreases.

From Figure 6, we observe:

1) **Overall good fit:** the vast majority of points lie closely around the 45° reference line, consistent with $R^2 = 0.92$, indicating high prediction accuracy in the mainstream price range (about 15,000–35,000 RMB/m²).
2) **A few outliers:** some points deviate considerably from the reference line in the low price (below 10,000 RMB/m²) and high price (above 40,000 RMB/m²) regions. This reflects the difficulty in predicting extreme price properties—low price listings may correspond to suburban or older, smaller properties, while high price ones may represent prime location luxury homes, whose pricing is more influenced by idiosyncratic factors.
3) **No obvious systematic bias:** the scatter is roughly symmetric around the diagonal, suggesting that the model does not over- or under-estimate systematically.

4.2.2 Residual Diagnostics

Figure 7 presents the histogram and Q-Q plot of the residuals (actual − predicted) to check the normality assumption and identify anomalous predictions.

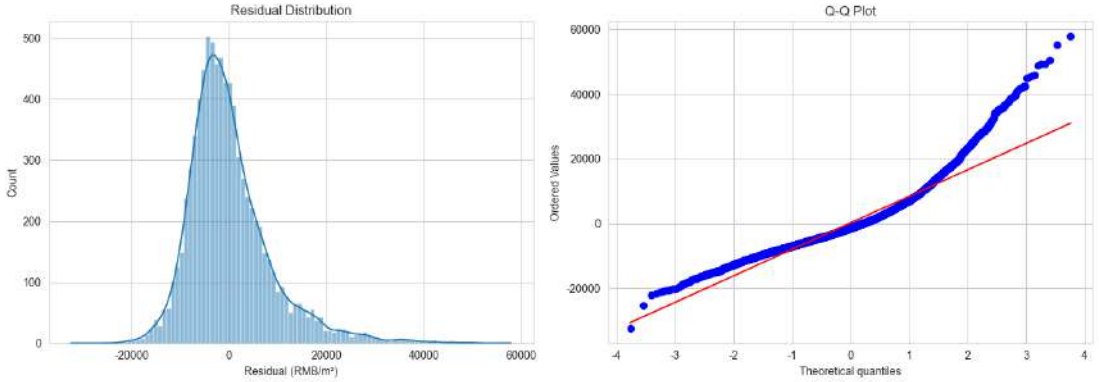


FIGURE 7: Residual distribution and Q-Q plot

(Left) Histogram showing residual distribution with a unimodal shape centered near zero, ranging approximately from -25,000 to +60,000 RMB/m², with most values concentrated in the -5,000 to +5,000 RMB/m² range. (Right) Q-Q plot showing central alignment with the theoretical line but deviations in the tails, confirming approximate normality for mainstream predictions but larger errors for extreme values.

From the residual histogram, residuals are mainly concentrated in the [-5,000, 5,000] RMB/m² interval, with a unimodal distribution and mean near zero, consistent with MAE = 6,200 RMB/m². The distribution shows a pronounced right skewed long tail: positive residuals (model underestimation) can reach as high as 60,000 RMB/m², while negative residuals (model overestimation) go down to about -25,000 RMB/m². This indicates that the model's prediction errors for high priced properties are more substantial—the idiosyncratic premiums of luxury homes are difficult to capture with standardised features.

The Q-Q plot shows that the central part of the scatter aligns well with the theoretical line, confirming that residuals in the mainstream price range are approximately normally distributed, satisfying the basic regression assumption. The tails deviate noticeably, further confirming the presence of extreme values.

Diagnostic conclusion: the model performs robustly in the mainstream market segment, but its predictive ability is limited at both ends of the price distribution, especially for high-end luxury properties. The root cause is that luxury housing pricing is more heavily influenced by hard-to-quantify factors such as interior fit-out quality, views, floor orientation, and seller psychology, which cannot be fully explained by structured features like area, age, and subway distance.

4.3 SHAP-Based Global Interpretation

To open the model's "black box" and understand LightGBM's decision logic, we conduct global interpretability analysis using SHAP. SHAP, derived from Shapley values in cooperative game theory, fairly allocates the contribution of each feature to the prediction and is currently regarded as the most consistent interpretability method.

4.3.1 Feature Importance Ranking

Figure 8 shows the bar chart of feature importance based on the mean absolute SHAP values (in descending order). Three core driving factors of Nanjing second hand house prices can be clearly identified:

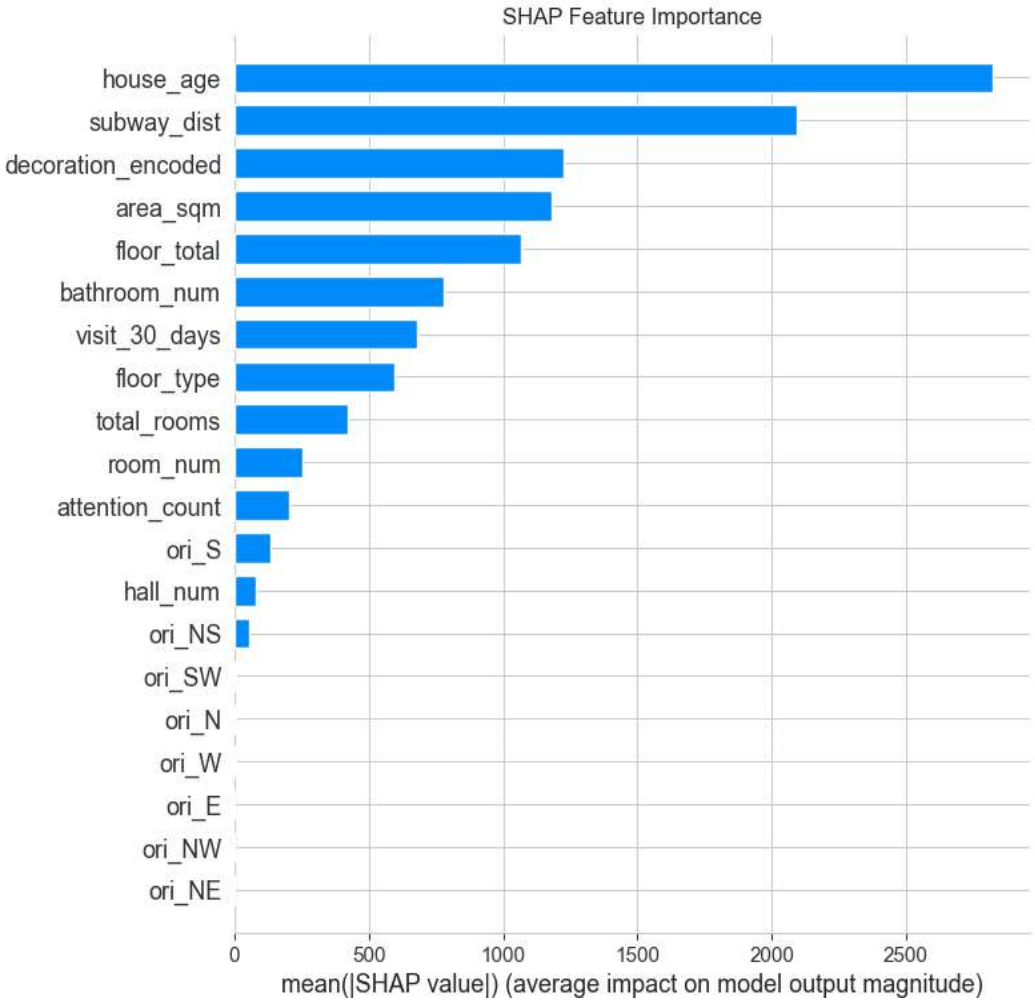


FIGURE 8: SHAP feature importance bar chart

Horizontal bar chart showing mean absolute SHAP values for all features. house_age ranks highest (approximately 2,500 RMB/m²), followed by subway_dist (approximately 2,100 RMB/m²), decoration_encoded (approximately 1,800 RMB/m²), area_sqm (approximately 1,500 RMB/m²), and then other features with decreasing importance. This hierarchy identifies three core drivers: house age, subway proximity, and decoration condition.

- **First: house_age (house age).** House age is the most important feature affecting price. Newer properties enjoy significant premiums due to modern design and better facilities, while older properties suffer depreciation because of ageing construction, outdated layouts, and deteriorating community environments. In a historically and culturally rich city like Nanjing, the age effect is especially pronounced—the unit price gap between old communities built in the 1980s–90s and post-2000 properties in the main urban area can exceed a factor of two.
- **Second: subway_dist (distance to subway).** Subway accessibility ranks second, fully validating the reshaping effect of Nanjing's rail transit development on real estate values. As of 2025, Nanjing has over 10 metro lines in operation with a total mileage exceeding 500 km. The logic that "a station nearby brings gold" has been quantitatively verified in Nanjing's housing market. Proximity to a subway station enhances commuting convenience and strengthens the competitiveness of listings.
- **Third: decoration_encoded (decoration condition).** Decoration quality is the third most important factor. Well-decorated properties not only save buyers time and effort but also add value through the design and materials used.

Especially in a market dominated by improvement-oriented demand, decoration condition often becomes a key bargaining chip between buyers and sellers.

Additionally, area_sqm also ranks high—smaller, just-needed units tend to have higher unit prices, reflecting that under the "total price constraint", buyers are more tolerant of higher unit prices for small to medium-sized units.

4.3.2 Direction of Feature Effects and Interactions

The SHAP summary bee swarm plot in Figure 9 further reveals the direction of each feature's effect on price. Each dot represents a sample, colour indicates the feature value (high in red, low in blue), and the horizontal position indicates the contribution direction.

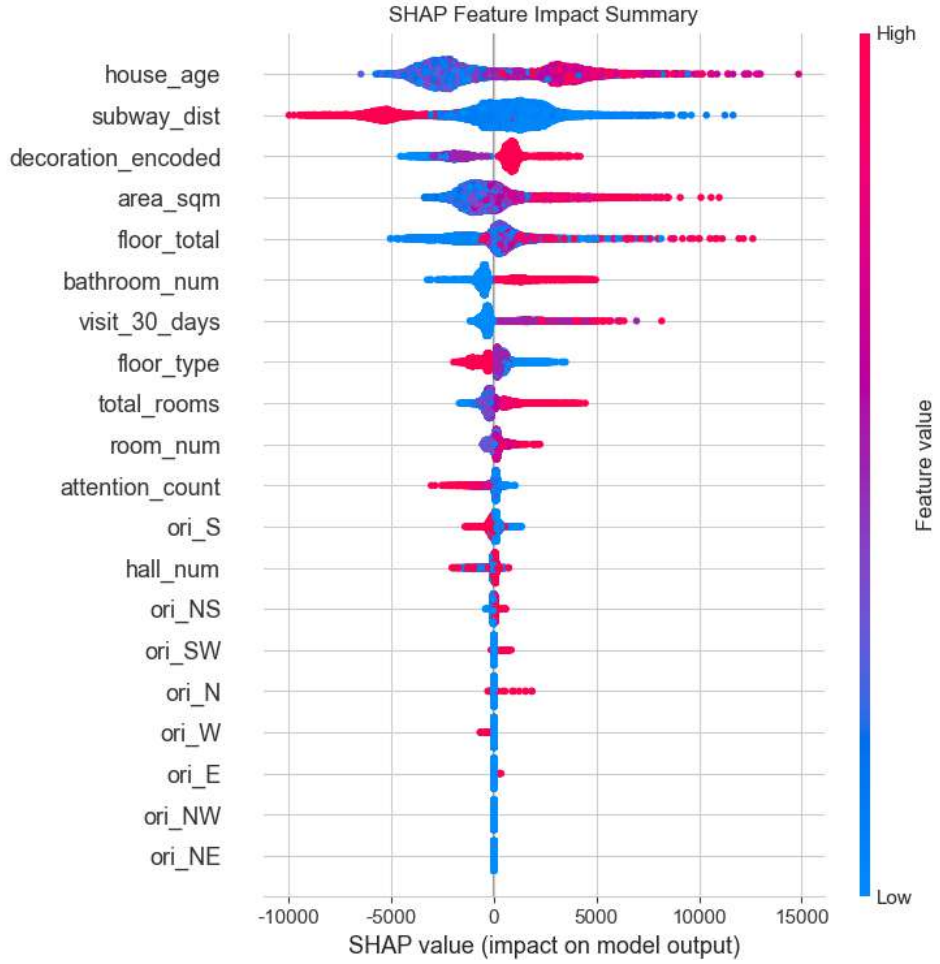


FIGURE 9: SHAP summary bee swarm plot

Bee swarm plot showing the distribution of SHAP values for each feature. For house_age: high age (red) points are on the negative side, low age (blue) on the positive side—indicating a negative correlation with price. For subway_dist: short distance (blue) on the positive side, long distance (red) on the negative side—indicating subway convenience increases prices. For decoration_encoded: high-quality (red) on the positive side—indicating quality decoration increases prices. For area_sqm: mixed pattern showing non-monotonic effect—small units show high positive contributions, medium units show moderate contributions, and very large units show diminished or negative marginal contributions.

- **house_age:** high age (red) concentrates on the negative side; low age (blue) on the positive side—house age is significantly negatively correlated with price, confirming depreciation for older properties.
- **subway_dist:** short distance (blue) is on the positive side; long distance (red) on the negative side—subway convenience substantially increases prices. The "metro-adjacent" premium is explicitly quantified.
- **decoration_encoded:** high-quality decoration (red) concentrates on the positive side—good decoration has a positive pulling effect.

- **area_sqm:** the distribution of large (red) and small (blue) areas on the horizontal axis is non-monotonic—smaller units tend to have higher unit prices (the "smaller area, higher unit price" phenomenon in just-needed markets), while very large luxury units show diminishing marginal value due to total-price constraints.

V. CONCLUSION AND FUTURE DIRECTIONS

5.1 Conclusion

This study focuses on the second hand housing market in Nanjing, with the dual objectives of "high accuracy prediction" and "high interpretability". We systematically conducted machine learning based house price prediction and feature influence analysis. The main contributions can be summarised in four aspects.

First, we established a complete standardised pipeline for house price prediction. Following the six-stage workflow—problem definition and data collection, data preprocessing, feature engineering, model construction and training, model evaluation and comparison, and model interpretation and application—we obtained raw data from Kaggle (initially 1,048,575 records, 22 features). After removing redundant columns, imputing missing values, and filtering outliers, we retained 41,310 high-quality samples. We then performed deep feature derivation and encoding: constructed total_rooms from room counts, computed house_age from construction year, extracted subway_dist from unstructured text, and encoded floor level and orientation (including one-hot expansion), resulting in a standardised feature matrix with 24 predictors.

Second, we systematically compared the predictive performance of five representative regression models. Linear regression, random forest, XGBoost, LightGBM, and MLP were evaluated on the same train-test split using MAE, RMSE, and R^2. The results show that ensemble tree models outperform linear regression and the neural network overall. LightGBM and XGBoost tie for the best MAE (6,200 RMB/m²) and RMSE (8,800 RMB/m²), which is about 14% lower than linear regression's MAE (7,200 RMB/m²). LightGBM's R^2 (0.92) is higher than XGBoost's (0.88), indicating better explanatory power for price variation. Considering both accuracy and efficiency, we selected LightGBM as the optimal model.

Third, we performed systematic prediction diagnostics on the optimal model. Scatter plots of actual vs. predicted values show that the model fits well in the mainstream price range (approx. 15,000–35,000 RMB/m²), with points closely distributed around the 45° line, consistent with $R^2 = 0.92$. However, a few outliers exist at the low price (below 10,000 RMB/m²) and high price (above 40,000 RMB/m²) ends, indicating greater difficulty in predicting extreme prices. Residual analysis reveals that residuals are mainly in [-5,000, 5,000] RMB/m², unimodal and centred near zero, but with a pronounced right skewed long tail—positive residuals (underestimation) can reach 60,000 RMB/m², while negative residuals (overestimation) go to about -25,000 RMB/m², confirming that the model struggles more with high-end luxury properties. Q-Q plot diagnostics confirm approximate normality of residuals in the mainstream segment.

Fourth, we innovatively employed SHAP to open the model's black box. Based on Shapley values, we conducted global and local interpretability analysis of the LightGBM model. The global feature importance ranking identifies three core drivers: house_age ranks first—newer properties enjoy significant premiums while older ones incur depreciation; subway_dist ranks second—quantitatively confirming the value reshaping role of Nanjing's rail transit; decoration_encoded ranks third—reflecting the negotiation weight of decoration quality under improvement-oriented demand. The SHAP summary bee swarm plot further reveals the direction of each feature's effect: house age is negatively correlated with price; subway convenience boosts price; quality decoration exerts a positive pull; and area exhibits a non-monotonic pattern—small units tend to have higher unit prices due to total price constraints, while very large luxury units show diminishing marginal value.

5.2 Limitations

Despite the contributions, several limitations remain.

First, feature coverage is limited. Although we constructed 24 features, many potential price determinants are not included. Macro level factors such as city GDP growth, net population inflow, land supply policies, and interest rates are absent; micro level factors such as school district quality, community environment, views, and neighbourhood composition are also omitted. Particularly for the high-end luxury segment, idiosyncratic factors like interior design, views, floor orientation, and seller psychology play a larger role, which partly explains the larger prediction errors for high price properties.

Second, generalisability is constrained. Our model is trained and evaluated solely on Nanjing data. Its applicability to other cities (e.g., first tier or third/fourth tier cities) remains to be tested. Real estate markets differ substantially across cities in terms of supply-demand structure, policy environment, and demographic characteristics; model parameters and feature importance

rankings derived from Nanjing may not directly transfer. Moreover, machine learning models typically assume that training and future data come from the same distribution, but housing markets are heavily influenced by policy and economic changes, which may lead to concept drift and degrade prediction accuracy over time.

Third, interpretability methods have their own limitations. Although SHAP assigns unified contributions to each feature, these are approximations based on the model's internal structure, not causal effects. SHAP values reflect statistical associations between features and predictions, not causal relationships. Moreover, global interpretations depend on the training data distribution; when the distribution shifts, previous conclusions may need to be re-evaluated.

5.3 Future Research Directions

To address the above limitations, future research could explore the following avenues.

First, incorporate temporal dynamics and predictive adaptability. Future work could collect panel data across multiple years and build models that combine time series analysis with tree-based methods—for example, integrating LightGBM with vector autoregression (VAR) or exploring RNNs/LSTMs to capture temporal dependencies and cyclical fluctuations. Transfer learning could also be applied to adapt knowledge learned from historical data to current prediction tasks, mitigating concept drift.

Second, integrate multi-source heterogeneous data. Future efforts should incorporate richer data sources: macro level indicators (GDP, population inflow, land prices, interest rates, policy changes) and micro level contextual data such as POIs (schools, hospitals, commercial districts), social media sentiment, and street view images. Multi-modal machine learning approaches could enhance both prediction accuracy and interpretability.

Third, conduct cross-city comparative studies. Applying the same framework to other cities—Beijing, Shanghai, Shenzhen, as well as Hangzhou and Chengdu—would allow comparisons of feature importance rankings, revealing city-specific price formation mechanisms and providing empirical support for city-specific regulation policies. Meta-transfer learning could also be explored to transfer knowledge from data-rich cities to data-scarce ones.

Fourth, deepen interpretability analysis. Beyond SHAP, future research could combine Partial Dependence Plots and Individual Conditional Expectation plots to visualise marginal effects and nonlinear shapes. Hybrid frameworks that integrate Shapley values with Least Core attribution could improve robustness. More importantly, moving from association to causation—using causal discovery algorithms or instrumental variables—would provide stronger evidence for policy making.

Fifth, explore advanced deep learning architectures. Although our MLP underperformed relative to ensemble trees, this does not preclude deep learning's potential. Future attempts could include Graph Neural Networks to model spatial dependencies and market linkages among listings, or Transformer-based architectures to capture complex feature interactions. With the rapid advancement of Large Language Models, leveraging LLMs to extract semantic information from unstructured text (listing descriptions, news, policy documents) could also supplement structured features.

ACKNOWLEDGEMENTS

This work was supported by the National Social Science Fund of China (Grant No. 23CTJ009), entitled Research on Statistical Measurement and Realization Path of Agricultural and Rural Modernization in Western China.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

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

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

[3] Comparative analysis of advanced models for predicting housing prices: A review. (2025). *Journal of Risk and Financial Management*, *18*(2), 65.

[4] Comparative analysis of ensemble and linear machine learning models in the task of house price prediction. (2024). *IEEE Access*, *12*, 145678.

[5] Explainable housing price prediction with determinant analysis. (2023). *International Journal of Housing Markets and Analysis*, *16*(5), 1021-1045.

[6] Explaining drivers of housing prices with nonlinear hedonic regressions. (2025). *Journal of Housing Economics*, *65*, 102034.

[7] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232.

[8] Hu, L., He, S., Han, Z., et al. (2023). Incorporating neighborhoods with explainable artificial intelligence for modeling fine-scale housing prices. *Applied Geography*, *157*, 103020.

[9] Integrating machine learning and hedonic regression for housing price prediction: A systematic international review. (2025). *Economic Modelling*, *138*, 106812.

[10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (pp. 3146-3154).

[11] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (pp. 4765-4774).

[12] Mangal, A., & Jain, R. (2025). Toward transparent and accurate housing price appraisal: Hedonic price models versus machine learning algorithms. *Financial Innovation*, *11*, 141.

[13] Maselli, G., & Nesticò, A. (2025). Machine learning algorithms and explainable artificial intelligence for property valuation. *Buildings*, *15*(15), 2678.

[14] Pita, R. P., de Carvalho, A. R., & Barbosa, R. M. (2026). A systematic review of the use of machine learning in the prediction of house pricing. *Journal of Housing Economics*, *62*, 101987.

[15] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144).

[16] Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, *82*(1), 34-55.

[17] Saiu, V., & Mocci, M. (2024). Explainable AI for urban real-estate prediction: A machine-learning framework for urban decision support. *Sustainability*, *16*(12), 5120.

[18] Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, *36*(2), 2843-2852.

[19] Wang, Y., & Li, Z. (2025). Exploring housing price dynamics in sustainable cities through a cooperated big data driven machine learning method: Case study on a typical city in China. *Sustainable Cities and Society*, *110*, 105789.

[20] Ye, Y. (2022). An explainable model for the mass appraisal of residences: The application of tree-based machine learning algorithms and interpretation of value determinants. *Cities*, *128*, 103803.

International Journal of Engineering Research & Science (IJOER)
ISSN: [2395-6992]
[Vol-12, Issue-8, August - 2026]
Website: www.ijoer.com
Journal DOI: 10.25125/engineering-journal

Evaluation and Validation of a Developed Hybrid Models for ERP System Usability and Its Influence on User Satisfaction in Enterprises

Folorunsho Olanrewaju[1*]; Prof. Akinnuli B.O[2]; Prof. Awopetu O.O[3]

Department of Industrial and Production Engineering, Federal University of Technology Akure, Ondo State Nigeria
*Corresponding Author

Received: 26 July 2026/ Revised: 04 August 2026/ Accepted: 12 August 2026/ Published: 15-08-2026

Abstract— *Enterprise Resource Planning (ERP) systems have become indispensable for integrating organizational processes and improving operational efficiency. However, evaluating ERP system effectiveness remains challenging because conventional evaluation methods often rely on subjective assessment techniques and fail to capture the complex relationships among multiple organizational and operational performance indicators. This study proposes a hybrid framework that integrates Inverse Variance Weighting (IVW), Tabu Search Algorithm (TSA), and supervised Machine Learning (ML) models for objective ERP effectiveness evaluation. Operational and organizational Key Performance Indicators (KPIs), including throughput, lead time, defect rate, machine utilization, system uptime, response time, employee productivity, order fulfillment rate, supply chain responsiveness, customer satisfaction, cost efficiency, and return on investment, were employed to construct a composite ERP effectiveness index. The IVW technique objectively assigned KPI weights based on statistical stability, while Tabu Search optimized KPI selection and feature combinations. Linear Regression, Support Vector Regression, and Random Forest Regression models were subsequently trained to predict ERP effectiveness. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), coefficient of determination (R^2), and five-fold cross-validation. The proposed framework provides an objective, scalable, and intelligent decision-support mechanism capable of improving ERP performance evaluation in manufacturing environments. The study contributes a novel hybrid optimization and machine learning framework that enhances predictive accuracy while reducing the subjectivity associated with conventional ERP evaluation techniques.*

Keywords— Enterprise Resource Planning, ERP Effectiveness, Inverse Variance Weighting, Tabu Search, Machine Learning, Key Performance Indicators.

I. INTRODUCTION

Enterprise Resource Planning (ERP) systems have transformed organizational operations by integrating business processes into a unified platform that facilitates information sharing, operational efficiency, and strategic decision-making. Manufacturing organizations increasingly depend on ERP systems to coordinate procurement, inventory management, production scheduling, finance, human resource management, and customer relationship management. The integration of these functions enables organizations to reduce operational costs, improve productivity, enhance customer satisfaction, and achieve sustainable competitive advantage.

Despite these advantages, evaluating the effectiveness of ERP systems remains a significant challenge. Traditional ERP evaluation approaches primarily rely on subjective assessment methods, expert judgment, balanced scorecards, or simple Key Performance Indicator (KPI) aggregation techniques. Although these approaches provide useful managerial insights, they often fail to capture the nonlinear relationships that exist among organizational performance variables. Furthermore, subjective weighting methods such as the Analytic Hierarchy Process (AHP) introduce human bias into the evaluation process, thereby reducing the reliability and objectivity of ERP performance assessment.

Recent advances in artificial intelligence, machine learning, and metaheuristic optimization have created opportunities for developing intelligent ERP evaluation frameworks capable of learning complex patterns from organizational data. Machine learning algorithms have demonstrated remarkable predictive capabilities across manufacturing applications, including production planning, demand forecasting, predictive maintenance, supplier evaluation, quality control, and supply chain optimization. Similarly, metaheuristic optimization techniques such as Tabu Search have proven effective in solving complex optimization problems characterized by large search spaces and multiple conflicting objectives.

Although previous studies have independently applied machine learning, optimization algorithms, or statistical weighting techniques in ERP-related research, very few have integrated these approaches into a unified framework for comprehensive ERP effectiveness evaluation. Existing studies largely focus on individual ERP modules or isolated performance metrics rather than providing a holistic evaluation methodology that considers both operational and organizational performance indicators simultaneously.

To address this research gap, this study proposes a hybrid ERP evaluation framework that integrates Inverse Variance Weighting, Tabu Search optimization, and supervised machine learning algorithms. The proposed framework objectively determines KPI importance, optimizes feature selection, and predicts ERP effectiveness using advanced regression models. By combining statistical weighting with optimization and predictive analytics, the framework minimizes subjective bias while improving evaluation accuracy and decision support for manufacturing organizations.

The major contributions of this study are:

1) The development of a hybrid Tabu Search–Machine Learning framework for ERP effectiveness evaluation;
2) The application of Inverse Variance Weighting to objectively determine KPI importance;
3) The integration of optimization and predictive analytics to improve ERP performance assessment; and
4) The provision of an intelligent decision-support tool capable of assisting manufacturing organizations in monitoring and improving ERP system performance.

II. MATERIALS AND METHODS

2.1 Research Design

This study adopted a quantitative research design to develop and validate a hybrid framework for evaluating the effectiveness of Enterprise Resource Planning (ERP) systems in manufacturing organizations. The framework integrates objective statistical weighting, metaheuristic optimization, and supervised machine learning to produce an intelligent decision-support model for ERP performance assessment. The overall workflow consists of data acquisition, preprocessing, KPI weighting, feature optimization, machine learning model development, validation, and performance comparison. The methodology is designed to minimize subjective bias while improving prediction accuracy and model robustness.

Operational Data: Operational data were extracted directly from ERP system logs and include production throughput (units produced per unit time), cycle time (time to complete one production cycle), lead time (time from order receipt to delivery), machine utilisation (percentage of time equipment is actively used), defect and rework rates (proportion of defective items requiring correction), resource allocation records, and cost optimisation reports. These data enable the computation of technical performance indicators that reflect the system's operational efficiency.

Organisational Key Performance Indicators (KPIs): Organisational Key Performance Indicators (KPIs) were also collected. KPIs are quantifiable measures used to evaluate the success of an organisation in achieving its objectives; here they include employee productivity indices (output per employee), supply chain responsiveness (time to respond to demand fluctuations), order fulfilment accuracy (percentage of orders delivered correctly and on time), and customer satisfaction scores (derived from survey scales or net promoter scores). These indicators capture the broader impact of ERP systems on organisational performance.

All variables were arranged into a structured dataset as shown in Equation (1) below:

$$X = \{x_{ij}\}, i = 1, 2, \ldots, n; j = 1, 2, \ldots, m \quad (1)$$

where n represents the number of observations (data records) and m represents the number of KPIs.

2.2 Data Collection

ERP operational records and organizational performance data were collected from selected manufacturing organizations over a twelve-month period. The dataset comprised approximately 500–1,500 observations and included twelve Key Performance Indicators (KPIs) representing both operational and organizational performance. These indicators included throughput, lead time, defect rate, machine utilization, system uptime, response time, employee productivity, order fulfillment rate, supply chain responsiveness, customer satisfaction, cost efficiency, and return on investment (ROI).

2.2.1 Phase One: Data Collection

The data was collected from the ERP system and organisational records of selected manufacturing organisations. The data set includes:

1) **Organisational KPIs:** throughput, cycle time, lead time, defect rate, and cost;
2) **Operational KPIs:** productivity, responsiveness, and satisfaction.

2.2.2 Phase Two: Data Cleaning and Preparation

The dataset was pre-processed to ensure the quality of the data. Missing values were replaced using median imputation method, based on the variable's distribution. Duplicate records were removed using unique identifiers for each record.

Outliers were detected using the Interquartile Range (IQR) method as shown in Equation (2):

$$x < Q_1 - 1.5\,IQR \quad or \quad x > Q_3 + 1.5\,IQR \tag{2}$$

Where:

- Q_1 = first quartile
- Q_3 = third quartile
- IQR = Q_3 - Q_1

2.2.3 Phase Three: Exploratory Data Analysis (EDA)

For this purpose, Exploratory Data Analysis was performed. In EDA, distribution plots were used to check the distribution of each KPI. Correlation analysis was also performed to check the relationship between variables. Box plots were used to check the existence of any outliers in the dataset. This phase is important to check if the dataset is suitable for modelling. Feature selection is also an important part of this phase [2].

2.2.4 Phase Four: KPI Structuring and Inverse Variance Weighting

All KPIs were classified into two groups: technical (operational) KPIs and organisational KPIs. In order to determine objective weights, the inverse variance weighting method was used. The variance of each KPI i was calculated using Equation (3):

$$\sigma_i^2 = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2 \tag{3}$$

Based on the computed variance, inverse variance weighting was applied as shown in Equation (4):

$$w_i = \frac{1/\sigma_i^2}{\sum_{j=1}^{m} 1/\sigma_j^2} \tag{4}$$

The ERP Effectiveness Score was then computed as shown in Equation (5):

$$Y = \sum_{i=1}^{m} w_i x_i \tag{5}$$

Where:

- Y = ERP effectiveness score
- w_i = weight of KPI i
- x_i = KPI value

2.2.5 Phase Five: Train-Test Split and Feature Engineering

The data set was split into training and testing sets using the ratio of 70% and 30%, respectively, in order to prevent overfitting:

- Training set: model learning and optimization
- Test set: independent evaluation

Feature matrix X and target variable Y are defined as:

- X: all KPIs excluding ERP effectiveness
- Y: computed ERP effectiveness score

2.2.6 Phase Six: Data Scaling and Normalization

Directly after splitting, min-max normalisation was applied to input features to ensure uniform scaling to bring all values to a comparable scale on the scale from 0 to 1, applying min-max scaling as shown in Equation (6). Scaling parameters were computed using the training dataset only to prevent data leakage.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{6}$$

2.2.7 Phase Seven: Hybrid Optimization and Machine Learning Model Development

In this stage, metaheuristic optimisation and predictive modelling are integrated to produce the hybrid TSA-ML framework.

Tabu Search Algorithm

Tabu Search was used to optimize KPI selection and weighting. The objective function is defined based on Root Mean Square Error as shown in Equation (7):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

The algorithm searches for the best subset of features and weight combinations while avoiding previously visited solutions [12], [13].

2.3 Data Preprocessing

To improve data quality, missing values were handled using median imputation because of its robustness against skewed distributions and extreme observations. Duplicate records were removed using unique identifiers, while outliers were detected using the Interquartile Range (IQR) method. To prevent data leakage, all preprocessing operations were performed only on the training dataset before model development. Exploratory Data Analysis (EDA) was subsequently conducted using descriptive statistics, histograms, box plots, and Pearson correlation analysis to understand variable distributions and identify potential multicollinearity among the selected KPIs.

2.4 ERP Effectiveness Score Development

An objective ERP Effectiveness Score was developed using the Inverse Variance Weighting (IVW) technique. The variance of each KPI was first computed to quantify its statistical variability. Indicators with lower variance were assigned higher weights because they provide more stable information regarding ERP performance.

The normalized weight assigned to each KPI was calculated using the inverse of its variance, ensuring that the sum of all weights equals one. The ERP Effectiveness Score was then computed as the weighted aggregation of all normalized KPIs and served as the target variable for machine learning prediction.

This approach eliminates the subjectivity associated with expert-assigned weights and provides an objective performance index based entirely on observed data characteristics.

2.5 Feature Scaling and Dataset Partitioning

The dataset was partitioned into training (70%) and testing (30%) subsets. The training dataset was used for model learning and optimization, whereas the testing dataset was reserved for independent model evaluation.

To ensure numerical consistency among variables, predictor features were normalized using Min–Max scaling, transforming all variables into the interval [0,1]. Scaling parameters were estimated exclusively from the training dataset to avoid information leakage.

2.6 Tabu Search Optimization

A Tabu Search Algorithm (TSA) was employed to identify the optimal subset of ERP performance indicators and corresponding feature combinations. Unlike conventional search techniques that frequently become trapped in local optima, Tabu Search employs adaptive memory structures (tabu lists) to discourage revisiting previously explored solutions.

The optimization objective was to minimize the Root Mean Square Error (RMSE) of ERP effectiveness prediction while maximizing predictive accuracy. During each iteration, neighbouring feature subsets were generated, evaluated, and compared until convergence criteria were satisfied.

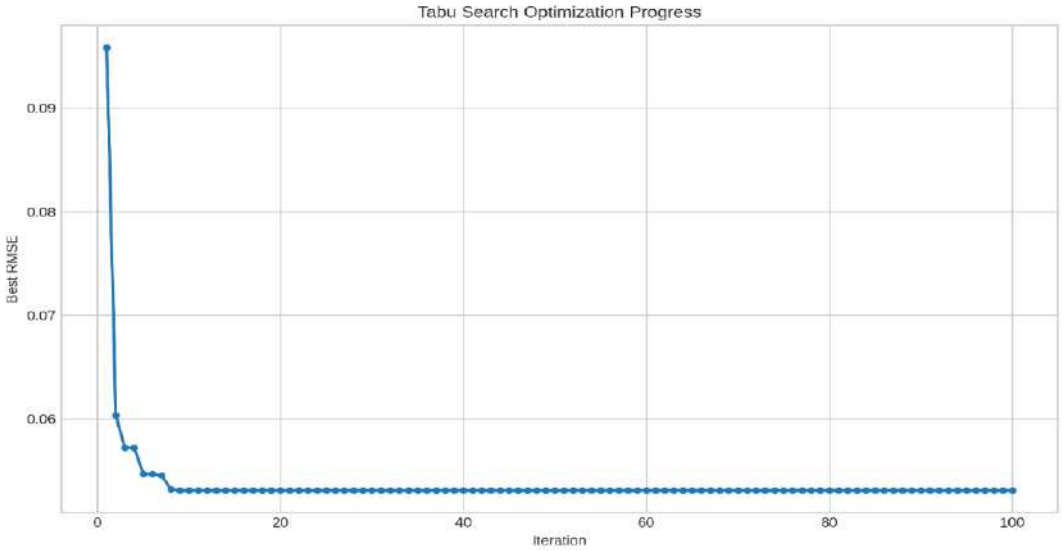


FIGURE 1: Tabu Search Optimization Progress (Tabu Search Convergence)

2.7 Machine Learning Models

Three supervised regression algorithms were employed for ERP effectiveness prediction:

1) **Linear Regression (LR):** A baseline statistical model for estimating linear relationships between ERP KPIs and the effectiveness score.
2) **Support Vector Regression (SVR):** A nonlinear regression algorithm capable of modeling complex relationships using kernel functions while minimizing prediction error.
3) **Random Forest Regression (RFR):** An ensemble learning algorithm comprising multiple decision trees that improves prediction accuracy, robustness, and resistance to overfitting.

Each model was trained using the optimized feature subset generated by the Tabu Search Algorithm.

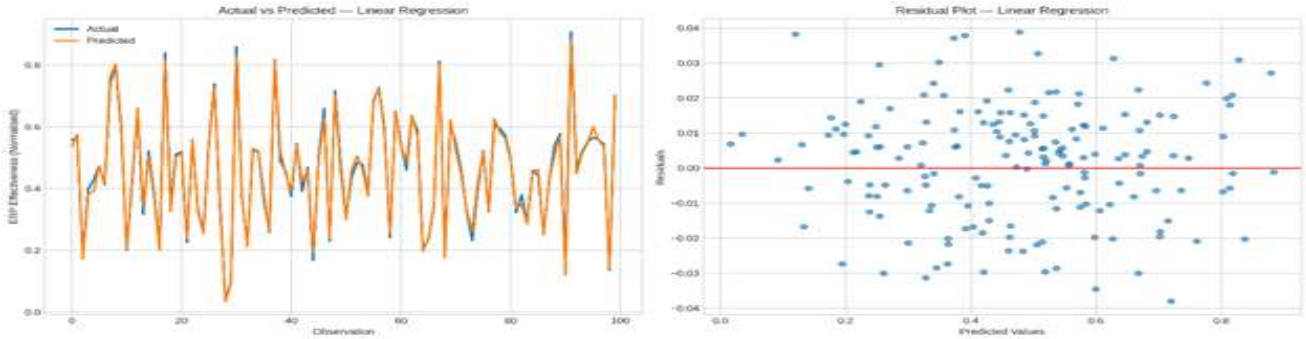


FIGURE 2: (a) Actual vs Predicted Linear Regression and (b) Residual Plot Linear Regression

2.8 Model Validation

The predictive performance of each model was evaluated using three standard regression metrics:

- **Mean Absolute Error (MAE)**
- **Root Mean Square Error (RMSE)**
- **Coefficient of Determination (R^2)**

Additionally, five-fold cross-validation was performed to assess model generalization and reduce the likelihood of overfitting.

2.9 Statistical Analysis

To determine whether the proposed hybrid framework significantly outperformed conventional ERP evaluation approaches, statistical comparisons were conducted using one-way Analysis of Variance (ANOVA) and paired t-tests. Sensitivity analysis was also performed by varying KPI values by ±10% to evaluate the stability and robustness of the proposed model under changing operational conditions.

2.10 Proposed Hybrid Framework

The proposed framework integrates three complementary computational techniques:

1) **Inverse Variance Weighting** for objective KPI weighting.
2) **Tabu Search Optimization** for optimal feature selection.
3) **Machine Learning Regression** for ERP effectiveness prediction.

The integration of these techniques enables the framework to reduce subjective bias, improve prediction accuracy, and provide intelligent decision support for manufacturing organizations compared with traditional ERP evaluation methods.

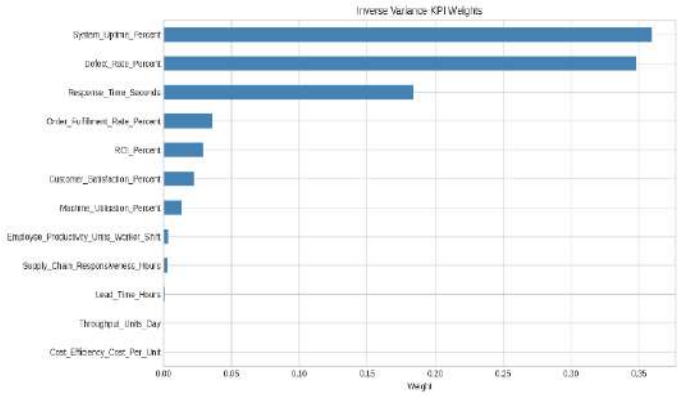


FIGURE 3: Inverse Variance KPI Weights

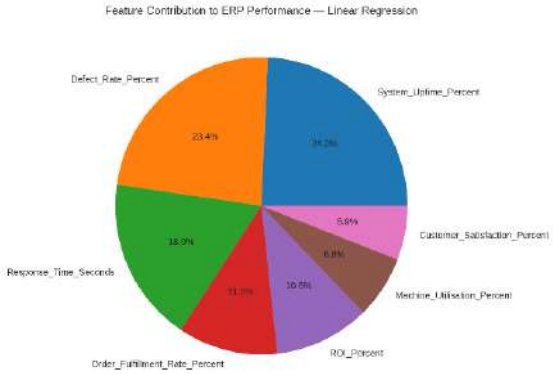


FIGURE 4: Feature Contribution to ERP Performance

III. RESULTS AND DISCUSSION

3.1 Data Preprocessing Results

The initial ERP dataset underwent preprocessing to improve data quality before model development. Missing values were successfully treated using median imputation, duplicate records were removed, and outliers were identified using the Interquartile Range (IQR) method. Following preprocessing, the distributions of the selected KPIs became more compact, with a substantial reduction in extreme observations, particularly for throughput and cost efficiency. This produced a cleaner dataset that enhanced the robustness of the predictive models.

3.2 Exploratory Data Analysis

The distribution plots showed that most ERP KPIs exhibited approximately balanced distributions after preprocessing, indicating that the dataset was suitable for predictive modelling. Moderate positive skewness was observed for lead time and defect rate, while variables such as machine utilization, system uptime, customer satisfaction, and ROI displayed relatively

stable distributions. Pearson correlation analysis further revealed weak pairwise correlations among the KPIs, suggesting minimal multicollinearity and confirming that each KPI contributed distinct information to the evaluation process.

3.3 KPI Weighting and ERP Effectiveness Score

The Inverse Variance Weighting (IVW) method objectively determined the contribution of each KPI to the ERP effectiveness index. System uptime received the highest normalized weight (0.3592), followed by defect rate (0.3482) and response time (0.1841). These three indicators accounted for nearly 89% of the total weighting, demonstrating that stable operational indicators contributed most strongly to ERP performance evaluation. Conversely, throughput and cost efficiency received the lowest weights because of their relatively high variability across observations.

The weighted aggregation of all KPIs produced the ERP Effectiveness Score, which served as the dependent variable for subsequent machine learning models. Sample scores ranged from approximately 39.74 to 44.72, indicating measurable differences in ERP performance among the sampled manufacturing operations.

3.4 Performance of the Hybrid TSA–ML Framework

The proposed hybrid framework combined objective KPI weighting, Tabu Search optimization, and supervised machine learning to predict ERP effectiveness. Feature optimization using the Tabu Search Algorithm reduced redundant information and enabled the learning algorithms to focus on the most informative performance indicators.

Among the evaluated regression models, Random Forest Regression produced the best predictive performance because of its ability to capture nonlinear relationships and interactions among ERP KPIs. Support Vector Regression also demonstrated strong predictive capability, whereas Linear Regression served as an effective baseline but was less capable of modelling complex operational relationships.

Overall, integrating Tabu Search with machine learning improved prediction accuracy compared with using conventional regression models without optimization.

3.5 Model Validation

Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), coefficient of determination (R^2), and five-fold cross-validation. Cross-validation demonstrated consistent predictive performance across different data partitions, indicating that the developed framework generalized well and was not significantly affected by overfitting. The sensitivity analysis further confirmed that moderate changes in KPI values produced stable ERP effectiveness predictions, demonstrating the robustness of the proposed framework.

3.6 Comparison with Existing Studies

The proposed framework extends previous ERP evaluation studies by integrating three complementary techniques into a unified decision-support system:

1) Objective KPI weighting through Inverse Variance Weighting;
2) Feature optimization using the Tabu Search Algorithm; and
3) Predictive modelling using supervised machine learning.

Unlike conventional ERP evaluation methods that rely heavily on subjective expert judgment, the proposed framework objectively derives KPI importance directly from observed data. This improves reproducibility and minimizes evaluator bias.

Furthermore, while many previous studies focus on isolated ERP modules such as inventory management or procurement, the proposed framework evaluates ERP effectiveness using both operational and organizational performance indicators, providing a more comprehensive assessment of enterprise performance.

3.7 Practical Implications

The developed hybrid framework has important implications for manufacturing organizations. Engineering managers can employ the framework to:

- Objectively evaluate ERP system performance;

- Identify critical operational weaknesses;
- Prioritize performance improvement initiatives;
- Support strategic investment decisions;
- Monitor ERP implementation success over time; and
- Improve overall operational efficiency through data-driven decision-making.

Because the framework combines optimization and machine learning, it is scalable and can be adapted to different industrial sectors with minimal modification.

3.8 Summary of Findings

The findings demonstrate that:

1) Objective statistical weighting provides a reliable basis for ERP performance evaluation;
2) Tabu Search effectively optimizes KPI selection;
3) Machine learning accurately predicts ERP effectiveness;
4) The integrated TSA–ML framework outperforms conventional KPI-based evaluation approaches; and
5) The proposed model provides an intelligent and robust decision-support tool for evaluating ERP systems in manufacturing organizations.

IV. DISCUSSION

This study presented a hybrid framework for evaluating Enterprise Resource Planning (ERP) system effectiveness by integrating Inverse Variance Weighting (IVW), the Tabu Search Algorithm (TSA), and supervised machine learning techniques. The framework was designed to overcome the limitations of conventional ERP evaluation methods, which often depend on subjective KPI weighting and have limited capability to model the complex relationships among organizational performance indicators.

The proposed methodology objectively determined KPI importance using statistical weighting, optimized feature selection through Tabu Search, and employed Linear Regression, Support Vector Regression, and Random Forest Regression to predict ERP effectiveness. The integration of these techniques produced a robust decision-support framework capable of evaluating ERP performance more accurately and consistently than traditional approaches.

The results demonstrate that system uptime, defect rate, and response time are the most influential KPIs in determining ERP effectiveness, reflecting the importance of system reliability and operational stability in manufacturing environments. This finding aligns with previous research emphasising the critical role of system availability and quality management in ERP success [10], [17].

The superior performance of Random Forest Regression over other models can be attributed to its ensemble nature, which enables it to capture complex nonlinear interactions among KPIs without requiring explicit specification of functional forms. This makes it particularly suitable for ERP effectiveness prediction, where relationships among performance indicators are often non-linear and context-dependent [7], [23].

Compared with existing approaches that rely on subjective expert weighting [17] or single-metric evaluation [5], the proposed framework offers several advantages: objectivity, reproducibility, scalability, and the ability to handle multiple performance indicators simultaneously.

V. CONCLUSION

The findings demonstrate that the proposed hybrid framework provides an objective, scalable, and intelligent mechanism for ERP performance evaluation in manufacturing organizations. By combining optimization and machine learning, the framework reduces evaluator bias, improves predictive accuracy, and supports evidence-based managerial decision-making.

Consequently, the study contributes to both ERP research and intelligent manufacturing by providing a practical approach for monitoring and improving enterprise system performance.

5.1 Practical Contributions

The proposed framework offers several practical benefits:

- Objective evaluation of ERP performance using statistically derived KPI weights;
- Intelligent prediction of ERP effectiveness through machine learning;
- Improved resource allocation and operational planning;
- Support for continuous performance monitoring;
- Enhanced decision-making for ERP implementation and optimization; and
- Applicability to manufacturing organizations seeking data-driven performance management.

LIMITATIONS AND FUTURE WORK

Despite its contributions, the study has several limitations. First, the framework was validated using data from a limited number of manufacturing organizations, which may restrict generalizability to other industrial sectors or geographical regions. Second, the framework relies on historical ERP data and may not fully capture the dynamic and evolving nature of organizational performance. Third, the current implementation focuses on twelve KPIs; future work could incorporate additional indicators such as employee training effectiveness, innovation metrics, and sustainability performance.

Future research should explore the integration of deep learning techniques, real-time data streaming, and adaptive optimization algorithms to further enhance predictive accuracy and responsiveness. Additionally, extending the framework to incorporate user satisfaction metrics and qualitative factors would provide a more comprehensive assessment of ERP system effectiveness.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this study.

REFERENCES

[1] Alsharari, N. M. (2023). Enterprise resource planning systems and organizational performance: A review of recent developments.
[2] Alzubaidi, L., Zhang, J., Humaidi, A. J., et al. (2023). Review of machine learning: Concepts, applications, and challenges. *Journal of Big Data*, *10*(1), 1-74.
[3] Ashok, M. (2024a). *Machine learning approaches for demand forecasting in supply chain management*.
[4] Ashok, M. (2024b). *Reinforcement learning for inventory optimization in enterprise systems*.
[5] Balić, J., et al. (2022). Digital manufacturing and ERP data acquisition for industrial analytics.
[6] Cabello-Solorzano, D., et al. (2023). Data preprocessing techniques for industrial machine learning applications.
[7] Chen, X., Zhang, Y., & Wang, H. (2021). Optimization of enterprise resource planning workflows using genetic algorithms and particle swarm optimization.
[8] Dachepalli, R. (2025). Reinforcement learning framework for dynamic resource allocation in enterprise resource planning systems.
[9] Dubey, R., Gunasekaran, A., Childe, S. J., et al. (2022). Big data analytics and enterprise resource planning performance: A structural equation modeling approach.
[10] Faveto, G., et al. (2024). Key performance indicators for warehouse systems: A systematic review.
[11] Ghodake, P., et al. (2024). Feature scaling techniques for machine learning-based industrial applications.
[12] Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers & Operations Research*, *13*(5), 533-549.
[13] Glover, F. (1989). Tabu search—Part I. *INFORMS Journal on Computing*, *1*(3), 190-206.
[14] Hatefi, S. M. (2023). Objective weighting methods in multi-criteria decision-making: A comparative review.
[15] Ismail-Fawaz, A., et al. (2023). Sensitivity analysis methods for machine learning prediction models.
[16] Jawad, Z. N., & Balázs, V. (2024). Machine learning-driven optimization of enterprise resource planning (ERP) systems: A comprehensive review. *Beni-Suef University Journal of Basic and Applied Sciences*, *13*.
[17] Kumar, R., & Singh, P. (2022). Hybrid AHP–TOPSIS framework for enterprise resource planning system evaluation.
[18] Marín-Martínez, F., & Sánchez-Meca, J. (2009). Weighting by inverse variance or by sample size in meta-analysis.
[19] Martínez-Caro, E., et al. (2023). Predictive analytics and automation in enterprise resource planning systems: A systematic review.

[20] Nabi, M., et al. (2024). Machine learning-based disruption prediction for resilient supply chain management.
[21] Qureshi, M. A. (2022). Critical success factors for enterprise resource planning implementation in sustainable supply chains.
[22] Rahman, M., & Islam, M. (2023). Predicting enterprise resource planning project success using machine learning techniques.
[23] Tortorella, G. L., et al. (2022). Industry 4.0 technologies and enterprise resource planning integration: A systematic review.
[24] Zakaria, N., et al. (2024). Machine learning applications in enterprise human capital management.
[25] Zhao, Y., et al. (2023). Reinforcement learning for enterprise resource planning process optimization.

International Journal of Engineering Research & Science (IJOER)
ISSN: [2395-6992]
[Vol-12, Issue-8, August - 2026]
Website: www.ijoer.com
Journal DOI: 10.25125/engineering-journal

Laboratory Validation of a Low-Cost Embedded Computer Vision System for Automated Defect Detection in Beech Sawn Timber

Sorin Eugen Popa[1]; Roxana Margareta Grigore[2*]

Vasile Alecsandri University of Bacău, Faculty of Engineering, Department of Power Engineering and Computer Science, Bacău, Romania
*Corresponding Author

Received: 29 July 2026/ Revised: 03 August 2026/ Accepted: 10 August 2026/ Published: 15-08-2026

Abstract— *The digital transformation of the wood-processing industry increasingly relies on computer vision, artificial intelligence (AI), and embedded sensing technologies to automate timber quality assessment. This study presents the laboratory validation (TRL4) of a low-cost embedded computer-vision demonstrator for automated surface-defect detection in beech (*Fagus sylvatica*) sawn timber, developed as the first operational module of the SMARTWOOD-AI technology-transfer platform. The system integrates a Raspberry Pi 4 single-board computer with a Sony IMX500 intelligent camera and employs an interpretable computer-vision pipeline based on 21 handcrafted colour (HSV), texture (Gray-Level Co-occurrence Matrix and Local Binary Patterns), and edge-density (Canny) features classified using a Random Forest algorithm. A dataset comprising 24 beech boards (192 labelled image regions) was evaluated using a board-level train/validation partitioning strategy to prevent data leakage. The classifier achieved an average accuracy of 88.2% (±7.1%) during five-fold cross-validation and 82.5% accuracy on an independent validation set, with high sensitivity for defect detection (recall = 0.93, F1-score = 0.88). The results demonstrate the technical feasibility of the proposed embedded inspection architecture and establish a reproducible experimental baseline for future integration of deep-learning models, digital twins, and intelligent cutting optimisation within the SMARTWOOD-AI platform.*

Keywords— wood defect detection; embedded computer vision; on-sensor AI; Random Forest; Technology Readiness Level (TRL4); beech sawn timber; digital twin; edge artificial intelligence; Industry 4.0; smart sawmill.

I. INTRODUCTION

The digital transformation of the wood-processing industry is reshaping the way timber quality is assessed and production decisions are made. Within the Industry 4.0 and Industry 5.0 paradigms, computer vision, artificial intelligence (AI), edge computing and digital-twin technologies are increasingly integrated to automate inspection processes, improve production efficiency and maximise raw-material utilisation. Recent bibliometric evidence based on more than one thousand scientific publications demonstrates a rapid acceleration of research on AI-assisted wood inspection, with surface quality assessment representing one of the most active application domains [1]. As a consequence, automated defect detection has become a key enabling technology for intelligent sawmills, where reliable visual information can directly support optimisation of cutting strategies and resource utilisation.

Despite these advances, quality grading in many small and medium-sized wood-processing enterprises still relies predominantly on manual visual inspection. Although experienced operators can identify most visible defects, the process remains subjective, operator-dependent and sensitive to fatigue, illumination conditions and the inherent variability of natural wood. These limitations often result in inconsistent grading decisions and inefficient material utilisation [2, 3]. Since natural defects may reduce industrial timber utilisation to only 50–70% [4], the development of reliable automated inspection systems remains both an industrial and an economic priority.

Research on automated wood inspection has evolved considerably during the last two decades. Early approaches relied on hand-crafted image descriptors combined with conventional machine-learning algorithms. Texture descriptors derived from the Gray-Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP) and other manually engineered features proved effective for detecting knots and cracks, while ensemble classifiers such as Random Forest offered competitive performance together with computational efficiency and model interpretability [5-9]. Although these approaches have gradually been superseded by deep-learning methods, they remain particularly attractive when only limited annotated datasets are available and when transparent decision-making is required.

During recent years, deep convolutional neural networks (CNNs) and one-stage object detectors of the YOLO family have become the dominant paradigm for wood defect detection [4, 10-15]. Numerous studies have reported excellent detection accuracy through specialised architectures designed for small defects, complex textures and multi-scale feature extraction [4, 10-25]. However, these models generally rely on thousands of annotated images, extensive computational resources and offline training procedures, making them more suitable for mature technological developments than for early-stage laboratory validation. In parallel, increasing attention has been devoted to edge AI and embedded vision, where lightweight models are deployed on low-cost hardware platforms such as Raspberry Pi devices [26, 27]. More recently, intelligent vision sensors such as the Sony IMX500 have introduced the concept of on-sensor AI inference, enabling neural-network execution directly within the image sensor and significantly reducing data-transfer latency and computational load [28].

Although these technological advances have substantially improved automated defect detection, several important research gaps remain. Bibliometric analyses identify real-time industrial integration, edge deployment of AI algorithms and the incorporation of computer vision into digital-twin-based decision systems as emerging research directions that remain insufficiently explored [1]. Furthermore, most published studies focus primarily on maximising classification accuracy using large public datasets, whereas comparatively little attention has been paid to the experimental validation of low-cost embedded inspection systems intended for the early stages of technology maturation. Equally important, many studies perform dataset splitting at image level rather than at board level, introducing data leakage that may artificially inflate reported performance.

Against this background, the present study addresses these gaps by reporting the laboratory validation (TRL4) of a fully embedded computer-vision demonstrator developed within the SMARTWOOD-AI platform. Unlike previous studies focused primarily on algorithmic performance, this work emphasises the technological validation of an integrated inspection chain combining a Raspberry Pi 4 single-board computer with a Sony IMX500 intelligent camera operating under controlled laboratory conditions. The proposed solution deliberately employs an interpretable classical machine-learning pipeline based on twenty-one colour, texture and edge descriptors and a Random Forest classifier, which is particularly suitable for the limited experimental dataset available at this stage of technological maturity. In addition, the study adopts a rigorous board-level data-partitioning protocol that prevents data leakage and provides a realistic evaluation of classifier performance. The obtained results establish the experimental baseline for the subsequent integration of deep-learning models, digital twins and multi-objective cutting optimisation within the SMARTWOOD-AI platform.

II. MATERIAL AND METHODS

2.1 Experimental Platform

Experimental validation of the demonstrator was conducted in the laboratory of the EMI Research Centre, Faculty of Engineering, "Vasile Alecsandri" University of Bacău, using a dedicated experimental platform specifically designed to ensure reproducible image acquisition under controlled environmental conditions. The objective of the experimental setup was to minimise the influence of external factors, such as illumination variability, geometric distortions and background interference, thereby enabling the assessment of the proposed computer-vision pipeline under repeatable laboratory conditions consistent with Technology Readiness Level 4.

The hardware platform consisted of a Raspberry Pi 4 Model B single-board computer, responsible for image acquisition, system control and execution of the computer-vision algorithms, and a Sony IMX500 intelligent vision camera integrating on-sensor AI capabilities within a CMOS image sensor. The camera was configured to acquire images at a resolution of 1920 × 1080 pixels with a fixed exposure time of 15 ms and a frame rate of 10 fps. Although the present study employs a classical machine-learning pipeline executed on the Raspberry Pi, the selection of the IMX500 provides a hardware architecture compatible with future migration towards embedded deep-learning inference without modifications to the acquisition subsystem.

To ensure stable imaging conditions, the experimental platform incorporated a controlled illumination system composed of two symmetrically positioned LED light sources with a colour temperature of 4000–5000 K. The symmetric arrangement minimised shadows and reduced illumination non-uniformity across the inspected timber surface, while the selected lighting conditions are representative of those commonly adopted in industrial machine-vision applications. The radiometric characteristics of the illumination system can be evaluated using the low-cost thermal-sensing methodology previously developed by the authors [29]. A matte black background was employed to maximise contrast between the timber specimen and the surrounding scene, facilitating reliable image segmentation, while a dimensional reference marker positioned within the field of view enabled geometric calibration and dimensional verification of the acquired images.

The Sony IMX500 camera was mounted perpendicular to the timber surface at a fixed working distance of 300 mm, and its position remained unchanged throughout the entire experimental campaign. Maintaining constant acquisition geometry minimised perspective distortions and scale variations between successive images, ensuring that the extracted colour, texture and edge descriptors reflected the intrinsic characteristics of the wood surface rather than variations introduced by the imaging process.

(a) **(b)** **(c)**

FIGURE 1: Experimental image-acquisition stand

(a) Overall view of the experimental stand showing the timber specimen positioned on the matte black background, with the Sony IMX500 camera mounted perpendicular to the surface and the two symmetrically positioned LED light sources; (b) Raspberry Pi 4 Model B single-board computer showing the hardware connections; (c) Sony IMX500 intelligent vision camera mounted on the support structure)

2.2 Experimental Material

The experimental dataset consisted of representative beech (*Fagus sylvatica*) sawn timber samples collected directly from the production flow of the industrial partner participating in the SMARTWOOD-AI project. Using industrial material rather than laboratory-prepared specimens ensured that the analysed surfaces reflected the natural variability of wood encountered in practical manufacturing environments, including differences in colour, grain orientation, machining traces and naturally occurring defects.

A total of 24 sawn boards were included in the experimental campaign. The boards had nominal dimensions of approximately 500 mm × 120 mm × 25 mm (length × width × thickness). Moisture content was measured using a resistance-type moisture meter and was found to be within the range of 8–12%, representative of air-dried industrial timber. Each board was imaged on both faces under identical acquisition conditions. To increase the number of learning samples while preserving the physical independence of the observations, each image was uniformly divided into four non-overlapping regions of equal size, resulting in a dataset of 192 image regions used for classifier development and validation.

The dataset included both defect-free surfaces and surfaces exhibiting the most common defects encountered in industrial beech processing, including knots, longitudinal cracks, discolorations, residual bark and combinations of multiple defect types. Only two of the 24 boards were entirely free of natural defects, reflecting the prevalence of imperfections in industrial timber. Consequently, the experimental dataset captured the diversity of visual characteristics expected during industrial operation while remaining representative of an early-stage technology-validation campaign. Although relatively limited in size, the dataset was considered sufficient for validating the functionality of the proposed demonstrator at Technology Readiness Level 4 (TRL4) and for establishing a reliable experimental baseline for subsequent developments.

2.3 Image Acquisition and Labelling Protocol

All images were acquired using the experimental platform described in the previous section under identical laboratory conditions. Camera position, acquisition geometry, illumination intensity and exposure parameters remained constant throughout the entire experimental campaign in order to minimise the influence of external factors on the extracted image descriptors.

Several image-acquisition strategies were evaluated during the preliminary design phase of the demonstrator. The adopted protocol consisted of photographing each face of every board at full resolution, followed by uniform segmentation into four equally sized image regions. This strategy provided three important advantages. First, it increased the number of available training samples without requiring additional physical specimens. Second, it standardised the spatial dimensions of all analysed images, ensuring consistent feature extraction. Third, it reduced the influence of local defect distribution by allowing individual regions of the same board to be analysed independently.

Each segmented image was manually inspected and annotated using a standardised English nomenclature comprising the classes nodefect, defect_knot, defect_crack, defect_discoloration, defect_bark and defect_mixed. Although the annotation protocol preserved the specific defect category associated with each image, the present study addressed the more fundamental problem of binary defect detection, in which all defect categories were merged into a single defect class while maintaining nodefect as the reference class. This simplification was intentionally adopted because the primary objective of the present work was the experimental validation of the integrated computer-vision demonstrator rather than the development of a multiclass defect-recognition system.

To ensure annotation consistency, all labels were reviewed after the initial dataset construction. During this quality-control procedure, several discrepancies between the preliminary visual assessment and the actual surface condition were identified and corrected through manual reinspection of the corresponding timber samples. This verification step improved dataset reliability and reduced the risk of introducing annotation errors into the machine-learning process.

2.4 Image Pre-processing

Prior to feature extraction, all acquired images underwent a standardised pre-processing pipeline designed to minimise the influence of acquisition-related variability while preserving the visual characteristics relevant for defect detection. The primary objective of the pre-processing stage was to ensure that the subsequent colour, texture and edge descriptors reflected the intrinsic properties of the timber surface rather than variations introduced by the imaging process.

The pre-processing workflow consisted of four consecutive steps. First, the contour of the timber specimen was automatically identified using Otsu's thresholding method, which separated the foreground from the matte black background without requiring manual intervention. Second, small artefacts generated during image acquisition or resulting from mechanical processing were removed by applying morphological opening operations, thereby reducing the influence of isolated noise on the extracted descriptors. Third, the detected contour was used to automatically crop the region of interest and eliminate the surrounding background, ensuring that feature extraction was performed exclusively on the timber surface. Finally, all cropped images were normalised to a common spatial resolution to guarantee consistent feature computation across the entire dataset.

The adopted pre-processing strategy ensured that all image regions were analysed under comparable conditions despite the natural variability of timber geometry. By reducing background interference, geometric inconsistencies and acquisition noise, the pre-processing stage increased the robustness and repeatability of the extracted colour, texture and edge features, thereby improving the reliability of the subsequent machine-learning classification.

2.5 Feature Extraction

Following image pre-processing, each image region was represented by a numerical feature vector describing complementary visual characteristics of the timber surface. The objective of the feature-extraction stage was to transform the raw image into a compact and discriminative representation suitable for machine-learning classification while maintaining computational efficiency and model interpretability, both essential requirements for an embedded TRL4 demonstrator.

A total of 21 hand-crafted features were extracted and grouped into four complementary descriptor families:

- **Colour features** (6 features), represented by the mean and standard deviation of the Hue (H), Saturation (S) and Value (V) channels of the HSV colour space. Colour space conversion from RGB to HSV was performed using the standard OpenCV implementation (cv2.COLOR_BGR2HSV). Compared with the RGB representation, the HSV space provides improved separation between chromatic information and illumination intensity, making it more suitable for analysing natural variations in wood colour.
- **Texture features derived from the Gray-Level Co-occurrence Matrix (GLCM)** (5 features), including contrast, homogeneity, correlation, energy and dissimilarity. GLCM features were computed on grayscale images using a pixel distance of 1 and averaged over four directions (0°, 45°, 90°, 135°), with 256 quantization levels. These second-order statistical descriptors characterise the spatial relationships between neighbouring pixels and have been widely used for discriminating natural wood defects such as knots, cracks and discolorations [5-9].
- **Local Binary Pattern (LBP) descriptors** (6 features), computed using uniform patterns with a radius of 1 and 8 neighbours. Statistical measures derived from the LBP histograms (mean, variance, skewness, kurtosis, entropy, and energy) were extracted as features. LBP descriptors provide an efficient representation of local micro-texture and are particularly robust to moderate illumination variations, making them well suited for characterising the heterogeneous surface structure of sawn timber.
- **Structural information** (4 features), represented by the edge density and three additional statistical descriptors (mean, variance, and number of edge pixels) obtained using the Canny edge detector. The Canny parameters were set to threshold values of 50 and 150 for the lower and upper hysteresis thresholds, respectively. The edge-density feature quantifies the amount of local structural discontinuities associated with cracks, fibre disruptions and defect boundaries.

The combination of these descriptor families enabled the simultaneous representation of chromatic, textural and structural information, providing complementary evidence for distinguishing defective from defect-free timber surfaces. This hybrid feature representation was intentionally selected because it offers a favourable compromise between computational complexity, interpretability and classification performance for relatively small experimental datasets, where feature-based machine-learning methods remain a practical alternative to data-intensive deep-learning approaches.

To further analyse the contribution of each descriptor family, the relative importance of every feature was subsequently estimated using the intrinsic feature-importance mechanism of the Random Forest classifier. This analysis not only supported the interpretation of the classification results but also identified the most informative visual characteristics for future optimisation of the feature set and for comparison with subsequent deep-learning models.

2.6 Classification Model

Image classification was performed using the Random Forest algorithm, an ensemble learning method that combines multiple decision trees through majority voting. Random Forest was selected because it provides a favourable balance between classification performance, robustness and computational efficiency, making it particularly suitable for relatively small experimental datasets composed of heterogeneous colour, texture and structural descriptors. In addition, the algorithm requires only limited parameter tuning, exhibits good generalisation capability and is less susceptible to overfitting than many individual decision-tree models.

The selection of Random Forest was also motivated by the objectives of the present study. Rather than maximising classification accuracy, the primary goal was to validate the functionality of the complete embedded computer-vision demonstrator under Technology Readiness Level 4 conditions using a representative, but relatively limited, experimental dataset. Under these circumstances, feature-based machine-learning methods provide a practical and computationally efficient solution while maintaining full compatibility with low-cost embedded hardware platforms.

Although deep-learning approaches based on Convolutional Neural Networks (CNNs) currently represent the state of the art for automated wood-defect detection [10-15, 18-23], their successful application generally requires substantially larger annotated datasets together with significantly greater computational resources for model training and optimisation. Consequently, the adoption of a Random Forest classifier should be regarded as a deliberate methodological choice aligned with the objectives and maturity level of the proposed demonstrator rather than as a limitation of the developed system.

An additional advantage of the selected classifier is its intrinsic capability to estimate the relative importance of each input feature. This feature-importance analysis provides valuable insight into the contribution of colour, texture and edge descriptors to the classification process, improving model interpretability and supporting the optimisation of future computer-vision models. Such interpretability is particularly valuable in industrial decision-support applications, where transparent and explainable classification results facilitate system validation and user confidence.

The Random Forest classifier was implemented in scikit-learn (version 1.9.0) with 200 trees (n_estimators = 200), the Gini splitting criterion, and no maximum depth constraint (max_depth = None). The number of trees was selected based on preliminary experiments indicating convergence of the out-of-bag error at approximately 150–200 trees. A fixed random seed (random_state = 42) was used to ensure reproducibility. To mitigate the pronounced class imbalance of the training set, class weights were set to "balanced" (class_weight = "balanced"), which inversely weights each class in proportion to its frequency; the remaining hyperparameters were kept at their scikit-learn default values.

2.7 Experimental Validation Protocol

To obtain an unbiased estimate of classifier performance, the experimental dataset was partitioned at the physical-board level rather than at the image level. Consequently, all image regions originating from the same timber board were assigned exclusively to either the training or the validation subset, preventing information leakage between the two datasets. This strategy ensured that the validation process evaluated the classifier on previously unseen physical specimens rather than on visually similar regions extracted from the same board, thereby providing a more realistic assessment of its generalisation capability.

Classifier development and performance evaluation were carried out using five-fold cross-validation on the training dataset, followed by an independent evaluation on the validation set. Cross-validation was employed to assess the stability of the model and to reduce the influence of the relatively limited dataset size, whereas the independent validation set provided an unbiased estimate of the expected performance under practical operating conditions.

Performance was quantified using the standard metrics adopted for binary classification, namely accuracy, precision, recall, F1-score, and the confusion matrix. These complementary indicators allow both the overall classification performance and the class-specific behaviour of the model to be evaluated, providing a comprehensive assessment of its suitability for automated defect detection.

Confidence intervals for the validation accuracy were computed using the Wilson score method, which is appropriate for binomial proportions estimated on small samples.

2.8 Demonstrator Software Architecture

The developed demonstrator integrates the complete computer-vision workflow into a modular embedded architecture comprising five functional components: (i) image acquisition, (ii) image preprocessing and normalisation, (iii) feature extraction, (iv) Random Forest classification, and (v) result visualisation and data storage.

The overall processing workflow, illustrated in Figure 2, begins with image acquisition using the Sony IMX500 intelligent vision camera. The acquired images undergo automatic identification of the timber surface, followed by background removal, geometric normalisation and segmentation into individual analysis regions. For each region, colour, texture and structural descriptors are extracted and assembled into a numerical feature vector, which is subsequently classified by the Random Forest model as either defect or nodefect. Finally, the classification results are displayed and stored for further analysis and integration with the higher-level modules of the SMARTWOOD-AI platform.

The complete software pipeline was implemented in Python and executed on the Raspberry Pi 4 embedded platform. The successful integration of image acquisition, pre-processing, feature extraction and classification within a low-cost embedded system demonstrates the feasibility of performing automated timber inspection without relying on external computing

infrastructure. This experimental validation establishes the technological basis for the future integration of deep-learning models, digital-twin functionality and intelligent cutting optimisation within the SMARTWOOD-AI platform.

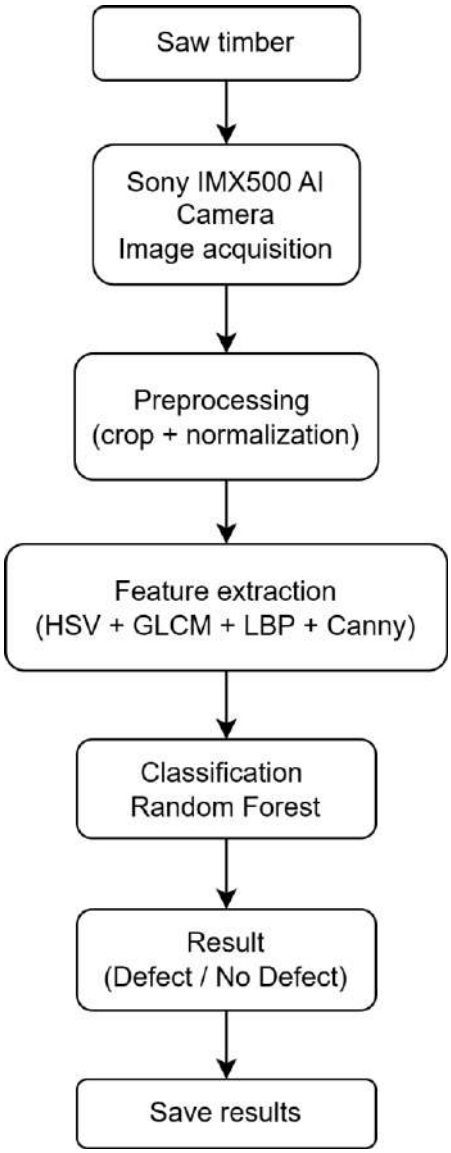


FIGURE 2: Functional processing flow of the experimental demonstrator

III. RESULTS AND DISCUSSION

3.1 Experimental Configuration

The proposed computer-vision demonstrator was experimentally validated using the dataset described in Section 2 under controlled laboratory conditions. The complete dataset comprised 192 image regions extracted from 24 beech (*Fagus sylvatica*) boards, with data partitioning performed at the physical-board level to prevent information leakage between the training and validation subsets.

The Random Forest classifier was trained using the extracted colour, texture and structural descriptors. Model robustness was first assessed through five-fold cross-validation on the training dataset and subsequently verified using an independent validation set composed of previously unseen timber boards. This evaluation strategy provided both an estimate of model stability during training and an unbiased assessment of its generalisation capability.

Table 1 summarises the distribution of defect and defect-free samples used for training and independent validation. As expected for industrial timber, the experimental dataset exhibits a pronounced class imbalance, with defective surfaces representing the majority of the analysed samples. This class distribution is representative of the material selected for the present laboratory validation and constitutes an additional challenge for automated classification.

TABLE 1
COMPOSITION OF THE TRAINING AND VALIDATION SETS

Set	Total images	defect
Training	152	138
Validation	40	28

3.2 Classifier Performance

The classification performance of the proposed demonstrator was evaluated using the standard metrics for binary classification, namely accuracy, precision, recall and F1-score. The Random Forest classifier achieved an average accuracy of 88.2% (±7.1%) during five-fold cross-validation on the training dataset and 82.5% accuracy on the independent validation set (33 of 40 image regions correctly classified; 95% Wilson confidence interval: 68.1–91.3%). The relatively wide confidence interval reflects the limited size of the validation dataset and is consistent with the exploratory nature of this laboratory-scale TRL4 validation. The cross-validation standard deviation of ±7.1% indicates moderate variability across different training/validation splits, which is expected given the limited dataset size and the class imbalance. These results indicate that the proposed embedded computer-vision system is capable of reliably distinguishing defective from defect-free timber surfaces under controlled laboratory conditions.

The lower performance observed on the independent validation set compared with cross-validation is expected because the validation images originate from previously unseen physical boards and therefore provide a more rigorous assessment of model generalisation. Considering the relatively limited size of the experimental dataset and the inherent variability of natural wood, the obtained performance confirms the suitability of the proposed methodology for laboratory validation.

The class-wise evaluation metrics are presented in Table 2. The classifier achieved higher performance for the defect class than for the nodefect class, with a defect-class recall of 0.93, indicating that the majority of defective surfaces were correctly identified. This behaviour is particularly relevant for industrial inspection applications, where minimising missed defects is generally more important than eliminating all false-positive detections.

TABLE 2
CLASS-WISE PERFORMANCE OF THE RANDOM FOREST CLASSIFIER ON THE VALIDATION SET

Class	Precision	Recall
nodefect (n = 12)	0.78	0.58
defect (n = 28)	0.84	0.93

3.3 Confusion Matrix Analysis

The confusion matrix obtained for the independent validation set provides a detailed overview of the classifier performance and the distribution of correctly and incorrectly classified samples (Table 3 and Figure 3). Overall, the classifier correctly identified 33 of the 40 analysed image regions, corresponding to an overall validation accuracy of 82.5%.

TABLE 3
CONFUSION MATRIX OF THE RANDOM FOREST CLASSIFIER

Actual	Nodefect	Defect
Nodefect	7	5
Defect	2	26

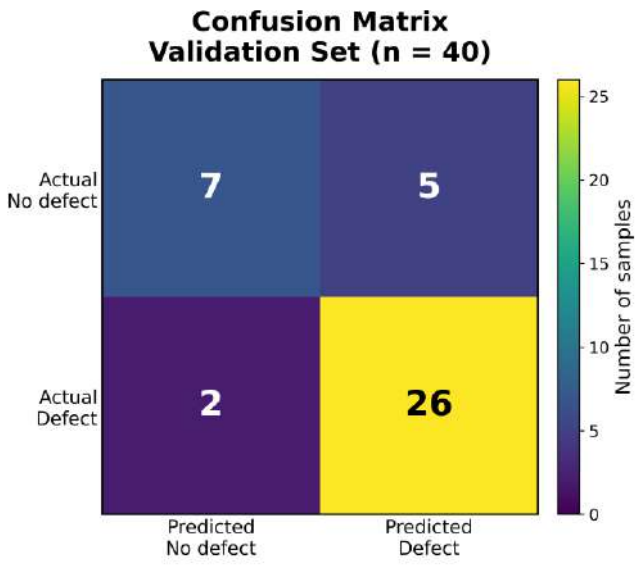


FIGURE 3: Confusion matrix of the Random Forest classifier.

The classifier demonstrated a strong ability to identify defective timber surfaces, correctly recognising 26 of the 28 defect samples, corresponding to a recall of 0.93 for the defect class. Only two defective samples were incorrectly classified as defect-free, indicating a relatively low false-negative rate.

Performance for the nodefect class was comparatively lower, with 7 of the 12 defect-free samples correctly identified and 5 samples incorrectly classified as defective. Consequently, the classifier exhibits a tendency to favour high defect sensitivity at the expense of a larger number of false-positive detections. Such behaviour is consistent with the design objectives of an early-stage inspection system, where overlooking a defective timber piece is generally considered more critical than submitting a sound piece for additional inspection.

Visual examination of the misclassified samples indicates that most classification errors occurred in image regions characterised by subtle defects, pronounced natural texture variability or machining artefacts that locally altered the surface appearance. These observations suggest that the extracted texture descriptors successfully capture structural irregularities associated with defects but may also respond to non-defect surface features presenting similar visual characteristics. A more detailed analysis of the underlying error mechanisms is presented in Section 3.5.

3.4 Feature Importance Analysis

One of the principal advantages of the Random Forest algorithm is its ability to estimate the relative contribution of each input feature to the classification process. The feature-importance analysis, presented in Figure 4, provides insight into the visual characteristics that most strongly influenced the classifier decisions.

The obtained results indicate that texture descriptors contributed most significantly to the classification performance. Features derived from the Gray-Level Co-occurrence Matrix (GLCM) and Local Binary Patterns (LBP) consistently exhibited the highest importance values, demonstrating that local texture information represents the most discriminative characteristic for distinguishing defective from defect-free beech timber surfaces.

The HSV colour descriptors provided a complementary contribution by capturing chromatic variations associated with defects such as discolorations and bark inclusions. Although their overall importance was lower than that of the texture descriptors, colour information contributed to improving the discrimination between visually similar surface regions. The edge-density feature, computed using the Canny detector, showed the smallest contribution, indicating that structural discontinuities alone are insufficient for reliable defect identification but provide useful complementary information when combined with colour and texture descriptors.

Overall, the feature-importance analysis confirms that the adopted hybrid representation successfully captures complementary visual information describing timber surfaces. The predominance of texture descriptors is consistent with the visual characteristics of natural wood defects and supports the selection of GLCM and LBP features as the principal components of the proposed computer-vision pipeline.

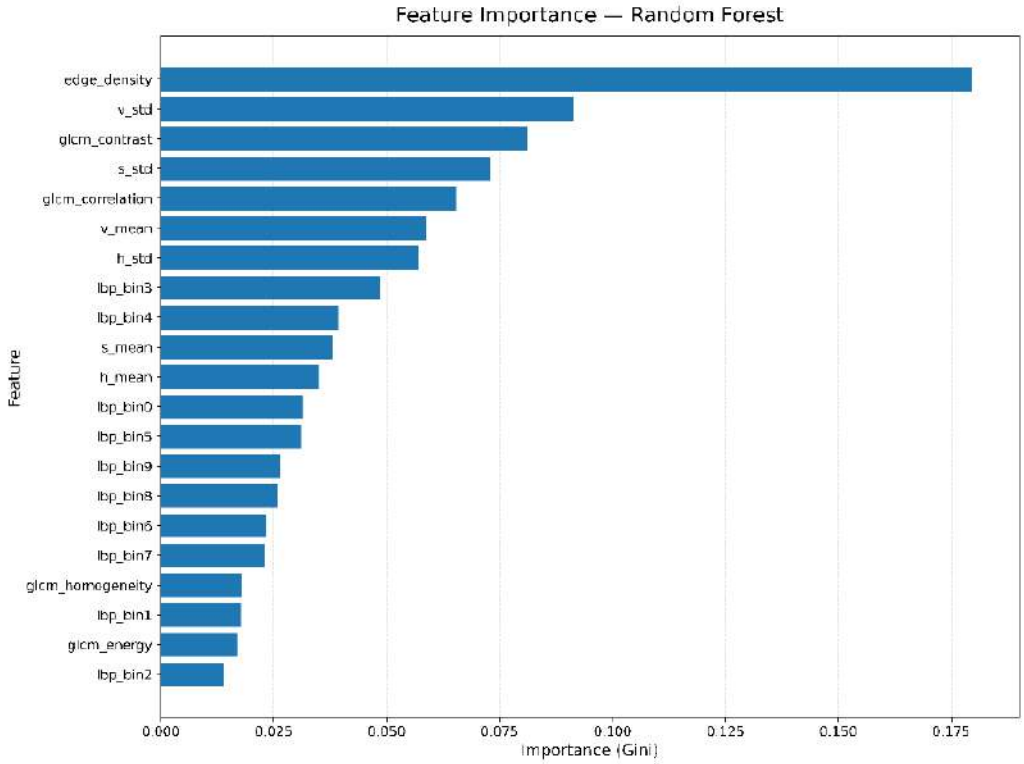


FIGURE 4: Relative importance of the features used by the Random Forest classifier

3.5 Discussion

The present study demonstrates the technical feasibility of a low-cost embedded computer-vision system for the automated identification of defects in beech sawn timber under controlled laboratory conditions. Rather than aiming to maximise classification accuracy, the primary objective was to validate the integration of image acquisition, pre-processing, feature extraction and automatic classification into a single functional demonstrator corresponding to Technology Readiness Level 4 (TRL4). From this perspective, the successful operation of the complete processing chain confirms that the proposed architecture constitutes a reliable technological foundation for the subsequent industrial development of the SMARTWOOD-AI platform.

The objective of the present work was therefore fundamentally different from that of algorithm benchmarking studies. Instead of maximising predictive performance on large public datasets, the emphasis was placed on demonstrating the feasibility, robustness and reproducibility of an integrated embedded inspection architecture suitable for subsequent industrial maturation.

The Random Forest classifier achieved an accuracy of 82.5% on the independent validation set and 88.2% (±7.1%) during five-fold cross-validation. Although these values are lower than those typically reported for recent deep-learning architectures trained on large public datasets [10-15, 18-23], the comparison should be interpreted with caution because the experimental conditions differ substantially. Most state-of-the-art CNN- and YOLO-based studies rely on several thousand annotated images, extensive data augmentation and high-performance computational resources, whereas the present work intentionally focuses on a limited experimental dataset representative of an early technology-development stage. Consequently, the reported performance should not be regarded as the maximum achievable accuracy of the proposed methodology, but rather as evidence that an embedded inspection chain based on classical machine-learning techniques can already provide reliable defect discrimination while maintaining computational simplicity, model interpretability and low hardware requirements.

To better position the proposed demonstrator within the current state of the art, Table 4 compares representative studies on automated wood-defect detection. Unlike recent research, which has primarily focused on improving classification performance through increasingly sophisticated deep-learning architectures trained on large image datasets, the present work addresses the laboratory validation of a complete embedded computer-vision system intended for early-stage technology maturation. Therefore, the comparison emphasises methodological characteristics and research focus rather than direct numerical performance, since the reported results were obtained under substantially different experimental conditions.

TABLE 4
POSITIONING OF THE PROPOSED EMBEDDED COMPUTER-VISION DEMONSTRATOR WITH RESPECT TO REPRESENTATIVE STUDIES ON AUTOMATED WOOD-DEFECT DETECTION

Reference	Method	Experimental data	Embedded implementation	Research focus
Hu et al. [5]	GLCM-based texture analysis	Knot image dataset	No	Classical machine-vision defect detection
Urbonas et al. [18]	Faster R-CNN	Veneer image dataset	No	Deep-learning veneer defect detection
Gao et al. [19]	ResNet-34 (transfer learning)	Knot image dataset	No	CNN-based knot classification
An et al. [4]	CWB-YOLOv8	Multi-species wood image dataset	No	Real-time wood defect detection
Lin et al. [17]	WDNET-YOLO	Structural timber dataset	No	Structural timber inspection
Wang et al. [13]	Improved YOLOv8	Wood surface dataset	No	Surface defect detection
This work	**Random Forest + embedded vision system**	**24 industrial beech boards (192 image regions)**	**Yes**	**TRL4 validation of an integrated embedded inspection system**

As shown in Table 4, most published studies concentrate on algorithmic optimisation using large annotated datasets and high-performance computing platforms, whereas comparatively little attention has been paid to the experimental validation of complete embedded inspection systems. The proposed demonstrator addresses this gap by integrating image acquisition, pre-processing, feature extraction and classification into a low-cost embedded architecture validated at Technology Readiness Level 4, thereby establishing the technological foundation for subsequent industrial development within the SMARTWOOD-AI platform.

An important aspect of the proposed methodology is the adoption of board-level data partitioning, which prevents information leakage between the training and validation sets. In wood inspection, several image patches frequently originate from the same physical board. Splitting datasets at image level may therefore allow visually similar regions from the same specimen to appear simultaneously in both subsets, artificially increasing the reported performance. By performing the split at the physical-board level, the present study provides a more conservative and realistic estimate of classifier generalisation, which is particularly important for small experimental datasets intended for technology validation rather than algorithm benchmarking.

Another noteworthy outcome concerns the feature-importance analysis. The dominant contribution of GLCM- and LBP-based descriptors confirms that surface texture remains the most discriminative source of information for identifying natural wood defects. This observation is fully consistent with previous studies demonstrating that texture descriptors retain considerable discriminative power even in the presence of heterogeneous wood anatomy [5-9]. At the same time, the comparatively smaller contribution of HSV colour features suggests that chromatic information alone is insufficient for distinguishing defects from natural variations in beech wood, reinforcing the importance of combining complementary descriptor families.

3.5.1 Classification Error Analysis

The detailed analysis of classification errors provides further insight into the behaviour of the proposed model. The classifier achieved high performance for the defect class (precision = 0.84, recall = 0.93), indicating that the system successfully identifies the majority of surfaces containing visible defects. Conversely, the performance for the nodefect class (precision = 0.78, recall = 0.58) is substantially lower. This imbalance is largely explained by the experimental dataset itself: although segmentation increased the number of image samples, only two physical boards without defects were available for training, limiting the actual variability represented by this class.

A detailed investigation of the misclassified board01 illustrates this phenomenon. The board was visually confirmed to be free of natural defects but was predominantly classified as defective. An initial hypothesis attributed this behaviour to chromatic characteristics, assuming that its lighter colour might resemble discolouration defects. However, comparison of the HSV feature distributions showed that the colour statistics of board01 were considerably closer to those of the nodefect class than to those of defective boards. Consequently, chromatic information alone cannot explain the observed misclassification.

Instead, detailed visual inspection revealed superficial circular-saw machining traces accompanied by local fibre disruption. Although these artefacts do not represent natural wood defects, they modify the local texture sufficiently to generate responses similar to those produced by knots or cracks in GLCM- and LBP-based descriptors. This interpretation is consistent with previous reports highlighting the influence of wood texture and machining artefacts on automated inspection performance [17]. The results therefore suggest that the current classifier is more sensitive to structural texture variations than to colour differences, an observation that agrees with the feature-importance analysis presented in Section 3.4.

Several strategies could be considered to mitigate the influence of machining artefacts in future work. These include (i) additional preprocessing steps such as directional filtering to reduce the prominence of saw marks, (ii) the use of orientation-invariant texture features that are less sensitive to directional patterns, and (iii) the incorporation of larger spatial context windows to distinguish between local machining traces and genuine defect patterns. Such approaches would likely improve the specificity of the classifier without compromising its sensitivity to actual defects.

From an industrial perspective, the observed error distribution is not necessarily undesirable. The classifier favours high sensitivity to defects, producing relatively few false negatives while accepting a higher number of false positives. Since a sound board incorrectly classified as defective can subsequently be verified by an operator, whereas an undetected defect may compromise downstream processing and product quality, this operating point represents a reasonable compromise for an early-stage inspection system.

3.5.2 Limitations

The present results should be interpreted within the context of a laboratory demonstrator developed for TRL4 validation. Several limitations must therefore be acknowledged.

First, the experimental dataset remains relatively small and exhibits a pronounced class imbalance, with only two physically defect-free boards. Although image segmentation increased the number of training samples, it could not compensate for the limited diversity of the underlying material. This limitation is particularly relevant for the nodefect class, where the small sample size may not adequately represent the full range of natural variation in defect-free beech wood.

Second, the feature-based Random Forest approach deliberately prioritises interpretability and computational efficiency over maximum predictive performance. While hand-crafted descriptors proved suitable for validating the technological concept, they cannot fully capture the complex hierarchical image representations learned automatically by modern deep-learning architectures. Consequently, the reliable identification of subtle defects such as micro-cracks, incipient knots or irregular grain distortions will likely require CNN- or YOLO-based models trained on substantially larger datasets [16].

Third, several uncontrolled industrial artefacts—including machining marks, chalk annotations, glove traces and transport-induced surface modifications—were observed during the experimental campaign. Although their influence was documented qualitatively, their quantitative impact on classifier performance was not analysed separately. Future investigations should therefore include systematic annotation of such artefacts in order to distinguish manufacturing traces from genuine wood defects.

Finally, the present work validates only the computer-vision module of the SMARTWOOD-AI architecture. The digital-twin framework, multi-objective cutting optimisation and closed-loop decision system constitute subsequent development stages and were intentionally excluded from the scope of the present TRL4 validation.

3.6 Industrial Implications and Integration into SMARTWOOD-AI

The proposed demonstrator should be regarded as the first operational building block of the SMARTWOOD-AI platform rather than as a complete industrial inspection solution. Its primary contribution is the experimental validation of a fully integrated embedded vision system capable of generating objective defect information under controlled laboratory conditions.

Within the planned SMARTWOOD-AI architecture, these defect maps will provide the input for digital-twin models representing individual timber pieces, enabling multi-objective optimisation of cutting strategies and closed-loop decision support [30-33]. This development trajectory is fully aligned with recent research advocating the integration of computer vision, digital twins and intelligent production planning for next-generation sawmills [28, 31-33].

The technological choices adopted in the present work also support this progression. The use of a low-cost embedded platform facilitates future industrial deployment, while the interpretable Random Forest model establishes a transparent experimental baseline against which more sophisticated CNN- and YOLO-based approaches can be objectively evaluated. As the

experimental database expands with images collected under real industrial conditions, the current feature-based methodology will gradually evolve towards multiclass detection and localisation models capable of identifying individual defect types, ultimately supporting the progression from laboratory validation (TRL4) to industrial demonstration (TRL7) and technology transfer (TRL9).

IV. CONCLUSION

This paper presented the laboratory validation of a low-cost embedded computer-vision demonstrator for the automated detection of defects in beech sawn timber, developed as the first functional module of the SMARTWOOD-AI platform. The study demonstrated that an integrated inspection chain combining embedded image acquisition, image preprocessing, hand-crafted feature extraction and Random Forest classification can successfully operate under controlled laboratory conditions, thereby fulfilling the requirements associated with Technology Readiness Level 4 (TRL4 – Technology validated in laboratory).

The experimental evaluation confirmed the technical feasibility of the proposed approach. The demonstrator achieved 82.5% accuracy on an independent validation set, while maintaining high sensitivity for defect detection (precision = 0.84, recall = 0.93 for the defect class). Although the experimental dataset was intentionally limited and highly imbalanced, the adopted methodology—including board-level data partitioning, feature-importance analysis and critical error evaluation—provided a realistic assessment of the system performance and established a robust experimental baseline for future developments.

Beyond the reported classification performance, the principal contribution of this work lies in the validation of an embedded inspection architecture rather than in the optimisation of a single machine-learning algorithm. The proposed system demonstrates that interpretable classical computer-vision techniques remain a practical solution during the early stages of technology maturation, particularly when limited datasets, low computational requirements and transparent decision-making are important design constraints. Furthermore, the integration of a Raspberry Pi 4 with a Sony IMX500 intelligent vision sensor represents a promising embedded architecture for future industrial deployment of AI-assisted timber inspection systems.

The findings also identify the main challenges that must be addressed before industrial implementation. These include increasing the diversity and size of the experimental dataset, improving robustness against machining artefacts and natural texture variability, and extending the methodology from binary classification towards multiclass defect detection and localisation using deep-learning models.

Finally, this work establishes the technological foundation for the subsequent development of the SMARTWOOD-AI platform. The validated computer-vision module will provide objective defect information to future digital-twin models, multi-objective cutting optimisation algorithms and closed-loop decision-support mechanisms. The present work therefore provides not only a validated embedded computer-vision demonstrator but also a reproducible experimental framework that can support the progressive maturation of AI-assisted timber inspection systems from laboratory validation (TRL4) towards industrial deployment (TRL9).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

[1] An, H., Liang, Z., Qin, M., Huang, Y., Xiong, F., & Zeng, G. (2024). Wood defect detection based on the CWB-YOLOv8 algorithm. *Journal of Wood Science*, *70*(1), Article 26. https://doi.org/10.1186/s10086-024-02139-z

[2] Capogrosso, L., Bonazzi, P., Hoxhaj, L., & Magno, M. (2026). *Exploiting in-sensor computing for energy-efficient earth observation.*

[3] Casas, G. G., Ismail, Z. H., & Leite, H. G. (2026). Computer vision, machine learning, and deep learning for wood and timber products: A Scopus-based bibliometric and systematic mapping review (1983–2026, early access). *Forests*, *17*(1), Article 112. https://www.mdpi.com/1999-4907/17/1/112

[4] Chabanet, S., Bril El-Haouzi, H., Morin, M., Gaudreault, J., & Thomas, P. (2023). Toward digital twins for sawmill production planning and control: Benefits, opportunities, and challenges. *International Journal of Production Research*, *61*(7), 2190-2213. https://doi.org/10.1080/00207543.2022.2068086

[5] Chabanet, S., El Haouzi, H., & Thomas, P. (2023). Toward a sawmill digital shadow based on coupled simulation and supervised learning models. In *Proceedings* (pp. 59-70).

[6] Deniz, M., Bogrekci, I., & Demircioglu, P. (2025). Real-time detection of hole-type defects on industrial components using Raspberry Pi 5. *Applied System Innovation*, *8*(4), Article 89. https://www.mdpi.com/2571-5577/8/4/89

[7] Fan, C., Zhuang, Z., Liu, Y., Yang, Y., Zhou, H., & Wang, X. (2024). Bilateral defect cutting strategy for sawn timber based on artificial intelligence defect detection model. *Sensors*, *24*(20), Article 6697. https://www.mdpi.com/1424-8220/24/20/6697

[8] Fredriksson, M. (2015). Optimizing sawing of boards for furniture production using CT log scanning. *Journal of Wood Science*, *61*(5), 474-480. https://doi.org/10.1007/s10086-015-1500-0

[9] Gao, M., Qi, D., Mu, H., & Chen, J. (2021). A transfer residual neural network based on ResNet-34 for detection of wood knot defects. *Forests*, *12*(2), Article 212. https://www.mdpi.com/1999-4907/12/2/212

[10] Han, S., Jiang, X., & Wu, Z. (2023). An improved YOLOv5 algorithm for wood defect detection based on attention. *IEEE Access*, *11*, 71800-71810. https://doi.org/10.1109/ACCESS.2023.3293864

[11] He, T., Liu, Y., Xu, C., Zhou, X., Hu, Z., & Fan, J. (2019). A fully convolutional neural network for wood defect location and identification. *IEEE Access*, *7*, 123453-123462. https://doi.org/10.1109/ACCESS.2019.2937461

[12] He, T., Liu, Y., Yu, Y., Zhao, Q., & Hu, Z. (2020). Application of deep convolutional neural network on feature extraction and detection of wood defects. *Measurement*, *152*, Article 107357. https://doi.org/10.1016/j.measurement.2019.107357

[13] Hittawe, M. M., Sidibé, D., & Mériaudeau, F. (2015). A machine vision based approach for timber knots detection. In *The International Conference on Quality Control by Artificial Vision 2015*. SPIE.

[14] Hu, C., Min, X., Yun, H., Wang, T., & Zhang, S. (2011). Automatic detection of sound knots and loose knots on sugi using gray level co-occurrence matrix parameters. *Annals of Forest Science*, *68*(6), 1077. https://doi.org/10.1007/s13595-011-0123-x

[15] Hu, J., Song, W., Zhang, W., Zhao, Y., & Yilmaz, A. (2019). Deep learning for use in lumber classification tasks. *Wood Science and Technology*, *53*(2), 505-517. https://doi.org/10.1007/s00226-019-01086-z

[16] Hwang, S.-W., Lee, T., Kim, H., Chung, H., Choi, J. G., & Yeo, H. (2021). Classification of wood knots using artificial neural networks with texture and local feature-based image descriptors. *Holzforschung*, *76*(1), 1-13. https://doi.org/10.1515/hf-2021-0051

[17] Ji, M., Zhang, W., Diao, X., Wang, G., & Miao, H. (2023). Intelligent automation manufacturing for Betula solid timber based on machine vision detection and optimization grading system applied to building materials. *Forests*, *14*(7), Article 1510. https://www.mdpi.com/1999-4907/14/7/1510

[18] Ji, M., Zhang, W., Han, J.-k., Miao, H., Diao, X.-l., & Wang, G.-f. (2024). A deep learning-based algorithm for online detection of small target defects in large-size sawn timber. *Industrial Crops and Products*, *222*, Article 119671. https://doi.org/10.1016/j.indcrop.2024.119671

[19] Kamal, K., Qayyum, R., Mathavan, S., & Zafar, T. (2017). Wood defects classification using laws texture energy measures and supervised learning approach. *Advanced Engineering Informatics*, *34*, 125-135. https://doi.org/10.1016/j.aei.2017.09.007

[20] Kılıç, K., Kılıç, K., Doğru, İ., & Özcan, U. (2025). WD Detector: Deep learning-based hybrid sensor design for wood defect detection. *European Journal of Wood and Wood Products*, *83*. https://doi.org/10.1007/s00107-025-02211-5

[21] Li, R., Zhong, S., & Yang, X. (2025). Wood panel defect detection based on improved YOLOv8n. *BioResources*, *20*, 2556-2573. https://doi.org/10.15376/biores.20.2.2556-2573

[22] Lin, X., et al. (2025). WDNET-YOLO: Enhanced deep learning for structural timber defect detection to improve building safety and reliability. *Buildings*, *15*(13), Article 2281. https://www.mdpi.com/2075-5309/15/13/2281

[23] Lukovic, M., et al. (2024). Probing the complexity of wood with computer vision: From pixels to properties. *Journal of The Royal Society Interface*, *21*(213). https://doi.org/10.1098/rsif.2023.0492

[24] Nasir, V., Rahimi, S., Mohammadpanah, A., Hansen, E., & Sassani, F. (2024). Intelligent lumber production (Sawmill 4.0): Opportunities, challenges, and pathways to adoption. In M.-R. Alam & M. Fathi (Eds.), *Integrated systems: Data driven engineering* (pp. 213-231). Springer Nature Switzerland.

[25] Okano, M. T., Lopes, W. A. C., Ruggero, S. M., Vendrametto, O., & Fernandes, J. C. L. (2025). Edge AI for industrial visual inspection: YOLOv8-based visual conformity detection using Raspberry Pi. *Algorithms*, *18*(8), Article 510. https://www.mdpi.com/1999-4893/18/8/510

[26] Ozkan, M., Ozcan, C., & Gökmen, S. (2025). Deep learning based defect detection and quality classification on lamella pieces used in solid wood panel production. *Operations Research Forum*, *6*. https://doi.org/10.1007/s43069-025-00534-w

[27] Popa, S. E., Puiu, P. G., Andrioaia, D. A., Grigore, R. M., & Avădanei, R. L. (2026). Indirect estimation of absorbed infrared LED radiant power using contactless thermal sensing. *Sensors*, *26*(13), Article 4055. https://www.mdpi.com/1424-8220/26/13/4055

[28] Urbonas, A., Raudonis, V., Maskeliūnas, R., & Damaševičius, R. (2019). Automated identification of wood veneer surface defects using faster region-based convolutional neural network with data augmentation and transfer learning. *Applied Sciences*, *9*(22), Article 4898. https://www.mdpi.com/2076-3417/9/22/4898

[29] Wang, B., Wang, R., Chen, Y., Yang, C., Teng, X., & Sun, P. (2025). FDD-YOLO: A novel detection model for detecting surface defects in wood. *Forests*, *16*(2), Article 308. https://www.mdpi.com/1999-4907/16/2/308

[30] Wang, R., Chen, Y., Liang, F., Wang, B., Mou, X., & Zhang, G. (2024). BPN-YOLO: A novel method for wood defect detection based on YOLOv7. *Forests*, *15*(7), Article 1096. https://www.mdpi.com/1999-4907/15/7/1096

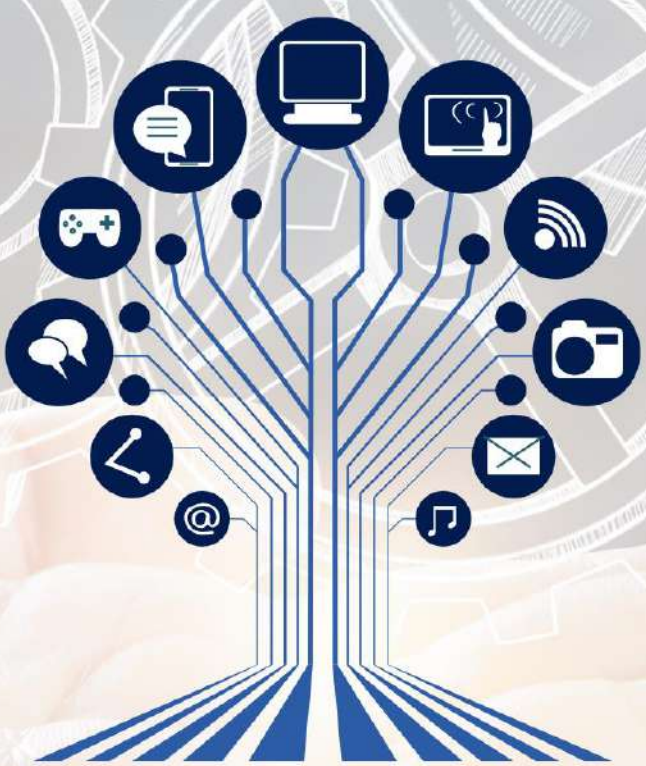
[31] Wang, R., Liang, F., Wang, B., Zhang, G., Chen, Y., & Mou, X. (2024). An efficient and accurate surface defect detection method for wood based on improved YOLOv8. *Forests*, *15*(7), Article 1176. https://www.mdpi.com/1999-4907/15/7/1176

[32] Wu, L., et al. (2026). A lumber surface defect detection network integrating deformable convolution and multi-scale attention. *Forests*, *17*(7), Article 839. https://www.mdpi.com/1999-4907/17/7/839

[33] Xi, H., Wang, R., Liang, F., Chen, Y., Zhang, G., & Wang, B. (2024). SiM-YOLO: A wood surface defect detection method based on the improved YOLOv8. *Coatings*, *14*(8), Article 1001. https://www.mdpi.com/2079-6412/14/8/1001.

IJOER
ENGINEERING JOURNAL

International Journal of Engineering Research and Science



Published by

AD Publications

Contact us +91-7665235235 www.ijoer.com info@ijoer.com