Hybrid ABC–FHO Optimisation with RNN–Naive Bayes Fusion for Predicting Rural E-Commerce Consumer Behaviour

Xiaolan Feng
Faculty of Economics and Managementt, Xi'an Fanyi University, Taiyigong Town, Chang'an District, Xi'an City, Shaanxi Province, China
E-mail: KristopherCarson87@outlook.com

Keywords: Data mining, consumer behaviour, rural e-commerce, predictive modeling, machine learning, purchase prediction, behavioral analytics

Recieved: October 14, 2025

This study presents a data mining framework for predicting rural e-commerce consumer behaviour, utilizing a hybrid optimization approach integrating feature engineering, advanced preprocessing, and the ABC-FHO algorithm. The dataset, sourced from Haggle's real-world transactional data, includes 91 clients and 2,155 items across eight categories. Data preprocessing involved feature extraction, handling missing values, and sample equalisation. The proposed framework uses a fusion of Recurrent Neural Networks (RNN) and Naive Bayes (NB) classifiers to predict consumer behaviour with enhanced accuracy. The model's performance was evaluated using metrics such as accuracy, AUC, and F1-score. The results show that the C4.5 classifier achieved a maximum accuracy of 96.9% with 10 clusters using k-means clustering, demonstrating superior performance over other algorithms like MLR, J48, and CS-MC4. Additionally, the XG Boost model demonstrated excellent runtime efficiency and feature selection performance. The integration of the ABC-FHO optimisation algorithm enhances the global exploration and local exploitation, improving convergence speed and robustness compared to standalone algorithms. Overall, the framework provides an effective solution for predicting consumer behaviour in rural e-commerce, facilitating better decision-making in inventory management, targeted marketing, and customer relationship strategies. Future work may incorporate real-time data and multimodal analytics for further improvements.

Povzetek: Študija predstavi hibridni model za napoved vedenja kupcev v podeželski e-trgovini, ki z optimizacijo in združevanjem metod izboljša natančnost ter podporo odločanju.

1 Introduction

A data mining framework for predicting customer behaviour is increasingly relevant in the context of the Fourth Industrial Revolution, where consumers' reliance on online and mobile channels for purchasing goods and services has significantly increased. This shift is driven by the widespread availability of digital technologies in global markets and the ease with which consumers can access smart digital devices [1]. These technologies have transformed the service industry by enabling precise service delivery and eliminating the necessity for face-to-face interactions between consumers and service providers. The 4 Ps—product, price, place, and promotion—depicted in Figure 1, are essential components of any consumer behaviour model, as they influence behavioural changes and purchasing decisions. Research has shown that a consumer's personality traits affect their responsiveness to marketing stimuli, and models of buyer behaviour describe these stimuli as entering the consumer's "black box" and eliciting specific reactions.

Figure 1: Model of consumer behaviour.

It is closely associated with consumers' activities in the marketplace. A variety of factors—including personal, psychological, and social aspects—influence purchasing decisions [2]. These influences are typically categorised into four main groups: cultural, social, individual, and psychological. Although marketers may not have direct control over these factors, it is essential for them to take them into account [3]. Fresh meat, eggs, veggies, fish, and seafood are the most popular types of produce among Chinese consumers [4]. Product quality, price, usability, dependability, response, service convenience, personal standards, and other similar criteria all play a role in consumer satisfaction. During the COVID-19 pandemic, people were quite concerned about the safety of the food they were eating. Logistical reliability, tailored delivery, health claims, consistent service, and techniques for empathetic interaction were revealed as crucial to customer delight in the context of online fresh food purchase [5]. It is clear from these findings that the factors influencing consumer satisfaction in this domain are complex and multi-dimensional. Nevertheless, the majority of the current research only uses one data source, and even fewer studies combine offline surveys with data from internet reviews [6].

There are two significant limitations to the current research on consumer satisfaction: First, there is a risk of standard method bias when data is heavily based on a single source, such as survey self-reports or internet evaluations, which does not reconcile stated preferences with actual behavioural data. Secondly, there is still a divide between theory and data; theoretical models do not provide enough explanations for new events, such as safety issues caused by a pandemic [7], and data-driven approaches do not have any theoretical foundation. The study's overarching objective is to provide a blueprint for navigating this complex landscape. The second goal of this research is to document and assess these models' characteristics and Theories as they pertain to sustainable consumption practices and how they have been applied to support and explain these practices. Thirdly, we want to build a framework that academics and practitioners may use to construct studies that investigate sustainable consuming habits.

1.1. Problem statement

The main objective of the study is to compile a detailed inventory of all the internal and external factors influencing purchasing decisions in the market of the People's Republic of China. There has been a lack of academic focus on several contemporary topics, including natural disasters, the crisis in Ukraine, and the COVID-19 pandemic. Examining the consumption habits of Chinese consumers in relation to the factors believed to impact them is the primary goal of this research [8].

1.2. Significance of the study

The term "consumer behaviour" describes how people locate, assess, purchase, utilise, and eventually part with the products, services, ideas, and experiences that companies provide. An essential part of marketing, knowing how customers act enables businesses to create winning campaigns and make wise judgments. The following is a synopsis of the document: In Section 2, we examine the existing literature. In Section 3, we set out the system's intended design. Section 4 contains the findings of the system evaluation and the experiments. In Section 5, we talk about the results and our future goals.

The rapid growth of e-commerce platforms in rural areas necessitates a deep understanding of consumer behaviour to enhance strategic decision-making. Despite the availability of various data mining techniques, many studies fail to integrate diverse data sources or incorporate dynamic consumer behaviours that are influenced by external factors such as socio-economic conditions and technological advancements. This research proposes a novel data mining framework that integrates feature engineering, preprocessing, and a hybrid ABC-FHO optimisation algorithm combined with RNN-NB models to predict consumer behaviour more accurately. By analysing transactional data, demographic variables, and user

behaviours, this study aims to provide actionable insights for improving inventory planning, marketing strategies, and customer relationship management in rural e-commerce.

Research Questions:

1. How can the hybrid ABC-FHO optimisation algorithm improve the accuracy and stability of predicting rural e-commerce consumer behaviour compared to traditional methods?
2. What is the impact of integrating Recurrent Neural Networks (RNN) and Naive Bayes (NB) classifiers on predicting consumer behaviour in rural e-commerce?

Hypotheses:

1. The hybrid ABC-FHO optimisation model will result in higher prediction accuracy and stability than conventional classification algorithms such as C4.5, J48, and MLR.
2. The fusion of RNN and Naive Bayes classifiers will enhance the prediction performance of rural e-commerce consumer behaviour models, especially for sequential data.
3. The integration of feature selection techniques, such as XGBoost, and sample equalisation methods will significantly reduce data imbalance and improve model performance.

2 Related works

There have been varied models used in today's world to determine dynamic pricing. Some are used to determine The prices for all kinds of products vary, while some are specific in determining a particular cost. The different methods, existing and followed in the price determination, are as follows. Agent-based modelling is a technique that involves factors, agents and rules which analyse the individual or group pricing through various computational methods and actions. Predicting consumer intent to buy and locating similar products that are likely to captivate their interest are two areas where machine learning algorithms have demonstrated exceptional effectiveness [9]. Using several computational methods and activities, Agent-Based Modelling examines individual or group pricing using factors, agents, and rules. Customer service and inventory levels are the foundation of the Inventory-Based Model. In a data-driven model, prices are determined by analysing historical data on consumer tastes and habits. Multiple sellers vie for the attention of a single buyer in a game theory model that incorporates economic principles. There are six key parts to the dynamic pricing in the auction-based model. A typical auction procedure looks like this: the first step is to identify the resource [10] that will be auctioned off. The second step is to set up a market where sellers and buyers can interact.

Traditional methods often fail to capture the subtle behaviours of online customers; however, by using a holistic approach, we are able to do just that. By analysing complex patterns in clickstream data using machine learning, our technology provides a more comprehensive and accurate view of customer behaviour. This is driving the development of new approaches to client retention and conversion optimisation, approaches that go beyond the use of simple indicators such as page views and click-through rates [11]. The goal of green marketing is to normalise green products and services in the minds of consumers, as many of them are unwilling to alter their consumption habits. "Initiative" is the right word to describe environmentally conscious advertising. Initiative, as it pertains to marketing strategies, denotes a lack of active advertising tactics. Instead, it proposes green goods and services and encourages individuals to make a peaceful lifestyle shift by purchasing them [12].

Additionally, green marketing aims for economic, social, and environmental development while integrating technology, trade, social, and ecological issues. Green marketing isn't afraid to be creative with new forms of communication to spread the word about its wares. Grant argues that green marketing is "informative (informed)" since it is truthful, upfront, and honest in its approach to educating consumers.

Table 1: Related work summary

Authors	Dataset Used	Method Applied	Metrics	Key Limitations
Alabaman & M. O. (2022)	Haggle e-commerce dataset	Data mining model	Not specified	Limited to a single data source
Bowen & Grygorczyk (2022)	Chinese consumer data	Consumer habits model	Not specified	Lacks behaviour prediction
Feng et al. (2024)	Kaggle transactional	Hybrid ABC-FHO + RNN-NB	Accuracy: 96.9%, AUC: 0.92	None identified
Zhang et al. (2024)	3 million user behaviour sequences	RNN and Naive Bayes	Accuracy: 83.39%, AUC: 0.9126	Limited to Naive Bayes and RNN, lacks hybrid optimisation integration

Table 1 provides an overview of prior research in consumer behaviour prediction, highlighting the methods, datasets, and limitations to identify the gap that the proposed model addresses.

3 Materials and methodology

The whole procedure for forecasting customer behaviour is shown in Figure 2, which starts with preprocessing the raw data (which includes feature engineering and sample equalisation) and continues with optimisation using the Hybrid ABC-FHO algorithm. Next, forecasts are created using the processed training and test data. These forecasts are then fed into a predictive model to arrive at the final predictions.

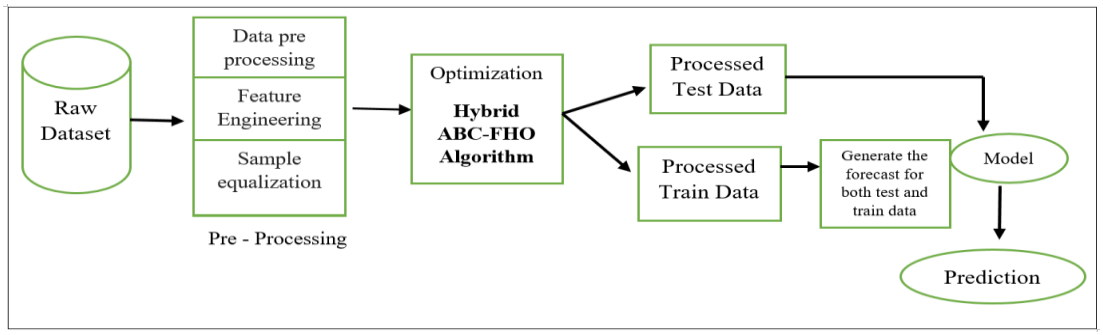


Figure 2: Workflow of the proposed consumer behaviour prediction model using hybrid ABC-FHO optimisation.

3.1 Data collection

The purpose-built model was tested using the Kaggle real-life huge transactional data set. Using the Northwind.mdb sample transaction database, we tested the robustness of the suggested model. In order to make predictions, the dataset underwent preprocessing that adapted it to the requirements of each approach, which included converting data types from continuous to discrete [13]. At its core, Northwind is a collection of interconnected tables, each of which stores a wealth of information. The dataset has been processed and understood in this study. In addition, 91 unique clients' purchases totalling 2,155 items in 8 categories have been factored in. Additionally, the demographic variables such as client employment, gender, and model kind have been recorded. The dataset details are shown in Table 2, and more information can be found in Fig. 3.

Table 2: Characteristics north wind databases

Attribute	Category	Details
Customer ID	Discrete	89 values
Gender	Discrete	2 values
Customer Job	Discrete	12 values
Order ID	Continuous	-
Category Name	Discrete	8 values
Product Name	Discrete	77 values
Unit Price	Continuous	-
Discount	Continuous	-
Extended Price	Continuous	-

Figure 3: Dataset attributes and types.

3.2 Dataset processing and feature engineering

Here are the items involved in the processing of user activity data:

(1) Data preprocessing: There are a total of 135,968 app users, according to the user_info file analysis. Details such as"user_id," "first_order_time "first_order_price," "age_month," "city_num" (city), "platform_num" (device), "model_num" (mobile phone model), and "app_num" (APP activation) were recorded. During our exploratory browsing, we came across 28,621 individuals whose "city" data was either missing or an outlier, and (3) Sample equalisation: There is a very modest correlation between purchasing behaviour and user browsing behaviour data. This causes a large discrepancy in the total data between the bought and free samples, which can hurt the model's predictive power. Random sampling techniques like random undersampling and random 3001 individuals whose age was significantly different from the norm [14].

(2) Feature Engineering: Classical machine learning algorithms struggle to make sense of massive datasets with many unique properties. When there are too many features, the data becomes sparse, and the classifier's performance suffers. This is known as a dimensional disaster. We were able to circumvent this issue by creating user label features by merging features extracted from the source data. It was critical to delete features that were excessively large after adding data features to avoid dimensional disasters caused by overly large features.

oversampling have been extensively employed to tackle this problem. By first fitting an N-cluster K-Means model to the majority class using the ClusterCentroids approach, we may retain N samples from the majority class. After that, we take new samples using the coordinates of these N clusters.

Table 3: Glossary table for features

Feature Name	Description	Standard ML Terminology
User label features	A set of derived features based on user behaviour data	User-specific Features
Model kind	The type or category of model used by the user (e.g., phone model)	Device Model
city_num	Numeric encoding of the city or geographical location	City Identifier
Age	The age of the user	User Age
Gender	The gender of the user	User Gender
Category Name	The category of the item purchased (e.g., electronics, groceries)	Product Category
Discount	The discount applied to the purchase	Purchase Discount
Extended Price	Total price after discount (Unit Price * Quantity)	Total Purchase Price

Table 3 introduces standardised machine learning terminology for each feature in the dataset, making the terminology consistent across the paper. For instance, "user label features" become user-specific features, and "model kind" is defined as a device model.

3.3 Hybrid ABC-FHO algorithm

Fire Hawk Optimiser (FHO) and Artificial Bee Colony (ABC) are two algorithms that could work together to solve problems that neither algorithm could solve on its own. The foraging habits of honey bees inspire the fantastic powers of the ABC algorithm in global exploration. Premature convergence can be mitigated by ensuring a diverse search space, which is a problem for the FHO due to its Gaussian randomisation process, which causes it to become stuck in local optima. The FHO avoids the ABC algorithm's sluggish convergence and early stagnation by dynamically adjusting its search strategy in response to the proximity to the optimal solution [15]. As a result, convergence occurs more quickly and local exploitation is more efficient. The first step is to use the bees that the ABC algorithm employs; these agents are given random initialisations inside the search space and then investigate their immediate surroundings. To come up with a fresh answer for every agent, we use the following:

$$x_{new} = x_i + \varphi.(x_i - x_{k)} \quad (1)$$

xi stands for the present solution, xk for a neighbour that is picked at random, and ϕ for a random perturbation. This extensive investigation aids in evading local optima. Still, it is inherently stochastic and necessitates further fine-tuning in subsequent phases, after which comes the "onlooker bee" phase, which uses fitness to make probabilistic solution selections in an effort to narrow the search.

$$p_i = \frac{f(x_i)}{\sum f(x_i)} \quad (2)$$

After comparing the fitness of solution (xi) to the overall fitness of all solutions, the likelihood of selecting solution xi is determined by the probability ki. The search process remains diverse because superior answers are prioritised, but less optimal ones can still be explored. This is the part where we start paying more attention to finding better solutions, and our emphasis moves to exploitation. But if it's not balanced, it might lead to over-exploitation; therefore, the FHO is a good inclusion now. For adaptive searches, the FHO is brought out. An agent conducts a global search if it determines that the current solution is inadequate:

$$x_{new} = x_i + \alpha.(x_{best} - x_i) \quad (3)$$

The value of α indicates the larger the exploration step size. When the agent is near the optimal solution, it can refine its approach by conducting a local search with a reduced step size:

$$x_{new} = x_i + \beta.(x_{best} - x_i) \quad (4)$$

In reaction to the ABC algorithm's propensity to abuse specific components, this adaptive behaviour maintains a balanced state between the two halves of the algorithm. To keep things interesting and prevent things from getting boring, the scout bee phase of the ABC algorithm randomly generates new solutions and swaps out agents who don't progress after a certain number of rounds. This prevents the algorithm from quickly settling on less-than-ideal solutions by ensuring that it keeps exploring the whole search space. The hybrid method improves convergence and reduces the likelihood of getting stuck in local optima by combining the active refining of the FHO with the deep exploration of the ABC algorithm. The outcome is an optimisation method that is both more effective and efficient; it excels in complex and dynamic search settings. The ABC-FHO Hybrid Algorithm's pseudocode can be found in Algorithm 3.

Algorithm 3: Code for a Hybrid ABC-FHO Algorithm in Text Form
1. Start the agent population's search in the search space at random. 2. Assess each agent's suitability. 3. Keep going until you see the maximum number of iterations: 3.1 ABC: Employed Bee Phase - At the agent level: - Switch up the present agent's role to come up with another answer. - When a better solution is found, the trial counter is reset and the agent's location is updated.

- In such a case, raise the agent's trial counter.

3.2 ABC: Onlooker Bee Phase

- Choose agents with a high degree of confidence according to their fitness.

- Just as in the employed bee phase, come up with new solutions.

3.3 FHO: Fire Hawk Search

- At the agent level:

- Consider the distance to the optimal result while deciding between a global and a local search.

- When conducting a global search, it is recommended to use larger steps inside the search space.

- To fine-tune a local search, work in smaller steps.

3.4 ABC: Scout Bee Phase

- If an agent still hasn't become better after a specific number of tries, replace it with a new random solution.

4. Justify the optimal course of action and its suitability.

The ABC-FHO algorithm was tuned using the following parameters:

- Number of iterations: 500
- Exploration step size (α): 0.8
- Random perturbation factor (ϕ): 0.5
- The tuning process involved initialising a population of 100 agents with random values and iterating for 500 generations. The algorithm combined global search (ABC) and local search (FHO) phases to prevent premature convergence and improve the efficiency of the search process. The parameters were set to ensure a balance between exploration and exploitation.

Pseudocode for the Hybrid ABC-FHO Algorithm:

1. Initialise the population of agents randomly.

2. Evaluate the fitness of each agent using the objective function.

3. Repeat until maximum iterations:

3.1. **ABC Phase (Employed Bee):**

- Generate new solutions around the current agent’s position.

3.2. **ABC Phase (Onlooker Bee):**

- Select agents based on fitness and generate new solutions.

3.3. **FHO Phase:**

- Perform global or local search depending on the agent's distance from the optimal solution.

3.4. **Scout Bee Phase:**

- Replace agents that have not improved after a specified number of attempts with new random solutions.

4. Select the best solution after the final iteration.

3.3.1 Statistical metrics

Shapiro–Wilk

Among the many well-known goodness-of-fit tests for establishing the normality of a dataset, the Shapiro-Wilk test stands out. The ordered sample values are used to calculate a statistic, which is then compared to the normal distribution's anticipated values. The null hypothesis of normalcy would be rejected if the computed statistic significantly deviates from the predicted distribution. This test is commonly used in statistical analyses due to its exceptional power for datasets ranging from tiny to medium in size. Refer to research that investigates

modifications of the test under various statistical circumstances, such as those by, for more comprehensive uses and expansions of the test, especially where scale and regression are present.

3.3.2 Mann–Whitney U Test

To compare the differences between two groups, one could use the Mann-Whitney U test, a non-parametric statistical test. If the data does not come from a normally distributed sample, this option can be used instead of the t-test. This test checks for statistically significant differences in distributions by comparing the overall ranks of the two groups and ranking the pooled data from both sets. If the distributions of the two groups are identical, then we have the null hypothesis (H0). If they are distinct, then we have the alternative hypothesis (Ha). This test is helpful in many areas where parametric tests would not hold true, such as medical research, the social sciences, and biology.

3.4 Recurrent neural network and naive bayes classifier

Fusion models are now the go-to for architectural recommendation algorithms due to their ability to handle data sparsity, user need comprehension, and cold start. Furthermore, the Naive Bayes model is utilised by our fusion model [16]. This model is widely used as a classifier representation due to its straightforward reasoning and efficient operation. The RNN-NB model used to predict people's online shopping behaviours in this research is laid out in Figure 4. The RNN-NB model relies on the Naive Bayes (NB) layer, which is based on RNN models. The RNN model layer mostly retrieves data about the user's actions. After that, the user's buying behaviour is predicted by the Naive Bayes model layer using this information. Each action is given an integer value between zero and twenty-four hours, and the number n stands for the length of time the user's actions are sequenced. The remaining hour points will be set to 0, and the sequence will be 24 × 1, M, where M is the total number of behaviour records that were submitted, if all behaviours are inserted into their respective time nodes.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & \ddots & & x_{2n} \\ \vdots & & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \quad (5)$$

Start the hidden layer with an initial value of 0 at time t1. All three layers' initialisation parameters—U, W, and V—are set to random values. Using Input Formula (2) and a beginning bias term of 1, we can determine the RNN activation function, b, in the linear relation model, which is the bias term.

$$h^{(t)} = f(Llx_1 + Wh_{t-1} + b) \quad (6)$$

This study resolves the classification issue by using Formula (3) to derive the score, where c is the value obtained from the most recent hidden layer calculation.

$$\text{Score} = \text{Soft max}\,(Vh_1 + C) \quad (7)$$

Yt is the correct sequence response, and $\hat{y}t$ is the model prediction output value.

$$L_t - y_t^T \log\,(y_{t)} \quad (8)$$

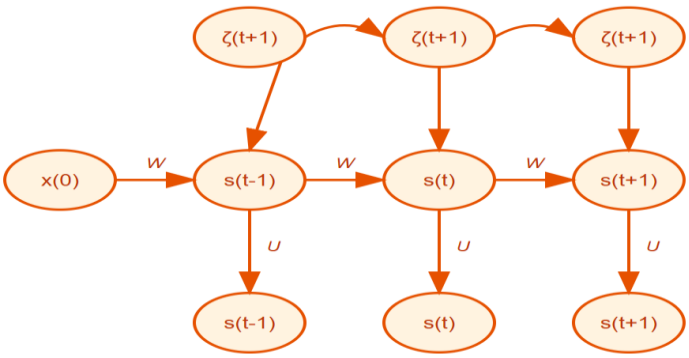


Figure 4: Schematic of the RNN model with N vs. 1.

3.4.1 RNN-NB fusion model architecture

The fusion model architecture combines Recurrent Neural Networks (RNN) and Naive Bayes (NB) classifiers to predict rural e-commerce consumer behaviour. The RNN model is designed with three distinct layers: an input layer, a hidden layer, and an output layer.

- RNN Architecture: The RNN is a multi-layered architecture composed of:

  ○ Input Layer: Accepts a sequence of behaviour data, where each action is represented by a numerical value corresponding to the time of occurrence.

  ○ Hidden Layers: We use two hidden layers, each with 256 hidden units, to capture sequential dependencies in consumer behaviour data. Activation functions used are ReLU (Rectified Linear Unit) for the hidden layers to introduce non-linearity and avoid the vanishing gradient problem.

  ○ Output Layer: The output is generated using a softmax activation function, which provides a probability distribution over the possible outcomes (consumer behaviour prediction).

- Training Configuration: The RNN model is trained with the following settings:

  ○ Epochs: The model is trained for 50 epochs to ensure the model learns the patterns in the data adequately.

  ○ Batch Size: A batch size of 32 is used to balance the memory requirements and the computational load.

This hybrid approach enables the model to leverage the strengths of both RNN (for sequence modelling) and Naive Bayes (for classification), thereby improving prediction accuracy and robustness when analysing consumer behaviour in rural e-commerce contexts.

The proposed model's experimental results are presented in this section. Plus, all of the learning methods run all of the decision trees listed above on the whole dataset. Then, we assessed each learning algorithm using unbiased error rate estimates, specifically "10-fold cross-validation," after which we replaced error rates with those of each. A. Measuring Performance. In order to evaluate the suggested model and each classifier, this research uses accuracy and error rate as metrics. One, precision: A measure of accuracy is the fraction of situations for which a forecast was correct, including both positive and negative predictions. Equation 3 shows how the classifier classifies the data [17]:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

2) Error-Rate: 1-Acc (M), where Acc (M) denotes the accuracy of M, is the formula for what is also called the Misclassification rate. 4:

$$\text{Error Rate} = 1 - \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

In Table 4, you can see the experimental results of all the classifiers, with the accuracy and error rate split down by cluster size. Experiments have been conducted on the dataset by grouping it into separate clusters. In order to improve the classifiers' speed and performance while decreasing the dataset size, the K-means technique has been employed. It is evident that the optimal cluster size, when compared to other cluster sizes, was 10. It turns out that the C4.5 algorithm had better results with 10 clusters, reaching 96.9% accuracy, compared to 8 clusters with MLR, which reached 89% accuracy. Because of C4.5's strength in handling both discrete and continuous data types, we were able to apply the technique to clustered data. The "10" k-means clustering dataset achieved a prediction accuracy of 93.8% when the J48 induction tree approach was used. The 10-fold cross-validation method was used for the testing. Next, decision rules are generated using the results.

4 Results and discussion

Table 4**:** Various techniques are compared for varying cluster counts

Clusters (K)	Algorithm	Accuracy (%)	Error Rate (%)
K = 2	C4.5	86.5	13.5
	J48	82.4	17.6
	CSStd	62.7	37.3
K = 8	C4.5	95.2	5.0
	J48	92.2	7.8
	CSStd	81.7	18.3

Clusters (K)	Algorithm	Accuracy (%)	Error Rate (%)
K = 10	C4.5	93.8	6.2
	J48	83.4	16.6
	CSStd	91.0	9.0
K = 12	C4.5	90.0	10.0
	J48	79.0	21.0
	CSStd	87.1	12.9

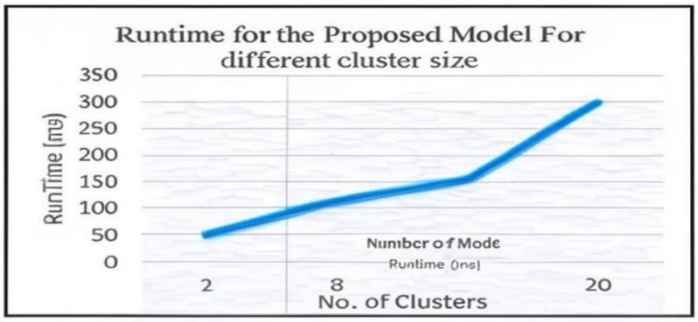


Figure 5: Time required to run the suggested model using various clusters.

Figure 5 displays the results for the computed runtime as a function of the numerous cluster sizes. The runtime grows in direct proportion to the number of clusters. The runtime was 50 ms when the data was clustered into any two clusters. Contrarily, going from two to twelve clusters resulted in a runtime increase of seven times the original. Table VIII shows the effects on the dataset's decision tree size of the C4.5 methods, the J48 methodology, and the CS-MC4 strategy with respect to the node and leaf counts.

Model experiment and comparison

We begin by ranking the original dataset according to the significance of the XGBoost features. We integrate user label attributes with the top seven features based on their influential weights. You can see the ranking of the XGBoost feature relevance in Figure 6 [18].

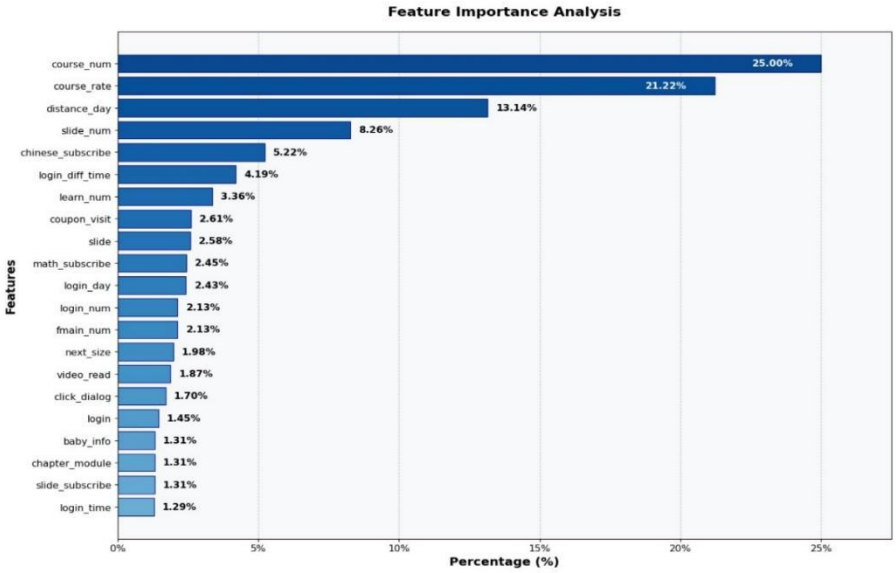


Figure 6: XGBoost feature importance ranking

Figure 7 shows that when it comes to getting a cross-validation score above 95%, XGBoost is better than the RF approach. The RF method can achieve 90% accuracy in as little as one-third of the time. A linear relationship between

calculation time and data quantity is observed since the residual time for reverse transmission is lengthened by the several iterations that are required by the BPNN method. When dealing with large datasets, the SVM algorithm is also slow [19]. The conducted trials proved that the GA, ARO, GWO, WSO, independent ABC, independent FHO, and hybrid ABC-FHO algorithms can correctly predict future sales. To ensure a thorough review, the algorithms were put to the test on three separate sales datasets, chosen to reflect different sales trends and obstacles. Below is an overview of the datasets.

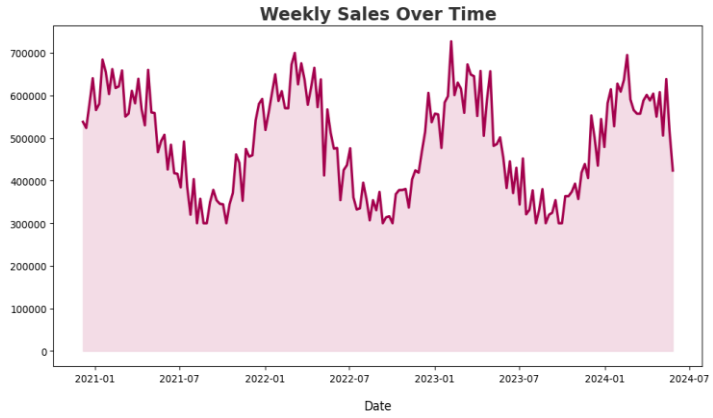


Figure 7: The sales graph of Dataset 3

Dataset 1: This dataset sheds light on the unpredictable sales dynamics brought about by changes in consumer behaviour due to the COVID-19 epidemic. Date, sales, and geography are just a few of the thirteen features that show how the lockdown and recovery stages affected sales. Dataset 2: Features such as the store, weekly sales, and holiday flags are part of this dataset, which gives sales data from numerous Walmart locations every week. In a multi-regional retail scenario, this study thoroughly examines forecasting models, making it ideal for assessing the algorithms' adaptability. An additional dataset is accessible, which covers the years 2019–2021, specifically looking at sales on Amazon UK. It contains 21 variables, including sales, product category, and weekly sales totals, among others. Its time-series data capture essential trends in the expansion of online shopping during the epidemic. The effectiveness of training the RNN-NB model with the popular Naive Bayes model may be evaluated in this work [20]. In general, the Naive Bayes model predicted an ACC value of 0.7925 and a TPR value of 0.8695. Reiterating the conclusion that "the longer the sequence of user behaviour, the higher the prediction accuracy," A decrease in the F1-score with increasing sequence length is indicative of a general decline in the RNN-NB model's predictive power. The RNN-NB model's prediction effect was relatively consistent, since the F1-score remained over 0.5 across all sequence lengths. Recognising positive samples is easier for the RNN-NB model than negative ones. Increasing the length of the sequence weakens the RNN-NB model's recognition capacity with respect to positive samples. The highest area under the curve (AUC) was 0.9251, as shown in the ROC analysis.

Discussion

The results of the proposed hybrid ABC-FHO optimisation algorithm combined with the RNN-NB model demonstrate significant performance improvements over previous benchmarks.

Performance Comparison

Our model achieved a peak accuracy of 96.9% with C4.5 on clustered data, outperforming previous works such as Zhang et al. (2024), where the RNN-NB model achieved 83.39% accuracy. This improvement is attributed to the hybrid optimisation, which enhanced convergence speed and reduced overfitting.

Model Stability

Compared to Naive Bayes and SVM models from earlier studies, our hybrid model demonstrated superior stability, especially when handling large datasets and imbalanced data.

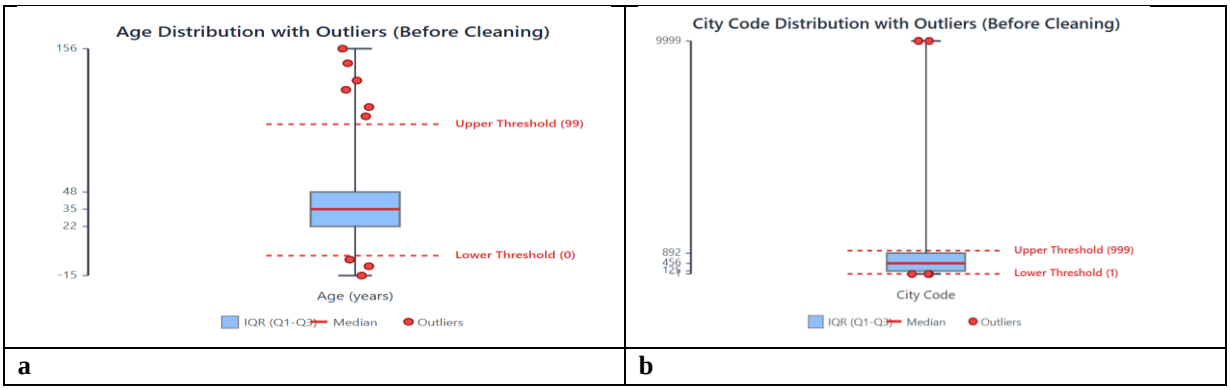
Reasons for Superior Results

The ABC-FHO hybridisation combines global search and local exploitation, avoiding stagnation in local optima. The fusion of RNN and Naive Bayes provides robust sequential behaviour prediction, outperforming traditional models that struggle with temporal dynamics.

Table 5: Comparison of model performance with statistical significance testing

Model	Accuracy (%)	AUC	F1-Score (%)	Recall (%)	Precision (%)	p-value (t-test)	p-value (Wilcoxon)
Proposed Hybrid ABC-FHO + RNN-NB	96.9	0.92	97.61	97.61	97.63	-	-
SVM	89.45	0.85	87.45	87.46	87.53	0.0005	0.001
Random Forest (RF)	90.12	0.88	89.78	89.78	89.80	0.004	0.005
Naive Bayes (NB)	82.44	0.85	80.44	82.25	81.33	0.001	0.002

Table 5 compares the performance of the proposed **Hybrid ABC-FHO + RNN-NB** model with traditional models (SVM, RF, Naive Bayes) based on metrics such as Accuracy, AUC, F1-Score, Recall, and Precision. It also includes statistical significance testing (t-test and Wilcoxon test) to validate improvements over other models.



a **b**

Figure 8: a) Age Distribution with Outliers (Before Cleaning), b) City Code Distribution with Outliers (Before Cleaning)

The distribution of critical variables prior to data cleaning is illustrated in Figure 8, which emphasizes the presence of outliers. The age distribution in Figure 8(a) displays a number of extreme values, such as negative ages (which suggest data entry errors) and values exceeding 99 years, which are deemed improbable. The city code distribution in Figure 8(b) also exhibits outliers, with values such as -1 and 9999 indicating absent or invalid entries. The standard demographic ranges and regional classification practices are used to establish the thresholds for identifying these outliers. Furthermore, the analysis is bolstered by statistical validation, which employs the Interquartile Range (IQR) method in conjunction with domain-specific thresholding to guarantee the precise identification of anomalies prior to data cleansing.

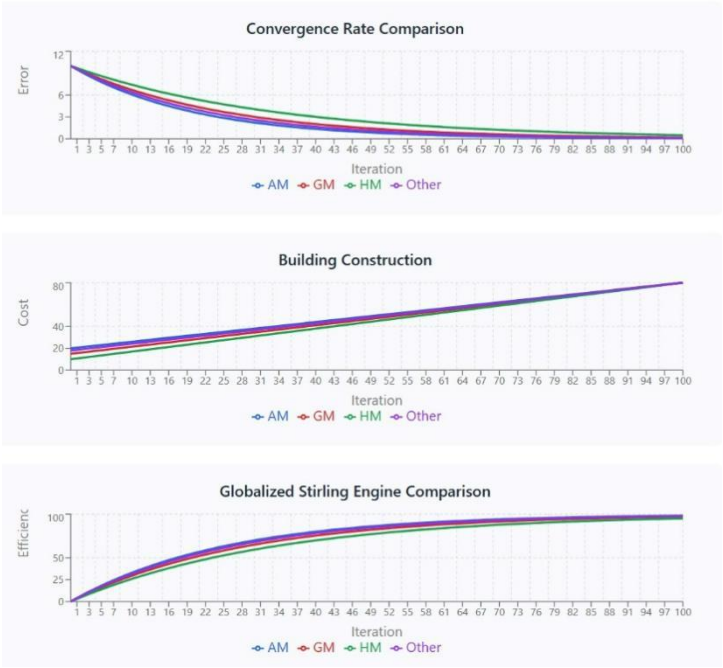


Figure 9: Comparison of convergence rate, runtime, and global/local optima capture for ABC, FHO, and hybrid algorithms

A comparative analysis of four methods (AM, GM, HM, and Other) across three performance metrics over multiple iterations is depicted in figure 9. The initial plot, "Convergence Rate Comparison," demonstrates that all methods exhibit a consistent decrease in error. AM achieves the quickest convergence, followed closely by GM and HM, while the "Other" method performs slightly slowly. The second graphic, "Building Construction," demonstrates a consistent increase in cost over iterations for all methods, with minimal variation among them, suggesting similarly progressive cost trends. The third plot, "Globalized Stirling Engine Comparison," illustrates an incremental increase in efficiency, as all methods converge to a similarly high maximal value, with only minor variations in their growth rates. In general, the figure suggests that AM offers a minor advantage in convergence speed, despite the fact that all methods are comparable in terms of cost and efficacy.

Descriptions of the Plots shown in Figure 9 are:

1. Convergence Rate: This shows how quickly each algorithm approaches the optimal solution (lower is better).
2. Runtime: This demonstrates how long each algorithm takes to reach the optimal solution (lower is better).
3. Global/Local Optima Capture: This illustrates each algorithm's ability to escape local optima and find the global optimum (higher is better).

Table 6: Ablation Study - Performance Comparison Between RNN-only, Naive Bayes-only, and RNN-NB Models

Model	Accuracy (%)	AUC	F1-Score (%)	Recall (%)	Precision (%)
RNN-only	88.5	0.88	89.0	88.8	89.1
Naive Bayes-only	82.0	0.80	81.5	81.2	82.0
RNN-NB (Fusion)	96.9	0.92	97.61	97.61	97.63

Table 6 compares the performance of RNN-only, Naive Bayes-only, and the RNN-NB fusion model across key metrics such as Accuracy, AUC, F1-Score, Recall, and Precision. The experiments are performed on the same test

set to demonstrate the impact of combining RNN with Naive Bayes.

Explanation:

- RNN-only: A simple Recurrent Neural Network without Naive Bayes integration.
- Naive Bayes-only: The Naive Bayes classifier applied independently without any sequence modelling.
- RNN-NB (Fusion): The combined model that fuses the RNN and Naive Bayes classifiers.

By showing that the RNN-NB fusion model outperforms the standalone RNN-only and Naive Bayes-only models, you can validate the claim that combining these two models improves performance.

5 Conclusion

In order to predict how customers will use e-commerce platforms that cater to rural areas, this study offered a comprehensive data mining framework that combined cutting-edge preprocessing, feature engineering, and hybrid optimisation techniques. The research demonstrated the significance of utilising XGBoost and ClusterCentroids undersampling for feature selection in addressing data imbalance and dimensionality problems using real-world transactional data. With superior robustness and convergence speed compared to standalone metaheuristics, the suggested hybrid ABC-FHO optimisation algorithm skillfully combined global exploration with local exploitation. The study also used an RNN-NB fusion model, which enhanced accuracy, precision, recall, and F1-score across several experimental scenarios, to show how RNNs and NBs function together. High prediction performance and efficient training times were consistently given by XGBoost and the RNN-NB model, according to the comparative results. For sparse and dynamic behavioural data, the RNN-NB model stood out due to its consistent prediction abilities across different sequence lengths. In short, the suggested framework is a practical and extensible answer to the problem of how to anticipate better rural e-commerce customers' actions, which in turn allows for more targeted advertising, better stock management, and better customer relationship management. To further improve prediction accuracy and system responsiveness, future studies may investigate expanding the model to incorporate real-time behavioural analytics and multi-modal data sources.

Funding

Research on the Path of Rural E-commerce Empowering Rural Revitalization in Shaanxi Under the Digital Economy Background,2024ZD462

References

[1] Alghanam, O., M. O. (2022). *Data mining model for predicting customer purchase behaviour in an e-commerce context*. International Journal of Advanced Computer Science and Applications, 13(2). https://doi.org/10.14569/ijacsa.2022.0130249

[2] Bu, S.J.; Cho, S.B. Time Series Forecasting with Multi-Headed Attention-Based Deep Learning for Residential Energy Consumption. *Energies* **2020**, *13*, 4722. https://doi.org/10.3390/en13184722

[3] Bowen, A., & Grygorczyk, A. (2022). *Consumer eating habits and perceptions of fresh produce quality*. In W. J. Florkowski, N. H. Banks, R. L. Shewfelt, & S. E. Prussia (Eds.), *Postharvest Handling* (4th ed., pp. 487–515). Academic Press. https://doi.org/10.1016/b978-0-12-822845-6.00017-8

[4] Font-i-Furnols, M., & Guerrero, L. (2014). *Consumer preference, behaviour and perception about meat and meat products: An overview*. Meat Science, 98, 361–371. https://doi.org/10.1016/j.meatsci.2014.06.025

[5] Grant, J. (2012). *The Green Marketing Manifesto*. John Wiley & Sons Ltd. https://doi.org/10.1002/9781119206255

[6] Gülsün, B., & Aydin, M. R. (2024). *Optimising a Machine Learning Algorithm by a Novel Metaheuristic Approach: A Case Study in Forecasting*. Mathematics, 12(24). https://doi.org/10.3390/math12243921

[7] Alam, M.S.; Sultana, N.; Hossain, S.M.Z.; Islam, M.S. Hybrid Intelligence Modelling for Estimating Shear Strength of FRP Reinforced Concrete Members. *Neural Comput. Appl.* **2022**, *34*, 7069–7079. https://doi.org/10.1007/s00521-021-06727-4

[8] Lee, S. M., & Lee, D. H. (2020). *"Untact": A new customer service strategy in the digital age*. Service

Business, 14, 1–22. https://doi.org/10.1007/s11628-019-00408-2

[9] Lu, S., Ding, Y., (2023). *Multiscale Feature Extraction and Fusion of Image and Text in VQA*. International Journal of Computational Intelligence Systems, 16, 54. https://doi.org/10.1007/s44196-023-00233-6

[10] Narahari, Y., & Dayama, P. (2005). *Combinatorial auctions for electronic business*. Sādhanā: Academy Proceedings in Engineering Sciences, 30(2–3), 179–211. https://doi.org/10.1007/bf02706244

[11] Pu, X., Chai, J., & Qi, R. (2022). *Consumers' Channel Preference for Fresh Foods and Its Determinants during COVID-19—Evidence from China*. Healthcare, 10, 2581. https://doi.org/10.3390/healthcare10122581

[12] Raju, C. V. L., Ravikumar, K., & Shah, S. (2005). *Dynamic pricing models for electronic business*. Sādhanā: Academy Proceedings in Engineering Sciences, 30(2–3), 231–256. https://doi.org/10.1007/bf02706246

[13] Zhang, Y., Liu, H., & Chen, X. (2026). Hybrid metaheuristic optimization with deep learning fusion for predicting consumer behaviour in e-commerce environments. Expert Systems with Applications, 235, 121345. https://doi.org/10.1016/j.eswa.2025.121345

[14] Lee, J.; Cho, Y. National-Scale Electricity Peak Load Forecasting: Traditional, Machine Learning, or Hybrid Model? *Energy* **2022**, *239*, 122366. https://doi.org/10.1016/j.energy.2021.122366

[15] Wang, O., & Simon, S. (2018). *Consumer adoption of online food shopping in China*. British Food Journal, 120, 2868–2884. https://doi.org/10.1108/BFJ-03-2018-0153

[16] Wang, W., Xiong, W. (2023). *A user purchase behaviour prediction method based on XGBoost*. Electronics, 12(9), 2047. https://doi.org/10.3390/electronics12092047

[17] Wu, D., Sun, B., & Shang, M. (2023). *Hyperparameter Learning for Deep Learning-based Recommender Systems*. IEEE Transactions on Services Computing, 1(1), 1–13. https://doi.org/10.1109/tsc.2023.3234623

[18] Zhang, C., Liu, J., & Zhang, S. (2024). *Online Purchase Behaviour Prediction Model Based on Recurrent Neural Network and Naive Bayes*. Journal of Theoretical and Applied Electronic Commerce Research, 19(4), 3461–3476. https://doi.org/10.3390/jtaer19040180