Optimizing E-Commerce with Ensemble Learning and Iterative Clustering for Superior Product Selection

Yuchen Liu[1,2], Meng Wang[1,2], Gangmin Li[3], Terry R. Payne[2], Yong Yue[1,2], and Ka Lok Man[1,2*]
[1] Department of Computing, Xi'an Jiaotong-Liverpool University
111 Ren'ai Rd, Suzhou, 215123, China
[e-mail: yuchen.liu21@student.xjtlu.edu.cn, meng.wang22@student.xjtlu.edu.cn, yong.yue@xjtlu.edu.cn, ka.man@xjtlu.edu.cn]
[2] Department of Computer Science, University of Liverpool
Brownlow Hill, Liverpool, L69 3BX, UK
[e-mail: t.r.payne@liverpool.ac.uk]
[3] HeXie Management Research Centre, College of Industry-Entrepreneurs (CIE),
Xi'an Jiaotong-Liverpool University
111 Ren'ai Rd, Suzhou, 215123, China
[e-mail: gangmin.li@xjtlu.edu.cn]
[*]Corresponding author: Ka Lok Man

Received February 21, 2024; revised May 30, 2024; revised August 15, 2024; accepted September 26, 2024; published October 31, 2024

Abstract

With the continuous growth of e-commerce sales, a robust product selection model is essential to maintain competitiveness and meet consumer demand. Current research primarily focuses on single models for sales prediction and lacks an integrated approach to sales forecasting and product selection. This paper proposes a comprehensive framework (VN-CPC) that combines sales forecasting with product selection to address these issues. We integrate a series of classical machine learning models, including Tree Models (XGBoost, LightGBM, CatBoost), Support Vector Machine (SVM), Bayesian Ridge, and Artificial Neural Networks (ANN), using a voting mechanism to determine the optimal weighting scheme. Our method demonstrates a lower Root Mean Square Error (RMSE) on collected Amazon data than individual models and other ensemble models. Furthermore, we employ a three-tiered clustering model: Initial Clustering, Refinement Clustering, and Final Clustering, based on our predictive model to refine product selection to specific categories. This integrated forecasting and selection framework can be more effectively applied in the dynamic e-commerce environment. It provides a robust tool for businesses to optimize their product offerings and stay ahead in a competitive market.

Keywords: E-commerce Sales Forecasting, Product Selection Strategy, Ensemble Machine Learning Models, Iterative Clustering Analysis, Product Attribute Extraction

1. Introduction

The e-commerce landscape is experiencing significant growth in platform expansion and sales volumes, transforming the global retail sector. On Black Friday 2023, consumers in the U.S. spent $9.8 billion in online sales, marking a 7.5% increase from the previous year [1]. Many offline retailers and manufacturers are entering the digital marketplace to capitalize on this e-commerce surge, aiming to enhance their sales volumes and market reach [2]. However, for newcomers aspiring to enter this digital marketplace, selecting the right products to sell presents a formidable challenge. Typically, these newcomers rely on their limited past sales experience, often making impulsive decisions about their product lines that may not align with market demands. Alternatively, some sellers analyze existing data on e-commerce platforms, opting to sell products with high sales volumes. However, these products have been on the market for an extended period, benefiting from well-established data. New sellers lack such established metrics, making direct comparisons with top-performing products misleading. Furthermore, the complexity of this research issue is compounded by the diverse nature of products on e-commerce platforms. The vast array of categories makes developing a universal model applicable across different product types exceedingly challenging. Each category exhibits unique characteristics, consumer preferences, and market dynamics, necessitating tailored approaches. Additionally, e-commerce landscapes are characterized by high uncertainty and dynamism. This environment makes it difficult for models trained on historical data to adapt effectively to market changes. The rapid evolution of consumer trends, competition dynamics, and economic factors often render static solutions obsolete soon after development. As a result, there is a pressing need for adaptive solutions that can evolve with the market, ensuring relevance and accuracy in their predictive capabilities.

Consequently, there has been an increased scholarly focus on product sales volume prediction. Current research, however, faces several limitations: 1) Models often focus on products within a single category, resulting in poor data generalization [3,4]. 2) Many studies employ essential linear regression networks or neural network models, which are only effective for predicting simple data sets and need help with more complex features [5,6,7,8]. 3) Most research has been confined to predicting sales volumes of existing products, with little exploration into forecasting new product sales or selecting new products for market entry [9,10]. In response to these gaps, our study makes several contributions. Firstly, our research encompasses a broad range of products, surveying 18 items across four major categories. This extensive coverage ensures a more comprehensive understanding of product lines (dataset comparison shown in **Table 1**). Secondly, we adopt an integrated model in the prediction part rather than relying on a single model. By amalgamating XGBoost, LightGBM, CatBoost, SVM, Bayesian Ridge, and ANN methodologies, our prediction model boasts enhanced generalization capabilities, making it suitable for predicting diverse products. Thirdly, we employ a three-tiered clustering methodology in the selection part to further refine and select products with specific attributes. Consequently, our VN-CPC (Voting for N base models and Clustering to Product Categories) effectively combines product forecasting and selection into a cohesive approach that can be adapted across various categories and products, assisting researchers and sellers in their endeavors. This paper is organized as follows: Section 2 delves into the existing literature on sales prediction and feature selection. Section 3 outlines the VN-CPC framework, and section 4 validates our framework's efficacy through empirical experiments on Amazon and extracting key product features. Finally, Section 5 concludes the paper, offering insights for future research direction.

Table 1. Sales Prediction Dataset Comparison

Comparison	BSP[3]	ESDP[4]	DSF[5]	M-GAN-XGBOOST[8]	VN-CPC
Categories	1	1	2	9	**18**
Samples	1,408	2,000	16,332	12,068	**24,834**
Features	8	10+	20+	30+	**60+**

2. Related Work

2.1 Machine Learning in Sales Forecasting

Sales Forecasting-based machine learning has emerged as a pivotal element in e-commerce, reshaping the landscape of business intelligence and strategic planning. As digital commerce experiences unprecedented growth, the need for sophisticated predictive models has become paramount, moving beyond traditional sales data analysis to encompass factors such as customer behaviors, product attributes, and promotional activities. This evolution underscores a significant shift towards leveraging advanced machine learning techniques, which have shown remarkable potential in deciphering complex sales patterns. Among these, neural networks, logistic regression, and support vector regression stand out for their ability to distill insights from vast e-commerce sales datasets. Deep neural framework for sales forecasting in e-commerce (DSF)'s [5] sequence-to-sequence learning approach uses an innovative product sales residual network to consider competition and promotional dynamics. Similarly, Bandara et al.'s [11] deployment of Long Short-term Memory (LSTM) neural networks is used to tailor forecasting to the given demands of e-commerce, while the work by Liu et al. [12] underscores the critical role that comprehensive feature engineering has improving sales prediction accuracy. The hybrid autoregressive integrated moving average and nonlinear autoregressive neural network (ARIMA-NARNN) model [13] further illustrate the effectiveness of combining traditional and neural network models to capture the multifaceted nature of e-commerce sales fluctuations. A similar work [14] forecasts the impact of COVID-19 on cigarette sales across Spanish provinces. The analysis compares sales data with model predictions to quantify the pandemic's effect. Petroșanu et al.'s [7] Directed Acyclic Graph (DAG) architecture represents cutting-edge advancements in sales forecasting, offering granular and long-term projections. The integration of deep learning for catalog management by Sales et al. [15] and the innovative Super Imperial Crown Model (SICM) for online agricultural trading by Mu et al. [16] illustrate the versatility of machine learning applications across diverse e-commerce segments. Moreover, Schneider and Gupta's [17] consumer review-based forecasting approach reflects the continuous quest for models that accommodate the intricate web of sales influencers. The explorations into e-commerce sales modes and information sharing by Yang et al. [18] and Makkar et al. 's [19] application of the autoregressive integrated moving average (ARIMA) model for forecasting further delineate the breadth of research on refining sales prediction mechanisms in the digital commerce domain.

2.2 Feature Selection and Multi-objective Optimization

Multi-objective optimization in sales forecasting is an essential component that enhances the predictive power of machine learning models, ensuring a delicate balance between model complexity and accuracy. Liu et al. [12] underscore the importance of meticulous feature

engineering in e-commerce. This demonstrates that model training and ensemble techniques can improve a classification model's efficacy for predicting repeat buyers. Several methods have emerged that enhance the use of feature selection. For example, Cheriyan et al. [20] demonstrate how the Gradient Boost Algorithm could maximize forecasting accuracy, thus illustrating the potential of sophisticated data mining techniques in refining prediction models. Likewise, the incorporation of online User-Generated Content (UGC) analysis by Chong et al. [21] represents a paradigm shift in feature selection, where the dynamics of online reviews and sentiments have become integral in forecasting e-commerce sales. The DSF model [5] illustrates the importance of using a diverse feature set, including user behavior and promotional campaigns, for navigating the e-commerce sector's ever-evolving landscape. The quest for optimal feature selection is further exemplified by Singh et al. [22], who advocate using forecasting methodologies based on the unique sales patterns observed in online retail, particularly during peak shopping seasons. Chen et al. [23] explore the challenges of price dispersion and forecasting within the niche baby food market in China's cross-border e-commerce by offering empirical insights into the synergies between price dispersion itself, seller reputation, and sales volume. This contrasted with the work of Krasonikolakis et al. [24], which offers novel insights into consumer behavior within virtual worlds, focusing on store selection criteria and sales prediction, diverging from traditional and online retail paradigms. The challenge of forecasting demand for short-life cycle products is explored by Afifi [25], who highlights the volatility and unpredictability inherent in industries such as fashion and technology. Recent research highlights that analyzing textual information, such as news sentiment, can enhance prediction accuracy [26]. Likewise, Thorleuchter et al. [27] investigate the impact of textual information on e-commerce websites, specifically looking at how it affects the success trajectory of top e-commerce entities. Shi et al. [28] focus on the complex inventory management landscape used by a Chinese fashion retailer engaged in cross-border e-commerce to understand the intricacies of managing a broad product offering and the associated risks. Cremer et al. [29] explore the role of scarcity messages in e-commerce, unraveling how such messages sway sales across different purchase stages, particularly for “long tail” goods. Flash sales in e-commerce can be characterized by ephemeral high discount offers that substantially alter user preferences and available product assortments. As such, this poses several unique challenges studied by Li et al. [30]. In contrast, Pålsson et al. [31] investigate to demystify the energy efficiency conundrum between e-commerce and traditional in-store shopping. This results in some conclusions on the factors influencing energy consumption across these retail modalities. In a strategic exploration, Zhang et al. [32] examine firms' manipulation of sales volumes on e-commerce platforms, using a game-theoretic framework to understand its repercussions on market dynamics and platform profitability. Finally, by studying the complexities of promotional strategies on e-commerce platforms, Tong et al. [33] examine the differential impacts of various promotional types on online sales and conversion rates within a hierarchical promotional structure. This diverse set of studies highlights the multifaceted nature of sales forecasting in e-commerce by addressing different approaches to feature selection and multi-objective optimization.

2.3 Customer Purchase Behavior Prediction in E-Commerce

Exploring customer purchase behavior prediction in e-commerce has been a focal point for researchers aiming to understand the intricate dynamics of online shopping. Cirqueira et al. [34] provide a comprehensive review of the literature on this subject, laying down an analytical framework that supports future investigations into consumer behavior online. Building on this

foundation, Gordini and Veglio [35] focus on the Business-to-Business (B2B) sector, introducing innovative methodologies for churn prediction, which is pivotal for sustaining long-term customer relationships and ensuring business continuity. Zhu et al. [36] address the challenge of predicting product returns, a critical aspect of e-commerce logistics, using a weighted hybrid graph (HyGraph) that aims to decode the complexities of online shopping behaviors and return patterns, offering e-tailers a tool to mitigate the impact of returns on their operations. The use of e-commerce analytics is illustrated by Chaudhuri et al. [37], where the impact of online customer engagement on e-commerce performance was examined, contrasting deep learning techniques with traditional machine learning models to predict purchase behaviors based on platform engagement and customer characteristics. This study exemplifies how advanced computational methods are increasingly employed to decipher consumer intent. Similarly, Vallés-Pérez et al. [38] and Liu et al. [39] focus on sales forecasting and search engine optimization in e-commerce, respectively. Vallés-Pérez et al. employ deep learning for fine-grained sales predictions, whereas Liu et al. propose a Cascade ranking model to enhance search efficiency and effectiveness on e-commerce platforms. These studies demonstrate the importance of integrating cutting-edge technologies to improve the digital marketplace's operational performance and customer experience. Peng et al. [40] and Xu et al. [41] explore the dynamics of online flash sales and live-streaming e-commerce, respectively, highlighting several factors influencing consumer purchase intentions and the pivotal role of anchor reputation in driving sales. These studies highlight the unique challenges and opportunities presented by emerging e-commerce formats and the need for tailored strategies to capitalize on them. Lastly, Mu et al. [42] examine the strategic interplay between online preferences, retail services, and dual-channel supply chain management within the context of sustainable development. Their research, grounded in consumer utility selection theory, offers valuable insights into optimizing business operations in a digitally transformed landscape while adhering to sustainable practices. Collectively, these studies provide several insights into customer purchase behavior prediction in e-commerce, each contributing unique perspectives and methodologies that enrich our understanding of the digital consumer psyche and the operational strategies that e-commerce platforms can employ to navigate the complexities of the online market.

Existing studies have concentrated on forecasting sales volumes or understanding consumer behaviors in isolation, often neglecting the critical decision-making phase that follows prediction. This results in a lack of comprehensive frameworks that integrate robust predictive analytics with actionable strategies for product selection and portfolio optimization. Our research advances the field of e-commerce sales forecasting by integrating an ensemble of machine-learning models for enhanced predictive accuracy. Also, it introduces a novel tiered clustering model for strategic decision-making. This approach differs from existing studies focusing solely on prediction by improving model generalizability across various product categories and facilitating informed product selection through a systematic filtering process. Our innovative framework bridges the gap between prediction and practical application, offering a comprehensive solution for dynamic decision-making in the e-commerce landscape.

3. Methodology

The VN-CPC framework (illustrated in **Fig. 1**) represents an innovative approach to forecasting sales and identifying key product attributes in the e-commerce domain. The framework consists of two main parts: an ensemble prediction model and a three-tiered

clustering model. The ensemble prediction model combines the strengths of six foundational models: three tree-based models (XGBoost [43], LightGBM [44], and CatBoost [45]), SVM [46], Bayesian Ridge [47], and ANN [48]. The tree-based models are effective for structured data, SVM handles high-dimensional classification, Bayesian Ridge provides probabilistic inference, and ANN models complex nonlinear relationships and then employs Optuna [49] to find the optimal combination of base model weights, ensuring robust sales forecasts for new products. The ensemble approach is designed to capture a broad spectrum of patterns and relationships within the data, thereby providing a robust and nuanced understanding of sales dynamics [50,51]. Building on the initial sales forecasts, the framework further refines its analysis through a systematic three-tier clustering process, updated from K-means [52]. This process narrows down product attributes to those most indicative of sales success, enhancing product selection strategies. The multi-stage clustering method filters vast product datasets, pinpointing essential features that guide product selection. Below are details of each part employed in our framework.

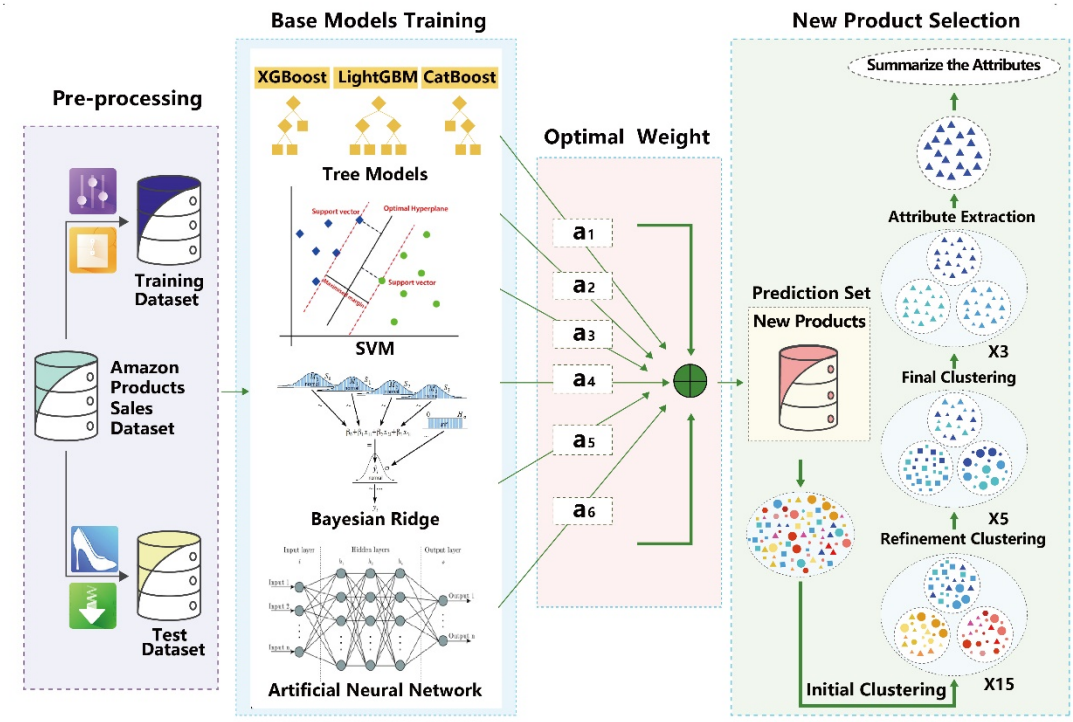

Fig. 1. The VN-CPC Framework

3.1 Ensemble Sales Prediction Model

The VN-CPC framework's first part integrates six foundational models to form an ensemble prediction model. Each model brings unique strengths, and their combined use ensures a comprehensive approach to sales forecasting.

3.1.1 XGBoost

XGBoost [43], an enhanced version of gradient-boosted decision trees, is tailored for speed and high performance. It is particularly notable for its proficiency in managing sparse datasets and broad applicability across various regression and classification challenges. The core algorithm can be represented as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \tag{1}$$

here $\hat{y}_i$ is the predicted sales value for the i-th instance, f_k are the decision trees based on products, x_i are the product features and $\mathcal{F}$ is the space of all possible trees.

3.1.2 LightGBM

A streamlined gradient boosting framework that employs tree-based learning algorithms. LightGBM [44] is notable for its remarkable efficiency in managing vast datasets and high computational speed, making it exceptionally suited for scenarios demanding rapid processing. The model can be mathematically described as:

$$\hat{y}_i = \gamma + \sum_{t=1}^{T} \eta \cdot h_t(x_i) \tag{2}$$

where $\hat{y}_i$ is the product sales prediction for the i-th instance, γ is the initial sales prediction at a random state, η is the learning rate between $[0,1]$, h_t represents the decision trees at iteration t based on products and x_i denotes the product feature set of the i-th instance.

3.1.3 CatBoost

CatBoost [45] builds an ensemble of decision trees to improve prediction accuracy. The final prediction is the sum of predictions from all individual sales prediction trees. One advantage of the approach is its ability to handle categorical features directly without requiring extensive preprocessing like one-hot encoding. Furthermore, it utilizes innovations such as Ordered Boosting and Target Statistics for categorical features that reduce overfitting and improve model generalization.

3.1.4 SVM (Support Vector Machine)

SVM [46] is a robust and adaptable machine learning model that excels in classification and regression tasks within high-dimensional spaces. It identifies the optimal hyperplane that segregates data points across different classes. The decision function for SVM, $f(x)$ can be formulated as follows:

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right) \tag{3}$$

where α_i are the Lagrange multipliers, y_i are the sales prediction class labels, $K(x_i, x)$ is the kernel function, and b is the bias.

3.1.5 Bayesian Ridge

Bayesian Ridge Regression [47] applies Bayesian inference to linear regression, introducing a prior distribution over the regression coefficients, typically Gaussian. This approach updates the coefficients' distribution based on the observed data, balancing the preceding belief and the data likelihood. The model is given by:

$$y = X\beta + \varepsilon \tag{4}$$

where y is the sales prediction for products, X represents the matrix of product features, β denotes the vector of regression coefficients and ε represents the error terms, assumed to be normally distributed. The fundamental equations include the posterior distribution of β, derived from Bayes' theorem:

$$P(\beta \mid X, y) \propto P(y \mid X, \beta)P(\beta) \tag{5}$$

where $P(\beta \mid X, y)$ is the posterior probability of the coefficients given the data and $\propto$ denotes proportionality. $P(y \mid X, \beta)$ is the likelihood of observing the data given the coefficients and $P(\beta)$ represents the prior probability of the coefficients.

3.1.6 ANN (Artificial Neural Networks)

Inspired by the structure and function of biological neural networks, ANN [48] captures complex nonlinear relationships through their layered architecture. They are highly adaptable and can be used for various tasks, from regression to classification.

$$y = \phi\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{6}$$

where ϕ is the activation function, w_i are the weights assigned to product features x_i, and b is the bias.

The VN-CPC framework uses Optuna [49], a hyperparameter optimization method, to optimize the predictive power of six foundational models. Optuna helps find the best base model weights, ensuring the ensemble model delivers robust and accurate sales forecasts for new products. This mechanism integrates the outputs from each model, leveraging their strengths and compensating for their weaknesses. Each model in the ensemble contributes to the final prediction, with weights assigned to reflect their relative importance and accuracy [53]. The ensemble is defined as :

$$\widehat{y_{ensemble}} = w_1 \cdot F_1(x) + w_2 \cdot F_2(x) + w_3 \cdot F_3(x) + w_4 \cdot F_4(x) + w_5 \cdot F_5(x) + w_6 \cdot F_6(x) \tag{7}$$

where $\widehat{y_{ensemble}}$ represents the output of the final ensemble sales prediction model in VN-CPC, which integrates the predictions from individual models—namely XGBoost, LightGBM, CatBoost, SVM, Bayesian Ridge, and ANN—through their respective prediction functions $F_1(x), F_2(x), F_3(x), F_4(x), F_5(x), F_6(x)$. The weights $w_1, w_2, w_3, w_4, w_5, w_6$ assigned from each corresponding model, the weighted sum equals 1.

3.2 Three-tiered Clustering Model

After completing the first part of a robust sales forecasting model, our next objective is to identify product attributes that influence sales performance. This endeavor extends beyond simple comparisons of predicted sales figures and requires a sophisticated approach to uncover the intrinsic qualities that drive sales success. To achieve this, we propose an iterative clustering-based model designed to systematically identify and highlight product attributes correlated with higher sales volumes.

Initially, we apply our forecasting model to a new sample batch, providing estimated sales values for each product. Products are initially grouped based on predicted sales and corresponding attributes, forming broad clusters that encapsulate many attributes. As we progress, each initial cluster undergoes further segmentation in subsequent iterations, focusing on products with higher predicted sales. These stages employ clustering algorithms that consider sales figures and delve into the multidimensional space of product attributes. This ensures that the clustering process is informed by both the potential for sales and the inherent characteristics of the products. Through this iterative refinement, we gradually eliminate products with lower sales forecasts and identify those with the highest sales potential. This methodical approach allows us to focus on products that promise high sales and share distinct attributes potentially indicative of their market success. The culmination of this process is extracting common characteristics from the final set of product clusters, which are then analyzed to gauge their impact on sales performance. By identifying recurrent attributes among the top-performing products, we can infer the features most likely to resonate with consumers and drive sales.

3.2.1 Initial Clustering

In the first phase, we employ the K-means [52] clustering method to group products into K_1 clusters based on their predicted sales and attributes:

$$C_{1j} = Cluster(P_i; K_1) \ \ for\ j = 1,2, \dots, K_1 \tag{8}$$

where C_{1j} denotes the j-th cluster in the initial round, P_i represents the i-th product, and K_1 is the predetermined number of clusters. We refine our clustering by adjusting initial conditions and performing 15 different samplings to address the potentially high cost of incorrect predictions at this stage. The optimal combination, selected based on frequency, is then advanced to the next stage.

3.2.2 Refinement Clustering

The initial clusters C_{1j} are further segmented into K_2 sub-clusters to refine the product groupings:

$$C_{2jk} = RefineCluster\left(C_{1j}; K_2\right) \ \ for\ k = 1,2, \dots, K_2 \tag{9}$$

where C_{2jk} indicates the k-th sub-cluster within C_{1j}, and K_2 is the number of sub-clusters. This stage employs 5 different samplings, with the optimal combination selected based on frequency to ensure robustness and accuracy before moving to the final stage.

3.2.3 Final Clustering for Attribute Extraction

The final round involves further analysis of each sub-cluster C_{2jk} to isolate products with the highest sales forecasts:

$$C_{3jkl} = FinalCluster(C_{2jk}; K_3) \ \ for\ l = 1,2, \dots, K_3 \tag{10}$$

where C_{3jkl} represents the final clusters focusing on top-performing products and K_3 is the number of final clusters isolating these products. This stage includes 3 different samplings, with the optimal combination selected based on frequency.

3.2.4 Attribute Extraction

After the final clustering round, we extract common attributes from top-performing products: where A_{top} denotes the attributes prevalent among products with the highest sales forecasts.

$$A_{top} = ExtractAttributes(C_{3jkl}) \tag{11}$$

This structured and iterative clustering framework enables a detailed analysis of product attributes associated with higher sales volumes, offering strategic insights for decision-making in e-commerce platforms. By focusing on attributes tied to higher sales predictions and employing a phased filtering approach, we can isolate critical factors contributing to product success within the e-commerce domain.

4. Experiment Analysis

4.1 Dataset

We collected product data from Amazon across 18 specific categories, ensuring each category contained 1,500 to 2,000 products for uniformity. The data was gathered from April 15, 2024, to May 15, 2024, resulting in 24,834 entries. We divided the dataset into three subsets: 14,898 entries for the training set, 4,968 for the test set, and another 4,968 for the prediction set. The split ratio of the dataset is 0.6:0.2:0.2. For product features, we categorized them into two types: common features and unique product-specific features, with 16 common features and 47 unique product-specific features, totaling 63 features. We processed these features using one-hot encoding. In the Appendix, **Table 4** provides examples of these product features. The training set was used to train our models, the test set was utilized to compare the performance between different models, and the prediction set was used in the subsequent three-tier clustering model. As seen in **Fig. 2**, the 18 specific product categories are evenly distributed among four major categories: Electronics, Home, Kitchen, and Personal Care. The graph shows the sales distribution per price unit, indicating that most products fall into the lower sales categories. As sales per price unit increase, the number of products in each category decreases. This trend is consistent across all four major categories, suggesting that it is impossible to select suitable categories directly based on the distribution alone.

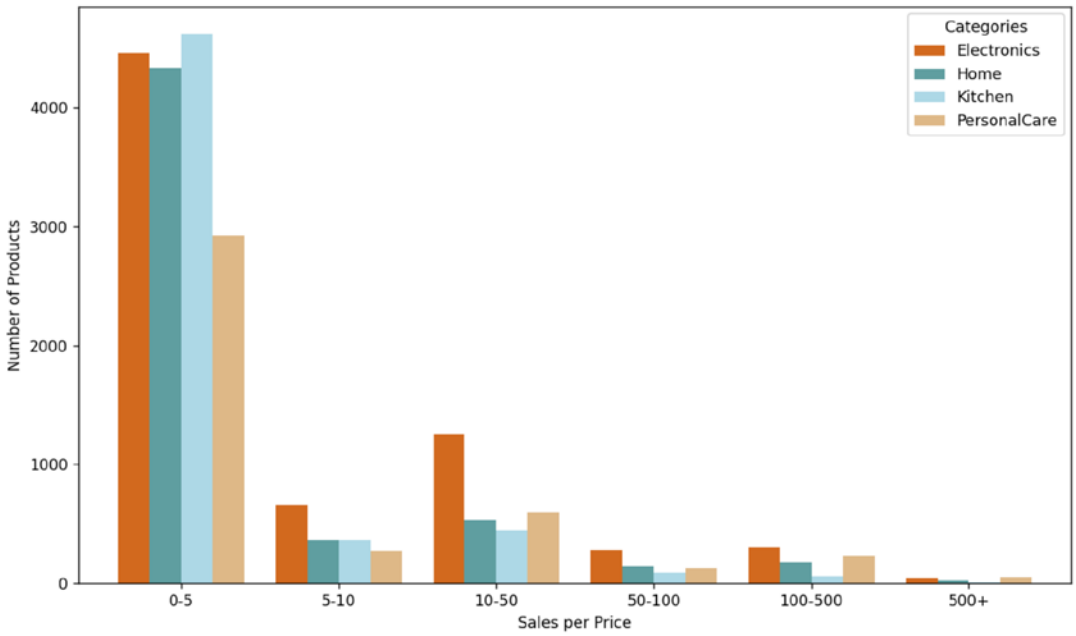

Fig. 2. Distribution of Sales per Price across Major Categories

4.2 Model Training and Convergence Analysis

We conducted comprehensive training on six base models using a dataset of 14,898 samples to analyze the VN-CPC framework empirically. Each model was trained for 100 epochs to observe the Root Mean Square Error (RMSE) convergence across the training sets. The hyperparameters for the tree-based models were set as follows: number of estimators = 100, maximum tree depth = 5, minimum samples per leaf = 20, and learning rate = 0.1. For the ANN model, we set the batch size = 32 and the learning rate = 0.001. As depicted in **Fig. 3**, tree-based models (XGBoost, LightGBM, and CatBoost) demonstrated superior convergence properties and performed commendably during the training phase. Traditional models such as SVM and Bayesian Ridge also achieved convergence but to a lesser extent than the tree-based models. The ANN model exhibited some fluctuations with minimal changes but reached a low point early in the training, indicating the fastest RMSE decline. This fluctuation might be due to a high learning rate, leading to instability. Subsequently, we utilized Optuna for hyperparameter tuning to optimize the weights within the VN-CPC sales prediction model. Initially, we set equal weights for the six base models, approximately 0.16 each. We then conducted the tuning in phases: Phase 1 at 25 iterations, Phase 2 at 50 iterations, and finally, at 100 iterations. As shown in **Fig. 4**, the RMSE progressively decreased with weight optimization. After 100 iterations (depicted by the purple line), the results outperformed those from Phase 1 and Phase 2, reaching the lowest RMSE. The optimized weight distribution at this point was XGBoost (0.41), LightGBM (0.19), CatBoost (0.17), SVM (0.07), Bayesian Ridge (0.05), and ANN (0.11). We also conducted an ablation study, maintaining the weight proportions of the three tree models from the optimal weights and removing SVM, Bayesian Ridge, and ANN. The convergence (indicated by the green line) was inferior to the optimal weight combination. This demonstrates that, despite their smaller weights, SVM, Bayesian Ridge, and ANN contributed to the overall optimization of the model. Additionally, the optimized weights of VN-CPC demonstrated superior performance compared to individual base models (indicated by the dark blue line in **Fig. 3**).

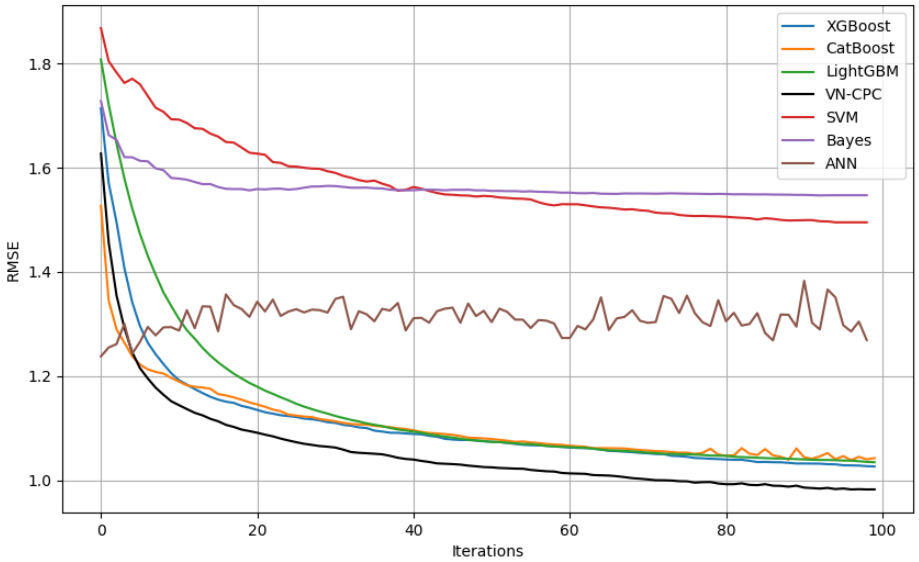

Fig. 3. Comparative Convergence Analysis of Different Models

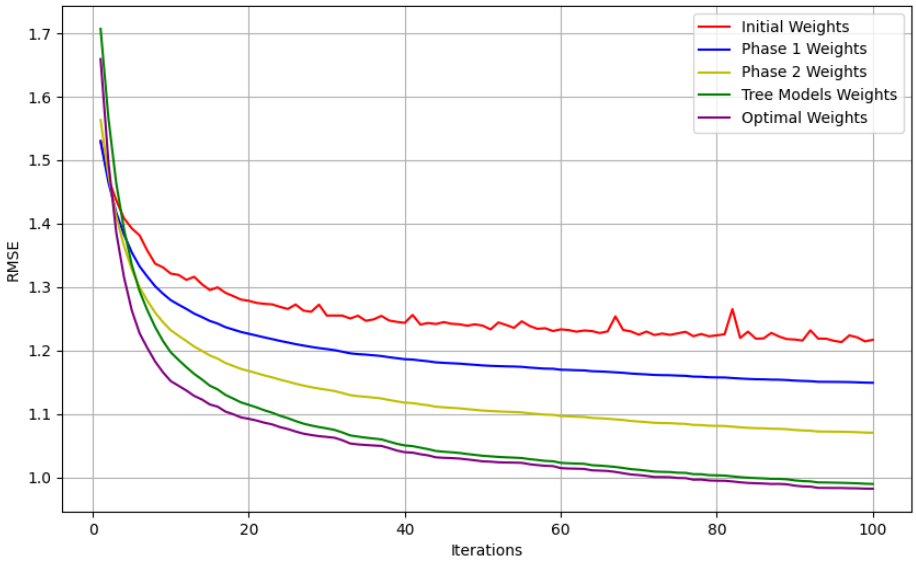

Fig. 4. Comparative Convergence Analysis of Different Weights Combination of VN-CPC

4.3 Test for Different Categories

We test predictive models for various appliance categories on Amazon's 4,968 samples. As shown in **Table 2**, “I” refers to “Consumer Electronics”, indicating the electronic devices typically used for daily activities such as entertainment, communication, and office work. “II” represents “Home Appliance”, which encompasses devices that assist in household functions such as cleaning, cooking, and home maintenance. “III” is designated for “Kitchen Appliance”, referring to the subset of home appliances used explicitly in the kitchen for food preparation

and storage. Lastly, “IV” denotes “Personal Care Appliance”, which includes electrical devices for personal grooming and healthcare purposes. The performance metrics used to evaluate the models were Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), which are standard in the field for quantifying prediction accuracy. For Consumer Electronics (I), VN-CPC achieved the lowest MSE (0.995) and RMSE (0.998), improving upon the next best model, XGBoost, by 9.0% and 4.5%, respectively. Its MAE (0.754) was also the lowest, outperforming XGBoost (0.792) by 4.8%, LightGBM (0.824) by 8.5%, and CatBoost (0.845) by 10.8%. For Home Appliance (II), VN-CPC demonstrated superior performance with an MSE of 0.892 and an RMSE of 0.944, improving upon XGBoost by 6.0% and 3.1%, respectively. Its MAE (0.646) was 3.6% better than XGBoost (0.670), 4.9% better than LightGBM (0.679), and 6.5% better than CatBoost (0.691). For Kitchen Appliance (III), VN-CPC's MSE (0.710) and RMSE (0.840) were the lowest, showing improvements of 1.5% and 1.1% over XGBoost. The MAE (0.537) of VN-CPC outperformed XGBoost (0.550) by 2.4%, LightGBM (0.557) by 3.6%, and CatBoost (0.584) by 8.1%. For Personal Care Appliance (IV), VN-CPC had the lowest MSE (1.213) and RMSE (1.101), outperforming the next best model, XGboost, by 7.0% and 3.6%, respectively. Its MAE (0.812) was 3.8% better than XGBoost (0.844), 6.7% better than LightGBM (0.870), and 8.1% better than CatBoost (0.884). Overall, VN-CPC demonstrated the lowest MSE, RMSE, and MAE values across all categories, indicating its superior predictive performance. It performed best in the Consumer Electronics category, with the most significant improvements over other models.

Table 2. Metrics of Different Models for E-commerce Sales Prediction

Model	MSE				RMSE				MAE			
	I	II	III	IV	I	II	III	IV	I	II	III	IV
XGboost	1.093	0.949	0.721	1.304	1.045	0.974	0.849	1.142	0.792	0.670	0.550	0.844
LightGBM	1.168	0.994	0.761	1.347	1.081	0.997	0.873	1.161	0.824	0.679	0.557	0.870
CatBoost	1.188	0.994	0.799	1.358	1.090	0.997	0.894	1.165	0.845	0.691	0.584	0.884
SVM	2.352	2.264	1.410	2.844	1.534	1.505	1.188	1.687	1.119	0.964	0.696	1.228
Bayes	2.457	2.218	1.456	3.450	1.567	1.489	1.207	1.858	1.241	1.118	0.840	1.464
ANN	1.650	1.281	1.297	1.866	1.284	1.132	1.139	1.366	0.968	0.817	0.673	1.055
VN-CPC	**0.995**	**0.892**	**0.710**	**1.213**	**0.998**	**0.944**	**0.840**	**1.101**	**0.754**	**0.646**	**0.537**	**0.812**

As seen in **Table 3**, the VN-CPC prediction model with optimal weights demonstrated the best performance, reducing RMSE and MSE values compared to the initial weights and Phase 1 and Phase 2 weights. For example, the optimal weights resulted in an RMSE of 0.998 for Consumer Electronics (I), compared to 1.141 with initial weights, 1.139 in Phase 1, and 1.076 in Phase 2. This trend was consistent across other categories, with the optimal weights consistently providing the best performance. Additionally, an ablation study was conducted where SVM, Bayesian Ridge, and ANN were removed while maintaining the tree models' weight proportions. This resulted in a performance decline (Ablation), as indicated by higher RMSE and MSE values than the optimal weights. For instance, the ablation study's RMSE for Consumer Electronics (I) was 1.011, higher than the optimal weight RMSE of 0.998. Similar trends were observed across other categories, demonstrating that although SVM, Bayesian Ridge, and ANN contributed smaller weights, their inclusion improved the model's overall performance and convergence stability. The VN-CPC model with optimal weights outperformed other weight configurations and individual base models, confirming the effectiveness of our ensemble approach and the importance of including a diverse set of base models to enhance predictive accuracy and stability. Meanwhile, we compared three common

ensemble models [54]: Bagging, Boosting, and Stacking with our VN-CPC model. As shown in **Fig. 5**, VN-CPC consistently outperformed the other models across all four product categories in both RMSE and MAE metrics. Notably, Bagging exhibited the weakest performance, suggesting that splitting the dataset for independent model training is not ideal for sales forecasting tasks. Although Boosting and Stacking produced similar results, both were surpassed by VN-CPC. This improvement can be attributed to VN-CPC's ability to combine the strengths of multiple base models effectively. Unlike Boosting, which may overly focus on reducing bias and can sometimes amplify noise, and Stacking, which may not fully capture the complexities of the data due to potential overfitting, VN-CPC uses a carefully balanced ensemble and optimized weighting mechanism to address these limitations, resulting in more accurate and stable predictions.

Table 3. Metrics of Different Weights Combination of VN-CPC

Weights	MSE				RMSE				MAE			
	I	II	III	IV	I	II	III	IV	I	II	III	IV
Initial	1.302	1.089	0.824	1.547	1.141	1.044	0.908	1.244	0.885	0.744	0.592	0.979
Phase 1	1.298	1.115	0.829	1.573	1.139	1.056	0.911	1.254	0.883	0.747	0.593	0.975
Phase 2	1.157	0.996	0.763	1.376	1.076	0.998	0.873	1.173	0.819	0.684	0.549	0.888
Ablation	1.022	0.894	0.710	1.234	1.011	0.946	0.842	1.111	0.768	0.646	0.537	0.825
Optimal	**0.995**	**0.892**	**0.710**	**1.213**	**0.998**	**0.944**	**0.840**	**1.101**	**0.754**	**0.646**	**0.537**	**0.812**

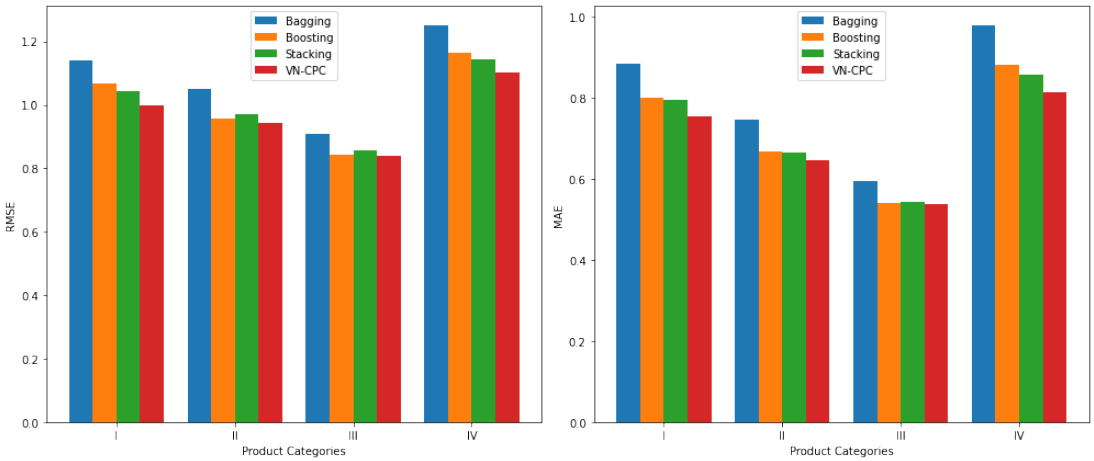

Fig. 5. Comparison of RMSE and MAE for Different Ensemble Models

4.4 Predictive Analysis of Emerging Product Sales

In the subsequent phase of our experiment, we used the remaining 4,968 new products, each in the nascent stage of their market lifecycle. Utilizing the optimal weight VN-CPC prediction model, we projected the potential sales for these products and initiated a three-tiered clustering model within VN-CPC. Referencing **Fig. 6**, our initial clustering identified three broad categories for the 4,968 newly introduced products, resulting in clusters containing 1,120, 212, and 3,636 samples, respectively. Upon careful inspection, we found that the third cluster, primarily home-based items (e.g., coffee machines, small humidifiers) and kitchen appliances (e.g., electric mixers, juicer machines, toasters), had a higher projected sales volume. In the second phase, Refinement Clustering, we focused on the 3,636 promising samples from the third cluster. These were further segmented into three refined sub-clusters of sizes 2,241, 1,121, and 274. This analysis highlighted the second sub-cluster (e.g., coffee machines, electric

mixers, juicer machines, air fryers) with the highest sales forecast. In the final stage, Final Clustering, we concentrated on the 1,121 samples from the selected sub-cluster, extracting specific product attributes. These were then subdivided into three more focused groups containing 280, 561, and 280 samples. This systematic approach to clustering culminated in identifying standout product attributes from the final clusters, thereby streamlining our product selection process. Our findings, particularly from the coffee machine category, highlighted vital attributes such as a milk frother, machine type, stainless steel construction, reusability, fully automatic operation, and classic black color. These insights inform our product selection strategy, ensuring our choices are firmly rooted in data-driven attributes poised to resonate well with the market and drive sales. The VN-CPC enhances our product selection strategy and sets a precedent for future expansions across various categories.

5. Conclusion

Our VN-CPC framework provides meaningful advancements in e-commerce sales forecasting and product selection. By integrating the strengths of various machine learning models through an ensemble mechanism, the VN-CPC prediction part demonstrates superior accuracy and reliability in predicting sales volumes across diverse product categories compared to individual prediction models and other ensemble methods. It effectively navigates the complexities of e-commerce data, providing valuable insights into future sales with minimized error margins, as reflected by the lower RMSE scores. This predictive proficiency is crucial for businesses aiming to optimize inventory management, tailor marketing strategies, and enhance customer satisfaction through improved product availability forecasting. Furthermore, the VN-CPC framework incorporates a three-tier clustering process: Initial Clustering, Refinement Clustering, and Final Clustering. This process enables explicit and informed decision-making for product selection, reducing the randomness often encountered in e-commerce inventory management. By systematically filtering through large product datasets, our framework identifies key product features at each clustering stage, ensuring that the selected products align with sales-boosting attributes.

However, the VN-CPC framework has its limitations. While the ensemble technique offers enhanced accuracy, it also introduces additional computational complexity, leading to increased processing time and resource demands, which may pose challenges for real-time analysis and scalability to larger datasets. Additionally, the dataset used in this study has limitations, particularly in the time frame of data collection, which could affect the generalizability of the results. Future research should focus on incorporating real-time market analytics to improve the model's responsiveness to immediate trends, exploring more computationally efficient algorithms to reduce resource demands, expanding the range of product categories, and extending the data collection period to enhance the robustness and applicability of the VN-CPC framework.

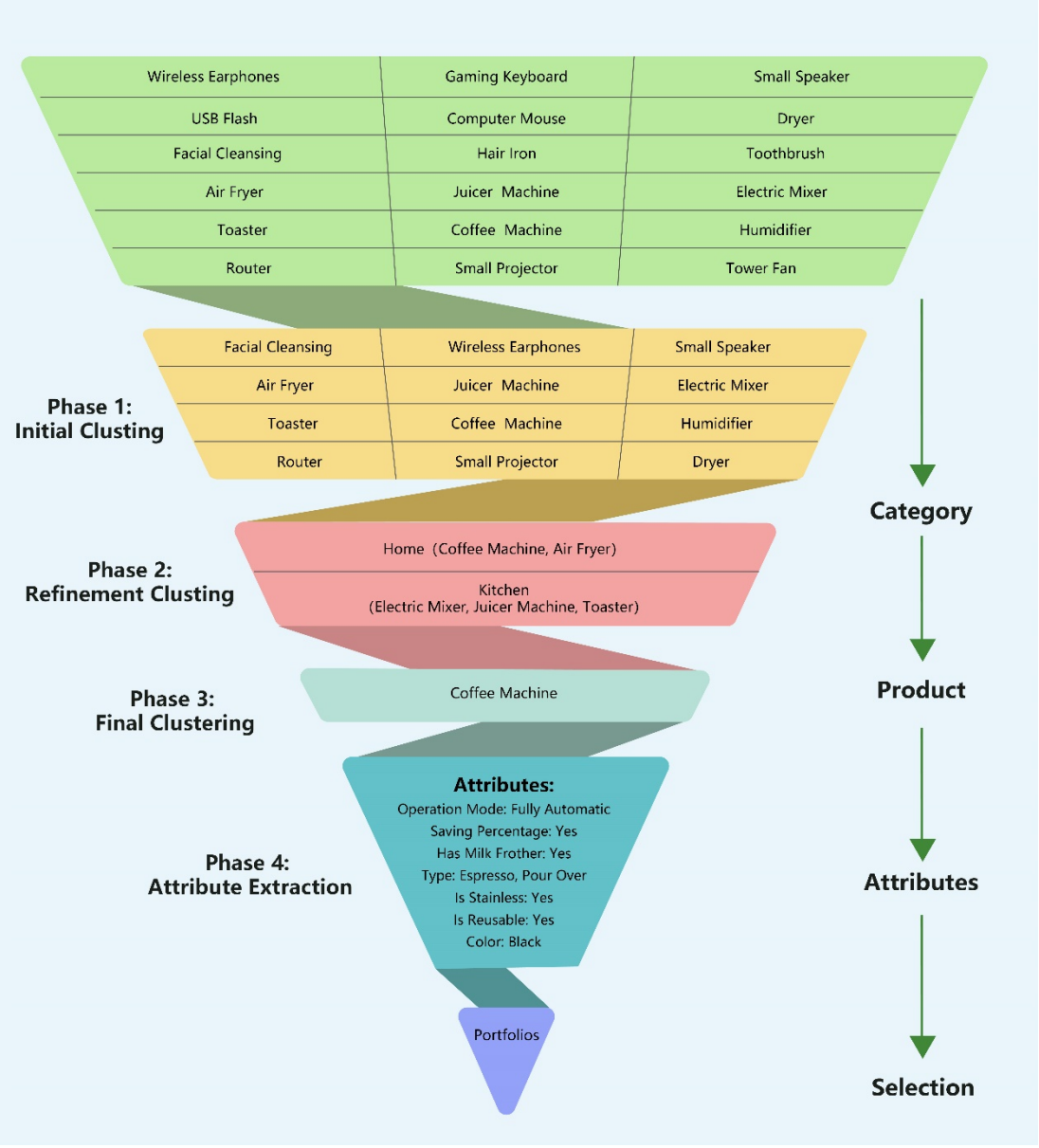

Fig. 6. Hierarchical Clustering Results

Appendix

As seen in **Table 4**, it provides examples of product features. We categorized the features into two types: common and unique product-specific features, with 16 common features and 47 unique product-specific features, totaling 63. We show some of the unique features below.

Table 4. Examples of Product Features

Common Features	Example	Unique Features	Description	Product Categories
Brand	Apple	Connectivity	Connect mode between the device and controller	Gaming keyboard; Computer mouse; USB flash; Wireless earphones; Small speaker
Color	Black	Storage	The memory size of the USB flash	USB flash
Manufacturer	Dell	Form	In-ear form	Wireless earphones
Price	15.99	Number of keys	Key count of keyboard	Gaming keyboard
Rating	4.4	Has milk frother	Whether it contains a milk frother	Coffee maker
Number of rating	64	Has 802.11ax	Whether it has 802.11ax	WIFI router
Price of delivery	1.99	Noise level	Runtime noise level	Tower fan
Days of delivery	5	Has brushing timer	Timer included or not	Electronic toothbrush
Launch time	2023	Is 1080P	1080P included or not	Small projector
Country	China	Is quiet	Whether it is quiet while running	Small humidifier; Juicer
Saving percentage	-0.2	Is dishwasher safe	Whether it is dishwasher-safe	Air fryer; Mixer; Juicer
Weight (pounds)	0.70625	Powersource	Charge mode	Hair dryer; Facial cleansing device
Length	4.1	Capacity	Equipment capacity	Air fryer; Mixer; Juicer; Small humidifier; Coffee maker
Width	7.9	Firmness description	The hardness of the brush	Electronic toothbrush
Height	1.4	Pack	Number of items in each package	USB flash
Rank	725	…	…	…

Acknowledgement

This work is partially supported by the XJTLU AI University Research Centre and Jiangsu Province Engineering Research Centre of Data Science and Cognitive Computation at XJTLU and SIP AI innovation platform (YZCXPT2022103). Also, it is partially funded by the Suzhou Municipal Key Laboratory for Intelligent Virtual Engineering (SZS2022004) as well as funding: XJTLU Key Program Special Fund (KSF-A-17).

References

[1] R. Picciotto, Black Friday shoppers spent a record $9.8 billion in U.S. online sales, up 7.5% from last year, Nov. 25, 2023. [Online]. Available: https://www.cnbc.com/2023/11/25/black-friday-shoppers-spent-a-record-9point8-billion-in-us-online-sales-up-7point5percent-from-last-year.html

[2] Y. Liu, K. L. Man, G. Li, T. Payne, and Y. Yue, "Dynamic Pricing Strategies on the Internet," in *Proc. of International Conference on Digital Contents: AICo (AI, IoT, and Contents) Technology*, 2022. Article (CrossRef Link)
[3] S. K. Sharma, S. Chakraborti, and T. Jha, "Analysis of book sales prediction at Amazon marketplace in India: a machine learning approach," *Information Systems and e-Business Management*, vol.17, no.2-4, pp.261-284, 2019. Article (CrossRef Link)
[4] Y. Liu, K. L. Man, G. Li, T. Payne, and Y. Yue, "Enhancing Sparse Data Performance in E-Commerce Dynamic Pricing with Reinforcement Learning and Pre-Trained Learning," in *Proc. of 2023 International Conference on Platform Technology and Service (PlatCon)*, pp.39-42, IEEE, 2023. Article (CrossRef Link)
[5] Y. Qi, C. Li, H. Deng, M. Cai, Y. Qi, and Y. Deng, "A Deep Neural Framework for Sales Forecasting in E-Commerce," in *Proc. of the 28th ACM International Conference on Information and Knowledge Management*, pp.299-308, 2019. Article (CrossRef Link)
[6] S. Neelakandan, V. Prakash, M. S. PranavKumar, and R. Balasubramaniam, "Forecasting of E-Commerce System for Sale Prediction Using Deep Learning Modified Neural Networks," in *Proc. of 2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)*, pp.1-5, IEEE, 2023. Article (CrossRef Link)
[7] D.-M. Petroșanu, A. Pîrjan, G. Căruțașu, A. Tăbușcă, D.-L. Zirra, and A. Perju-Mitran, "E-Commerce Sales Revenues Forecasting by Means of Dynamically Designing, Developing and Validating a Directed Acyclic Graph (DAG) Network for Deep Learning," *Electronics*, vol.11, no.18, 2022. Article (CrossRef Link)
[8] S. Wang and Y. Yang, "M-GAN-XGBOOST model for sales prediction and precision marketing strategy making of each product in online stores," *Data Technologies and Applications*, vol.55, no.5, pp.749-770, 2021. Article (CrossRef Link)
[9] Y. Liu, K. L. Man, G. Li, T. R. Payne, and Y. Yue, "Evaluating and Selecting Deep Reinforcement Learning Models for OptimalDynamic Pricing: A Systematic Comparison of PPO, DDPG, and SAC," in *Proc. of the 2024 8th International Conference on Control Engineering and Artificial Intelligence*, pp.215-219, 2024. Article (CrossRef Link)
[10] G. Tsoumakas, "A survey of machine learning techniques for food sales prediction," *Artificial Intelligence Review*, vol.52, no.1, pp.441-447, 2019. Article (CrossRef Link)
[11] K. Bandara, P. Shi, C. Bergmeir, H. Hewamalage, Q. Tran, and B. Seaman, "Sales Demand Forecast in E-commerce Using a Long Short-Term Memory Neural Network Methodology," in *Proc. of Neural Information Processing: 26th International Conference, ICONIP 2019, Part III*, LNTCS, vol.11955, pp.462-474, Springer, Sydney, NSW, Australia, Dec. 12-15, 2019. Article (CrossRef Link)
[12] G. Liu, T. T. Nguyen, G. Zhao, W. Zha, J. Yang, J. Cao, M. Wu, P. Zhao, and W. Chen, "Repeat Buyer Prediction for E-Commerce," in *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.155-164, 2016. Article (CrossRef Link)
[13] M. Li, S. Ji, and G. Liu, "Forecasting of Chinese E-Commerce Sales: An Empirical Comparison of ARIMA, Nonlinear Autoregressive Neural Network, and a Combined ARIMA-NARNN Model," *Mathematical Problems in Engineering*, vol.2018, no.2, pp.1-12, 2018. Article (CrossRef Link)
[14] A. Andueza, M. Á. D. Arco-Osuna, B. Fornés, R. González-Crespo, and J. M. Martín-Álvarez, "Using the statistical machine learning models ARIMA and SARIMA to measure the impact of Covid-19 on official provincial sales of cigarettes in Spain," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol.8, no.1, pp.73-87, 2023. Article (CrossRef Link)
[15] L. F. Sales, A. Pereira, T. Vieira, and E. de B. Costa, "Multimodal deep neural networks for attribute prediction and applications to e-commerce catalogs enhancement," *Multimedia Tools and Applications*, vol.80, no.17, pp.25851-25873, 2021. Article (CrossRef Link)
[16] S. Mu, Y. Wang, F. Wang, and L. Ogiela, "Transformative computing for products sales forecast based on SCIM," *Applied Soft Computing*, vol.109, 2021. Article (CrossRef Link)

[17] M. J. Schneider and S. Gupta, "Forecasting sales of new and existing products using consumer reviews: A random projections approach," *International Journal of Forecasting*, vol.32, no.2, pp.243-256, 2016. Article (CrossRef Link)
[18] M. Yang, T. Zhang, and C.-x. Wang, "The optimal e-commerce sales mode selection and information sharing strategy under demand uncertainty," *Computers & Industrial Engineering*, vol.162, 2021. Article (CrossRef Link)
[19] S. Makkar and S. Jaiswal, "Predictive Analytics on E-commerce Annual Sales," in *Proc. of Data Analytics and Management: ICDAM 2021*, vol.1, pp.557-567, Springer, Singapore, 2022. Article (CrossRef Link)
[20] S. Cheriyan, S. Ibrahim, S. Mohanan, and S. Treesa, "Intelligent Sales Prediction Using Machine Learning Techniques," in *Proc. of 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, pp.53-58, IEEE, 2018. Article (CrossRef Link)
[21] A. Y. L. Chong, B. Li, E. W. T. Ngai, E. Ch'Ng, and F. Lee, "Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach," *International Journal of Operations & Production Management*, vol.36, no.4, pp.358-383, 2016. Article (CrossRef Link)
[22] B. Singh, P. Kumar, N. Sharma, and K. P. Sharma, "Sales Forecast for Amazon Sales with Time Series Modeling," in *Proc. of 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)*, pp.38-43, IEEE, 2020. Article (CrossRef Link)
[23] J. Chen, N. Tournois, and Q. Fu, "Price and its forecasting of Chinese cross-border E-commerce," *Journal of Business & Industrial Marketing*, vol.35, no.10, pp.1605-1618, 2020. Article (CrossRef Link)
[24] I. Krasonikolakis, A. Vrechopoulos, and A. Pouloudi, "Store selection criteria and sales prediction in virtual worlds," *Information & Management*, vol.51, no.6, pp.641-652, 2014. Article (CrossRef Link)
[25] A. A. Afifi, "Demand Forecasting of Short Life Cycle Products Using Data Mining Techniques," in *Proc. of Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Part I*, IFIPAICT, vol.583, pp.151-162, Springer International Publishing, Neos Marmaras, Greece, Jun. 5-7, 2020. Article (CrossRef Link)
[26] C.-H. Chen, P.-Y. Chen, and J. C.-W. Lin, "An Ensemble Classifier for Stock Trend Prediction Using Sentence-Level Chinese News Sentiment and Technical Indicators," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol.7, no.3, pp.53-64, 2022. Article (CrossRef Link)
[27] D. Thorleuchter and D. Van den Poel, "Predicting e-commerce company success by mining the text of its publicly-accessible website," *Expert Systems with Applications*, vol.39, no.17, pp.13026-13034, 2012. Article (CrossRef Link)
[28] Y. Shi, T. Wang, and L. C. Alwan, "Analytics for Cross-Border E-Commerce: Inventory Risk Management of an Online Fashion Retailer," *Decision Sciences*, vol.51, no.6, pp.1347-1376, 2020. Article (CrossRef Link)
[29] S. Cremer and C. Loebbecke, "Selling goods on e-commerce platforms: The impact of scarcity messages," *Electronic Commerce Research and Applications*, vol.47, 2021. Article (CrossRef Link)
[30] Z. Li, D. Amagata, Y. Zhang, T. Maekawa, T. Hara, K. Yonekawa, and M. Kurokawa, "HML4Rec: Hierarchical meta-learning for cold-start recommendation in flash sale e-commerce," *Knowledge-Based Systems*, vol.255, 2022. Article (CrossRef Link)
[31] H. Pålsson, F. Pettersson, and L. W. Hiselius, "Energy consumption in e-commerce versus conventional trade channels - Insights into packaging, the last mile, unsold products and product returns," *Journal of Cleaner Production*, vol.164, pp.765-778, 2017. Article (CrossRef Link)
[32] Q. Zhang, J. Li, and T. Xiao, "Sales manipulation strategies of competitive firms on an e-commerce platform: Beneficial or harmful?," *Decision Sciences*, 2023. Article (CrossRef Link)
[33] T. Tong, X. Xu, N. Yan, and J. Xu, "Impact of different platform promotions on online sales and conversion rate: The role of business model and product line length," *Decision Support Systems*, vol.156, 2022. Article (CrossRef Link)

[34] D. Cirqueira, M. Hofer, D. Nedbal, M. Helfert, and M. Bezbradica, "Customer Purchase Behavior Prediction in E-commerce: A Conceptual Framework and Research Agenda," in *Proc. of 8th International Workshop on New Frontiers in Mining Complex Patterns*, LNAI, vol.11948, pp.119-136, Cham: Springer International Publishing, 2020. Article (CrossRef Link)
[35] N. Gordini and V. Veglio, "Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry," *Industrial Marketing Management*, vol.62, pp.100-107, 2017. Article (CrossRef Link)
[36] Y. Zhu, J. Li, J. He, B. L. Quanz, and A. A. Deshpande, "A Local Algorithm for Product Return Prediction in E-Commerce," in *Proc. of 27th International Joint Conference on Artificial Intelligence (IJCAI 2018)*, pp.3718-3724, 2018. Article (CrossRef Link)
[37] N. Chaudhuri, G. Gupta, V. Vamsi, and I. Bose, "On the platform but will they buy? Predicting customers' purchase behavior using deep learning," *Decision Support Systems*, vol.149, 2021. Article (CrossRef Link)
[38] I. Vallés-Pérez, E. Soria-Olivas, M. Martínez-Sober, A. J. Serrano-López, J. Gómez-Sanchís, and F. Mateo, "Approaching sales forecasting using recurrent neural networks and transformers," *Expert Systems with Applications*, vol.201, 2022. Article (CrossRef Link)
[39] S. Liu, F. Xiao, W. Ou, and L. Si, "Cascade Ranking for Operational E-commerce Search," in *Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.1557-1565, 2017. Article (CrossRef Link)
[40] L. Peng, W. Zhang, X. Wang, and S. Liang, "Moderating effects of time pressure on the relationship between perceived value and purchase intention in social E-commerce sales promotion: Considering the impact of product involvement," *Information & Management*, vol.56, no.2, pp.317-328, 2019. Article (CrossRef Link)
[41] W. Xu, Y. Cao, and R. Chen, "A multimodal analytics framework for product sales prediction with the reputation of anchors in live streaming e-commerce," *Decision Support Systems*, vol.177, 2024. Article (CrossRef Link)
[42] Z. Mu, X. Liu, and K. Li, "Optimizing Operating Parameters of a Dual E-Commerce-Retail Sales Channel in a Closed-Loop Supply Chain," *IEEE Access*, vol.8, pp.180352-180369, 2020. Article (CrossRef Link)
[43] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785-794, 2016. Article (CrossRef Link)
[44] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Advances in Neural Information Processing Systems*, vol.30, 2017. Article (CrossRef Link)
[45] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Advances in Neural Information Processing Systems*, vol.31, 2018. Article (CrossRef Link)
[46] A. Kurani, P. Doshi, A. Vakharia, and M. Shah, "A Comprehensive Comparative Study of Artificial Neural Network (ANN) and Support Vector Machines (SVM) on Stock Forecasting," *Annals of Data Science*, vol.10, no.1, pp.183-208, 2023. Article (CrossRef Link)
[47] A. Bedoui and N. A. Lazar, "Bayesian empirical likelihood for ridge and lasso regressions," *Computational Statistics & Data Analysis*, vol.145, 2020. Article (CrossRef Link)
[48] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol.521, pp.436-444, 2015. Article (CrossRef Link)
[49] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proc. of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp.2623-2631, 2019. Article (CrossRef Link)

[50] Y. Liu, D. Mikriukov, O. C. Tjahyadi, G. Li, T. R. Payne, Y. Yue, K. Siddique, and K. L. Man, "Revolutionising Financial Portfolio Management: The Non-Stationary Transformer's Fusion of Macroeconomic Indicators and Sentiment Analysis in a Deep Reinforcement Learning Framework," *Applied Sciences*, vol.14, no.1, 2023. Article (CrossRef Link)

[51] Y. Liu, G. Li, T. R. Payne, Y. Yue, and K. L. Man, "Non-Stationary Transformer Architecture: A Versatile Framework for Recommendation Systems," *Electronics*, vol.13, no.11, 2024. Article (CrossRef Link)

[52] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol.622, pp.178-210, 2023. Article (CrossRef Link)

[53] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol.14, pp.241-258, 2020. Article (CrossRef Link)

[54] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *Journal of King Saud University - Computer and Information Sciences*, vol.35, no.2, pp.757-774, 2023. Article (CrossRef Link)

Yuchen Liu received the M.S. degree from the Guangdong University of Foreign Studies, Guangzhou, China, in 2016. He is currently pursuing the Ph.D. degree with University of Liverpool, Liverpool, United Kingdom. His research interests include reinforcement learning, recommendation systems, and dynamic pricing.

Meng Wang received the B.S. degree from the Macau University of Science and Technology, China, and his M.S. degrees from the Queen Mary University of London, United Kingdom, and University of Liverpool, United Kingdom. He is currently pursuing the Ph.D. degree with University of Liverpool, Liverpool, United Kingdom. His research interests include machine learning, natural language processing, supply chain risk, and green finance.

Gangmin Li is currently a Senior Associate Professor with Xi'an Jiaotong-Liverpool University, Suzhou, China. He is a highly educated academic with over 40 years of research and teaching experience in five British and two Chinese universities. He has obtained over 200K in funding from the government and industry. He has published over 100 journals and conference papers. His research interests include data science, distributed AI, knowledge engineering (KE), agent and multi-agent systems, grid computing, and HCI. He has served as many academic journal editors and international conference technical committee members.

Terry R. Payne has been actively working in distributed AI and multi-agent systems research, since 1994. He is currently a Senior Lecturer with the University of Liverpool, where he heads the Robotics and Autonomous Systems Research Group. In addition to teaching final year research-led subjects (on robotics and multi-agent systems), he is the Academic Lead for Student Recruitment, Admissions, and Widening Participation with the Faculty of Science and Engineering. His research interests include ontological knowledge to support agent/service discovery and provision. He has also worked on coalition formation, self-organizing agent communities, and machine learning for adaptive agents. He is the winner of several research and teaching awards.

Yong Yue (BEng Northeastern China, PhD Heriot-Watt UK, CEng, FIET, FIMechE, FHEA) is a Professor at the Department of Computing, and Director of the Virtual Engineering Centre (VEC) and Suzhou Municipal Key Lab for Intelligent Virtual Engineering. He was Head of the Department of Computer Science and Software Engineering (2013-2019). Prior to joining XJTLU, he had held various positions in industry and academia in China and the UK, including Engineer, Project Manager, Professor, Director of Research and Head of Department. Professor Yue has experience in learning and teaching, research and enterprise as well as management. His current research interests are virtual reality, computer vision, robot applications and operations research. He has led a variety of research and professional projects supported by major funding bodies and industry. He has over 250 peer-reviewed publications and supervised 27 PhD projects to successful completion.

Ka Lok Man (Member, IEEE), received the Dr. Eng. degree in electronic engineering from Politecnico di Torino, Turin, Italy, in 1998 and the Ph.D. degree in computer science from Technische Universiteit Eindhoven, Eindhoven, The Netherlands, in 2006. He is currently a Professor of Computer Science and Software Engineering with Xi'an Jiaotong-Liverpool University, Suzhou, China. His research interests include formal methods and process algebras, embedded system design and testing, and photovoltaics.