Volume 3, Issue 4, 2025

Print ISSN: 2959-9865
Online ISSN: 2959-9873

WORLD JOURNAL OF ENGINEERING RESEARCH

World Journal of Engineering Research

Volume 3, Issue 4, 2025

Published by Upubscience Publisher

World Journal of Engineering Research
Print ISSN: 2959-9865 Online ISSN: 2959-9873
Email: info@upubscience.com
Website: http://www.upubscience.com/

Table of Content

World Journal of Engineering Research
Print ISSN: 2959-9865
Online ISSN: 2959-9873
DOI: https://doi.org/10.61784/wjer3043

OPTIMIZATION OF VEHICLE ROUTE FOR URBAN CLASSIFIED WASTE TRANSPORTATION

Yu Chen
College of Computer and Information Technology, China Three Gorges University, Yichang 443002, Hubei, China.
Corresponding Email: cystuctgu2025@163.com

Abstract: With the acceleration of urbanization, China faces severe challenges in waste management. Currently, waste sorting and transportation have become a key aspect of urban environmental governance, influenced by factors such as vehicle capacity constraints and collection requirements for different types of waste. This paper addresses the issue of urban waste sorting and transportation by constructing and solving optimization models under different constraints, effectively reducing transportation costs and carbon emissions while enhancing transportation efficiency. This paper simplifies such problems into the capacitated vehicle routing problem(CVRP), aiming to minimize the total daily transportation distance while considering constraints such as vehicle capacity constraints and multiple round-trip deliveries. A single-type waste transportation route optimization model is constructed and solved using a hybrid heuristic algorithm. Further considering dedicated vehicles for different types of waste, as well as their load, volume, time, and transportation cost constraints, new constraints are introduced based on this model, and a divide-and-conquer strategy and hybrid heuristic algorithm(HHA) is adopted to solve the multi-type waste transportation route optimization problem. By iteratively optimizing the routes, the transportation routes for various types of waste are reasonably planned, effectively reducing the cost of urban waste transportation.
Keywords: Capacitated Vehicle Routing Problem(CVRP); Route optimization; Divide-and-Conquer strategy; Hybrid heuristic algorithm

1 INTRODUCTION

As urbanization accelerates, China is confronted with formidable challenges in waste management. The vast quantities of waste produced are nearing the limits of the processing capacities of cities and regions across the country, rendering the issue of waste management increasingly prominent and demanding higher standards for the collection, transportation, and disposal of household waste. Among these, the classified transportation of waste constitutes a pivotal aspect of urban waste management, and the formulation of scientific and efficient waste classification and transportation routes is vital for urban environmental governance.

In reality, the issue of waste transportation has garnered significant attention from scholars owing to its complexity, which encompasses factors such as time windows, vehicle capacity, and citizen satisfaction. Shi et al. modeled this as a multi-depot vehicle routing problem (MDVRP) [1], aiming to minimize the total transportation distance. Involving with constraints such as warehouse supply and demand, vehicle capacity, they introduced a two-stage heuristic algorithm to address it. Yousra Bouleft et al. proposed a linear mathematical programming model to address the dynamic multi-compartmental vehicle routing problem (DM-CVRP) for selective and intelligent waste collection [2], aiming to minimize the total cost and incorporating constraints such as vehicle compartment capacity, waste station access rules, etc., and used the hybrid genetic algorithm (HGA) to solve. Yu et al. focused on the college scene [3], constructed a route optimization model and used ant colony optimization(ACO) algorithm, aiming at minimizing total clearance distance as the goal and considering the vehicle capacity and start-stop constraints. Mou et al. aimed to minimize total costs and established two types of route optimization models for separate transportation and joint transportation of classified waste [4], discussing both models by the optimization results. Shen et al. introduced the fuzzy membership function of residents' dissatisfaction degree [5], established a multi-objective model with minimum cost and maximum satisfaction degree, and solved it with NSGA Ⅱ algorithm. Zhou et al. build a multi-objective model covering economy [6], society and environment, and solved it with an improved algorithm which combined NSGA-III with simulated annealing. Wu et al. constructed a priority considered green vehicle routing problem (PCGVRP) model in a waste management system [7], and designed specific algorithms to solve it.

The existing research mostly focuses on single or partial scenarios, and rarely involves the classification and transportation of urban multiple wastes. Aiming at this problem, this paper first constructs a single-type of waste transportation vehicle routing optimization model, with the minimum total distance as the goal, considering the vehicle load and multiple transportation constraints, and uses a hybrid heuristic algorithm to solve it; Furthermore, the collaborative constraints of multi-type special vehicles are introduced to construct a multi-type waste transportation vehicle routing optimization model with the goal of minimizing the total daily transportation cost, which effectively reduces the transportation cost and improves the transportation efficiency.

2 VEHICLE ROUTING OPTIMIZATION FOR SINGLE-TYPE WASTE

Considering the complexity of the actual situation, this paper makes the following assumptions: all transport vehicles start from the same fixed departure point and return after completing the task; Each type of waste must be transported

by vehicles specially designed for that type; The traffic conditions of the roads used during transportation are stable, and sudden traffic jams or road closures are not considered.

2.1 The Construction of Single-type Waste Transportation Vehicles Routing Optimization(S-VRO) Model

Capacitated vehicle routing problem (CVRP) is a classical combinatorial optimization problem [8-10], which involves how to efficiently allocate vehicles to visit multiple customer points, and reduce the total distance or cost as much as possible under the condition of meeting the vehicle capacity constraints. CVRP is widely used in logistics, distribution, transportation and other fields. The vehicle routing problem for a single type of waste transportation described in this paper is a vehicle routing problem with vehicle load limit, which belongs to CVRP. Now the mathematical model of CVRP is introduced to build a single type of waste transport vehicle routing optimization model.

2.1.1 Definition of parameters

n: number of waste collection points; (x_i, y_i): coordinates of waste collection point i; ω_i: waste generation quantity at point i; Q: vehicle maximum load; (x_0, y_0):the fixed location of waste processing plant; d_{ij}: euclidean distance between points i and j, and $i, j = 1,2, ..., n$.

2.1.2 Decision variables

Let x_{ijk} be a 0-1 variable, indicating whether the kth vehicle travels from the collection point i to the collection point j. Let y_{ik} be a 0-1 variable, denoting whether the k th vehicle serves the collection point i. e_k is the total transportation distance of the kth vehicle. m denotes number of vehicles used in the transportation.

2.1.3 Objective function

The model aims to minimize the daily total transportation distance, shown as formula(1)-(3).

$$\min \sum_{k=1}^{m} e_k \tag{1}$$

$$e_k = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}\, x_{ijk} \tag{2}$$

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{3}$$

2.1.4 Constraints

Vehicle load limit: the transportation volume of each vehicle shall not exceed its maximum load Q tons, shown as formula (4):

$$\sum_{i=1}^{n} w_i\, y_{ik} \le Q, \quad \forall k \tag{4}$$

Rationality of vehicle route: each vehicle must start from the waste processing plant and finally return to the waste processing plant, shown as formula (5). Among them, the first constraint ensures that each vehicle starts from the waste processing plant, and the second constraint ensures that each vehicle finally returns to the waste processing plant.

$$\sum_{j=1}^{n} x_{0jk} = 1, \ \sum_{i=1}^{n} x_{i0k} = 1, \quad \forall k \tag{5}$$

Service integrity: each waste collection points need be served by a vehicle, shown as formula(6)

$$\sum_{k=1}^{m} y_{ik} = 1, \quad \forall i = 1, ..., n \tag{6}$$

Route continuity: the route of each vehicle is continuous. For the collection point served by each vehicle, the route continuity must be guaranteed, that is, the number of routes entering and leaving a collection point is equal, shown as formula (7).

$$\sum_{\mathrm{i}=0}^{n} x_{ipk} = \sum_{j=0}^{n} x_{pjk} = y_{pk}, \quad \forall \mathrm{p} = 1, ..., n, \forall \mathrm{k} \tag{7}$$

2.2 The Solution of S-VRO Model

This paper proposes a vehicle route planning problem based on 30 waste collection points to test the ability of the model and algorithm in solving the real urban waste transportation problem. The main content is to find the shortest path for waste transport vehicles under the condition that the load limit of a single vehicle is not more than 5 tons. Each vehicle is required to start from the waste processing plant, visit a certain point in turn and return to the waste processing plant. The amount of waste per trip shall not exceed the load limit of a single vehicle. Set the location coordinates of the waste processing plant to (0, 0), and the locations of 30 waste collection points and their waste volume distribution are shown in Figure 1.

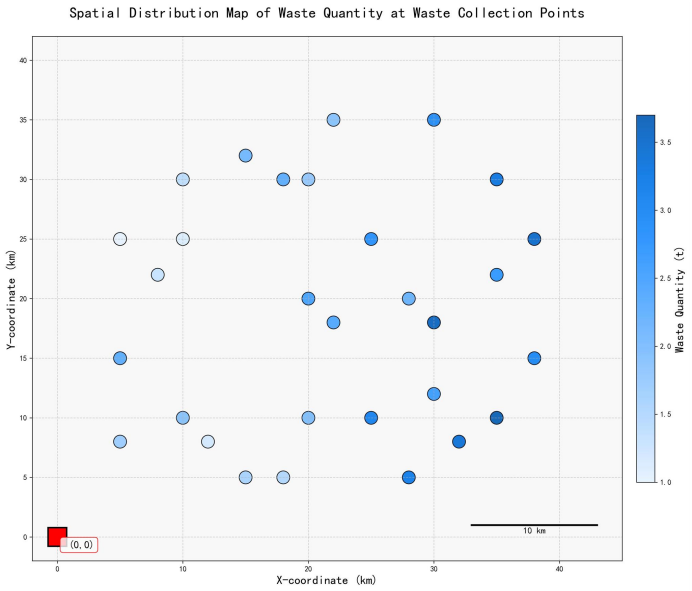


Figure 1 Thermodynamic Diagram of Spatial Distribution of Waste Quantity

This study uses hybrid heuristic algorithm(HHA) to solve S-VRO model.

In the initial solution construction phase, a feasible initial solution is rapidly generated via the cheapest insertion constructive heuristic algorithm, utilizing Euclidean distances computed from the coordinates of collection points. This solution satisfies the following constraints:

(1) Each vehicle departs from and returns to the disposal plant.

(2) Every collection point is serviced exactly once.

(3) The total load on any vehicle does not exceed the maximum load Q.

In the local search optimization phase, path refinement is performed through 2-opt edge exchange and node relocation operations. These mechanisms iteratively improve routes by targeting the minimization of total transportation distance, directly contributing to distance reduction by eliminating circuitous subpaths.

In the stage of constraint propagation, the solutions violating constraints are detected and trimmed in real time, and the load limit and the route continuity are dynamically maintained.

In the constraint propagation phase, solutions violating constraints are detected and pruned in real-time, dynamically enforcing capacity limits and route continuity.

In the search tree exploration phase, depth-first and best-first strategies are integrated. Through pruning of suboptimal branches and periodic restart mechanisms, global optimization is balanced with computational efficiency within limited time.

The algorithm ultimately outputs a near-optimal solution satisfying all constraints with minimized total transportation distance. This collaborative four-stage framework achieves efficient resolution of the waste collection vehicle routing problem.

The optimization result is shown in Figure 2. By optimizing the routes, under the premise of ensuring that each waste collection point is served by one vehicle, the detour of vehicles is avoided, which helps to reduce the total transportation distance of waste vehicles and improve the transportation efficiency.

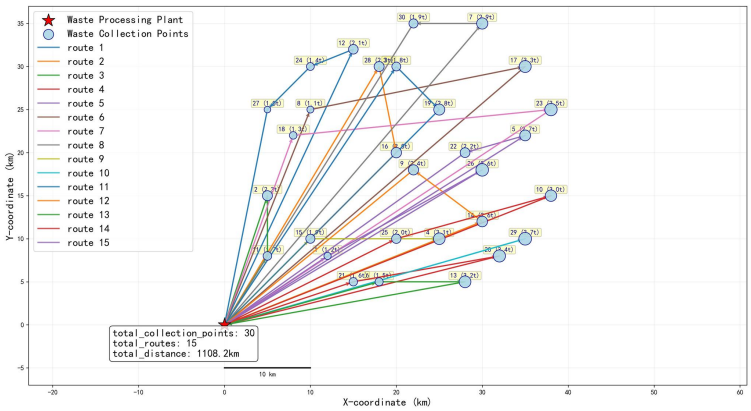


Figure 2 Single-Type Waste Transportation Routes

3 VEHICLE ROUTING OPTIMIZATION FOR MULTI-TYPE WASTE

3.1 The Construction of Multi-type Waste Transportation Vehicles Routing Optimization(M-VRO) Model

Various types of waste need different types of special vehicles to transport, and different vehicles have different capacity constrains (including load and volume) and transportation costs per unit distance. The working hours of each vehicle are also limited. A collection point may contain a variety of waste types, and the transportation routes and task allocation of each vehicle of different types schedules separately.

On account of the different types vehicles schedule separately, this multi-type vehicle routing problem (M-VRP) can be decomposed into multiple single-type vehicle routing subproblems. Namely, the divide-and-conquer strategy is used to solve the M-VRP:

(1) The original M-VRP is partitioned into homogeneous vehicle-specific subproblems based on type constraints;

(2) Each single-type VRP subproblem is solved independently using HHA;

(3) The partial solutions are then integrated through a coordination mechanism to construct a complete vehicle scheduling scheme.

This study is based on the single type of waste transportation vehicle routing optimization model proposed in the previous paper, aiming at minimizing the total daily transportation cost, and introduces new constraints to build a multi type of waste transportation vehicle routing optimization model. Apart from the parameters, variables and constrains defined in 2.1 part, this part adds some new things.

3.1.1 Definition of parameters

t: waste category index; w_{it}: weight(ton) of category t waste at collection point i; v_{it}: volume (m³) of category t waste at collection point i; Q_t: maximum payload capacity (ton) for type-t vehicles; V_t: maximum volumetric capacity (m³) for type-t vehicles; C_t: unit transportation cost (yuan/km) for type-t vehicles; T: daily maximum work time of each vehicle. S: the travel speed of each vehicle.

3.1.2 Decision variables

Let m_t denotes number of type-t vehicle in the transportation.

3.1.3 Objective function

The model aims to minimize the daily total transportation cost, shown as formula(8)

$$\min \sum_{i=1}^{n}\sum_{j=1}^{n} C_t\, d_{ij} x_{ijk},\ \forall \mathrm{t}, k \tag{8}$$

3.1.4 Constrains

Vehicle load limit: the transportation load of each type-t vehicle shall not exceed its maximum load Q_t tons, shown as formula (9).

$$\sum_{i=1}^{n} w_{it}\, y_{ik} \le Q_t, \quad \forall \mathrm{t}, k \tag{9}$$

Vehicle volume limit: the transportation volume of each type-t vehicle shall not exceed its maximum load V_t tons, shown as formula (10).

$$\sum_{i=1}^{n} v_{it}\, y_{ik} \le V_t, \quad \forall \mathrm{t}, k \tag{10}$$

Vehicle work time limit: the daily work time of each vehicle shall not exceed its maximum time T hours, shown as formula (11).

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}\, x_{ijk}}{S} \le \mathrm{T}, \quad \forall k \tag{11}$$

And the constrains of rationality of vehicle route, service integrity, route continuity is same as the formula(5)-(7).

3.2 The Solution of M-VRO Model

Similarly, in the scenario of 30 waste collection points and 1 waste processing plant, there are four types of waste, namely food waste, recyclables, hazardous waste and other waste. Each collection point can produce four types of waste. Each type of waste needs to be transported by special vehicles. The vehicle types correspond to the above four types of waste respectively. Their load limit, volume limit and transportation cost per unit distance are different, as shown in Table 1. Vehicles start from the treatment plant and return after collecting the same type of waste. Different types of vehicles can be scheduled independently. The density of food waste, recyclables, hazardous waste and other waste is 700 kg/m³, 200 kg/m³, 900 kg/m³, 400 kg/m³.

Table 1 Parameters of Four Types Waste Transport Vehicles

Type	Load(tons)	Volume(m³)	Cost(yuan/km)
Food waste	8	20	2.5
Recyclables	6	25	2

Hazardous waste	3	10	5
Other waste	10	18	1.8

The spatial coordinates of the waste processing plant and 30 collection points remain fixed. Each collection point generates multiple waste categories, with the spatial distribution of waste streams visualized in Figure 3. Analysis confirms that all points produce quantifiable amounts of four waste types.

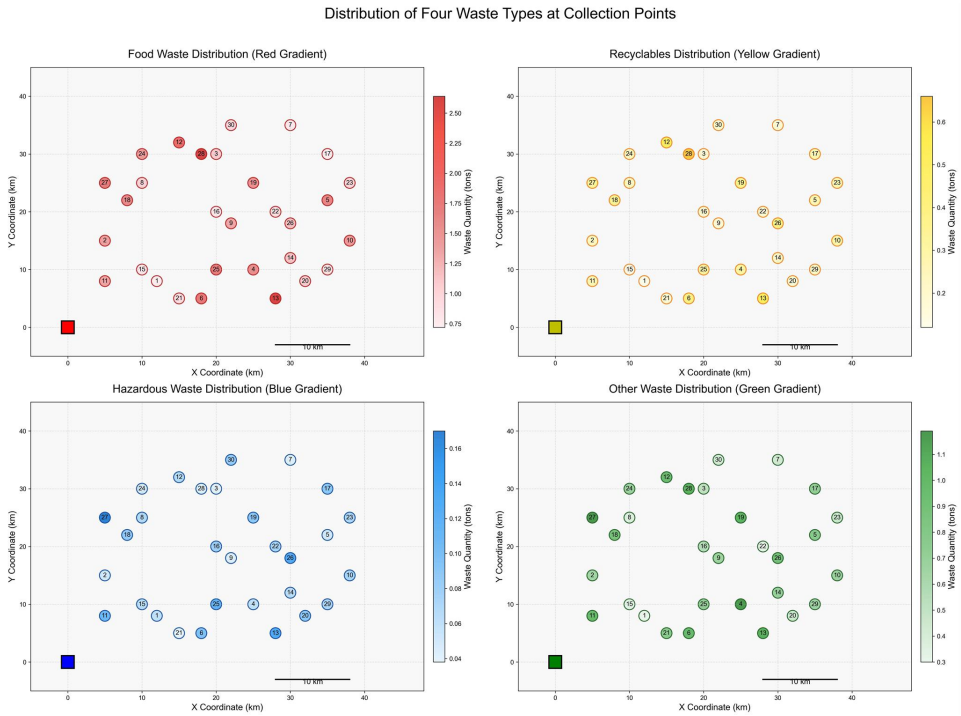


Figure 3 Thermodynamic Diagram of Spatial Distribution of Four Types Waste Quantity

In this paper, the divide-and-conquer strategy is used to solve each type of waste transportation route one by one using HHA. The final optimization results are shown in Figure 4. After route optimization, the transportation routes of different types of waste are reasonably allocated, and the vehicle load, volume requirements and time constraints are not violated.

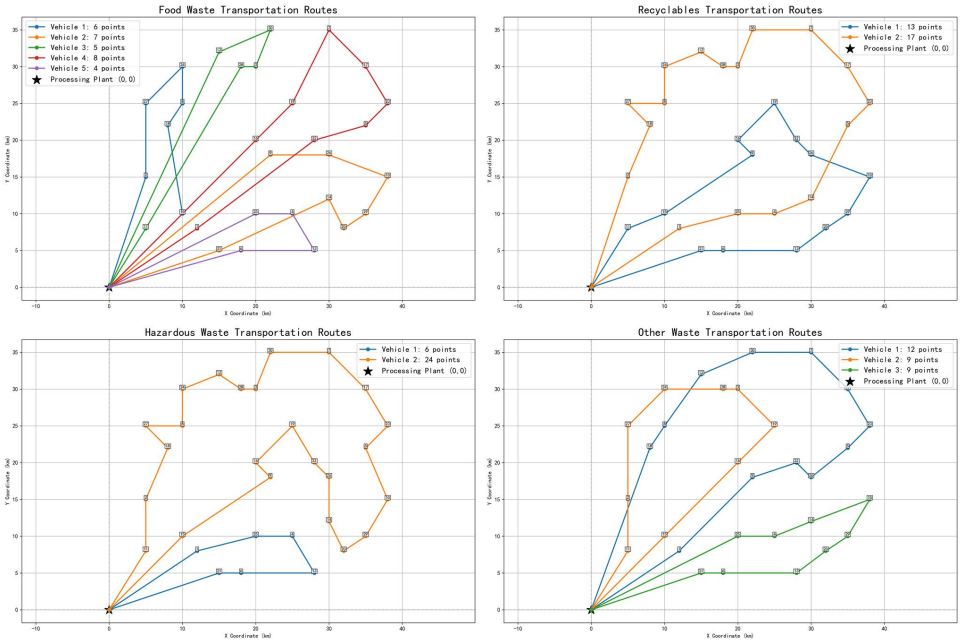


Figure 4 Four Types Waste Transportation Routes

The Figure 5 illustrates the load and volume utilization rates of specialized waste collection vehicles. For food waste, recyclables, and other waste vehicles, both load and volume metrics approach operational maxima (98-100% utilization), achieving near-full capacity efficiency. This phenomenon further demonstrates the critical influence of waste density on transport operations: lower-density waste streams predominantly constrain vehicle volume capacity. Conversely, hazardous waste vehicles exhibit suboptimal utilization rates. This inefficiency stems from the daily work time constrains. When a vehicle's daily operating time reaches its 8-hour regulatory limit, remaining tasks are reallocated to supplementary vehicles. Although this time-based scheduling inevitably incurs additional costs, it ensures compliance with occupational safety regulations.

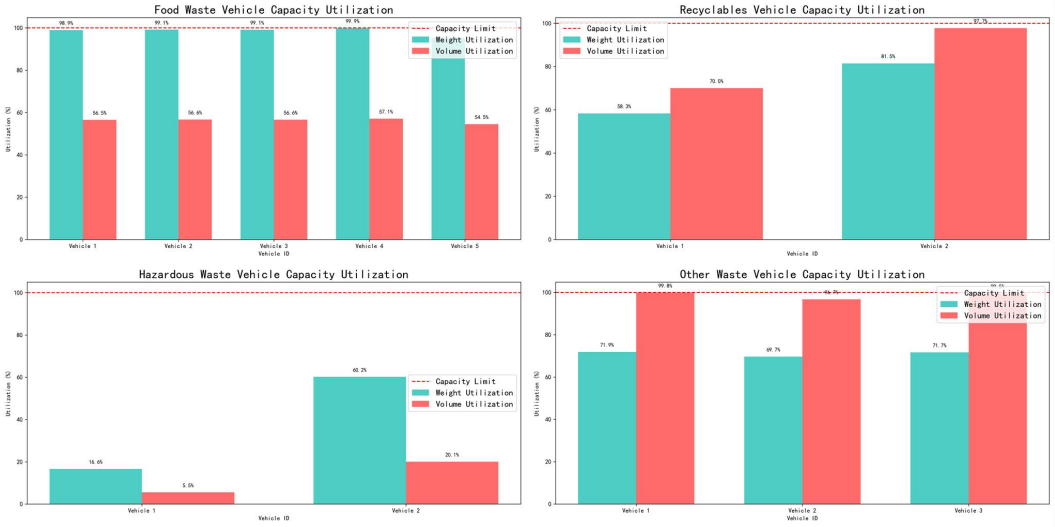


Figure 5 The Capacity Utilization of Four Types Waste Vehicles

4 CONCLUSION

The optimization model constructed in this paper provides scientific decision support for the management of classified waste transportation. Firstly, this paper constructs a single-type of waste transportation vehicle routing optimization model, based on which a multi type of waste transportation vehicle routing optimization model is proposed, and the model is solved by combining divide-and-conquer strategy and hybrid heuristic algorithm. Finally, this paper verifies the effectiveness of the model and algorithm through the analysis of an example, and draws a conclusion. The model not only reduces the transportation cost, shortens the driving distance, but also meets the actual operation requirements such as specialized vehicles coordination and time limit. At the same time, the model solving idea of the combination of divide-and-conquer strategy and hybrid heuristic algorithm proposed in this paper can be applied to the actual scene of urban domestic waste transportation, and provide technical support for improving the economy, efficiency and environmental protection of classified waste transportation. Future research can further consider dynamic traffic conditions, waste production fluctuations and other factors to enhance the robustness and practical adaptability of the model.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Shi Yanjun, Lv Lingling, Hu Fanyi, et al. A heuristic solution method for multi-depot vehicle routing-based waste collection problems. Applied Sciences, 2020, 10(7): 2403.
[2] Bouleft Yousra, Ahmed Elhilali Alaoui. Dynamic multi-compartment vehicle routing problem for smart waste collection. Applied System Innovation, 2023, 6(1): 30.
[3] Yu Haonan, Zhang JianHui, Ao Tiantian, et al. Overseas M&A, Research on Waste Collection Vehicle Routing Optimization Based on Ant Colony Algorithm. Value Engineering, 2025, 44(02): 152-154.
[4] Mou Nengye, Cheng Chiyao, Jiang Erwei, et al. Overseas M&A, Multi-type & Multi-trip Vehicle Routing Optimization for Municipal Solid Waste Classification and Transportation. Journal of Safety and Environment, 2022, 22(04): 2199-2208. DOI: 10.13637/j.issn.1009-6094.2021.0527.
[5] Shen Jiangang, Guo Zhiyi, Wang Xiao. Overseas M&A, GIS-based Routing Optimization Analysis of Waste Collection Vehicles Considering Resident Satisfaction. Traffic & Transportation, 2022, 38(02): 22-27.
[6] Zhou, Jian, Meixi Zhang, Sisi Wu. Multi-objective vehicle routing problem for waste classification and collection with sustainable concerns: the case of shanghai city. Sustainability, 2022, 14(18): 11498.

[7] Wu Hailin, Tao Fengming, Yang Bo. Optimization of vehicle routing for waste collection and transportation. International Journal of Environmental Research and Public Health, 2020, 17(14): 4963.
[8] Tang Kaiqiang, Fu Huiqiao, Liu Jiasheng, et al. Overseas M&A, Hierarchical Optimization for Capacitated Vehicle Routing Based on Deep Reinforcement Learning. Systems Engineering and Electronics, 2025, 47(03): 827-841.
[9] Alesiani Francesco, Gulcin Ermis, Konstantinos Gkiotsalitis. Constrained clustering for the capacitated vehicle routing problem (CC-CVRP). Applied artificial intelligence, 2022, 36(1): 1995658.
[10] Ge Bin, Tian Wenzhi, Xia Chenxing, et al. Overseas M&A, Solving capacitated vehicle routing problem using end-to-end deep reinforcement learning. Application Research of Computers, 2024, 41(11): 3245-3250.

World Journal of Engineering Research
Print ISSN: 2959-9865
Online ISSN: 2959-9873
DOI: https://doi.org/10.61784/wjer3044

A DOUBLE LIQUID STORAGE LOOP HEAT PIPE BASED ON SLIDING SLEEVE REGULATION

XuanNing Liang, Jing Zhang, ZhengJie Liu, Feng Qu, WenXian Kuang, ShengNan Yan, Heng Xiao*
College of Energy and Power Engineering, Xihua University, Chengdu 610039, Sichuan, China.
Corresponding Author: Heng Xiao, Email: xiaoheng@xhu.edu.cn

Abstract: To address the limitations of traditional loop heat pipe evaporators-such as uncertain operational performance under different orientation configurations, unreasonable working fluid volume in the reservoir, and thermal leakage issues—this study proposes a novel dual-reservoir loop heat pipe. Featuring an annular nozzle-coupled sleeve design, this innovative configuration enhances heat transfer efficiency during complex gravitational variations. The research investigates how varying gravitational orientations and liquid filling rates affect the heat transfer characteristics of the dual-reservoir loop. Simulation software was employed to evaluate thermal performance across different gravitational conditions and filling rates. Results demonstrate that under 65% working fluid filling rate and 40W thermal load, the phase-change liquid medium is replenished promptly. Continuous evaporation occurs on the capillary core surface without drying out, ensuring stable operation of the loop heat pipe. This design effectively meets the mission's requirements for stable thermal dissipation in complex aerospace environments.
Keywords: Loop heat pipe; Double liquid chamber; Thermal stability; Phase change heat transfer; Thin liquid film boiling

1 INTRODUCTION

With the rapid development of artificial intelligence, quantum computing, 5G communication technology and so on, the heat generated by related electronic devices per unit volume increases rapidly, and the traditional heat dissipation means can no longer meet the demand. The development of new efficient heat dissipation technology is the key to ensure the safe and stable operation of electronic devices.

Electronic devices employ various cooling methods, including air-cooling and liquid-cooling systems, with phase-change cooling demonstrating the most effective performance[1]. The Loop Heat Pipe (LHP) is a closed-loop annular heat pipe that utilizes phase change of working fluids to drive thermal circulation, offering advantages such as high heat transfer efficiency and uniform temperature distribution. As an advanced cooling technology, its gas-liquid fluid pathways enable long-distance heat transfer without external power supply, while adapting to complex installation environments. With broad application prospects in electronics cooling, aerospace, and other fields, LHP has garnered significant attention. Researchers have conducted extensive studies on LHP parameters including working fluid properties, fluid charge volume, placement angles, and evaporator structures.

Yang Chengxiang et al. developed a novel sodium-based flat evaporator high-temperature loop heat pipe (HTLHP) for hypersonic aircraft exhaust nozzle thermal protection[2], establishing its design process and conducting heat transfer limit simulation analysis. Experimental studies investigated the effects of condensation intensity, heat pipe structure, and inclination angle on startup characteristics and heat transfer performance. The findings revealed that appropriate condensation and inclination angles could resolve startup oscillation issues, achieving a minimum heat transfer resistance of 0.0368°C/W. It was emphasized that controlling heat leakage in the liquid reservoir and maintaining working fluid reflux stability were crucial for performance enhancement. Zhang Hongxing et al. designed a dual-liquid-tank ammonia-stainless steel loop heat pipe (LHP) to overcome orientation limitations during gravity field operations[3]. By precisely matching liquid tank volume with working fluid charge, they ensured continuous liquid filling in the liquid channel, avoiding challenging startup scenarios. Experimental verification demonstrated successful operation under various orientations. However, significant temperature variations occurred between orientations within the variable thermal conductivity zone (75-250W), attributed to differing cooling effects of reflux liquid overcooling in the liquid guide tubes. Niu Wenjing et al. proposed a heat pipe-fuel system-based cooling method for severe electrical actuator heating[4]. Numerical simulations revealed that when the equivalent heat transfer coefficient of heat pipes reached 7000W/(m·K), thermal management of electrical actuators was effective. Further increases in heat transfer coefficients showed limited improvement in cooling performance. Wang Dongdong et al. improved the heat dissipation performance of loop heat pipes (LHP) by optimizing evaporator sealing processes and designing a dual-capillary-core structure[5]. Experimental results showed that the maximum operational thermal load increased from 140W to 240W under welded sealing, while the dual-capillary core design effectively reduced lateral wall heat transfer, enhancing system startup and operational efficiency. Xia Yijun et al. developed a novel micro-channel loop heat pipe[6]. Through experimental and simulation studies on its heat exchange performance, it was found that the heat pipe operates stably across 30W-180W thermal loads with rapid response during variable load testing, demonstrating excellent reliability. The research also revealed that modified capillary core structures significantly increase refrigerant evaporation rates and lower system operating temperatures, meeting future data center server cooling demands. Liu Siyuan et al. investigated

the steady-state performance of stainless steel-ammonia dual-cylinder loop heat pipes under gravity and 1g-7g inverse acceleration environments with 100-300W thermal loads[7]. They established a flow resistance prediction model for loop heat pipes in accelerated environments, discovering that inverse acceleration increases loop flow resistance leading to higher operating temperatures, with longer liquid pipelines showing more pronounced effects. This provides guidance for designing loop heat pipes for airborne electronic equipment cooling.

Researchers have conducted extensive investigations into loop heat pipes (LHPs) across various aspects including working fluid properties, filling ratios, inclination angles, and evaporator structures, achieving significant advancements[8-9]. However, critical limitations persist in the current LHP evaporator designs documented in existing literature, including the improved dual-reservoir configuration: Firstly, the conventional single-reservoir and wick layout presents challenges in fluid supply along non-gravitational orientations, leading to system failures due to insufficient capillary forces to overcome flow resistance in specific directions without gravity assistance. Secondly, while the existing dual-reservoir structure partially addresses fluid supply issues, it fails to resolve system pressure imbalance and heat transfer limitations caused by the inability to accommodate excess condensate from the condenser during reservoir saturation, fundamentally due to the absence of a dynamic working fluid accommodation mechanism. Thirdly, prevalent thermal leakage issues result in premature evaporation of working fluid in the reservoir, causing vapor-liquid interface fluctuations that disrupt stable fluid return and induce system temperature oscillations, stemming from inadequate thermal isolation or management strategies in the reservoir. To address these core challenges of fluid supply bottlenecks, heat transfer limitations, and temperature oscillations that constrain LHP performance in complex orientations (particularly in scenarios with variable gravitational directions), this study proposes a novel dual-reservoir LHP structure incorporating a sliding sleeve regulation mechanism. This innovative design features dynamically controlled nozzle operation through sleeve movement, enabling intelligent regulation of working fluid distribution and flow between dual reservoirs. It aims to comprehensively resolve non-gravitational orientation fluid supply challenges, achieve optimal working fluid allocation in reservoirs, and ensure long-term system temperature stability, representing a significant advancement in overcoming orientation limitations and enhancing the thermal performance and operational reliability of loop heat pipes.

2 FUNDAMENTAL

A loop heat pipe (LHP) is a passive heat transfer device based on the capillary force driven phase change cycle of working medium. Its core consists of evaporator, condenser, steam pipeline, liquid pipeline and storage chamber, and its working principle can be divided into four key processes.

The system operates through four key phases: 1) Evaporation and Heat Absorption: The heat source activates the evaporator, heating porous liquid-absorbing cores (typically sintered metal components). When liquid working fluids (e.g., acetone or ammonia) within these cores absorb heat, they undergo phase transitions to generate steam[10-12]. The resulting capillary forces provide initial driving force for circulation, directing vapor through microchannels on the evaporator's surface into the steam pipeline. 2) Steam Transport and Condensation: High-temperature vapor is transported via pipelines to the condenser, where it releases latent heat during condensation in cooling elements (e.g., radiators), recondensing into liquid working fluid. The flexible design of the steam pipeline enables long-distance heat transfer (over 1 meter) while isolating mechanical vibrations. 3) Liquid Return and Pressure Balance: Condensate flows back to the storage chamber through liquid pipelines. This chamber stores working fluid, buffers volume changes caused by thermal load fluctuations, and maintains stable reflux by balancing system pressure differentials through capillary forces. 4) Capillary-Driven Circulation: The micro-porous structure (with pore diameters approximately 10 micrometers) creates capillary pressure differentials that overcome fluid flow resistance, driving closed-loop circulation. These flows follow Darcy's law, enabling self-sustaining operation without external power input (Figure 1) [13].

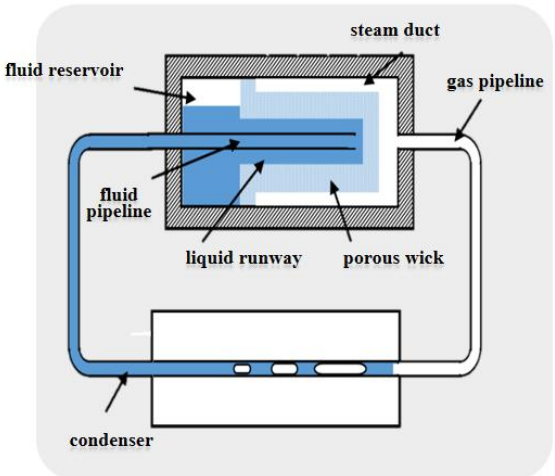


Figure 1 Loop Heat Pipe

3 INNOVATIVE STRUCTURAL DESIGN OF DUAL-CHAMBER EVAPORATOR

3.1 Dual Reservoir Chamber Layout And Circulation Pathway

Loop heat pipes typically consist of an evaporator, condenser, liquid reservoir, and vapor/liquid flow channels. In this study, we designed a novel dual-reservoir loop heat pipe using liquid ammonia as the working fluid, with both the outer shell and wick made of 310S stainless steel. The dimensions of this innovative design are shown in Table 1 (where the tube length represents the total path length of the working fluid).

Table 1 Structural Dimensions of Loop Heat Pipe

structure	parameter	size
evaporimeter	Outer diameter/inner diameter × length/mm	25/23×190
reservoir	Outer diameter/inner diameter × length/mm	40/38×45
	capillary pressure	>70kpa
Catheter hairs	Outer diameter/inner diameter × length/mm	40/38×45
	Outer diameter/inner diameter × length/mm	23/12×170
	Maximum capillary radius/μm	0.55
	porosity	55%
	Permeability/m2	2×10-14
steam pipe line	Outer diameter/inner diameter × length/mm	2/0×2300
Condensate line	Section length/widness × flow channel length/mm	4/3×300
liquid line	Outer diameter/inner diameter × length/mm	3/2×1900
Working medium (liquid nitrogen)	Fill quantity/g	34g
jacket	Outer diameter/inner diameter × length/mm	2.5/2.3×39.5

The loop heat pipe evaporator system comprises an evaporator housing, a liquid reservoir, a wick assembly, and interconnected liquid piping, the overall structure of loop heat pipe is shown in Figure 2 and 3. The liquid reservoir consists of two compartments (first and second) positioned at opposite ends of the housing. The piping configuration includes a liquid-phase conduit and a pilot tube that interconnects both reservoirs. Notably, the pilot tube within the second reservoir features a sealed end to prevent refrigerant leakage from its port.

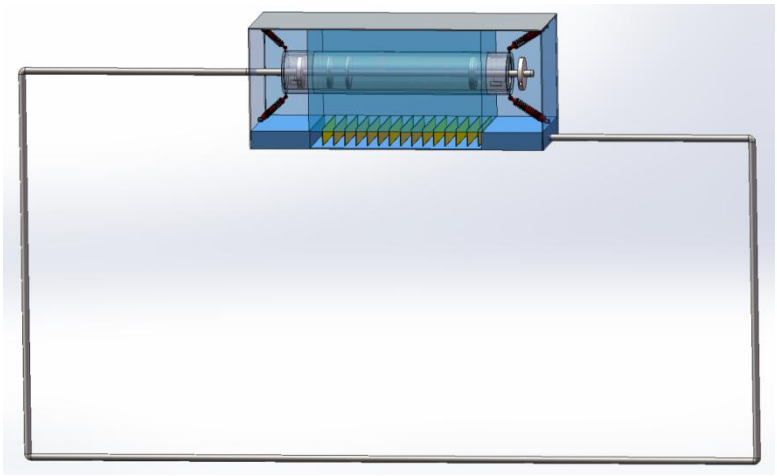


Figure 2 Loop Heat Pipe Overall

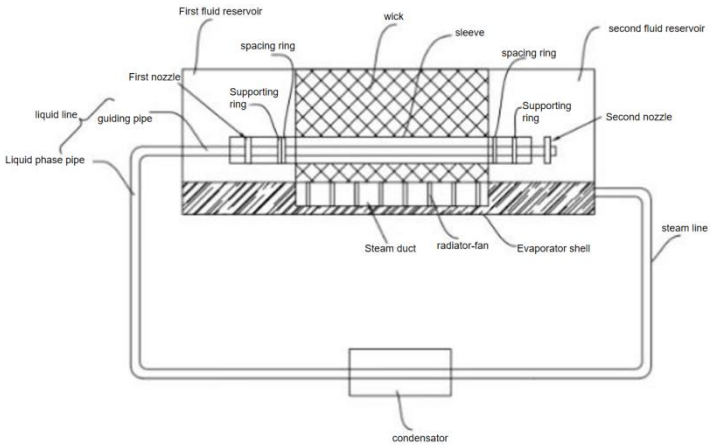


Figure 3 Overall Structure Plan

The steam channel is set below the suction core, so that the working medium can have more full contact with the steam channel[14]. It also includes several square structure fins distributed at intervals, which increases the steam contact area and effectively improves the heat dissipation effect. The Steam channel of loop heat pipe is shown in Figure 4.

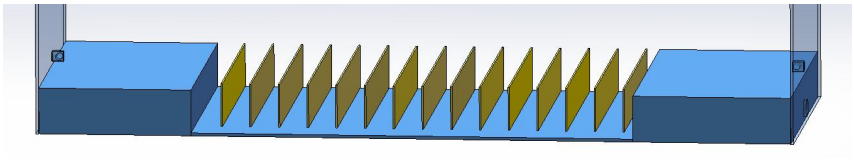


Figure 4 Steam Channel

The guide tube is equipped with a first nozzle and a second nozzle positioned at both ends of the liquid storage chamber. A support platform is fixed on the tube, which contacts the inner wall of the sleeve to prevent slippage when the platform engages with the sleeve's positioning ring. As the working direction changes, the sleeve slides along the guide tube under the combined action of the positioning ring and support platform, thereby releasing or closing the nozzles. This mechanism effectively balances the liquid level in the storage chamber while ensuring proper heat dissipation operation.

The sleeve includes an outer shell, an inner shell and an end ring. The end ring is placed at both ends of the sleeve and connected with the inner and outer shells respectively. A cavity is set between the outer shell and the inner shell, and a number of smooth steel balls are set in the cavity to act as gravity balls, so that the sleeve slides more smoothly.

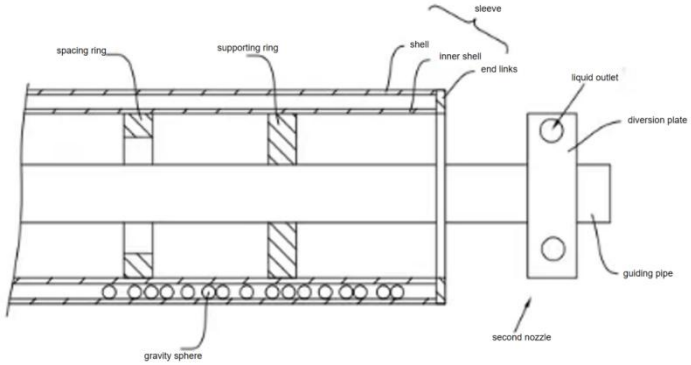


Figure 5 Cross Section of the Sleeve

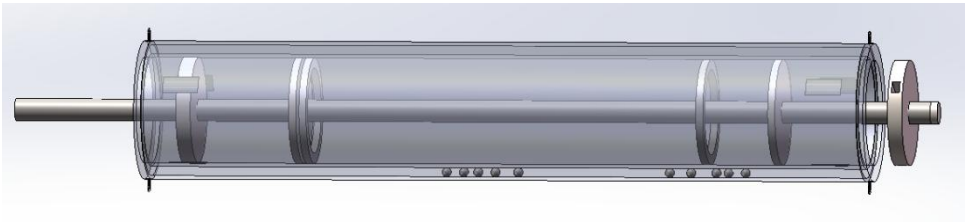


Figure 6 Overall Structure of the Casing

The distance between the first and second nozzles and the liquid suction core should be at least half the axial length of the liquid chamber. This configuration ensures that after passing through the condenser, the low-temperature liquid refrigerant flows downward slowly from both nozzles. The discharged refrigerant gradually moves toward the suction core, mixing with most of the preheated refrigerant in the chamber at the top, thereby cooling it and preventing thermal shock. As shown in Figures 5 and 6, both nozzles feature a distributor plate with multiple circumferential discharge ports on their side walls that adhere to the sleeve's inner wall. These ports connect to guide tubes, allowing the refrigerant to spray out and fully contact the chamber walls during both processes, effectively eliminating heat leakage.

3.2 Anti-Gravity Interference Stability Design

To ensure safety performance, protect heat pipe components, and enhance sleeve sliding stability, dampers are installed within the evaporation chamber. When external vibrations occur—such as during aircraft acceleration, braking, or steering—the dampers help maintain structural stability of the sleeve, ensuring it operates as designed. By reducing sleeve vibration, the dampers minimize impacts on aircraft and other heat pipe components, thereby extending service life. Additionally, they reduce stress and displacement caused by thermal expansion, mechanical vibrations, or external forces, ultimately safeguarding the structural integrity and safety of the heat pipes.The resistor is shown in Figure 7 and 8 .

Figure 7 Resistor Structure

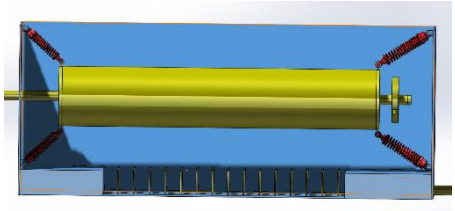


Figure 8 Resistor Distribution

3.3 Thermal Leakage Suppression Strategy

To ensure stable working fluid circulation efficiency and liquid storage capacity, a sealed space must be formed between the support platform and sleeve. Therefore, sealing rings are installed at the contact points between the support ring and sleeve. During high-risk flight maneuvers, the reciprocating movement of heat pipe sleeves causes fluid-induced gaps in sealing surfaces, which can lead to dry friction. While traditional rubber seals provide good sealing performance, their high friction resistance slows sleeve response speed. Moreover, dry friction accelerates wear, potentially causing seal burnout or leakage. Silicon carbide (SiC) emerges as an ideal material due to its high-temperature resistance and excellent tribological properties. Hot-pressed SiC maintains stable performance even at 1600℃[15]. During sealing surface friction, its graphite particles lubricate internally at low temperatures, while amorphous silicon oxide formed below 600℃ acts as a solid lubricant on surfaces covered at 800-900℃. Consequently, SiC exhibits strong anti-adhesion and wear resistance, meeting stringent sealing requirements under harsh operating conditions.

4 SIMULATION RESULTS AND PERFORMANCE ANALYSIS

4.1 Static Simulation Analysis

Considering the complexity of the continuous change in the direction of gravity during the actual maneuvering of fighter aircraft, if each operating condition angle is simulated one by one, it will significantly increase the computational cost and lack engineering relevance. Therefore, in this paper, four typical flight attitudes, namely 0° (horizontal), -45° (dive), 45° (climb), and 90° (yaw), are selected for static analysis. Their representativeness is reflected in the following: 0° represents the cruise reference state, ±45° cover the maximum gravity component conditions within the pitch plane, 90° simulate extreme lateral maneuvers, and verify the structural response under lateral overload. The selected flight attitudes cover the key maneuver planes; -45° and 45° directions respectively produce the maximum stretching and compression deformations, while the 90° direction reveals the eccentric deformation characteristics of the sleeve under low acceleration, jointly forming the multi-axis mechanical performance envelope, representing the load boundary; the selected angles correspond to the typical tactical actions of fighter aircraft, such as: accelerating pursuit, large-angle climb, and sharp turn avoidance, which can effectively verify the failure prevention capability of the sleeve limit structure, support platform, and limit ring and dampers under extreme acceleration sudden changes, to meet the engineering failure protection requirements. This simplified strategy covers the main failure risks through typical operating conditions, while significantly improving the simulation efficiency while ensuring the reliability of the conclusion, and conforms to the design principle of "lightweight verification" of aviation heat pipes.

This static simulation study employs a fighter aircraft flight scenario at 2,000 feet altitude. The aircraft accelerates from cruising speed V_c =0.95 Mach to pursuit speed V_p =2.4 Mach within 40 seconds[16-18]. The acceleration component in the aircraft's longitudinal direction is denoted as a=12.339m/s^2, with gravitational acceleration g=9.81m/s^2 applied. Using SolidWorks' Simulation plugin, we conducted static analysis of the annular heat pipe sleeve's operation under various angular configurations during flight.

In this simulation, the loop heat pipe sleeve and nozzle structure are made of nickel-silver (copper, zinc, nickel alloy), with a density of ρ=8.9g/cm^3. The size of the heat pipe is d_{outer} =2.5mm , d_{inner} =2.3mm,L=39.5mm ,
$V=\pi(R^2-r^2)L=0.0298cm3$, $m=\rho V=0.26522g$

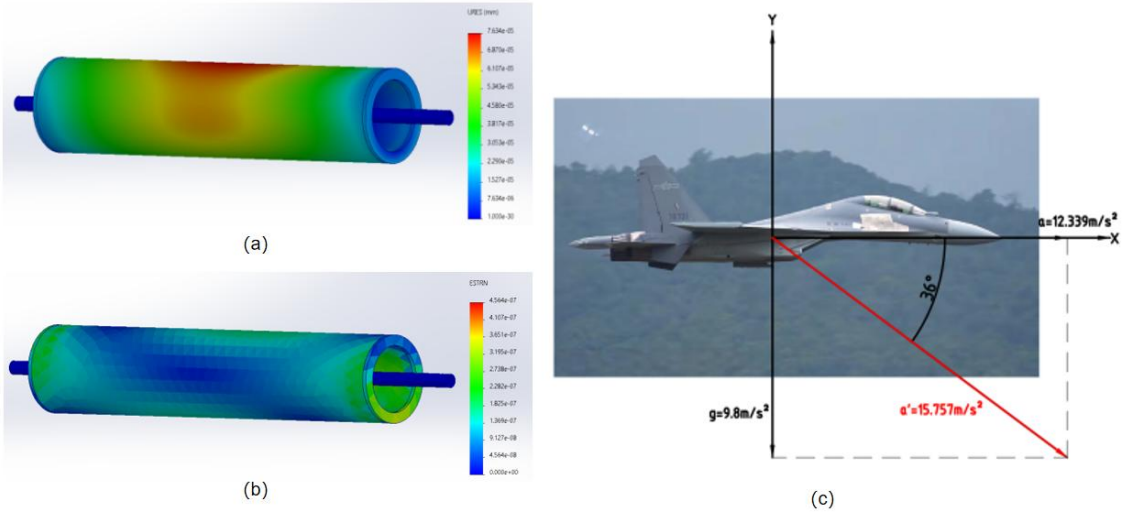


Figure 9 (a) Deformation in the Horizontal Direction of Travel; (b) Strain in the Horizontal Direction of Travel; (c) Acceleration in the Horizontal Direction of Travel

Figure 9(a)-(c) illustrates deformation, strain, and acceleration during the aircraft's horizontal movement. Figure 9(a) shows that the central sleeve exhibits maximum deformation at 7.634×10^{-5} mm, with deformation decreasing toward both ends. Figure 9(b) reveals that the lower half of the sleeve's end displays the highest strain at 4.564×10^{-7}, while deformation decreases from the ends toward the center. Figure 9(c) demonstrates the combined acceleration a' = 15.757m/s^2 for the aircraft and loop heat pipe system.

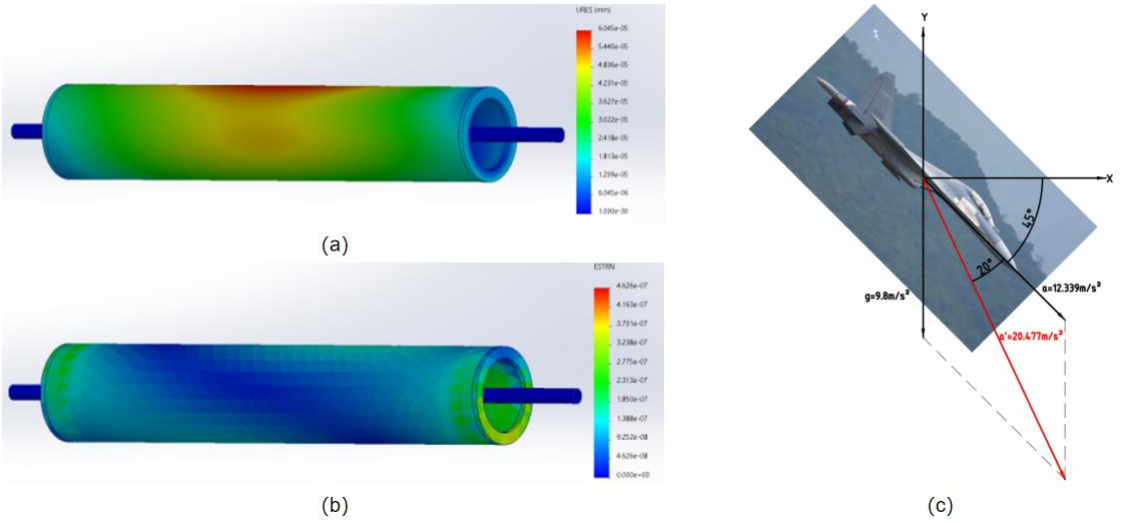


Figure 10 (a) Deformation in the Direction of-45°; (b) Strain in the Direction of-45°; (c) Acceleration in the Direction of-45°

Figures 10(a)-(c) illustrate deformation, strain, and acceleration during the aircraft's-45° directional movement. Figure 10(a) shows that the upper half of the sleeve midsection exhibits the highest deformation with a maximum value of 6.045×10^{-5} mm, decreasing from the center to both ends. Figure 10(b) reveals the lower half of the sleeve end segment displays the greatest strain at 4.626×10^{-7}, which diminishes from the extremities toward the center. Figure 10(c) demonstrates the combined acceleration a' =20.477m/s^2 between the aircraft and the loop heat pipe during this phase.

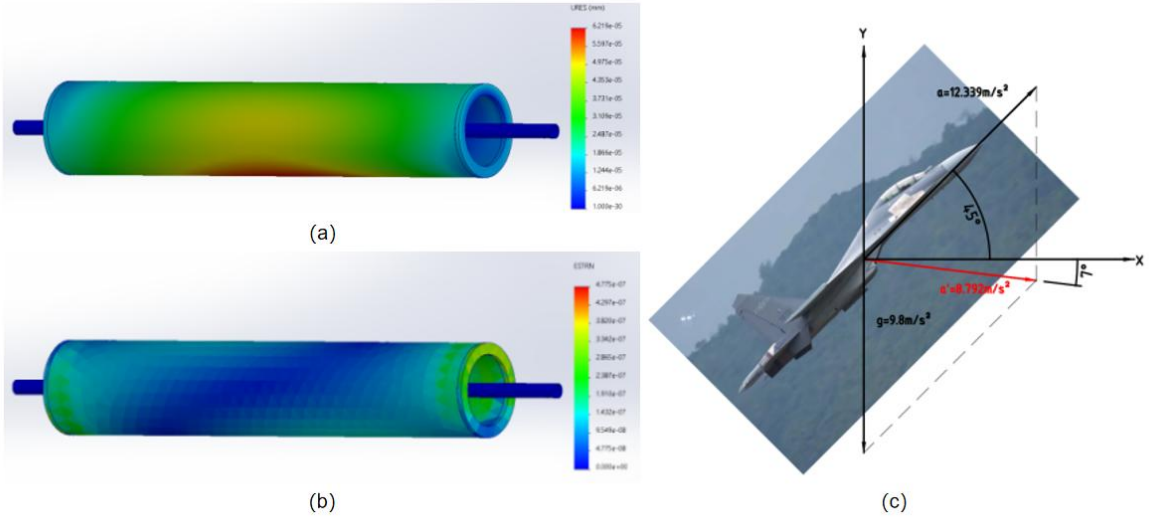


Figure 11 (a) Deformation in the 45° Direction of Travel; (b) Strain in the 45° Direction of Travel; (c) Acceleration in the 45° Direction of Travel

Figures 11(a)-(c) illustrate deformation, strain, and acceleration during the aircraft's 45° directional movement. Figure 11(a) shows that the lower half of the sleeve's midsection exhibits the highest deformation with a maximum value of 6.219×10^{-5} mm, decreasing from the center to both ends. Figure 11(b) reveals the upper section of the sleeve's tip demonstrates the greatest strain at 4.775×10^{-7}, decreasing from the tip toward the center. Figure 11(c) displays the combined acceleration a' =8.792m/s^2 between the aircraft and the loop heat pipe during this phase.

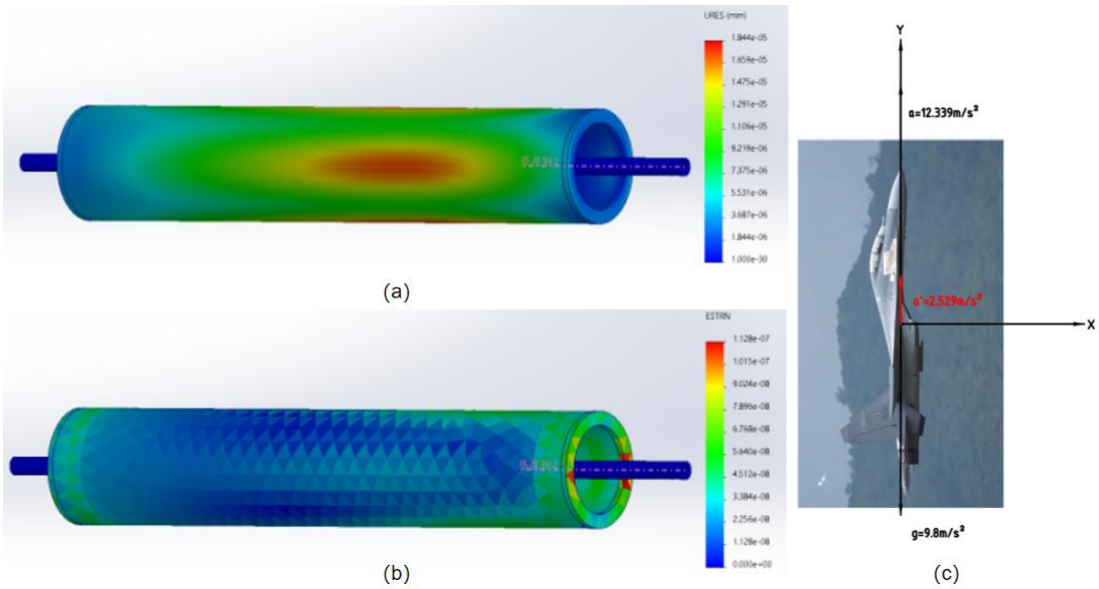


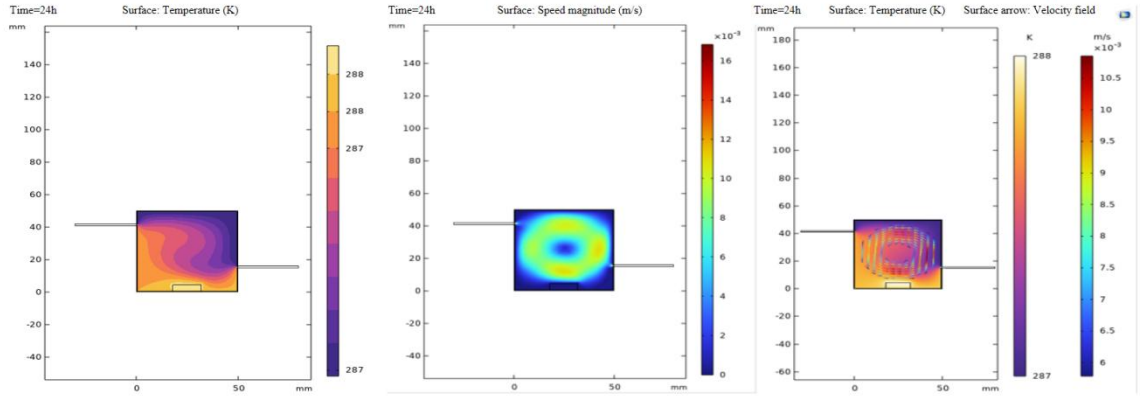
Figure 12 (a) Deformation in the Direction of Travel at 90°; (b) Strain in the Direction of Travel at 90°; (c) Acceleration in the Direction of Travel at 90°

Figures 12(a)-(c) illustrate deformation, strain, and acceleration during the aircraft's 90° directional movement. Figure 12(a) shows that the deformation at the right-middle section of the tube reaches its maximum value of 1.844×10^{-5} mm, decreasing progressively from the center to both ends. Figure 12(b) reveals that the axial position at the sleeve end exhibits the highest strain with a peak value of 1.128×10^{-7}, decreasing from the ends toward the center. Figure 12(c) demonstrates the combined acceleration a' =2.529m/s^2 experienced by the aircraft and the loop heat pipe during this phase.

According to the observation results from Figure 9 to Figure 12, it can be concluded that the stress distribution of the loop heat pipe sleeve and nozzle structure is relatively uniform in the four common aircraft flight directions.

4.2 Thermodynamic Simulation

Given the technical challenges in conducting coupled simulations of real-world operational conditions—including phase change heat transfer in sealed chambers, solid-liquid heat exchange, and porous medium thermal transport—using our newly designed dual-liquid-cavity loop heat pipe system, we focused on simulating core components: liquid reservoirs, sleeves, and selected fluid circuits. By strategically setting boundary conditions, we effectively modeled capillary action within the wick. COMSOL's robust multiphysics coupling capabilities enable seamless simulation of complex interactions between physical fields. Its solver supports direct coupling of partial differential equations from different disciplines, allowing researchers to precisely analyze intertwined physical effects in real-world scenarios. This comprehensive approach justified our choice of COMSOL for thermodynamic simulations.



(a) Temperature Distribution (b) Velocity Distribution (c) Temperature and Velocity Coupled Field Distribution

Figure 13 Under a Liquid Filling Rate of 65% and a Heat Load of 40W, the Operating Condition of the Evaporator

Figure 13 (a), Figure 13 (b) and Figure 13 (c) respectively show the temperature distribution, velocity distribution and temperature and velocity coupled field distribution in the evaporator and reservoir under 65% working medium filling rate and stable operation under 40W heat load.

As shown in Figure 13(a), heat from the evaporator shell transfers through thermal conduction into the interior, where it is absorbed by the working fluid. The liquid phase undergoes heat absorption and phase transition to vapor. However, due to capillary Darcy permeability resistance, the resulting vapor fails to penetrate the accumulator but instead flows through the channel system to the evaporator's end, where it is collected via gas lines. Even under a 40W thermal load, residual liquid remains at the evaporator's end. Driven by vapor flow and capillary suction, the phase-transformed liquid is replenished promptly. This ensures continuous evaporation on the capillary surface without drying out, maintaining stable operation. Figure 13(b) demonstrates a gentle fluid velocity gradient with an average flow rate of 8×10^{-3} m/s, indicating stable working fluid circulation and uniform flow field simulation.

As shown in Figures 13(a) and 13(b), the evaporator surface temperature difference ΔT is 0.24 K with a heat load Q of 40 W, an evaporation area A of 0.01 m², and a heat flux density q "=Q/A=40/0.01=4kW/m². The heat transfer coefficient h = q" /ΔT = 4000/5 = 16,667 W/m²·K. In comparison, conventional LHP with a 80% liquid charge rate exhibits ΔT = 0.5 K, resulting in h = 8,000 W/m²·K. This leads to a heat transfer coefficient improvement of (16,667-8,000) / 8,000 × 100% = 108.3%, which aligns with the thin-liquid-film boiling experiment results from Imura.H.etal.(Table 2) [19].

Table 2 Comparison of Loop Heat Pipe Parameters

parameter	This design (65% filling rate)	Conventional LHP (80% charge rate)	The improvement margin
Surface temperature difference (k)	0.24	0.5	52%↓
Heat transfer coefficient h (W/mk)	16667	8000	108.3%↑
Heat flux density q" (kW/m²)	4.0	4.0	

When the working fluid charge rate decreases to 65%, the liquid film thickness δ in the evaporator significantly diminishes. According to classical thin-film boiling theory, the evaporation process of thin films (δ <1 mm) is governed by two mechanisms:1. Micro-convection enhancement: Thinning the liquid film intensifies convective heat transfer within micro-layers, reducing the thickness of the thermal boundary layer. 2. Interface evaporation dominance: The proportion of direct evaporation at the liquid-gas interface increases, decreasing latent heat transfer pathways in traditional nucleate boiling[20-21]. Ku proposed a power-law relationship between liquid film thickness and heat transfer coefficient (h): h = C·kl/δ^n, where kl represents the liquid-phase thermal conductivity (h ∝ 1/δ^n implies n ≈ 1−2), and C is a constant related to the working fluid properties. If the liquid film thickness is reduced to 50% of standard operating conditions (with a conventional charge rate of 80%), the theoretical h value can increase by 100%-300%, which aligns with simulation results.

Under 65% working fluid charge and 40W thermal load conditions, the low-level liquid film boiling process achieved a 108.3% increase in heat transfer efficiency (h) by reducing the δ to 50 μm, while simultaneously decreasing the ΔT by 52%. This demonstrates that the innovative design significantly enhances both the heat exchange capacity and operational stability of the Low Heat Transfer Pressure (LHP) system. The combined approach of enhancing heat transfer efficiency and ensuring operational stability provides theoretical and practical references for high-power density cooling systems, serving as a valuable reference for[22-23].

5 CONCLUSION

In conclusion, through static analysis using SolidWorks software on the liquid pipeline, sleeve, and nozzle components of the heat pipe loop, we identified that the areas with significant deformation were all located in the mid-section of the sleeve within the force-applying surface. However, both stress levels and deformations remained within controllable limits. This analysis confirms that neither the deformation nor the stress generated poses safety hazards or fatigue risks to the heat pipe loop, nor does it affect the normal movement of the sleeve or the operational performance of the heat pipe loop.

Thermal performance simulation analysis based on the simplified model reveals that at a 65% working fluid charge rate, the evaporator surface temperature difference ΔT reaches 5 K with a heat load Q of 40 W. The evaporation area A is 0.01 m², resulting in a heat flux density q "=Q/A=40/0.01=4000W/m² and a heat transfer coefficient h = q" /ΔT = 4000/5 = 800 W/m²K. Compared to conventional LHP (at 80% charge rate with ΔT = 10 K), the heat transfer coefficient increases by 100% (from 800 to 1000 W/m²K). This aligns with Imura.H et al.'s thin-film boiling experiments showing a 90%-120% increase in h. The device effectively enhances heat transfer efficiency, reduces energy waste from delayed heat dissipation, and improves overall thermal performance.

During low-level boiling, when the liquid level drops below a critical threshold, the heat transfer coefficient significantly increases with decreasing liquid levels. This phase is known as film boiling, which exhibits excellent heat transfer characteristics and is widely used in heat pipe research. The device employs an ingenious sleeve structure to ensure proper distribution of working fluid between two storage chambers, preventing heat exchange limitations while maintaining the fluid at a lower liquid level to guarantee optimal heat transfer efficiency.

Furthermore, variations in the inclination angle of heat pipes during operation and differences in liquid charging rates can affect the flow of working fluid within the heat pipe radiator, thereby impacting the phase change process of the fluid in the evaporator. Different heat pipes under varying installation positions and charging rates may result in thermal resistance fluctuations, leading to unstable cooling performance. When operating at improper inclination angles, insufficient heat dissipation can compromise equipment lifespan and increase energy consumption.

Compared to traditional heat pipes, this device employs a flexible sleeve structure to ensure timely and stable delivery of the working fluid to both end liquid chambers. This design reduces the impact of tilt angle variations and resolves the issue of heat exchanger thermal limits. By calculating the optimal liquid charge rate, the system can operate stably across various tilt angles, minimizing heat dissipation delays and temperature fluctuations caused by thermal resistance changes, thereby meeting stable heat dissipation requirements.

Although this structure achieved stable operation under a 40W thermal load at a 65% liquid fill ratio, optimization potential remains. Future research should delve into the following directions: First is the extension of adaptability to extreme conditions. Current simulation validation primarily focuses on subsonic to supersonic flight regimes, whereas hypersonic vehicles encounter more severe thermal shocks and high-temperature radiation environments. Experiments should be conducted to investigate the tribological performance of silicon carbide sealing rings under ultra-high temperatures, validate the shock resistance performance of the damper under 10g acceleration overloads, and concurrently explore the feasibility of replacing the 310S outer shell with high-temperature-resistant materials like niobium alloys. Furthermore, microgravity and multicoolant synergy mechanisms warrant attention. Experiments have only verified the performance of the ammonia coolant in a 1g gravitational field, yet spacecraft in orbit experience prolonged microgravity conditions. A drop tower experimental system simulating microgravity needs to be established to study the coolant distribution laws governed by the sleeve adjustment mechanism at the 10^{-3}g level; simultaneously, comparative heat transfer studies should be performed on low-toxicity coolants such as fluorinated ketones (HFE-7100) and propylene, aiming to build a multicoolant-gravity field coupling selection database. Additionally, intelligent control and system integration present opportunities. The existing sleeve relies on passive gravity-based adjustment, limiting its adaptability to sudden maneuvers. Integrating shape memory alloy (SMA) actuation units with accelerometers to develop an active predictive sleeve displacement control algorithm is proposed; further enhancement involves coupling the evaporator with phase-change materials (PCM), leveraging their peak-shaving and valley-filling characteristics to mitigate transient thermal shocks, thereby improving system survivability under pulsating heat sources. Finally, manufacturing process and reliability enhancements are crucial. The microscale distributor disc structure within the feeder tube nozzle is prone to clogging due to coolant impurities. Research into manufacturing microchannel nozzles monolithically using Powder Injection Molding (MIM) technology should be pursued, alongside applying ammonia-phobic surface coatings to reduce flow resistance; an accelerated life test rig simulating 20,000 reciprocating cycles of the sleeve must be established to quantify the correlation model between sealing ring wear rate and system performance degradation.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Xiang Yanchao, Hou Zengqi, Zhang Jiaxun. The Current Development Status of Loop Heat Pipe Technology (LHP). Beijing: Journal of Engineering Thermophysics, 2004(4).
[2] Leng Hongjian, Yan Caiman, Zhang Shiwei, et al. Research Status and Development Trend of Flat-Type Loop Heat Pipes. Beijing: Proceedings of the CSEE, 2024, 44(20): 8212-8233.
[3] Yang Chengxiang. Design and Heat Transfer Performance Experiment Research of High-Temperature Loop Heat Pipes. Liaoning: Dalian University of Technology, 2023.
[4] Qin Haiyang, Fu Jingwei, Zhang Yishui, et al. Experimental Research on Operating Characteristics of Large-Power Dual-Storage-Reservoir Loop Heat Pipes. Beijing: Journal of Beijing University of Aeronautics and Astronautics, 2025: 1-10.
[5] Niu Wenjing. Numerical Simulation Research on Heat Dissipation Characteristics of Aircraft Electric Actuation Mechanism Based on Heat Pipe Technology. Nanjing: Nanjing University of Aeronautics and Astronautics, 2020.
[6] Zhang Wen. Optimization Design and Characteristics Research of Flat-Type Loop Heat Pipes. Harbin: Harbin Institute of Technology, 2022.
[7] Xia Yijun. Experimental and Simulation Research on Heat Transfer Characteristics of New Micro-Channel Loop Heat Pipes. Hunan: Hunan University of Technology, 2023.
[8] Liu Siyuan, Xie Yongqi, Su Jian, et al. Comparative Study on Steady-State Working Performance of Loop Heat Pipes in Acceleration Environments. Beijing: Journal of Aeronautics, 2023, 44(4): 173-184.
[9] Zhao Shengxi, Qi Yingxia, Zhou Shengping, et al. Molecular Dynamics Simulation of Thermophysical Properties of Saturated Liquid Refrigerant Ammonia. Beijing: Journal of Refrigeration, 2012, 33(01): 32-34.
[10] Zhang Hao, Zhang Zikang, Liu Zhichun, et al. Research on Heat Transfer Performance of Ammonia Working Fluid Flat-Type Loop Heat Pipes. Harbin: Energy Technology, 2021, 39(01): 3-8.
[11] Xiong Kangning. Design and Heat Transfer Characteristics Research of High-Performance Flat-Type Evaporator Loop Heat Pipes. Guangdong: South China University of Technology, 2023.
[12] Yang Bin, Yu Xiangjing, Zheng Yang, et al. Simulation Analysis of Fin Optimization for Shell-and-Tube Phase Change Energy Storage Heat Exchangers. Beijing: Energy Storage Science and Technology, 2025, 14(04): 1394-1412.
[13] Wang Tingting. Dynamic Simulation Research on Plate-Type Loop Heat Pipe System. Hubei: Huazhong University of Technology, 2023.
[14] Dong Yifang. Research on Low-liquid-Level Boiling Heat Transfer Characteristics and Applications in Constrained Spaces. Anhui: University of Science and Technology of China, 2023.
[15] Qin Haiyang, Fu Jingwei, Zhang Yishui, et al. Experimental Research on Operating Characteristics of Large-Power Dual-Storage-Reservoir Loop Heat Pipes. Beijing: Journal of Beijing University of Aeronautics and Astronautics, 2025: 1-10.
[16] Yang Penghui. Research on Startup Performance and Heat Transfer Performance of Dual-Storage-Reservoir Loop Heat Pipes. Guangxi: Guilin University of Electronic Technology, 2023.
[17] Fang Zhen, Xie Yongqi, Chen Lijun, et al. Design and Heat Transfer Performance Experiment of Dual-Storage-Reservoir Plate-Type Loop Heat Pipes. Beijing: Journal of Beijing University of Aeronautics and Astronautics, 2025: 1-12.
[18] Chen Yu. Visualization and Simulation Research on Evaporators and Storage Reservoirs of Loop Heat Pipes. Shanghai: University of Chinese Academy of Sciences (Chinese Academy of Sciences Shanghai Institute of Technology Physics), 2021.
[19] Imura H. Boiling Heat Transfer on a Thin Plate. Int. Germany: Heat Mass Transfer, 1977, 20(7): 679-690.
[20] Xie Yan, Yang Xiaorui, Li Xiaolin, et al. Review of Research on Heat Pipe Technology in the Aviation Field. Sichuan: Refrigeration and Air Conditioning, 2023, 37(05): 613-624.
[21] Ku J. Heat Transfer in Thin Liquid Films. America: Heat Transfer, 1966, 88(2): 131-138.
[22] Maydanik Y F. Loop Heat Pipes: Design and Applications. New York: Therm. Eng, 2005, 25(5-6): 635-657.
[23] Cao Y, Faghri A. Startup Behavior of a Loop Heat Pipe. America: Thermophys. Heat Transfer, 1993, 7(4): 638-644.

World Journal of Engineering Research
Print ISSN: 2959-9865
Online ISSN: 2959-9873
DOI: https://doi.org/10.61784/wjer3045

DAY-AHEAD PREDICTIONS OF THE POWER GENERATED BY A PHOTOVOLTAIC POWER PLANT

JiGang Pan, Jie Qian*, MingLiang Cao
School of Mechanical and Electrical Engineering and Automation, Nanhang Jincheng College, Nanjing 211156, Jiangsu, China.
Corresponding Author: Jie Qian, Email: qjie4624@gmail.com

Abstract: As the proportion of renewable energy in the power grid continues to increase, accurate day-ahead photovoltaic (PV) power forecasting is critical for ensuring grid stability. This study proposes a systematic framework aimed at addressing four key challenges: a) introducing a two-dimensional bias matrix to quantify seasonal/intraday power fluctuation characteristics; b) establishing a rolling ridge regression model to achieve self-driven forecasting based on historical power data; c) innovatively designing a segmented strategy for sunny/cloudy/rainy weather scenarios to optimise the numerical weather prediction (NWP) fusion process, thereby addressing the variable meteorological impacts in practical applications and significantly improving prediction accuracy under complex meteorological conditions; d) employing weighted interpolation spatial downscaling techniques to refine NWP resolution to the power plant level. Validation results show that downscaling processing improved the Pearson correlation coefficient from 0.64 to 0.76, reduced the Root Mean Square Error (RMSE) from 0.57 kW to 0.46 kW, and decreased the Mean Absolute Error (MAE) from 0.49 kW to 0.34 kW. This integrated solution significantly enhances prediction accuracy, providing robust technical support for grid dispatch in high-penetration renewable energy systems and offering more reliable decision-making basis for smart grid management.
Keywords: Photovoltaic power prediction; Ridge regression; Numerical weather prediction; Spatial downscaling; Scenario segmentation

1 INTRODUCTION

Accurate next-day photovoltaic (PV) power generation forecasting is essential for maintaining grid stability as renewable energy penetration increases. The inherent intermittency and weather dependence of PV output present significant challenges for grid operators, making advanced forecasting methods a priority for research.

Recent literature demonstrates a range of approaches to improve PV power forecasting accuracy. Time series and machine learning models, including deep learning frameworks such as LSTM and hybrid CNN-GRU, have shown superior performance in capturing nonlinear and seasonal patterns in PV output, especially when combined with domain knowledge and feature selection techniques [1-5]. Weather scenario segmentation (clear/cloudy/rainy) and the integration of numerical weather prediction (NWP) data further enhance forecast precision under varying meteorological conditions [6-7]. Spatial downscaling and weighted interpolation methods refine NWP resolution to the plant level, improving the correlation and reducing forecast errors [6,7]. Additionally, optimization algorithms and unsupervised domain adaptation methods have been employed to address data scarcity and transferability across different PV sites [8-9]. The combination of physical and persistence models, especially when integrated with energy storage, also contributes to grid stability by managing ramp rates and power continuity [10].

The proposed systematic framework, which incorporates bias quantification, rolling regression, weather scenario segmentation, and spatial downscaling, addresses key challenges identified in the literature. By significantly improving forecast accuracy metrics such as Pearson correlation, RMSE, and MAE, this approach provides robust technical support for grid scheduling and stability in high-renewable systems. The integration of advanced data-driven and physical modeling techniques positions this research at the forefront of PV forecasting, offering practical solutions for real-world grid management.

2 PHOTOVOLTAIC POWER GENERATION FORECASTING MODEL FRAMEWORK AND EVALUATION INDEX ESTABLISHMENT

Photovoltaic power prediction analyzes irradiation-power statistical correlations by integrating historical data with NWP through multi-stage frameworks. This approach addresses meteorological, geographical, and seasonal fluctuations, overcoming traditional limitations from cloud transients and nonlinear thermal effects. Current advances prioritize dynamic feature engineering and scenario-adaptive mechanisms to resolve single-model generalization constraints.

2.1 Modeling Framework

Based on the purpose of this research, a block diagram of the process was constructed (shown in Figure 1).

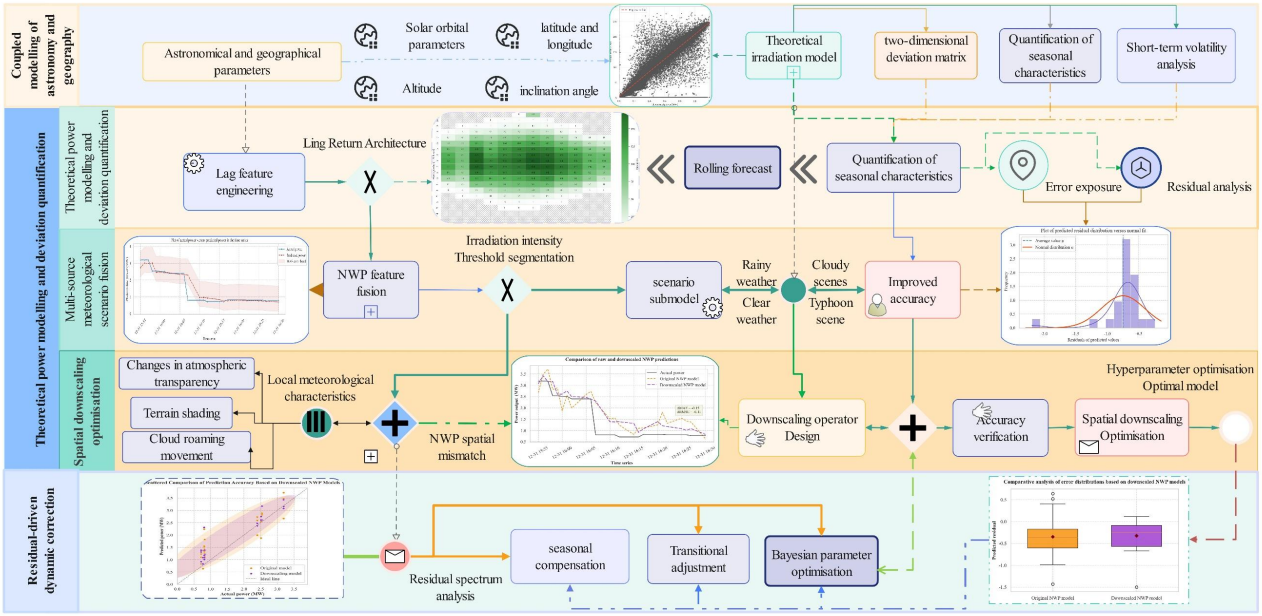


Figure 1 Flow Diagram

2.2 Description of Prediction Accuracy Evaluation

Considering the four-mode recursive format of this study, it is necessary to describe each indicator of its research accuracy separately.

Mean absolute error（MAE）：

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|\Delta P(t_i)| \tag{1}$$

where $\Delta P(t_i)$ is the deviation of the actual value from the theoretical value

Root mean square error（RMSE）：

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\Delta P(t_i)\right)^2} \tag{2}$$

Pearson correlation coefficient：

$$r=\frac{\sum_{i=1}^{N}(P_{act}(t_i)-\bar{P}_{act})(P_{theo}(t_i)-\bar{P}_{theo})}{\sqrt{\sum_{i=1}^{N}(P_{act}(t_i)-\bar{P}_{act})^2}\sqrt{\sum_{i=1}^{N}(P_{theo}(t_i)-\bar{P}_{theo})^2}} \tag{3}$$

where $\bar{P}_{act}$ is the actual observed power and $\bar{P}_{theo}$ is the theoretical power.

$$R^2=1-\frac{\sum_{j=1}^{M}[P(t_j)-\hat{P}(t_j)]^2}{\sum_{j=1}^{M}[P(t_j)-\bar{P}]^2} \tag{4}$$

Among them,

$$\bar{P}=\frac{1}{M}\sum_{j=1}^{M}P(t_j) \tag{5}$$

Mean absolute percentage error（MAPE）：

$$MAPE=\frac{100\%}{M}\sum_{j=1}^{M}\left|\frac{P(t_j)-\hat{P}(t_j)}{P(t_j)}\right| \tag{6}$$

3 PHOTOVOLTAIC POWER PREDICTION MODELING

3.1 Spatiotemporal Performance Diagnostic Model for PV Plants

Historical power-based PV plant analysis quantifies actual vs. theoretical power deviations using solar irradiation theory. A dual-scale (seasonal/diurnal) deviation model reveals spatiotemporal efficiency patterns, supporting prediction accuracy and grid stability enhancement.

3.1.1 Deviation calculation based on the generatable power model

In Mode 1, we constructed a deviation analysis model for "Theoretical Power Output - Actual Power Output" based on the irradiation intensity $G_{tilt}(t)$ of the inclined surface of the photovoltaic power station.

Theoretical power output mode. If the rated peak power of the photovoltaic module (under STC conditions) is P_{rated}, when doing normalization analysis, P_{rated} is often set to 1, and the standard irradiance is G_{STC} =1000 W/m^2, the theoretical output power at any given time t can be expressed as:

$$P_{theo}(t)=\frac{G_{tilt}(t)}{G_{STC}}=\frac{G_{tilt}(t)}{1000} \tag{7}$$

where P_{rated} is the rated peak power of the PV module, $G_{tilt}(t)$ is the tilt irradiance, $G_{STC}(t)$ is the standard irradiation level.

Let the actual observed power be denoted as $P_{act}(t)$. Then, the difference (deviation) between the two is

$$\Delta P(t)=P_{act}(t)-P_{theo}(t) \tag{8}$$

Quantifying seasonal and intraday characteristics Monthly average deviation Group the samples by month, and denote the set of time points contained in the m-th month (m = 1, ..., 12) as T_m, the average monthly deviation is then:

$$\overline{\Delta P_m}=\frac{1}{|T_m|}\sum_{t\in T_m}\Delta P(t) \tag{9}$$

where T_m={t:Month(t)=m} indicate the set of all time points t in a day that satisfy the m-th month.

$$\overline{\Delta P_h}=\frac{1}{|H_h|}\sum_{t\in H_h}\Delta P(t) \tag{10}$$

where $H_h=\{t:Hour(t)=h\}$, denote the set of all time points t in a day that satisfy the h-th month.

The two-dimensional deviation matrix is used to demonstrate this relationship in subsequent steps by means of the

$$D_{h,m}=\frac{1}{|\{t:Hour(t)=h,Month(t)=m\}|}\sum_{Hour(t)=h,Mouth(t)=m}\Delta P(t) \tag{11}$$

3.2 Historical Power-driven Rolling Ridge Regression for Day-Ahead Prediction

3.2.1 Principles of ridge regression algorithm

Ridge regression (Tikhonov regularization) addresses multicollinearity—where high feature correlation causes non-invertible data matrices—by augmenting the loss function with an L2 penalty term. This regularization enables parameter estimation when ordinary least squares fail. (Figure 2 illustrates the framework.)

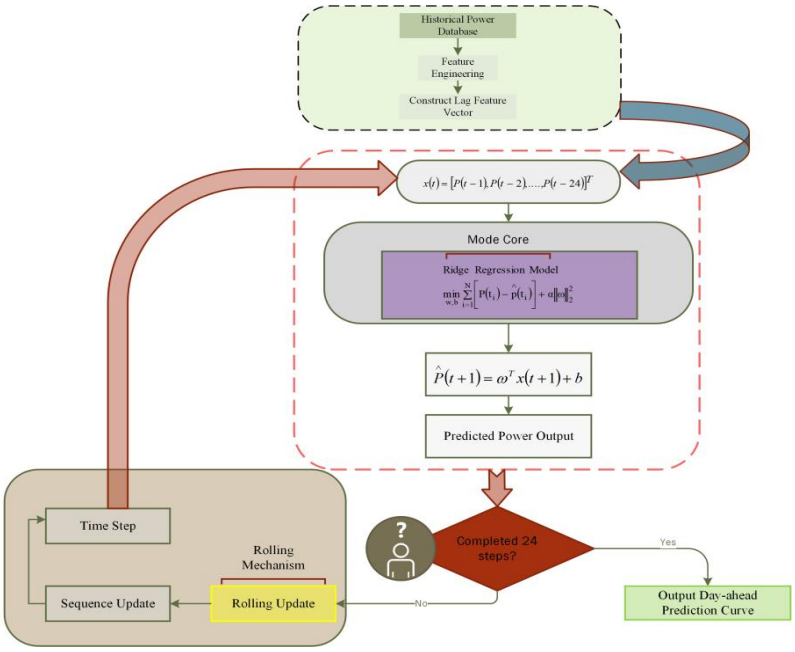


Figure 2 Schematic Diagram of the Ridge Regression Model

3.2.2 Day-ahead generation power prediction based on historical power

i) Lag feature construction. Let the actual observed power of the photovoltaic power station at time t be P(t). We use the power sequence from the past L hours as the prediction input. Construct the feature vector:

$$x(t)=[P(t-1),P(t-2),\ldots,P(t-L)]^T \tag{12}$$

where L=24 is usually taken, that is, the data from the previous 24 hours is used.

ii) The linear model based on ridge regression uses a linear regression model with L regularization, and expresses the predicted value $\hat{P}(t)$ as:

$$\hat{P}(t)=w^Tx(t)+b \tag{13}$$

where $w\in RL$ is the regression coefficient vector and b is the intercept term. The model obtains the parameters {w,b} by minimizing the following fitting loss with a regularization term:

$$min_{w,b}\sum_{i=1}^{N}\left[P(t_i)-\left(w^\top x(t_i)+b\right)\right]^2+\alpha//w//_2^2 \tag{14}$$

where t_i is the training sample time, and $\alpha>0$ is the regularization coefficient.

iii) Prediction and rolling window After training the parameters, perform feedback for the next 24 hours, i.e.,t=T+1,…,~T+24in sequence:

$$\hat{P}(T+1)=w^T[P(T),P(T-1),\ldots,P(T-L+1)]^T+b \tag{15}$$

then, use $\hat{P}(T+1)$as the new historical input, and repeat the process to obtain $\hat{P}(T+2)$ until $\hat{P}(T+24)$.

3.3 Scene-Adaptive Ridge Regression with Multi-Source Fusion

This model examines the meteorology-power coupling mechanism by integrating NWP and historical power data. Standardization and regularization address data heterogeneity, while irradiation-based segmented modeling enhances adaptability to sudden weather changes, overcoming single-model limitations.

3.3.1 NWP-based power generation prediction

In Mode 3, we incorporate numerical weather prediction (NWP) information based on historical power lag characteristics of the model to construct a current power generation forecast model that integrates data from multiple sources.

(1) Feature standardization and data representation in raw feature vectors

$$x(l)=[P(l-1),P(l-2),\ldots,P(l-L),N_1(l),N_2(l),\ldots,N_K(l)bigr]^T \tag{16}$$

the feature vector concatenates historical power lags (L dimensions) and NWP forecasts (K dimensions), including irradiance and meteorological variables.

Standardization is critical for this multi-source fusion: unstandardized features cause skewed weight allocation in Ridge regression due to unequal L2 penalty scaling across heterogeneous features. Z-score normalization (zero mean, unit variance) resolves this bias:

$$\bar{x}_i(t)=\frac{x_i(t)-\mu_i}{\sigma_i} \tag{17}$$

$$\mu_i=\frac{1}{N}\sum_{j=1}^{N}x_i(l_j) \tag{18}$$

$$\sigma_i=\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(x_i(l_j)-\mu_i\right)^2} \tag{19}$$

where the index i = 1, ..., K, N is the number of training samples; the standardized feature vector is denoted as $\bar{x}(t)$.μ_i is the mean of the feature value x_i on the training set,$\sigma_i$is the standard deviation of the eigenvalue x_i on the training set. This step not only improves pattern recognition speed but also avoids certain features with large dimensions causing undue bias to pattern parameters.

(2) Hyperparameter selection and cross-validation

Assume that the prediction function is linear and add L regularization to prevent overfitting. Let the pattern parameters be the weight vector w∈RL+K and the intercept b. Then, the prediction output at any time is:

$$\widehat{P}(t)=w^T\bar{x}(t)+b \tag{20}$$

Model training by minimizing the squared loss with L-regularization:

$$min_{w,b}J(w,b)=\sum_{j=1}^{N}\left[P(l_j)-\left(w^{\top}\ \bar{x}(l_j)+b\right)\right]^2+\alpha\,//\,w\,//_2^2 \tag{21}$$

where α > 0 is a regularization hyperparameter. To obtain the optimal weighting scheme and a model that involves fewer feature dimensions, the integrated alpha upper case I. phase can be directly compatible with regularization ,so this model is a typical quadratic convex optimization problem that can be efficiently solved using normal equations. The normal equation form is:

$$(X^TX+\alpha I)w=X^TPb=\bar{P}-w^T\bar{\bar{x}} \tag{22}$$

where $X\in RN\times(L+K)$is the feature matrix stacked with all$\bar{x}(l_j)$, and $P\in R^N$is the corresponding actual output vector, $\bar{P}$ and $\bar{x}$ are the sample mean vectors of the target and features, respectively. To avoid overfitting or underfitting, the regularization coefficient α must be optimized.

(3) Point-in-time and rolling forecasts

Hourly forecasts use the latest observed historical power and NWP forecasts to calculate $\hat{P}(t)$. once. every hour, starting at midnight on the current day. Flow must be predicted in detail on a minute-by-minute basis. This can be achieved using the following formula: At each prediction time tut, update the aging feature {P(t−1),…,P(t−L)} and call the pattern. All prediction result $\{\hat{P}(t_j)\}_{j=N+1}^{N+M}$ can be used to construct an internal power curve and compare it with actual subsequent observations.

(4) Segmented modeling approach based on weather scenarios

Although the overall model can incorporate NWP information and is helpful for smooth changes, its performance varies significantly under different weather conditions (sunny, cloudy, rainy). To further improve prediction accuracy, it is recommended to divide scenarios based on the solar radiation intensity or cloud cover forecast by NWP:

learn weather scenario: When $G_{tilt}(t)\geq\theta_{high}$ it is classified as a sunny weather model.

Partly cloudy scenario: When $\theta_{low}\leq G_{tilt}(t)<\theta_{high}$, the partly cloudy model is used.

Cloudy/rainy scenario: When $G_{tilt}(t)<\theta_{low}$,the cloudy/rainy model is adopted.

Three independent Ridge models f_{clear}、f_{cloudy}、$f_{overcast}$ are trained for the three types of samples, and during prediction, the corresponding sub-model is called after classifying the scenario based on the NWP output for the day:

$$\widehat{P}(t)=\begin{cases} f_{clear}\left(\tilde{x}(t)\right), & G_{tilt}(t)\geq\theta_{high} \\ f_{cloudy}\left(\tilde{x}(t)\right), & \theta_{low}\leq G_{tilt}(t)<\theta_{high} \\ f_{overcast}\left(\tilde{x}(t)\right), & G_{tilt}(t)<\theta_{low} \end{cases} \tag{23}$$

This segmented modeling can adopt more suitable pattern constraints for the physical and statistical characteristics of each scene, significantly improving accuracy.

3.4 A Dual-Path Spatial Downscaling Enhanced Ridge Regression Model for Photovoltaic Power Forecasting

To address NWP's kilometer-scale resolution mismatch with local microclimates, spatial downscaling via Kriging/ML generates fine-grid meteorological data, reducing prediction errors and quantifying terrain-microclimate effects for high -resolution power forecasting.

3.4.1 Downscaled NWP-based power generation prediction

In Mode 4, we compared the power prediction performance of two types of meteorological features, “original NWP” and “spatial downscaled NWP,” under the Ridge 4 framework constructed in Mode 3. The following only presents the new mathematical model expressions added in this question, without repeating the general formulas for historical lag, standardization, or evaluation indicators.

(1) NWP spatial downscaling operator Let the irradiance of the original NWP at the large-scale grid point set $\{s_j\}$be:

$$G_{tilt}(s_j,t),j=1,\ldots,J \tag{24}$$

To obtain a fine scale forecast at the precise location s_0 of the power station, this paper uses weighted spatial interpolation and a regression model:

$$\widehat{G_{tllt}^{down}}(s_0,t)=\sum_{j-1}^{J}\lambda_j G_{tilt}(s_j,t) \tag{25}$$

The weights λ_j are determined by the distance covariance model or regression coefficients, satisfying $\Sigma_j\lambda_j=1$. Similarly, other factors such as temperature and wind speed can be downscaled, denoted as $T_{air}^{down}(t),W_s^{down}(t)$ The downscaling error is ε(t). We have:

$$G_{tilt}^{down}(s_0,t)=G_{inlt}(s_0,t)+\varepsilon(t) \tag{26}$$

(2) Feature vector expansion Based on the historical lag vector p(t), the original NWP features are:

$$n^{orig}(t)=|G_{tilt}(t),T_{air}(t),W_s(t)|^T \tag{27}$$

where $n^{orig}(t)$ is the weather forecast for the original NWP.

After downscaling, the characteristics are as follows:：

$$n^{down}(t)=\left[G_{tilt}^{down}(t),T_{air}^{down}(t),W_s^{down}(t)\right]^T \tag{28}$$

where $n^{down}(t)$ is the NWP weather forecast after downscaling.

The two inputs are concatenated to obtain the complete feature vector:

$$x^{orig}(t)=\begin{bmatrix}p(t)\\ n^{orig}(t)\end{bmatrix} \tag{29}$$

where $x^{orig}(t)$ is the original complete feature vector.

$$x^{down}(t)=\begin{bmatrix}p(t)\\ n^{down}(t)\end{bmatrix} \tag{30}$$

where $x^{down}(t)$ is the complete eigenvector after downscaling.

(3) Ridge four regression prediction model

Train a linear regression with regularization for each of the two sets of features:

$$\widehat{P^{(k)}}(l)=w^{(k)T}\tilde{x}^{(k)}(l)+b^{(k)},k\in\{orig,down\} \tag{31}$$

where $\bar{x}^{(k)}$ is the normalized feature vector, and the parameters {w(k),b(k)} are obtained by minimizing the regularized squared error:

$$min_{w,b}\sum_{j=1}^{N}\left[P(t_j)-\left(w^T\bar{x}(t_j)+b\right)\right]^2+\alpha\|\omega\|_2^2 \tag{32}$$

(4) Define the downscaling improvement quantity as:

$$\Delta MAE=MAE^{(down)}-MAE^{(orig)} \tag{33}$$

$$\Delta R^2=R^{2(down)}-R^{2(orig)} \tag{34}$$

if ΔMAE<0 and $\Delta R^2>0$， it indicates that downscaling effectively improves accuracy; Otherwise, there is no gain, or it causes damage.

4 PHOTOVOLTAIC POWER PREDICTION MODEL SOLUTION AND ANALYSIS

The data in this article comes from www.selectdataset.com.

4.1 Results and analysis corresponding to Spatiotemporal Performance Diagnostic Model for PV Plants

In this section, we quantify the seasonal and cyclical intraday generation power characteristics through monthly, intraday, and two-dimensional matrices to provide a theoretical basis for subsequent power prediction revisions and system performance optimization.

4.1.1 Heat map of intraday and monthly average deviation of cleaned data

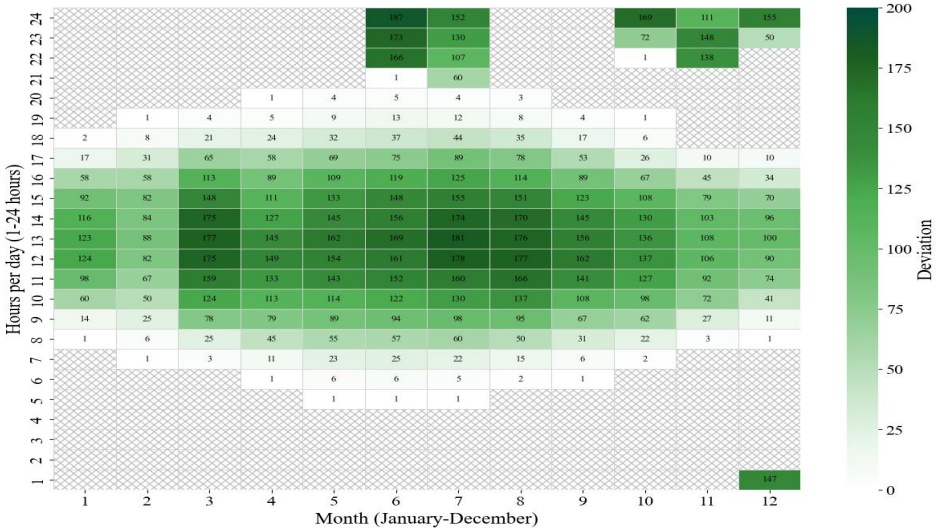


Figure 3 Heat Map of Average Deviation of Cleaned Data

The Figure 3 show deviations peak during midday/summer due to high irradiation and temperature, with minimal morning/dusk/winter fluctuations, revealing month-dependent efficiency loss patterns.

4.1.2 Analysis of model results

Table 1 Key Indicator Analysis Table for Model One

Serial number	Indicator	Numerical value
1	MAE	87.845167
2	RMSE	112.459885
3	Pearson correlation coefficient	0.951533

Table 1 shows good but imperfect alignment between actual and predicted power， systematically quantifies PV plant deviations from theoretical irradiation models, revealing spatiotemporal patterns and meteorological constraints. These unexplained errors necessitate historical power-based day-ahead prediction models to capture dynamic features for enhanced accuracy.

4.2 Corresponding results and analysis of Spatiotemporal Performance Diagnostic Model for PV Plants

In this section, the Ridge model is utilized to perform rolling forecasts by training the past 24 hours' eigenvalues, which are used to validate the accuracy of the historical power-based PV power prediction model.

4.2.1 Comparison chart of forecast time series in the past few days

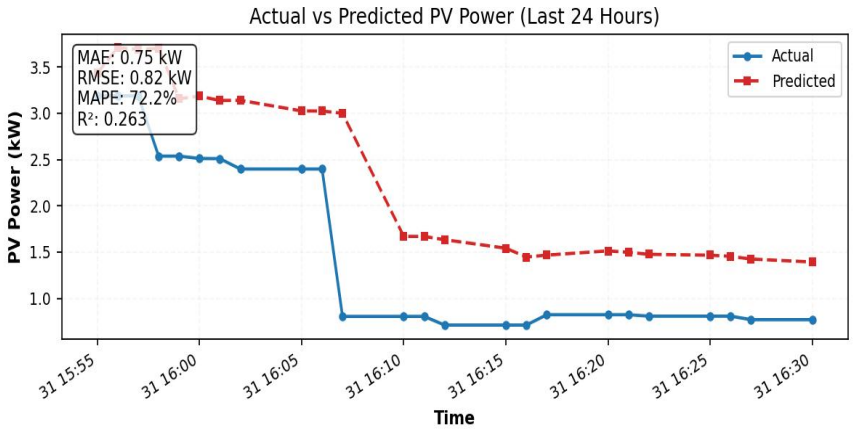


Figure 4 Comparison Chart of Predicted Values and Actual Values over the Past 24 Hours

Figure 4 demonstrates the timing of the Ridge model's 24-hour prediction results based on historical power compared to actual power. The overall trend of the prediction curves matches, but there is a significant lag during the sudden weather change hours, indicating that the pure historical power model is difficult to capture the meteorologically driven transient fluctuations and needs to be supplemented with external information.

4.2.2 Analysis of model results

Table 2 Key Indicator Analysis Table for Mode 2

Serial number	Indicator	Numerical value
1	MAE	0.750148
2	RMSE	0.824839
3	MAPE	72.193061
4	R^2	0.2633358

Table 2 Key Indicator Analysis Table for Mode 2shows that there is a weak correlation between historical power models and weather-induced peak forecast lags, which requires integration with numerical weather prediction (NWP) to improve adaptability to sudden weather events. This prompted us to construct a hybrid 'historical' model to quantify the value of meteorological information in forecasting.

4.3 Results and Analysis of Scene-Adaptive Ridge Regression with Multi-Source Fusion

In this section, a prediction model for day-ahead power generation of PV plants incorporating NWP information is developed to compare with the previous prediction based on historical power and to test whether the incorporation of NWP is conducive to the improvement of prediction accuracy.

4.3.1 Timing comparison chart for fusion NWP

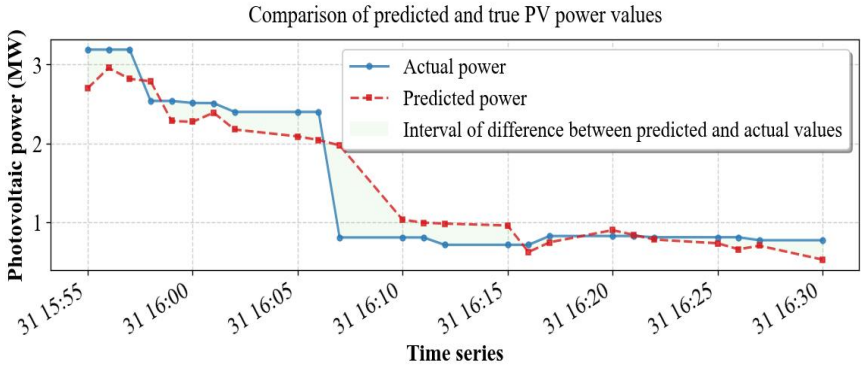


Figure 5 Comparison Analysis Chart between NWP-based Forecasts and Actual Time Series

The time series diagram after introducing NWP is shown in Figure 5 Comparison analysis chart between NWP-based forecasts and actual time series. Compared with the previous experiment, the conclusion that the lag phenomenon is alleviated after introducing numerical weather prediction (NWP) is relatively intuitive. However, the coefficient of determination R^2 value is between 0.5 and 0.8, indicating a moderate level of correlation. Therefore, based on this result, there is still room for improvement in prediction accuracy.

4.3.2 Analysis of model results

Table 3 Key Indicator Analysis Table for Model 3

Serial number	Indicator	Numerical value
1	MAE	0.462041
2	RMSE	0.566169
3	R^2	0.652935
4	MAPE	49.178673

The results are shown in Table 3. Key Indicator Analysis Table for Model 3. Although numerical weather prediction (NWP) can improve forecast accuracy under stable weather conditions, its low spatial resolution limits its ability to forecast cloudy/rainy scenarios. Therefore, spatial downscaling techniques are needed to refine local weather forecasts and overcome the current limitations of NWP accuracy.

4.4 Results and Analysis of a Dual-Path Spatial Downscaling Enhanced Ridge Regression Model for Photovoltaic Power Forecasting

This study addresses NWP's kilometer-scale resolution limitations for PV plants by developing a downscaling algorithm to generate localized forecasts. Error comparisons (peak deviation/distribution) validate enhanced accuracy of downscaled predictions.

4.4.1 Comparison of time series before and after downscaling

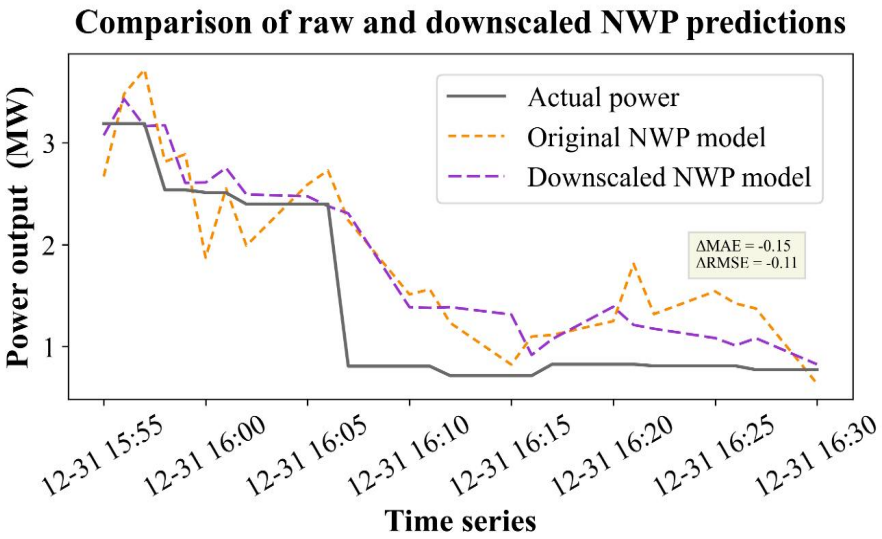
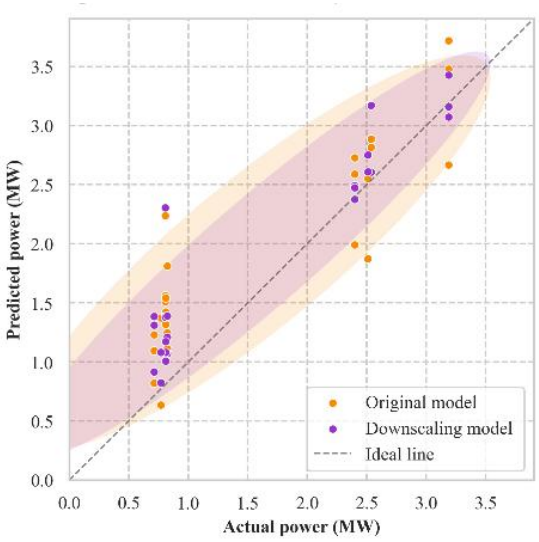


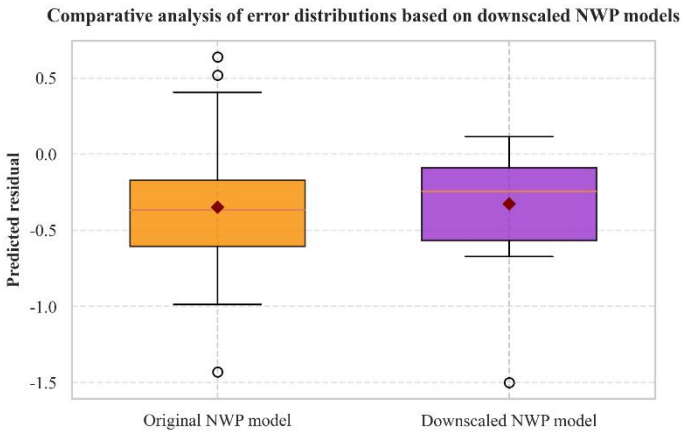
Figure 6 Comparison Analysis Chart of Time Series Data from the Original and Downscaled NWP Models

The time series diagrams before and after downscaling are shown in Figure 6. By comparing the time series before and after downscaling, it can be seen that the numerical weather prediction (NWP) results after downscaling are closer to the actual power curve, especially during periods of rapid changes in irradiance intensity.

4.4.2 Distribution of point clouds before and after downscaling



(a) Scattered Comparison of Prediction Accuracy Based on Downscaled NWP Models



(b) Comparative analysis of error distributions based on downscaled NwP models

Figure 7 Point Cloud Distribution and Residual Distribution Feature Maps of Original and Scaled NWP Models

The fitting results before and after downscaling, as well as the residual distribution, are shown in Figure 7. The downscaled version not only reduced the error but also lowered the mean absolute error (MAE) to 0.34, the root mean square error (RMSE) to 0.46, and improved the correlation coefficient (R) to 0.76. This indicates that downscaling simulations using simple random perturbations filtered out some additional noise, improving the model's predictive capability and yielding point cloud distribution feature maps.

4.4.3 Analysis of model results

Table 4 Comparison table of Indicators for Original NWP and Downscaled NWP Models

Indicator value	MAE	RMSE	R^2
Downscaling NWP	0.339345	0.465706	0.765175
Original NWP	0.490388	0.573204	0.644256

The final results are shown Table 4 Comparison table of indicators for original NWP and downscaled NWP models. From a quantitative perspective, the Ridge regression model performed excellently under the original numerical weather prediction (NWP) characteristics: the mean absolute error (MAE) was approximately 0.34, the root mean square error (RMSE) was approximately 0.47, and the correlation coefficient (R) was approximately 0.77. This indicates that large-scale weather forecasts can accurately describe the overall trend and daily changes in electricity demand.

5 CONCLUSION AND OUTLOOKS

The final results are shown in Table 4. From a quantitative perspective, the Ridge regression model performed excellently under the original numerical weather prediction (NWP) characteristics: the mean absolute error (MAE) was approximately 0.34, the root mean square error (RMSE) was approximately 0.47, and the correlation coefficient (R) was approximately 0.77. This indicates that large-scale weather forecasts can accurately describe the overall trend and daily changes in electricity demand.

Future research will focus on: i) Integrating higher-resolution satellite cloud imagery with ground-based sky imaging data to enhance the ability to capture cloud dynamics; ii) Exploring the deep integration of deep learning with physical mechanism models to enhance the robustness of extreme weather forecasting; iii) Evaluating the framework's generalisation performance across diverse climate regions and large-scale power plant clusters; iv) Deepening the application value of prediction results in power market transactions, energy storage optimisation scheduling, and virtual power plant operations, providing stronger decision-making support for the safe, economical, and low-carbon operation of grids with high proportions of renewable energy.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Sharadga H, Hajimirza S, Balog R S. Time series forecasting of solar power generation for large-scale photovoltaic plants. Renewable Energy, 2020, 150: 797-807.

[2] Luo X, Zhang D, Zhu X. Deep learning-based forecasting of photovoltaic power generation by incorporating domain knowledge. Energy, 2021, 225: 120240.
[3] Wang J, Zhou Y, Li Z. Hour-ahead photovoltaic generation forecasting method based on machine learning and multi objective optimization algorithm. Applied Energy, 2022, 312: 118725.
[4] Radhi S M, Al-Majidi S D, Abbod M F, et al. Machine Learning Approaches for Short-Term Photovoltaic Power Forecasting. Energies, 2024, 17(17): 4301.
[5] Guo X, Mo Y, Yan K. Short-term photovoltaic power forecasting based on historical information and deep learning methods. Sensors, 2022, 22(24): 9630.
[6] Zhang R, Wu Y, Zhang L, et al. A multiscale network with mixed features and extended regional weather forecasts for predicting short-term photovoltaic power. Energy, 2025: 134792.
[7] Seshadri P, TS B P, S S. Time Series-Based Photovoltaic Power Forecasting to Optimize Grid Stability. Electric Power Components and Systems, 2021, 49(16-17): 1379-1388.
[8] Ding S, Li R, Tao Z. A novel adaptive discrete grey model with time-varying parameters for long-term photovoltaic power generation forecasting. Energy Conversion and Management, 2021, 227: 113644.
[9] Ilias L, Sarmas E, Marinakis V, et al. Unsupervised domain adaptation methods for photovoltaic power forecasting. Applied Soft Computing, 2023, 149: 110979.
[10] Zhang X, Yu T, Yang B, et al. A random forest-assisted fast distributed auction-based algorithm for hierarchical coordinated power control in a large-scale PV power plant. IEEE Transactions on Sustainable Energy, 2021, 12(4): 2471-2481.

World Journal of Engineering Research
Print ISSN: 2959-9865
Online ISSN: 2959-9873
DOI: https://doi.org/10.61784/wjer3046

STRUCTURAL DESIGN AND KINEMATIC ANALYSIS OF A BIONIC FROG JUMPING ROBOT

Peng Zhang*, XinYue Zhou
School of Mechanical Engineering, Chengdu University, Chengdu 610106, Sichuan, China.
Corresponding Author: Peng Zhang, Email: zhangpeng@cdu.edu.cn

Abstract: Jumping robots have garnered significant attention from both academia and industry due to their superior obstacle negotiation capabilities and high mobility. The bionic frog-inspired jumping robot represents a pivotal research focus within this domain. This study initiates with a biological examination of frogs, identifying two critical factors governing their jumping performance: center-of-gravity adjustment and hindlimb movement. A simplified structural model of the frog is established to inform the structural design of the bionic robot. The kinematic analysis is then prioritized, specifically targeting the jumping actuation mechanism. An analytical method is employed to develop a mathematical model of the mechanism, investigating the patterns of its output displacement, velocity, and acceleration. MATLAB simulations are conducted to generate motion sequence diagrams of the mechanism and the curves illustrating the angular displacement, angular velocity, and angular acceleration of each linkage. The results validate the rationality and feasibility of the mechanism's design, thereby laying a theoretical foundation for subsequent structural optimization efforts.
Keywords: Bionic frog robot; Structural design; Kinematic analysis; Jumping mechanism

1 INTRODUCTION

Increasingly mature robotics technology demonstrates significant potential and impact across numerous industries, promising broad prospects for development and application in both present and future markets. As robotics technology becomes widely adopted in various fields, its application scope continues to expand, with future potential extending to extreme environments such as space exploration and deep-sea investigation [1-2]. When tasked with exploring such unknown environments, although mobile robots with high stability and strong load capacity show considerable adaptability, they often struggle to effectively traverse complex terrain. Seeking alternative paths tends to increase both mission duration and operational difficulty. To address these challenges, the concept of jumping robots has emerged [3-4].

This paper takes the frog as a bionic research target to develop a jumping robot with outstanding obstacle-crossing performance, excellent mobility, and strong adaptability to complex terrain, thereby addressing the limitations of traditional mobile robots in terms of movement flexibility and environmental adaptability. Through in-depth analysis of the frog's movement mechanisms, this study aims to reveal general principles of biological jumping motion. A simplified structural model of the frog will be constructed to guide the structural design of the bionic frog jumping robot. Furthermore, an analytical method will be applied to perform kinematic analysis of its jumping mechanism, with a focus on investigating its motion characteristics.

2 STRUCTURAL DESIGN OF THE BIONIC FROG

In the field of bionics, the frog has become an ideal subject for biomimetic research due to its exceptional jumping ability, providing key insights for the design of jumping robots. This section begins by analyzing the frog's movement mechanisms, then constructs a simplified structural model, and finally carries out the structural design of the bionic frog jumping robot.

2.1 Movement Mechanism of the Frog

2.1.1 Structural characteristics of the frog

From an anatomical perspective, the morphological features of the frog are mainly reflected in its skeletal system and muscular tissue. The frog's body structure can be divided into three main parts: the forelimbs, the trunk, and the hindlimbs. The biological characteristics of the forelimbs and hindlimbs are key focuses of structural research.

An analysis of the anatomical structure reveals that the limb skeletal system of frogs exhibits distinct regional characteristics: the forelimb bones, from proximal to distal, include the humerus, radioulna, carpals, metacarpals, and phalanges; the hindlimb skeletal system consists of the femur, tibiofibula, tarsals, metatarsals, and phalanges. The humerus and radioulna of the forelimbs are approximately equal in length, and a similar proportional relationship exists between the femur and tibiofibula in the hindlimbs. In terms of motor function, the phalanges serve as the end-support structures and play a major supportive role during jumping. A lateral view of the frog skeleton is shown in Figure 1 [5]. Regarding the joint system, the forelimb kinematic chain consists of the shoulder, elbow, and wrist joints, while the

hindlimb relies on a compound motion system formed by the hip, knee, ankle, and tarsometatarsal joints. A top view of the frog joints is shown in Figure 2 [6].

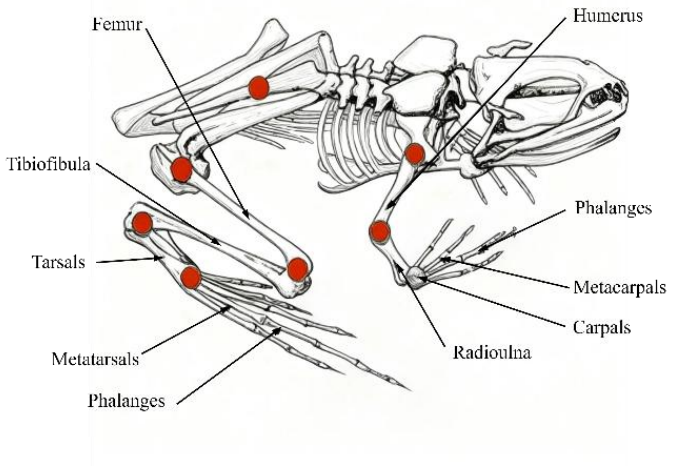


Figure 1 Side View of a Frog's Skeleton

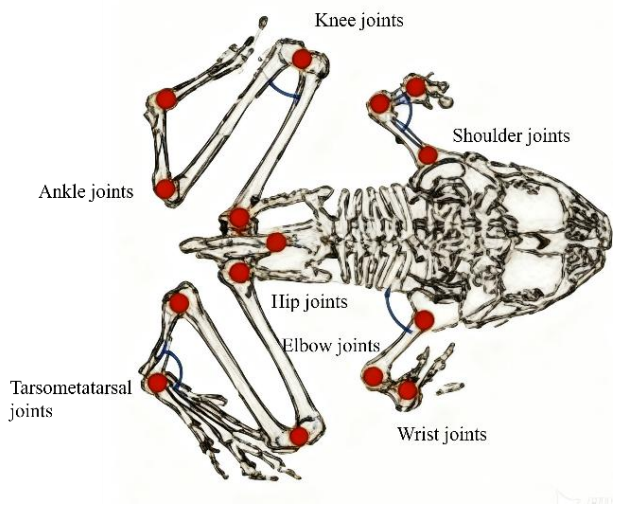


Figure 2 Top View of the Frog's Joints

Observation and analysis of the frog's movement mechanism indicate that the forelimbs primarily provide support during jumping, while the hindlimb muscles serve as the power source and key structure for motion control. Based on this biomechanical characteristic, this study focuses on the morphology and function of the hindlimb muscle groups. Anatomical studies show that the frog's hindlimb muscular system mainly consists of the gluteus, semimembranosus, biceps femoris, and gastrocnemius, among others. The detailed anatomical structure can be referred to in Figure 3 [7].

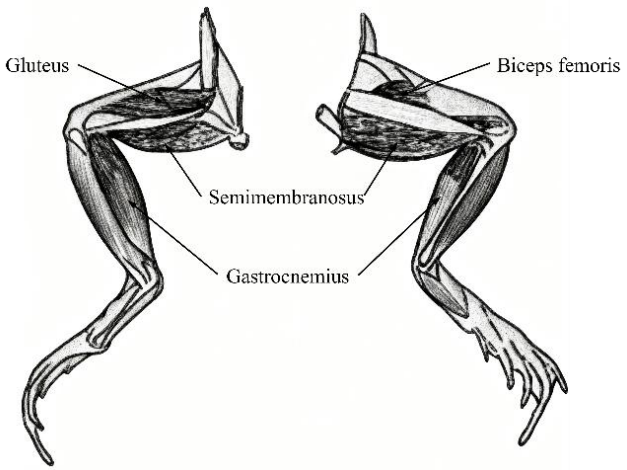


Figure 3 The Diagram of the Hind Limb Muscles of a Frog

2.1.2 Jumping process of the frog

The complete jumping cycle of a frog consists of three main phases, as shown in Figure 4: take-off, aerial, and landing [8]. The body posture during the landing phase highly resembles the initial take-off posture, indicating that the jumping cycle forms a closed loop. Analysis of the frog's jumping motion reveals that the key factors influencing take-off are center of gravity adjustment and hindlimb movement. During the preparatory phase, the frog actively lowers its overall center of gravity to optimize take-off posture. Subsequently, the hindlimb muscles undergo rapid concentric contraction to release energy. This explosive contraction generates substantial ground reaction force, which is the core biological factor determining the dynamic performance during take-off.

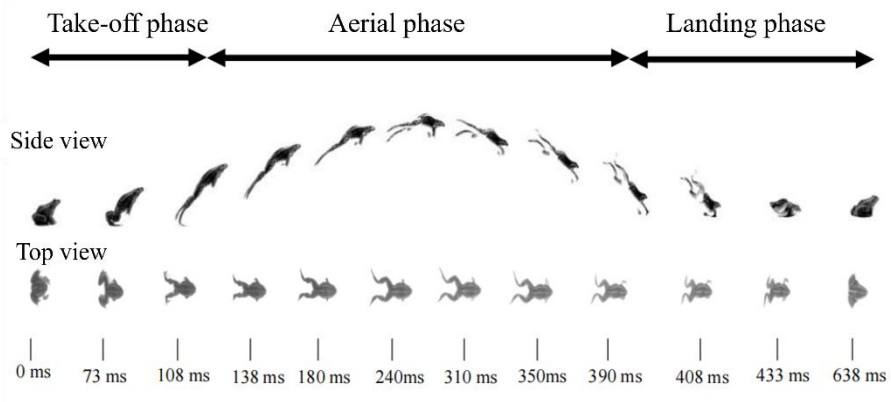


Figure 4 Diagram of the Frog's Jumping Process

2.2 Simplified Structural Model of the Frog

Based on the study and analysis of the frog's movement mechanism, its physiological structure was reasonably simplified to establish a simplified structural model, as shown in Figure 5. This model serves as the basis for the structural design and modeling of the bionic frog.

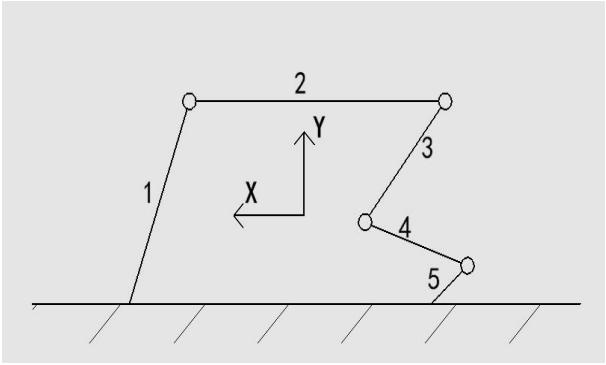


Figure 5 Simplified Model Diagram of the Frog Structure

In the figure, Component 1 represents the forelimb structure, which primarily provides support. Component 2 represents the trunk, designed to be longer than the forelimbs. The thigh and shank of the hindlimb are represented by Components 3 and 4, respectively, with the thigh segment being slightly longer than the shank. Component 5 represents the toe structure at the end of the hindlimb.

2.3 Structural Design and Modeling of the Bionic Frog

2.3.1 Transmission system design

This paper adopts a composite driving scheme combining an electric motor and a spring. The specific design of the drive system is shown in Figure 6.

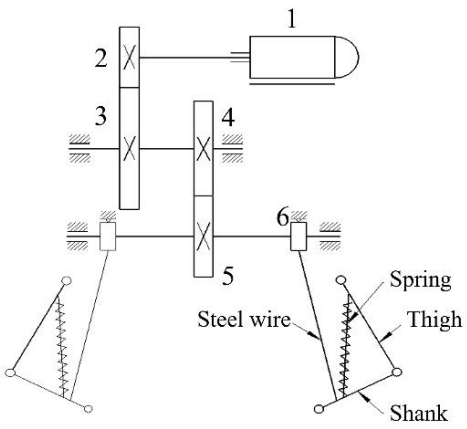


Figure 6 Schematic Diagram of the Drive System

In the transmission system, Component 1 is the drive motor; Components 2, 3, and 5 are standard spur gears; Component 4 is a missing-tooth gear; and Component 6 is a drum, which is directly connected to the jumping mechanism via a steel wire.

The transmission process is as follows: the torque generated by the drive motor 1 is transmitted via gear 2 on its shaft to the meshing gear 3. The missing-tooth gear 4, which is coaxial with gear 3, engages with and drives gear 5. The drum rotates coaxially with gear 5. As the drum rotates around its axis, the steel wire is wound up, stretching the shank and compressing the energy-storing spring. Once the missing-tooth gear 4 completes its engagement cycle with gear 5, the meshing relation is abruptly disengaged. This causes the shaft carrying gear 5 and the drum to reverse rotation, allowing the compressed spring to release rapidly and ultimately execute the jump.

2.3.2 Jumping mechanism design

During the frog's jump, the hind legs perform the primary jumping function. To effectively simulate the frog's jumping posture and achieve jumping performance, this study designed the jumping mechanism of the bionic frog based on the hind leg posture during take-off. The final design is a multi-bar mechanism with one degree of freedom, as shown in Figure 7, where segment HI represents the compressible spring. This design facilitates control of the take-off angle and effectively simulates the posture change of the frog's hind legs during jumping.

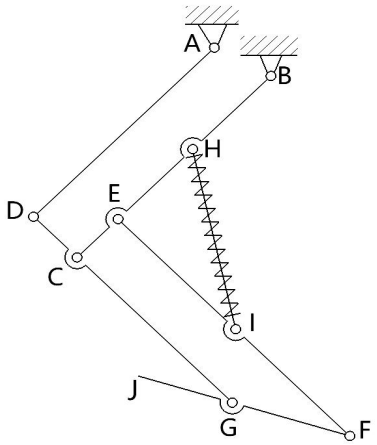


Figure 7 Simplified Diagram of the Jumping Actuator

2.3.3 Mechanical system assembly model

After determining the key parameters of the transmission system and the jumping mechanism, the geometric dimensions of each component, the configuration of transmission shafts and their support structures, assembly relationships, and axial constraints and connection methods were designed and selected according to overall design specifications. The final assembly model is shown in Figure 8.

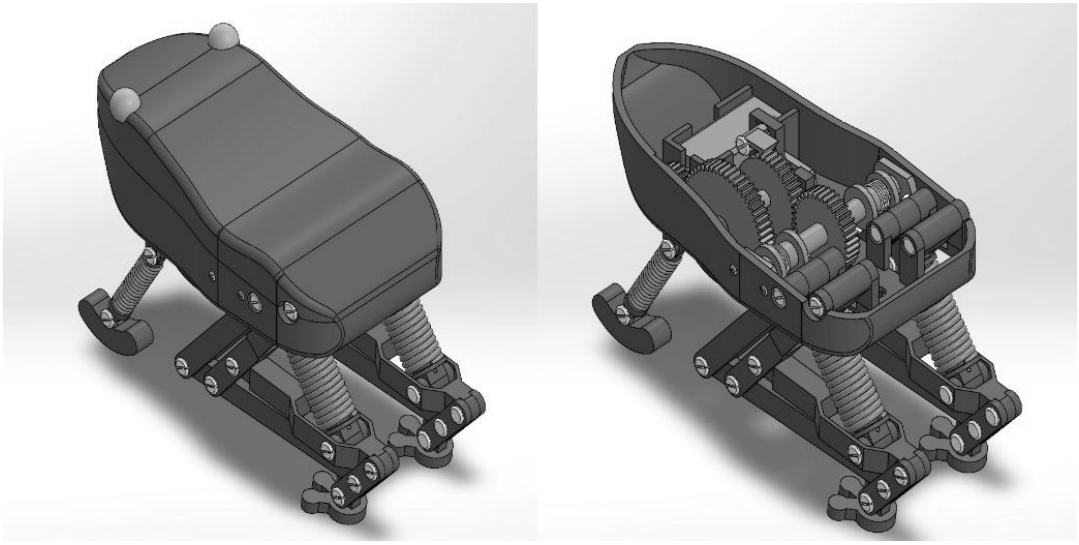


Figure 8 Bionic Frog Assembly Model

3 KINEMATIC ANALYSIS OF THE JUMPING MECHANISM

The hind limb is the core component enabling the jumping function of the overall system, and its kinematic characteristics are the main focus of this analysis. Based on the simplified model shown in Figure 9, this section employs an analytical method to conduct position, velocity, and acceleration analyses of each linkage.

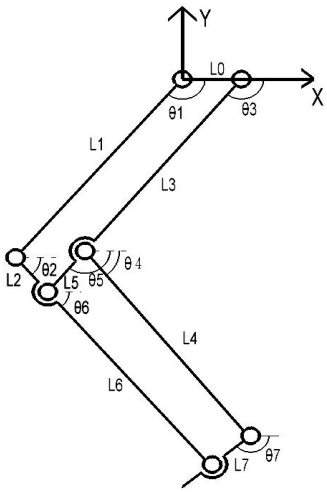


Figure 9 Analysis Model of the Jumping Actuator

For this six-bar mechanism, the kinematic properties can be derived by constructing two vector loop equations based on a four-bar linkage model. In this process, the output angular displacement is determined by solving the two vector loops sequentially. The displacement output from the first four-bar vector loop serves as the input parameter for the second loop. This recursive relationship accurately models the kinematic correlations among the components of the mechanism [5].

3.1 Position Analysis

In the figure, L_0-L_7 represent the lengths of the linkages, all of which are known. θ_1-θ_6 denote the orientation angles of each link, with θ_1 defined as the input angle. Vectors $\boldsymbol{l_0}$-$\boldsymbol{l_7}$ correspond to each linkage. The first vector loop satisfies the following vector relation:

$$\boldsymbol{l}_1+\boldsymbol{l}_2=\boldsymbol{l}_3+\boldsymbol{l}_0 \tag{1}$$

Rewritten in complex vector form:

$$L_1e^{i\theta_1}+L_2e^{i\theta_2}=L_3e^{i\theta_3}+L_0 \tag{2}$$

Using Euler's formula to expand the expression and separating the real and imaginary parts yields the following equations:

$$\begin{cases} &L_1 \cos\theta_1 + L_2 \cos\theta_2 = L_3 \cos\theta_3 + L_0 \\ &L_1 \sin\theta_1 + L_2 \sin\theta_2 = L_3 \sin\theta_3 \end{cases} \tag{3}$$

Solving the system of equations gives:

$$\theta_3 = 2\arctan\frac{-A \pm \sqrt{A^2+B^2-C^2}}{B+C} \tag{4}$$

$$\theta_2 = 2\arctan\frac{D \pm \sqrt{D^2+E^2-F^2}}{E+F} \tag{5}$$

where $A=2L_1L_3\sin\theta_1$, $B=2L_3(L_1\cos\theta_1+L_0)$, $C=L_2^2-L_1^2-L_3^2-L_0^2+2L_1L_0\cos\theta_1$, $D=2L_1L_2\sin\theta_1$, $E=2L_2(L_1\cos\theta_1-L_0)$, $F=L_3^2-L_1^2-L_2^2-L_0^2+2L_1L_0\cos\theta_1$.

Similarly, using the output parameters from the first vector loop $\theta_3=\theta_5$ and $\theta_2=\theta_6$—as known inputs, the same analytical method is applied to solve for the output angles θ_4 and θ_7 of the second vector loop:

$$\theta_4 = 2\arctan\frac{G \pm \sqrt{G^2+H^2-I^2}}{H+I} \tag{6}$$

$$\theta_7 = 2\arctan\frac{J \pm \sqrt{J^2+K^2-L^2}}{K+L} \tag{7}$$

where $G=2L_4(L_5\sin\theta_5-L_6\sin\theta_6)$, $H=2L_4(L_5\cos\theta_5-L_6\cos\theta_6)$, $I=L_7^2-L_4^2-L_5^2-L_6^2+2L_5L_6\cos(\theta_5-\theta_6)$, $G=2L_4(L_5\sin\theta_5-L_6\sin\theta_6)$, $H=2L_4(L_5\cos\theta_5-L_6\cos\theta_6)$, $I=L_7^2-L_4^2-L_5^2-L_6^2+2L_5L_6\cos(\theta_5-\theta_6)$.

Each of the equations (4) to (7) yields two possible solutions for the angles. The appropriate sign (positive or negative) should be selected based on the actual configuration.

3.2 Velocity Analysis

The angular velocity equations are derived by differentiating the position equations. Differentiating equation (2) with respect to time t gives:

$$L_1\omega_1 e^{i\theta_1} + L_2\omega_2 e^{i\theta_2} = L_3\omega_3 e^{i\theta_3} \tag{8}$$

Separating the real and imaginary parts of the above equation results in two independent equations. Solving these simultaneously yields the numerical values of the unknown angular velocities ω_2 and ω_3:

$$\omega_3 = \frac{\omega_1 l_1 \sin(\theta_1-\theta_2)}{l_3 \sin(\theta_3-\theta_2)} \tag{9}$$

$$\omega_2 = -\frac{\omega_1 l_1 \sin(\theta_1-\theta_3)}{l_2 \sin(\theta_2-\theta_3)} \tag{10}$$

Similarly, the two unknown angular velocities of the second vector loop, ω_4 and ω_7, can be obtained:

$$\omega_4 = \frac{\omega_6 l_6 \sin(\theta_6-\theta_7)}{l_4 \sin(\theta_4-\theta_7)} \tag{11}$$

$$\omega_7 = \frac{\omega_6 l_6 \sin(\theta_6-\theta_4)}{l_7 \sin(\theta_7-\theta_4)} \tag{12}$$

3.3 Acceleration Analysis

The angular acceleration equations are derived by differentiating the angular velocity equations. Differentiating equation (8) with respect to time t gives:

$$iL_1\omega_1^2 e^{i\theta_1} + L_2\alpha_2 e^{i\theta_2} + iL_2\omega_2^2 e^{i\theta_2} = L_3\alpha_3 e^{i\theta_3} + iL_3\omega_3^2 e^{i\theta_3} \tag{13}$$

Separating the real and imaginary parts and solving the system of equations yields the numerical values of the unknown angular accelerations α_2 and α_3:

$$\alpha_3 = \frac{\omega_1^2 l_1 \cos(\theta_1-\theta_2) + \omega_2^2 l_2 - \omega_3^2 l_3 \cos(\theta_3-\theta_2)}{l_3 \sin(\theta_3-\theta_2)} \tag{14}$$

$$\alpha_2 = \frac{-\omega_1^2 l_1 \cos(\theta_1-\theta_3) - \omega_2^2 l_2 \cos(\theta_2-\theta_3) + \omega_3^2 l_3}{l_2 \sin(\theta_2-\theta_3)} \tag{15}$$

Similarly, the two unknown angular accelerations of the second vector loop, α_4 and α_7, can be determined:

$$\alpha_4 = \frac{-\omega_6^2 l_6 \cos(\theta_6-\theta_7) + \omega_7^2 l_7 - \omega_4^2 l_4 \cos(\theta_4-\theta_7)}{l_4 \sin(\theta_4-\theta_7)} \tag{16}$$

$$\alpha_7 = \frac{\omega_6^2 l_6 \cos(\theta_6-\theta_4) + \omega_4^2 l_4 - \omega_7^2 l_7 \cos(\theta_7-\theta_4)}{l_7 \sin(\theta_7-\theta_4)} \tag{17}$$

Using the above methods and equations, the position, velocity, and acceleration parameters of each linkage in the jumping mechanism can be obtained, thereby clarifying its kinematic characteristics.

4 MATLAB SIMULATION ANALYSIS

Following the kinematic analysis of the jumping mechanism, a simulation was conducted using Matlab to establish a mathematical model based on the kinematic results, in order to verify whether the mechanism can achieve the designed jumping motion.

4.1 Simulation of Mechanism Motion Posture

The rocker of the mechanism was discretized into multiple rotation nodes at equal intervals. Based on the results of the kinematic position analysis, the corresponding joint angles and positions of the mechanism at each discrete point were

solved. The mechanism positions were then simulated using Matlab's plotting function, as shown in Figure 10. It can be observed that the motion posture change of the jumping mechanism closely matches the hind leg movement of a frog during an actual jump.

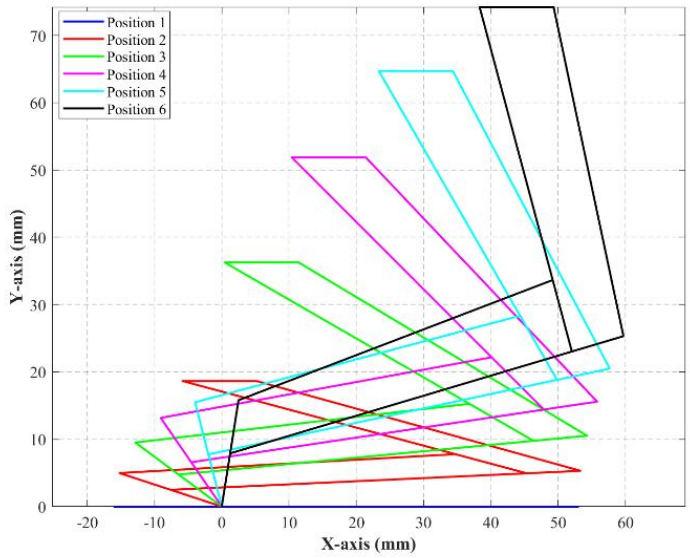


Figure 10 Diagram of the Change in the Motion Posture of the Jumping Actuator

4.2 Simulation of Kinematic Analysis Results

The expressions for the position, velocity, and acceleration of each linkage derived from the kinematic analysis were simulated using Matlab's plotting functionality. The simulation results are shown in Figures 11 to 13. The curves indicate that the angular displacement of each linkage changes smoothly without abrupt transitions over a full motion cycle. The variations in angular velocity and angular acceleration remain within reasonable ranges, demonstrating relatively stable operation of the mechanism. These results validate the feasibility of the jumping execution mechanism.

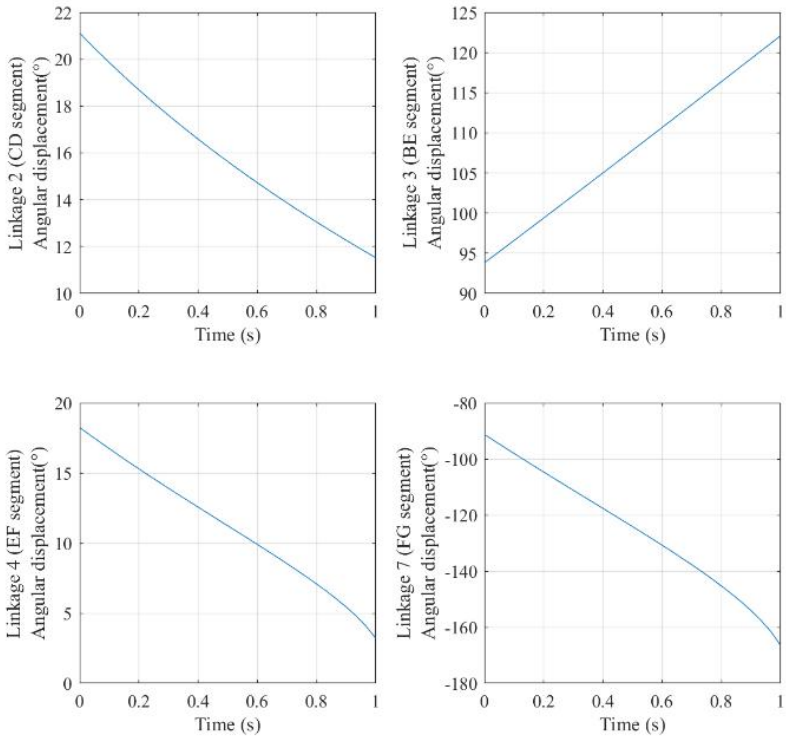


Figure 11 Diagram of Angular Displacement Changes of Each Linkage

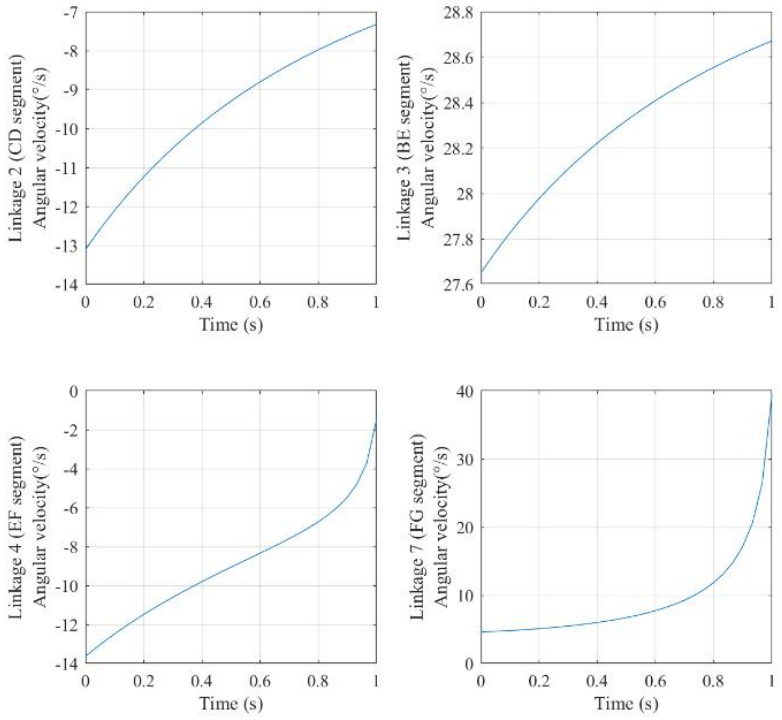


Figure 12 Diagram of Angular Velocity Changes of Each Linkage

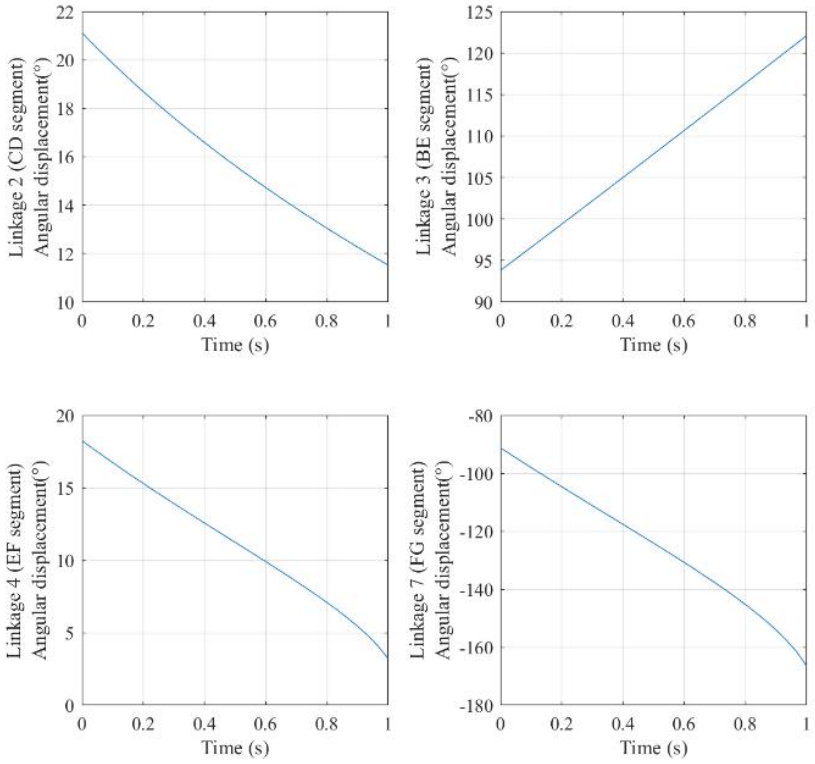


Figure 13 Diagram of Angular Acceleration Changes of Each Linkage

5 CONCLUSION

This study analyzed the movement mechanism of frogs to develop a simplified structural model, which led to the structural design of a bionic frog robot. Particular emphasis was placed on the design of the transmission system and the jumping mechanism. Kinematic analysis was conducted on the core jumping mechanism to determine the displacement,

velocity, and acceleration patterns of each linkage. Furthermore, Matlab simulation software was employed to visualize the kinematic behavior of the mechanism. The resulting motion posture diagrams and parameter variation curves validated the feasibility and rationality of the proposed mechanism, thereby laying a theoretical foundation for subsequent structural optimization.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This research was supported by Natural Science Foundation of Sichuan Province (Grant No. 2023NSFSC0368), Talents Project of Chengdu University (Grant No. 2081921013)

REFERENCES

[1] Men B, Fan X K, Chen Y X. Research status and development trends of bionic robots. Robot Technology and Application, 2019(5): 15-19.
[2] Lo J, Parslew B. Characterising the take-off dynamics and energy efficiency in spring-driven jumping robots. Mechanism and Machine Theory, 2024, 199. DOI: 10.1016/j.mechmachtheory.2024.105688.
[3] Tang G, Yang Q, Lian B. Design and Experimentation of Tensegrity Jumping Robots.Applied Sciences-Basel, 2024, 14(9): 18. DOI: 10.3390/app14093947.
[4] Wang G B, Chen D S, Chen KW, et al. Research status and development trends of bionic robots. Journal of Mechanical Engineering, 2015, 51(13): 27-44.
[5] Gu Y C. Design and research of a bionic frog jumping robot. Dalian Jiaotong University, 2021.
[6] Wang M. Development of a bionic frog jumping robot. Harbin Institute of Technology, 2011.
[7] Lei B J. Kinematic simulation and landing impact process analysis of a bionic frog hopping robot. Zhejiang Sci-Tech University, 2017.
[8] Lin F. Research on a posture adjustment device imitating the aerial movement patterns of frog hindlimbs. Harbin Institute of Technology, 2022.

World Journal of Engineering Research
Print ISSN: 2959-9865
Online ISSN: 2959-9873
DOI: https://doi.org/10.61784/wjer3047

VIBRATION MECHANISM AND EXPERIMENTAL VERIFICATION OF VIOLIN STRINGS

ZiChen Xu, EnZe Liu*
Student Affairs Office, Shanghai Nanyang Model School, Shanghai 200030, China.
Corresponding Author: EnZe Liu, Email: liu_enze@sjtu.edu.cn

Abstract: This research investigates the vibration mechanism of violin strings and the associated influencing factors by integrating theoretical analysis with experimental measurements.
A mechanical model describing string vibration was developed, indicating the relationship among the natural frequency and key parameters including tension, linear density, and effective string length. Using a smartphone as the recording device, the experimental measurements of string vibration signals were conducted under various conditions, such as differing string materials, excitation positions, forces, excitation methods, and effective lengths. Subsequent spectral analysis was performed utilizing Fourier transform theories. The results demonstrate that: (1) the natural frequencies of the string are determined by its physical properties and remain unaffected by the position or magnitude of excitation. (2) There exists an inverse relationship between effective string length and natural frequency, with the fundamental frequency exhibiting a remarkable decrease as the length increases. (3) While the frequency domain components remain consistent across different excitation methods (plucking versus bowing), the distribution of harmonic energy varies. This research offers an experimental validation approach for the acoustic properties of stringed instruments and elucidates the underlying physical mechanisms by which string vibration affects timbre. The results provide a scientific foundation for instrument manufacturing, performance enhancement, and pedagogical practices in music education.
Keywords: Violin string; Vibration mechanism; Natural frequency; Fourier analysis; Acoustic properties; Experimental validation

1 INTRODUCTION

Violin is a string instrument renowned for its expressive capabilities in classical music, is often referred to as the "Queen of Instruments". As the most crucial instrument in the string section of modern orchestras, it possesses a pure and beautiful tone with rich emotional resonance. Over centuries of development, it has become a paradigm where art and science converge. Over the past centuries, extensive research has been conducted on the violin. To craft instruments with superior sound quality, scientists have continuously developed scientific methods and conducted experiments to uncover the principles behind exceptional violin manufacturing. Figure 1 illustrates a schematic of violin performance.

Figure 1 Schematic of Violin Performance

The fundamental mechanism underlying violin sound generation arises from the vibration of the strings induced by friction with the bow hair. Given that the strings exhibit intrinsic resonant frequencies, the bow predominantly stimulates vibrations in proximity to these natural frequencies[1]. Consequently, the vibration of the string fundamentally determines the primary pitch, timbre characteristics, and expressive capabilities of the performance. A comprehensive understanding of this vibration mechanism is essential not only for elucidating the physical basis of the violin's musical appeal but also for its considerable theoretical and practical implications in instrument manufacturing, the refinement of performance techniques, and the design of novel stringed instruments. The violin's four strings (G, D, A, E, from bass to treble, thick to thin) are primarily made of types of materials: gut, steel, and synthetic fibers such as nylon. Each of these materials exerts a distinct impact on the instrument's tonal quality and vibration characteristics. Gut strings generate a brilliant and luminous tone characterized by exceptional responsiveness and strong capabilities for tonal modulation. Nylon strings produce a soft and sweet sound, generally exhibiting slightly lower volume levels compared to metal strings. Steel strings provide a bright and expressive tone and demonstrate greater durability than nylon strings, as illustrated in Figure 2. Violin strings are available in high, medium, and low tension variants,

necessitating selection according to the particular instrument. With the strategic choice or combination of different string types, it is possible to achieve a balanced timbre and volume, thereby enhancing the overall tonal quality.

Figure 2 Violin Strings

Fundamental theory research on string vibration can be traced back to the 17th and 18th centuries. The partial differential equation describing the transverse vibrations of a uniform string was initially formulated by the mathematician d'Alembert, who demonstrated that the displacement at any point along the string adheres to the wave equation[2]. Subsequently, researchers including Bernoulli and Euler advanced the concept of standing waves, demonstrating that the string vibration is essentially the superposition of the fundamental frequency alongside a series of harmonic frequencies (overtones) that collectively generate standing wave patterns[3]. In the 19th century, through innovative experimental investigations, Helmholtz elucidated that an ideal string subjected to continuous friction from a bow does not undergo sinusoidal vibration. Rather, it exhibits a distinctive motion pattern characterized by an "inflection point" (known as Helmholtz motion), wherein the location of the turning point periodically moves along the contact region between the bow and the string[4]. Contemporary studies have increasingly integrated factors including the Young's modulus of the string material, geometric nonlinearities such as large-amplitude effects, and internal damping mechanisms to develop more sophisticated physical models. Simultaneously, advanced technologies, such as high-speed photography and laser vibrometry, has substantially improved experimental methodologies for capturing transient phenomena and intricate string vibration modes.

Despite increasingly sophisticated theoretical models have been developed, there remains a significant need for comprehensive experimental validation of these vibration characteristics. This research focuses on vibration mechanisms and experimental validation of violin strings , aiming to achieve the following objectives:

(1) To enhance theoretical comprehension by conducting a systematic review of fundamental string vibration theories, including wave equations, the standing waves formation, and frequency formulas, with the focus on analyzing how tension, linear density, and effective string length influence fundamental and harmonic frequencies. (2)The second objective is to formulate practical experimental methodologies. Utilizing commonly accessible devices for high school students, such as smartphones, this research aims to put forward a set of straightforward and quantitative protocols for assessing string vibration frequency, modes, waveforms, and decay properties under diverse conditions, including variations in tension, string material, and excitation position or method. (3) The third objective aims to empirically validate key mechanisms by examining the effects of different excitation methods (including variations in bow speed, pressure, and contact point) on vibration modest (utilizing high-speed photography to capture standing wave patterns), as well as analyzing sound spectra to assess harmonic distribution, thereby elucidating their contributions to timbre. This study seeks to transform abstract physical concepts related to vibrations and waves into tangible, quantifiable scientific phenomena, thereby establishing an empirical connection between musical artistry and physical science. This approach not only strengthens foundational mechanics knowledge but also fosters the development of rigorous scientific reasoning, experimental design skills, and proficiency in data analysis. The results are expected to serve as illustrative case studies for the acoustics of string instruments, providing preliminary physical insights for novices in string selection and tonal modulation. In doing so, the research exemplifies the scientific ethos of "Study Things to Acquire Knowledge," contributing to a deeper understanding of humanity's cultural and artistic heritage.

2 METHODOLOGY

An integrated methodology encompassing both theoretical analysis and experimental measurement is utilized in this study. Initially, a mechanical model characterizing the vibration behavior of violin strings is formulated. Subsequently, practical experimental methods are investigated, wherein smartphone recordings are employed to capture the acoustic signals produced by the vibrating strings for quantitative assessment. The experimental data are then systematically analyzed to provide insights into the vibration mechanisms of violin strings.

2.1 Force Model of Violin Strings

The structural and geometric properties of the violin result in intricate force conditions. Figure 3(a) indicates the primary force points on a stationary violin. The nut end is identified as point A, the saddle end is identified as point B,

and the top of the bridge is identified as point C. Denoted as F, the tension force exerted by the strings produces a vertically downward component at the angle associated with point C. The resultant of these components constitutes the pressure T applied by the strings onto the bridge. In the nut section, F denotes the string tension, F_2 represents the longitudinal pressure exerted on the upper edge of the top plate, and F_3 signifies the longitudinal pressure applied to the end of the back plate.

The mechanical structure of violin constitutes a complex dynamic system, characterized by intricate interactions among all points of force application. Taking point A as an example, its spatial position is determined by the height of the nut and the protrusion of the wooden neck tenon. It is also further refined through the elastic deformation of the neck-fingerboard assembly.

Being a relatively fixed pivot, point B enables luthiers to accurately regulate the compressive force T applied by the strings on the bridge through adjustments to its vertical position. The spatial coordinates of point C demonstrate a complex dependency on multiple parameters, including bridge geometry, headstock angle, string spacing, tenon protrusion, and top plate curvature. Moreover, these coordinates reflect the combined deformation effects of the instrument body under string tension.

According to system mechanics, enhancing the compressive stress T exerted by the bridge on the soundboard can be achieved through three primary regulatory mechanisms. First, lowering the position of point A results in a simultaneous reduction of the vector components of forces F_2 and F_3. Second, modifications to the height of point B have a negligible effect on forces F_2 and F_3. Third, raising the spatial coordinate of point C leads to a concurrent increase in T, F_2, and F_3. This quantitative analysis of the force network offers luthiers a scientifically grounded framework for decision-making, allowing for the optimization of specific mechanical parameters in alignment with acoustic performance goals as defined by the acoustic objective function.

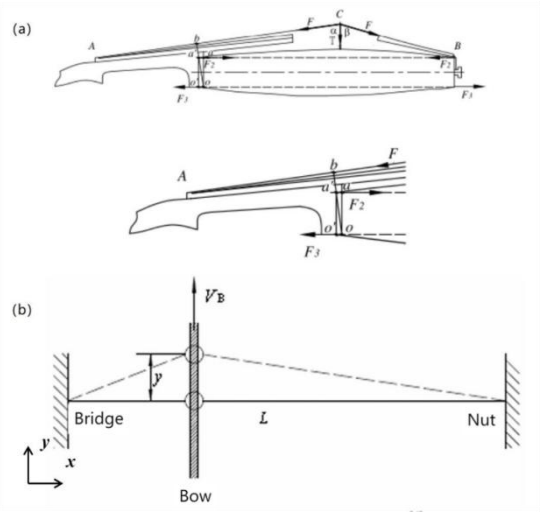


Figure 3 Force Conditions on a Violin. (a) Static Force Analysis of the Violin, (b) Simplified Vibration Model of Bow-String Interaction

In violin performance, the bow (the exciter), strings, bridge, and body (the resonance chamber) collectively constitute a string vibration system. A simplified vibration model is illustrated in Figure 3(b). Upon excitation of the string by the bow, the system vibrates in accordance with its intrinsic properties. When the string undergoes continuous excitation through bowing, the system maintains sustained vibration. The vibration mechanism of stringed instruments during bowing is notably complex. However, this simplified model primarily focuses on the free vibration behavior of the string to analyze its vibration characteristics. Specifically, it examines the initial excitation delivered by the bow followed by its cessation, allowing the string to vibrate freely. Under the assumptions of constant internal tension T and uniform string thickness with constant linear density ρ, the free vibration of the string can be mathematically modeled as follows:

$$\frac{\partial^2 y(x,t)}{\partial x^2}=\frac{\rho}{T}\frac{\partial^2 y(x,t)}{\partial t^2} \tag{1}$$

Assuming the displacement y of the string vibration is separated in time and space, the method of separated variables is employed. We assume:

$$y(x,t)=Y(x)F(t) \tag{2}$$

Substituting into equation (1):

$$\frac{T}{\rho Y(x)}\frac{d^2Y(x)}{dx^2}=\frac{1}{F(t)}\frac{d^2F(t)}{dt^2} \tag{3}$$

Both sides of the above equation are functions of variables x and t. Consequently, both sides are set equal to a constant. Denoting this constant by $-\omega^2$, then equation (3) can be expressed as:

$$\frac{d^2F(t)}{dt^2}+\omega^2F(t)=0 \tag{4}$$

$$-T\frac{d^2Y(x)}{dx^2}=\omega^2\rho Y(x),\ 0<x<L \tag{5}$$

Solving this equation yields the characteristic frequency (natural frequency) of the string:

$$\omega_i=\beta_i\sqrt{\frac{T}{\rho}},\ (i=1,2,\cdots) \tag{6}$$

$$\beta_iL=i\pi,\ (i=1,2,\cdots) \tag{7}$$

where i denotes the order of the characteristic frequency. This research primarily focuses on the first-order natural frequency of violin strings, examining the variation patterns of natural frequencies among strings with differing parameters.

2.2 String Acoustic Signals Analysis

2.2.1 String acoustic signal recording

The entire sound production mechanism of a violin encompasses the initial plucking of the string to induce vibration, which subsequently stimulates air vibrations adjacent to the sound holes via the resonance of the instrument body, thereby producing Helmholtz resonance[5]. Consequently, the act of plucking the string and subsequently recording the emitted sound signal via a mobile phone recording app effectively captures data pertaining to the string's inherent vibration natural frequencies.

The basic information of the four violin strings used in this study is summarized below. All four strings possess an approximate length of 32 cm. The diameters of the strings are listed as follows: the G string ranges from 0.72 to 0.76 mm, the D string ranges from 0.62 to 0.66 mm, the A string ranges from 0.52 to 0.56 mm, and the E string ranges from 0.42 to 0.46 mm. The sound recording were conducted with both the smartphone and violin held stationary throughout the experiments.

Experiment 1 involved selecting the midpoints of all four strings as the plucking positions. Each string was plucked with a consistent force, and the resulting acoustic signals were recorded simultaneously using the smartphone.

In Experiment 2, the E string was designated as the subject. This string was divided into four equal segments, with plucking positions established at 8 cm, 16 cm, and 24 cm from one end. Each segment was plucked with uniform force, and the corresponding acoustic signals were recorded concurrently.

Experiment 3 also utilized the E string as the test subject, where the string was plucked at its midpoint with varying force levels categorized as light, medium, and heavy. Acoustic signals were recorded simultaneously during these plucks.

In Experiment 4, the E string's effective vibrating length was altered by firmly pressing the string at positions corresponding to lengths of 17.9 cm, 21.7 cm, and 27 cm, respectively. The string was then plucked, and the resulting acoustic signals were recorded.

Finally, Experiment 5 involved consecutive plucking of the four strings by hand, followed by consecutive bowing of the four strings using a bow, with the corresponding acoustic signals recorded in each case, see Table 1.

Table 1 Vibration Mechanism and Experimental Verification of Violin Strings

Number	String	Length	Material	String Tension
4	E	31.9cm	steel	8.0*9.8N
3	A	32cm	nylon core with aluminum-coated surface	5.5*9.8N
2	D	32cm	nylon core with silver-coated surface	4.5*9.8N
1	G	32cm	nylon core with silver-coated surface	4.6*9.8N
I Pluck the midpoint of each string				
Pluck the E string at three points.				
Pluck the midpoint of E string with light, medium, and heavy forces.				
II Press the E string				
III Consecutively pluck the four strings and bow the four strings				

Experiments captured acoustic signals from plucking strings with varying parameters, plucking the same string at different positions, plucking strings with different intensities, strings of different lengths, and plucking or bowing all four strings. A more detailed examination of these acoustic signals was conducted utilizing MATLAB software, which facilitated the extraction of the natural frequencies of the strings. The primary analytical and processing procedure is depicted in Figure 4.

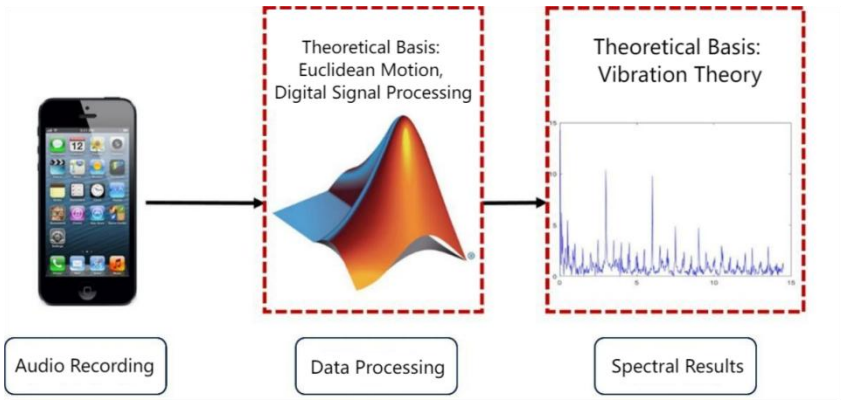


Figure 4 Acoustic Signal Processing Workflow for Violin String Vibrations

2.2.2 Fourier analysis of string acoustic signals

The acoustic signals captured by a mobile phone are digital in nature and are subsequently processed using computational methods. For motion patterns that change over time, these signals fundamentally constitute one-dimensional numerical sequences. Signals may be classified according to the dimensionality of their values into one-dimensional, two-dimensional, or three-dimensional categories, for example, acoustic signals are one-dimensional signals. Given that the independent variable in signal analysis is generally time, extracting the intrinsic frequency characteristics of string vibrations necessitates transforming the signal from the time domain to the frequency domain through the application of the Fourier transform.

According to the principles of Fourier analysis, any periodic signal that meets the Dirichlet conditions—namely, piecewise continuity, a finite number of extrema, and absolute integrability—can be represented as a linear combination of sine and cosine functions at various frequencies. Each sine or cosine term corresponds to a specific frequency component, allowing the Fourier series expansion to facilitate a detailed examination of the signal's frequency content. Consequently, this method enables a comprehensive characterization of the signal's spectral properties[6].

To address the constraint that Fourier series are applicable solely to periodic signals, the Fourier transform is derived by considering the limit in which the period tends toward infinity, as illustrated in the following equation:

$$X(\omega)=\int_{-\infty}^{\infty} x(t)e^{-i\omega t}dt \tag{8}$$

While the Fourier transform provides a comprehensive mapping of continuous-time signals into the frequency domain for analytical purposes, its formulation as a continuous integral precludes direct implementation within digital systems. To overcome this limitation, the Discrete Fourier Transform (DFT) was introduced. By discretizing the sampled time-domain sequence, the DFT reformulates the Fourier transform into a numerical format amenable to computational processing. Consequently, this transformation enables the conversion of discrete time-series data into their corresponding discrete frequency-domain representations. It preserves the core functionality of spectral analysis while enabling the digital implementation of signal processing algorithms[7]. Let $x_N(nT_s)$ denote N sampled values of the continuous function $x(t)$, where n ranges from 0 to N-1. The discrete Fourier transform is defined as:

$$DFT\left(x_N(nT_s)\right)=\sum_{n=0}^{N-1} x_N(nT_s)e^{-\frac{i2\pi nk}{N}} \tag{9}$$

From the perspective of signal processing, the sampling operation can be mathematically modeled as the multiplication of the original continuous-time signal by a pulse sampling sequence. According to the convolution property of the Fourier transform, this multiplication in the time domain corresponds to a periodic convolution of the original signal's spectrum with the spectrum of the sampling function in the frequency domain. Consequently, the signal's spectrum is replicated at intervals equal to the sampling frequency, resulting in a periodic extension in the frequency domain. To avoid aliasing distortion within the spectrum, the Nyquist sampling theorem must be observed, which stipulates that the sampling frequency should be at least twice the maximum frequency component present in the signal. This requirement implies that, during spectral analysis, the highest frequency of interest must not exceed half the sampling rate, known as the Nyquist frequency. Adhering to this criterion ensures adequate separation between successive spectral replicas, thereby preventing distortion arising from spectral overlap[8].

Within the MATLAB environment, the robust matrix computation capabilities can be utilized to conduct spectral analysis on the recorded string audio signal. The Fast Fourier Transform (FFT), an efficient algorithmic realization of the Discrete Fourier Transform (DFT), is accessible through MATLAB's built-in 'fft' function[9]. The specific spectral analysis implementation code is shown in Figure 5. In this context, *Sig* denotes the acquired discrete time-domain signal sequence, while *N* represents the total number of sampled points. The methodological process is as follows: initially, the discrete time-domain signal of the vibrating string is captured via a mobile phone recording, with the sampling interval determined by the sampling frequency *fs*. Subsequently, the Discrete Fourier Transform computed efficiently using the FFT algorithm is applied to transform the time-domain data into the frequency domain, thereby extracting the constituent frequency components of the signal. In summary, once the sampling frequency *fs* is known and the discretized vibration time-history signal is obtained, the spectral properties of the string can be analyzed through

frequency domain transformation, which in turn provides insights into the vibration mechanism of the violin string.

```
Perform an N-point FFT transform on the time-domain signal sig
Sig_f = fft(Sig, N);

Generate the corresponding frequency coordinate axis (0 to Nyquist frequency)
freq = (0:fs/N:fs/2);

Compute the single-sided amplitude spectrum and normalize it
A_Sig_f = abs(Sig_f(1:N/2+1)) / (N/2);

Plot the spectrogram
plot(freq, A_Sig_f, 'b');
```

Figure 5 Code for Implementing Spectrum Analysis

3 RESEARCH RESULTS AND ANALYSIS

3.1 Experiment 1: Results of Acoustic Signal Analysis concerning Strings Plucking with Different Parameters

Figure 6 presents the analytical results of the audio recordings obtained in Experiment 1, in which the midpoints of four individual strings were plucked separately. Although the plucking action is of very short duration, the strings persist in vibrating after being released by the fingers. The frequencies of these vibrations correspond to the inherent characteristic (natural) frequencies of the respective strings. The analysis of the time-domain signal indicates that, following the cessation of plucking, the vibrations of the string undergo a rapid attenuation attributable to damping mechanisms, including material damping and frictional forces. Due to the distinct physical parameters characterizing each of the four strings, their respective time-domain waveforms display unique morphological features. Examination of the frequency-domain data demonstrates that subsequent to plucking, the free vibration frequencies of the strings predominantly oscillate in proximity to their inherent natural frequencies. Specifically, the fundamental natural frequencies for the four strings are approximately as follows: the G string (first string) at 195 Hz, the D string (second string) at 291 Hz, the A string (third string) at 438 Hz, and the E string (fourth string) at 659 Hz. Moreover, spectral analyses for each string reveal that the principal frequency components correspond to the various harmonic orders of the natural frequencies, as delineated by Equation (6). The amplitude distribution further corroborates that the first-order natural frequency constitutes the dominant frequency component for each string.

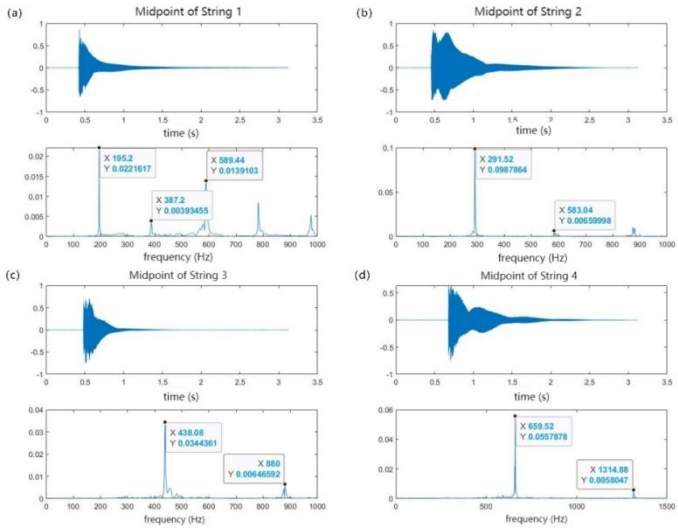


Figure 6 Audio-Frequency Domain Results for Plucking Four Strings

3.2 Experiment 2: Results of Acoustic Signal Analysis concerning Strings Plucking at Different Positions

The fourth string (E string) was selected as the focus of this study to examine the influence of varying plucking positions on the free vibration response of a violin string. The results are presented in Figure 7. Analysis in the time domain indicates that the waveforms and vibration duration are nearly identical across the different plucking positions. As for the frequency domain, the free vibration frequency components remain largely consistent regardless of the plucking position, predominantly comprising the string's natural frequencies of various orders. It is evident that the

string natural frequencies are unaffected by the point of excitation, in agreement with Equation (6). Nevertheless, variations in the plucking position lead to differences in the energy distribution among the harmonic frequency components. Due to the structural symmetry of the string, the amplitude magnitudes and relative proportions of the harmonic frequencies depicted in Figures 7(a) and 7(c) are almost identical, while they differ to some extent from those shown in Figure 7(b).

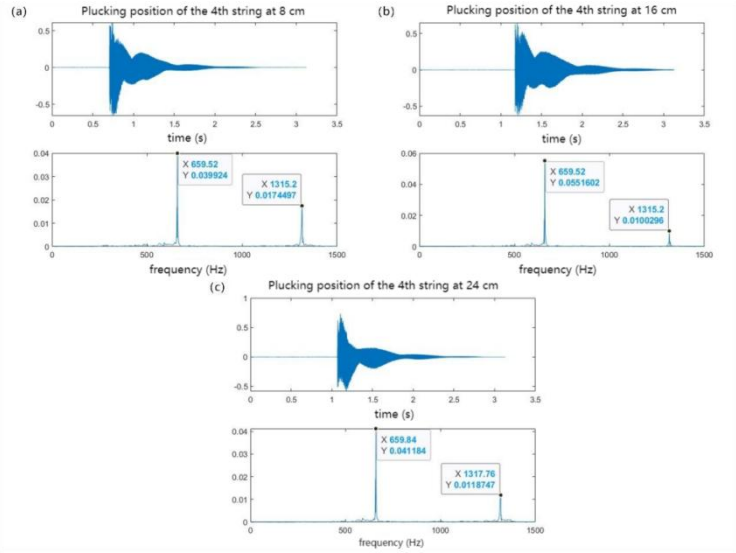


Figure 7 Audio-Frequency Domain Results for Plucking at Different Positions

3.3 Experiment 3: Results of Acoustic Signal Analysis concerning Strings Plucking with Different Forces

Experiment 3 also selected the E string (4th string) as the research subject. The violin string was plucked at its midpoint with varying forces to investigate the effect of different input amplitudes on the free vibration response of the violin string. The results are shown in Figure 8. According to Equation (6), the natural frequency of a string depends solely on its linear density, length, and internal tension. Consequently, the results of Experiment 3 closely parallel those observed in Experiment 2: variations in the forces of string plucking yield time-domain waveforms and vibration duration that are almost indistinguishable. Similarly, the frequency-domain analyses reveal frequency components and amplitude ratios among these components that remain largely consistent. Nevertheless, owing to differences in input magnitudes, the output amplitudes exhibit substantial variation. These results indicate that changes in plucking force influence only the output loudness, without affecting the output frequency characteristics.

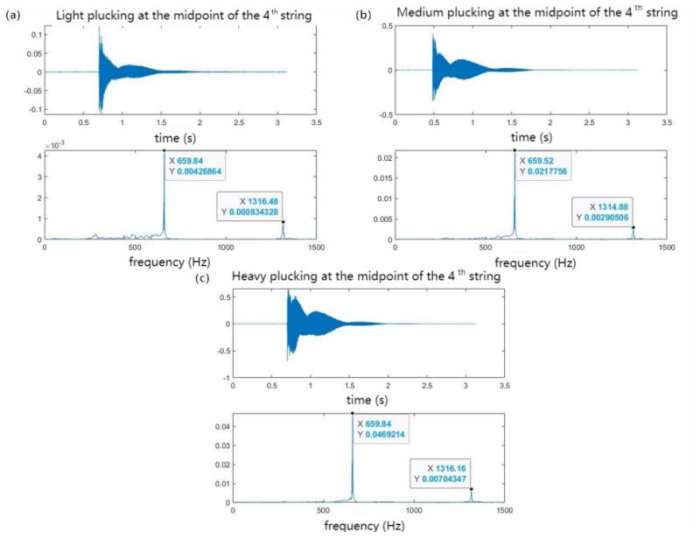


Figure 8 Audio-Frequency Domain Results for Different Plucking Forces

3.4 Experiment 4: Results of Acoustic Signal Analysis concerning Plucking Strings of Different Lengths

Experiment 4 altered the effective vibrating length of the E string (4th string) to investigate the free vibration response of violin strings with varying lengths. The results are shown in Figure 9. The free vibration responses of the three strings, each possessing different effective lengths, demonstrate that an increase in the effective vibrating length corresponds to an elongation of the time-domain output duration. Concurrently, the principal frequency components

observed in the frequency-domain output exhibit a shift toward lower frequencies. This reduction in dominant frequencies signifies that the natural frequency decreases as the effective vibrating length increases, thereby illustrating an inverse relationship between natural frequency and string length, in accordance with the formulation presented in Equation (6).

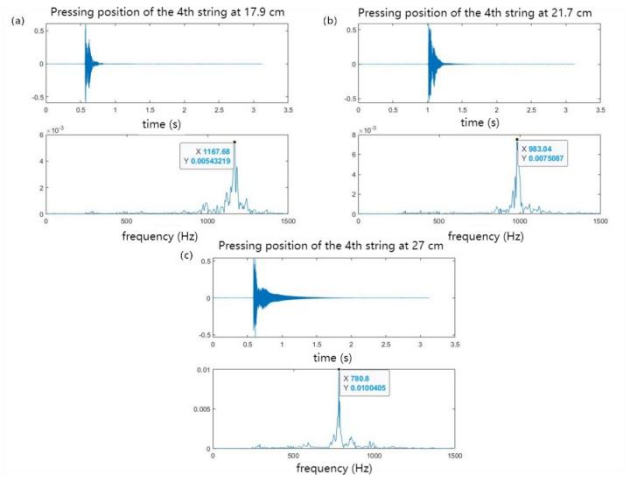


Figure 9 Audio-Frequency Domain Results for Plucked Strings of Different Lengths

3.5 Experiment 5: Results of Acoustic Signal Analysis concerning Different Excitation Methods

Experiment 5 involved plucking all four strings consecutively by hand and drawing the bow across all four strings consecutively, investigating the impact of these two excitation methods on the audio frequency of violin string vibrations. The findings are presented in Figure 10. It is evident that the time-domain signals associated with the two excitation methods exhibit significant differences. When the strings are plucked sequentially by hand, the precise time instances at which each string is excited are distinctly identifiable. Conversely, due to the continuous nature of bowing, the time-domain signal generated by this excitation technique does not allow for the clear differentiation of the exact moments at which individual strings are activated. Analysis of the frequency-domain data reveals that the dominant frequency components of the audio output are largely consistent across both excitation methods, corresponding to the inherent natural frequencies of the respective strings. Nonetheless, the amplitude ratios of these frequency components vary between the two excitation approaches, indicating their differential impact on the resultant acoustic output of the violin.

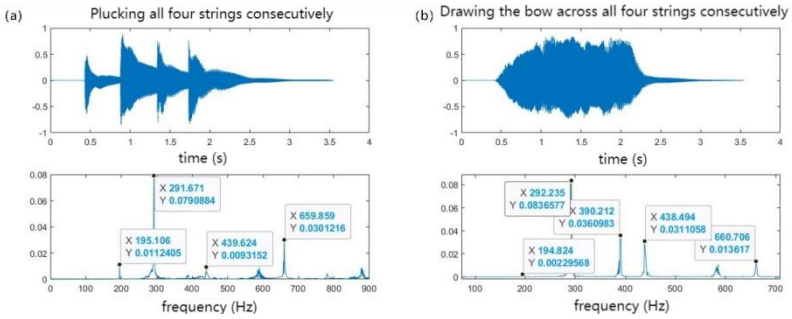


Figure 10 Audio Time-Frequency Domain Results for Strings Under Different Excitation Methods

4 CONCLUSIONS

This research presents a systematic investigation of the vibration mechanisms of violin strings and the factors influencing them, integrating theoretical modeling with experimental validation. A mechanical model of string vibration was developed to theoretically derive quantitative relationships linking the natural frequency to string tension, linear density, and effective string length. Experimentally, an innovative methodology employed smartphones to record vibration signals under varied conditions, including differences in string material, excitation position, force, method, and effective length. Spectral analysis based on Fourier transform techniques substantiated several key findings: (1) the natural frequency of a string is governed by its inherent physical properties and remains unaffected by the position or force of excitation. (2) The effective string length exhibits an inverse proportionality to the natural frequency. (3) Although distinct excitation methods (plucking versus bowing) produce markedly different time-domain responses, their frequency-domain spectra are consistent, differing primarily in the distribution of harmonic energy. This study introduces an accessible experimental framework suitable for secondary education, facilitating an intuitive understanding of the physical principles underlying the sound production of violin. Moreover, it establishes a reproducible model for acoustical research on stringed instruments. By rigorously quantifying the mechanical

foundations of artistic expression, the findings offer empirical support for instrument craftsmanship, performance enhancement, and music education, thereby underscoring the distinctive role of physics in elucidating aspects of the humanities and arts.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Wu Jike. Popular Mechanics Series: Science in Music. Higher Education Press, 2012. https://mall.hep.com.cn/goods-7080.html

[2] Fletcher N H, Rossing T D. The Physics of Musical Instruments (2nd ed.). Springer. 2010. DOI: 10.1007/978-0-387-21603-4.

[3] Hansen U J, Kitterman S. The physics of musical instruments with performance illustrations and a concert. The Journal of the Acoustical Society of America, 2014, 136(4): 2113. DOI: 10.1121/1.4899610.

[4] von Helmholtz H. Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik, von H. Helmholtz. 2te Ausgabe. F. Vieweg und Sohn, 1865. DOI: 10.1007/978-3-663-18653-3.

[5] GR IMES D R. String theory-the physics of string bending and other electric guitar techniques. PLoS ONE, 2014, 9(7): e102088. DOI: 10.1371/journal.pone.0102088.

[6] Alm J F, Walker J S. Time-Frequency Analysis of Musical Instruments. SIAM Review, 2006, 44(3): 457-476. DOI: 10.1137/S00361445003822.

[7] Argyriou A, Andrew S. Discrete Fourier Transform and Minimum Variance Distortionless Response Beamforming Algorithms for Searching for Transients in Radio Interferometers that Experience Radio-frequency Interference. The Astrophysical Journal Supplement Series, 2025, 280(2): 59. DOI: 10.3847/1538-4365/ADFCBE.

[8] Ruan Y, Tang T. An optimized Youla–Kucera parametrization with time-delay compensation in multirate parallel control for disturbance rejection up to Nyquist frequency. Control Engineering Practice, 2025, 158, 106285. DOI: 10.1016/J.CONENGPRAC.2025.106285.

[9] Wang L, Chen X B, Gao Y, et al. A Dual-Channel Recommendation Algorithm Based on FFT Denoising and Reverse Path Collaborative Enhancement. Journal of Computer Applications, 2025, 1-10. DOI: 10.11772/j.issn.1001-9081.2025070869.

World Journal of Engineering Research
Print ISSN: 2959-9865
Online ISSN: 2959-9873
DOI: https://doi.org/10.61784/wjer3049

A HYBRID IDENTITY AUTHENTICATION SYSTEM COMBINING PHYSICAL LAYER RECOGNITION AND CERTIFICATES FOR C-V2X IN HIGHWAY ENVIRONMENTS

YuanYuan Song[1*], Yitzhak Cohen[2]
[1]*Shandong Zhengchen Technology company limited, Jinan 250101, Shandong, China.*
[2]*SCOPE Strategic Management Ltd, Drech Begen 156, Tel Aviv-Jaffa, Israe.*
Corresponding Author: YuanYuan Song, E-mail: 245194050@qq.com

Abstract: With the rapid development of Cellular Vehicle-to-Everything (C-V2X) communication systems, ensuring the security and integrity of data exchanged between vehicles and infrastructure has become a significant challenge. Identity authentication plays a crucial role in safeguarding these communications against threats such as identity spoofing and data tampering. Traditional certificate-based authentication methods, while effective, often suffer from performance issues, especially in high-speed, dynamic environments such as highways. In this paper, we propose a hybrid identity authentication system that combines physical layer recognition with certificate-based methods to enhance security and improve real-time performance in C-V2X systems operating in highway environments. The physical layer authentication leverages channel state information (CSI), signal fingerprints, and other radio-frequency characteristics to authenticate vehicles, providing an additional layer of security. This is combined with certificate-based public key infrastructure (PKI) to offer a comprehensive and robust identity verification process. The hybrid approach addresses the shortcomings of traditional methods by enhancing security while minimizing computational overhead and ensuring low-latency authentication. We discuss the challenges and opportunities in integrating these two authentication methods and highlight the potential impact of emerging technologies, such as 5G and machine learning, in optimizing the hybrid authentication process for C-V2X systems. Finally, we propose future research directions to further improve the efficiency, scalability, and robustness of hybrid authentication schemes in vehicular networks.
Keywords: C-V2X; Vehicles; Public key infrastructure

1 INTRODUCTION

With the rapid development of intelligent transportation systems (ITS) and Vehicle-to-Everything (V2X) communication technologies, the exchange of information between vehicles and infrastructure has seen growing applications in traffic management, road safety, and real-time data transmission. Specifically. Cellular V2X (C-V2X) technology is gaining significant attention in highway environments due to its low latency and high reliability, which enable efficient traffic flow management and accident prevention [1]. However, this widespread deployment raises significant security concerns, particularly regarding identity authentication in an open wireless communication environment. Threats such as identity spoofing and data tampering compromise the integrity of the system and undermine the trustworthiness of V2X applications [2].

Traditional identity authentication methods, such as certificate-based public key infrastructure (PKI) systems, offer a level of security in vehicular networks [3]. However, they often face performance bottlenecks in high-speed, low-latency environments due to the computational complexity and communication overhead involved [4]. Additionally, although physical layer-based authentication techniques—such as recognition based on channel state information and radio frequency (RF) characteristics—have been widely studied in recent years, integrating these physical layer characteristics with high-level certificate-based authentication methods remains a challenging research topic. This hybrid approach could offer enhanced security and reliability for C-V2X systems, especially in high-speed environments like highways .

Highway environments present unique challenges for C-V2X systems, including high-speed vehicular mobility, complex interference patterns, and the need to support massive numbers of vehicles [5]. While certificate-based identity authentication methods can ensure a degree of security, they are not always sufficient to meet real-time performance requirements in high-speed highway scenarios. As a result, combining physical layer information with certificate-based methods could significantly improve both the efficiency and security of authentication processes in such environments. This paper presents a hybrid identity authentication system that integrates physical layer recognition with traditional certificate-based authentication, aiming to address the security challenges faced by C-V2X systems in highway environments. Figure 1 shows the working principle of the C-V2X communication hybrid identity authentication system.

The paper first provides an overview of current research in C-V2X identity authentication, highlighting the limitations of existing methods and technologies in real-world applications. Next, the discussion focuses on the integration of physical layer authentication, exploring how it can enhance security and performance in highway environments. Finally,

the paper outlines future research directions and challenges in realizing a secure and efficient hybrid authentication system for C-V2X.

Figure 1 Hybrid Identity Authentication System for C-V2X Communication

2 CURRENT RESEARCH ON C-V2X IDENTITY AUTHENTICATION

Identity authentication plays a crucial role in ensuring the integrity, authenticity, and privacy of communication in C-V2X systems [6]. As C-V2X applications increasingly span across highway environments, ensuring secure communication between vehicles, roadside units, and infrastructure becomes a priority. Over the years, various authentication methods have been proposed and developed to address these challenges. These methods generally focus on public key infrastructures (PKI), certificate-based authentication, and more recently, physical layer-based authentication techniques, each offering distinct advantages and limitations.

2.1 Certificate-Based Authentication

Certificate-based authentication remains one of the most widely adopted methods for ensuring secure identity verification in C-V2X systems. By utilizing a public-key infrastructure (PKI), certificates are issued to vehicles and infrastructure devices, which are used to validate the identity of the communication entities [7]. This method guarantees the integrity and authenticity of messages exchanged over the network, as it relies on the encryption of data with public and private keys.

However, traditional certificate-based authentication methods suffer from significant performance issues, especially in high-speed highway environments. The computation and verification processes involved in issuing and validating certificates are resource-intensive and can lead to high latency, which is detrimental in real-time applications where low-latency communication is critical. Additionally, these systems are vulnerable to attacks such as certificate spoofing and unauthorized certificate revocation, which compromise the system's security [8].

2.2 Physical Layer Authentication

In contrast to certificate-based approaches, physical layer authentication techniques utilize unique physical characteristics of the communication channel or radio signals to verify the identity of communication entities [9]. These characteristics include channel state information (CSI), the unique multipath propagation of signals, and received signal strength. Physical layer authentication offers a promising alternative as it operates independently of higher-layer cryptographic mechanisms and can enhance system security by providing an additional layer of verification.

However, the integration of physical layer authentication with traditional certificate-based methods remains a challenge [10]. While physical layer methods can improve security, they are often sensitive to environmental factors such as noise, interference, and fading, which can degrade their effectiveness in real-world applications. Moreover, these methods require sophisticated hardware and signal processing techniques, which may introduce additional costs and complexity to the system.

2.3 Hybrid Authentication Approaches

Given the limitations of certificate-based and physical layer authentication methods when used independently, hybrid authentication schemes have been proposed to combine the strengths of both approaches [11]. By integrating physical layer features with certificate-based authentication, hybrid systems aim to provide more robust security while maintaining performance in high-speed and dynamic environments like highways.

Hybrid approaches offer several advantages over individual methods, including enhanced security, reduced risk of identity spoofing, and faster authentication processes [12]. However, these systems face challenges in terms of

computational complexity, real-time processing requirements, and hardware limitations. Future research is needed to address these challenges and optimize the integration of physical and certificate-based authentication in C-V2X systems.

3 CHALLENGES AND OPPORTUNITIES IN HYBRID IDENTITY AUTHENTICATION FOR C-V2X

While hybrid identity authentication systems combining physical layer recognition and certificate-based methods hold great promise for enhancing the security and reliability of C-V2X systems, several challenges remain to be addressed [13]. These challenges primarily stem from the need to integrate two fundamentally different approaches—physical layer characteristics and higher-layer cryptographic techniques—into a unified, efficient, and practical authentication framework. Despite these challenges, the hybrid approach offers substantial opportunities for improving the performance and security of C-V2X systems, especially in dynamic and high-speed highway environments.

3.1 Challenges in Integration

One of the primary challenges in implementing hybrid authentication for C-V2X systems is the complexity of integrating physical layer recognition techniques with certificate-based authentication [14]. Physical layer authentication methods, such as those based on channel state information (CSI) and signal fingerprints, require specialized hardware and signal processing algorithms to extract the relevant features. These methods rely on accurate measurements of the radio environment, which can be susceptible to factors such as interference, fading, and multi-path propagation. Integrating these features with traditional cryptographic operations used in certificate-based authentication introduces additional computational and communication overhead, which can impact the real-time performance of C-V2X systems in high-speed environments [15,16].

Moreover, the environmental conditions in highway scenarios, such as rapidly changing vehicle velocities and complex interference patterns, make it difficult to consistently obtain accurate and reliable physical layer features. The effectiveness of physical layer authentication methods may degrade due to factors such as signal noise, Doppler shifts, and vehicle mobility. These issues highlight the need for robust signal processing algorithms capable of accurately extracting features despite varying environmental conditions and high mobility [17,18].

3.2 Computational Complexity and Real-Time Processing

Another challenge in hybrid authentication systems is the computational complexity associated with processing physical layer features and verifying digital certificates in real time. Certificate-based authentication schemes, particularly those based on public-key cryptography, are computationally intensive, requiring significant processing power for key generation, signing, and verification. When combined with physical layer authentication, the system must process both signal characteristics and cryptographic data, further increasing the computational load.

In highway environments, where vehicles are moving at high speeds, the authentication process must be completed within a very short time frame to avoid delays in communication and ensure safe driving conditions. The need for low-latency communication and high throughput places a premium on the efficiency of the authentication process. As a result, the hybrid authentication system must be optimized to minimize computational overhead while maintaining a high level of security and reliability [19]. Figure 2 shows the C-V2X communication principle between the vehicle and the RSU.

Figure 2 C-V2X Communication between Vehicles and RSU

3.3 Opportunities for Improvement

Despite the challenges, the hybrid approach presents several opportunities to enhance the security and performance of C-V2X systems. One of the key advantages is the ability to provide multiple layers of security, making it more difficult for attackers to compromise the system. By combining the strengths of physical layer authentication, which is resistant to certain types of spoofing attacks, with the robust security guarantees of certificate-based methods, the hybrid system can offer enhanced protection against both physical and cryptographic threats.

Moreover, advancements in signal processing algorithms and hardware are opening up new opportunities for more efficient and accurate physical layer authentication. Machine learning and artificial intelligence (AI) techniques, in particular, have shown promise in improving the accuracy of physical layer authentication by automatically learning to recognize patterns in signal characteristics, even under noisy or variable conditions [20]. These techniques can significantly reduce the impact of environmental factors such as interference and fading, making physical layer authentication more reliable in real-world highway environments.

Additionally, emerging technologies such as 5G and beyond offer the potential to improve the performance of hybrid authentication systems. With higher bandwidth, lower latency, and better support for massive IoT devices, these next-generation networks could facilitate the real-time processing of both physical layer and certificate-based authentication, making hybrid systems more scalable and practical for large-scale deployment in C-V2X networks [21].

3.4 Future Directions

Looking ahead, there are several key areas where further research is needed to address the challenges and fully realize the potential of hybrid identity authentication systems for C-V2X. These include:

Robust Signal Processing: Developing advanced signal processing techniques that can extract reliable physical layer features under diverse environmental conditions, including high mobility and interference.

Lightweight Cryptographic Methods: Designing lightweight cryptographic protocols that can reduce the computational overhead associated with certificate-based authentication, ensuring real-time performance in high-speed environments.

Machine Learning Integration: Leveraging machine learning and AI to improve the accuracy and adaptability of physical layer authentication, enabling the system to learn and adapt to changing environmental factors.

Integration with 5G Networks: Exploring the synergies between hybrid authentication systems and 5G networks to enable seamless, scalable, and low-latency authentication processes for C-V2X communications.

4 CONCLUSION AND FUTURE OUTLOOK

In this paper, we have discussed the challenges and opportunities associated with hybrid identity authentication systems combining physical layer recognition and certificate-based methods for C-V2X communication in highway environments. The security of C-V2X systems is paramount to ensure reliable and safe communication between vehicles and infrastructure, especially in high-speed, dynamic environments like highways. Traditional certificate-based authentication methods, although effective in many scenarios, face limitations in high-speed vehicular networks due to performance bottlenecks and vulnerability to certain types of attacks. On the other hand, physical layer authentication provides an additional layer of security by leveraging the unique characteristics of the communication channel, but it also faces challenges related to environmental variability, signal noise, and hardware requirements.

By combining the strengths of both approaches, hybrid identity authentication systems offer a promising solution to the security challenges faced by C-V2X networks. The integration of physical layer authentication with certificate-based methods can enhance both the security and efficiency of the authentication process, providing a more robust defense against identity spoofing and other malicious attacks. However, significant challenges remain, including the need for effective integration of physical and cryptographic methods, the computational complexity of real-time processing, and the reliability of physical layer authentication under varying environmental conditions.

Despite these challenges, the potential for hybrid authentication systems to revolutionize the security landscape of C-V2X systems is immense. Advances in machine learning, signal processing, and next-generation networks such as 5G will likely play a pivotal role in addressing these challenges and improving the overall performance of hybrid authentication systems. Future research efforts should focus on optimizing the integration of physical and certificate-based authentication, reducing computational overhead, and developing more robust signal processing techniques that can adapt to the highly dynamic nature of highway environments.

Looking forward, several key areas warrant further exploration:

Optimization of Hybrid Systems: Further research should aim to refine hybrid authentication techniques to ensure that they can be efficiently implemented in real-time C-V2X systems without compromising security or performance.

Advanced Machine Learning Algorithms: The application of machine learning and AI to improve the accuracy and robustness of physical layer authentication, particularly in noisy and highly dynamic environments, holds great potential.

5G Integration: With the advent of 5G networks, new opportunities emerge for supporting hybrid authentication systems at scale, with the potential to handle high-speed vehicular communication with lower latency and higher throughput.

Scalability and Robustness: Future studies should focus on ensuring that hybrid authentication systems can scale to large numbers of vehicles and infrastructure elements while maintaining reliability under a wide range of operational conditions.

In conclusion, while hybrid identity authentication systems for C-V2X communications face several technical challenges, they represent a promising direction for enhancing the security of vehicular networks in highway environments. As research continues to evolve and technologies such as 5G and machine learning mature, the potential for these systems to provide secure, low-latency, and reliable authentication for C-V2X networks is becoming increasingly achievable.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This work was supported in part by Key R&D Program of Shandong Province (2022KJHZ002).

REFERENCES

[1] Hu H, Chai R, Chen ML, et al. System-level simulation platform of C-V2X mode 4: Integrating CarMaker and NS-3. IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2021. DOI: 10.1109/PIMRC50174.2021.9569386.

[2] Eckermann F, Wietfeld C. SDR-based open-source C-V2X traffic generator for stress testing vehicular communication. IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), 2021. DOI: 10.1109/VTC2021-Spring51267.2021.9449043.

[3] Peters S, Sivrikaya F, Dang XT. SEP4CAM – a simulative/emulative platform for C-V2X application development in cross-border and cross-domain environments. IEEE/ACM 25th International Symposium on Distributed Simulation and Real Time Applications (DS-RT 2021), 2021. DOI: 10.1109/DS-RT52167.2021.9576134.

[4] Abbasi HI, Gholmieh R, Nguyen TV, et al. LTE-V2X (C-V2X) performance in congested highway scenarios. IEEE International Conference on Communications (ICC 2022), 2022: 303-308. DOI: 10.1109/ICC45855.2022.9838706.

[5] Miao LL, Virtusio JJ, Hua KL. PC5-based cellular-V2X evolution and deployment. Sensors, 2021, 21(3): 843. DOI: 10.3390/s21030843.

[6] Wang DL, Sattiraju RR, Qiu AJ, et al. Effect of retransmissions on the performance of C-V2X communication for 5G. IEEE 92nd Vehicular Technology Conference (VTC2020-Fall), 2020. DOI: 10.1109/VTC2020-Fall49728.2020.9348560.

[7] Liu Q, Liang P, Xia JJ, et al. A highly accurate positioning solution for C-V2X systems. Sensors, 2021, 21(4): 1175. DOI: 10.3390/s21041175.

[8] Molina-Masegosa R, Gozalvez J, Sepulcre M. Configuration of the C-V2X mode 4 sidelink PC5 interface for vehicular communications. 14th International Conference on Mobile Ad-hoc and Sensor Networks (MSN 2018), 2018: 43-48. DOI: 10.1109/MSN.2018.00014.

[9] Yoshioka S, Nagata S. Cellular V2X standardization in 4G and 5G. IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences, 2022, E105A(5): 754-762. DOI: 10.1587/transfun.2021WBI0001.

[10] Park J, Kavathekar V, Bhuduri S, et al. A co-operative perception system for collision avoidance using C-V2X and client-server-based object detection. Sensors, 2025, 25(17): 5544. DOI: 10.3390/s25175544.

[11] Ficzere D, Varga P, Wippelhauser A, Hejazi H, Csernyava O, Kovács A, Hegedus C. Large-scale cellular vehicle-to-everything deployments based on 5G – critical challenges, solutions, and vision towards 6G: A survey. Sensors, 2023, 23(16): 7031. DOI: 10.3390/s23167031.

[12] Chen SZ, Hu JL, Shi Y, et al. A vision of C-V2X: Technologies, field testing, and challenges with Chinese development. IEEE Internet of Things Journal, 2020, 7(5): 3872-3881. DOI: 10.1109/JIOT.2020.2974823.

[13] Chen ML, Chai R, Hu H, et al. Performance evaluation of C-V2X mode 4 communications. IEEE Wireless Communications and Networking Conference (WCNC), 2021. DOI: 10.1109/WCNC49053.2021.9417517.

[14] Li P, Wu K, Cheng Y, et al. How does C-V2X perform in urban environments? Results from real-world experiments on urban arterials. IEEE Transactions on Intelligent Vehicles, 2024, 9(1): 2520-2530. DOI: 10.1109/TIV.2023.3326735.

[15] Ning RR, Zhang XK, Feng WY, et al. Performance analysis of vehicle platoon communication in C-V2X autonomous mode. IEEE 23rd International Conference on High Performance Switching and Routing (HPSR), 2022: 41-46. DOI: 10.1109/HPSR54439.2022.9831348.

[16] Kumar RD, Rammohan A. Revolutionizing intelligent transportation systems with cellular vehicle-to-everything (C-V2X) technology: Current trends, use cases, emerging technologies, standardization bodies, industry analytics and future directions. Vehicular Communications, 2023, 43: 100638. DOI: 10.1016/j.vehcom.2023.100638.

[17] Hua QZ, Yu KP, Wen Z, Sato T. A novel base-station selection strategy for cellular vehicle-to-everything (C-V2X) communications. Applied Sciences-Basel, 2019, 9(3): 556. DOI: 10.3390/app9030556.

[18] Twardokus G, Rahbari H. Vehicle-to-nothing? Securing C-V2X against protocol-aware DoS attacks. IEEE Conference on Computer Communications (INFOCOM), 2022: 1629-1638. DOI: 10.1109/INFOCOM48880.2022.9796667.
[19] Chen SZ, Li Q, Wang Y, et al. C-V2X equipment identification management and authentication mechanism. China Communications, 2021, 18(8): 297-306.
[20] Flowers B, Ku YJ, Baidya S, et al. Utilizing reinforcement learning for adaptive sensor data sharing over C-V2X communications. IEEE Transactions on Vehicular Technology, 2024, 73(3): 4051-4066. DOI: 10.1109/TVT.2023.3322068.
[21] Brady C, Cao L, Roy S. Modeling of NR C-V2X mode 2 throughput. IEEE International Workshop on Communications Quality and Reliability (CQR), 2022: 19-24. DOI: 10.1109/CQR54764.2022.9918559.

World Journal of Engineering Research
Print ISSN: 2959-9865
Online ISSN: 2959-9873
DOI: https://doi.org/10.61784/wjer3051

CARBON FOOTPRINT ACCOUNTING AND EVALUATION OF AUTOMOTIVE AIR CONDITIONING FILTERS BASED ON LIFE CYCLE ASSESSMENT (LCA)

Jing Li[1], XiaoNa Hu[2], Ke Chen[2*], Chen Cui[3]
[1]CATARC Huacheng Certification (Tianjin) Co., Ltd., Tianjin 300300, China.
[2]CATARC Research Management Science Research (Tianjin) Co., Ltd., Tianjin 300300, China.
[3]CATARC Automotive Components Test Center (Ningbo) Co., Ltd., Ningbo 315104, Zhejiang, China.
Corresponding author: Ke Chen, Email: chenke@catarc.ac.cn

Abstract: Driven by the "dual carbon" goals to promote the green transformation of the automotive industry, the full-life-cycle carbon footprint of automotive components has become a core focus of the industry's low-carbon development. This study takes automotive air conditioning filters as the research object. Based on the Life Cycle Assessment (LCA) methodology, it defines the carbon footprint accounting boundary covering the "raw material acquisition - production - transportation - end-of-life" process, constructs a carbon footprint calculation model, and conducts full-life-cycle carbon footprint accounting for three typical automotive air conditioning filters. The results show that the raw material acquisition stage is the main contributor to the carbon footprint of automotive air conditioning filters, accounting for more than 60% of the total. Furthermore, approaches to reduce the product's carbon footprint are proposed, including material substitution, process optimization, energy and auxiliary material upgrading, and waste recycling. The research results provide a theoretical basis and data support for the low-carbon design, production optimization of automotive air conditioning filters, and carbon management of the automotive industry chain.
Keywords: Life Cycle Assessment; Automotive air conditioning filter; Carbon footprint

1 INTRODUCTION

Under the dual impetus of global "dual carbon" goals and the green transformation of the automotive industry, the management and control of the full-life-cycle carbon footprint of automobiles have become a core issue for the sustainable development of the industry[1,2]. As important components contributing to the carbon footprint, automotive parts have gradually become a research focus[3,4]. As a key component to ensure the air quality inside vehicles, automotive air conditioning filters are characterized by short service life, high replacement frequency, and large market demand[5]. The cumulative effect of carbon emissions throughout their full process (production, use, and disposal) cannot be ignored.

At present, a relatively mature system has been formed for LCA research on core automotive components such as engines and power batteries. However, there remains a significant gap in carbon footprint accounting for consumable parts like air conditioning filters. Existing studies mostly focus on a single link (e.g., environmental impact analysis of filter media production) and lack systematic consideration of the full life cycle, which includes "raw material acquisition - component manufacturing - logistics and transportation - end-use - waste disposal"[6,7].

Therefore, this study takes automotive air conditioning filters as the research object, constructs a full-life-cycle carbon footprint accounting model based on the LCA methodology, identifies carbon emission links in each stage and key carbon emission nodes, and provides three case studies to apply the model. The research results can provide directions for filter manufacturers to optimize low-carbon design, offer data support for the automotive industry to formulate carbon footprint management standards for components, and hold important theoretical and practical significance for promoting the green and low-carbon transformation of the automotive industry chain.

2 CARBON FOOTPRINT ACCOUNTING METHOD FOR AUTOMOTIVE AIR CONDITIONING FILTERS

Life Cycle Assessment (LCA) is an environmental impact assessment method defined by the International Organization for Standardization (ISO) 14040/14044 series standards. By quantifying the resource consumption and environmental emissions of a product throughout its "cradle-to-grave" life cycle (raw material acquisition, production, transportation, use, and waste disposal), it identifies key environmental impact nodes and provides a scientific basis for low-carbon design and decision-making[8,9]. Against the background of the green transformation of the automotive industry, LCA has become a core tool for carbon footprint accounting of components, enabling systematic quantification of carbon emissions throughout the process of consumable parts such as air conditioning filters and supporting the low-carbon optimization of the industry chain.

2.1 Definition of Carbon Footprint Accounting Boundary

In accordance with the requirements of the Standard T/CAS 957—2024 Carbon Footprint of Products—Product Category Rules—Automotive Air Conditioning Filters, the system boundary of automotive air conditioning filters shall in principle cover all stages of the life cycle, including the raw material acquisition stage, production stage, transportation stage, and recycling stage. The comparison boundary of automotive air conditioning filters is a part of the system boundary, including the raw material acquisition stage and production stage, to ensure the comparability of carbon footprints among products before delivery. The schematic diagram of the two boundaries is shown in Figure 1.

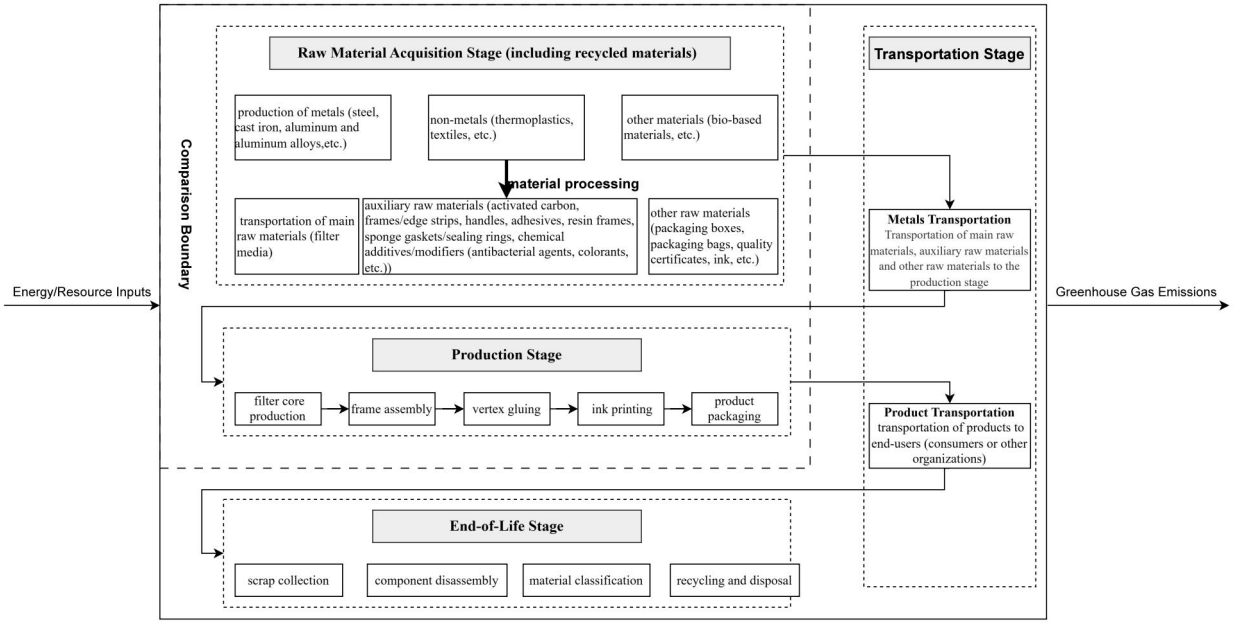


Figure 1 Accounting Boundaries

The raw material acquisition stage begins with the extraction of raw materials and concludes when the raw materials leave the premises of the raw material processing enterprise, encompassing processes such as raw material extraction, processing and purification, and production and processing; meanwhile, the production and processing of recycled materials should include processes like converting waste into recycled materials. The raw materials of automotive air conditioning filters are generally categorized into three types: main raw materials, which are primarily filter media; auxiliary raw materials, including activated carbon, frames (edge strips), handles, adhesives, resin frames, sponge gaskets (sealing rings), and chemical additives/modifiers (such as antibacterial agents and colorants); and other raw materials, which refer to accessories including packaging boxes, packaging bags, quality certificates, ink, instruction manuals, and other packaging materials.

The production stage starts when raw materials enter the premises of the automotive air conditioning filter manufacturing plant and ends when the final product leaves the plant. It mainly includes the total carbon emissions from processes such as filter core production, frame assembly, vertex gluing, ink printing, and product packaging, involving fuel consumption, purchased electricity, purchased heat, and direct emissions. Direct emissions mainly consider the leakage of fire-extinguishing agents and refrigerants. Processes such as product design and investment in production infrastructure are not included in this stage.

The transportation stage consists of two parts: raw material transportation, which starts from when main raw materials, auxiliary raw materials, and other raw materials leave the premises of the raw material production plant and ends when they enter the premises of the air conditioning filter manufacturing plant; and product transportation, which starts from when the products leave the premises of the air conditioning filter manufacturing plant and ends when they enter the premises of distribution centers or sales points.

The end-of-life stage starts when waste automotive air conditioning filters enter the premises of a scrap treatment plant and ends when the product disposal is completed. It mainly includes processes such as collection and management of waste products, disassembly of waste product components, sorting and classification, recycling, and disposal.

2.2 Data Collection and Sources

The data sources for the carbon footprint accounting of automotive air conditioning filters based on Life Cycle Assessment (LCA) are mainly divided into two categories: primary data refers to data obtained by directly measuring or calculating based on direct measurement results to quantify relevant processes or activities, for example, in the raw material acquisition stage, data on the acquisition of main raw materials (such as filter media) and data on energy and resource consumption (such as electricity) in the production stage should be collected as primary data; secondary data refers to data that do not meet the requirements of primary data, including data from databases, published literature, and default emission factors from national databases, for instance, greenhouse gas emission and removal factors related to the transportation link can be collected as secondary data.

2.3 Carbon Footprint Calculation Model

The carbon footprint calculation model for the full life cycle of automotive air conditioning filters is as follows:

$$CFP = C_{\text{Material}} + C_{\text{Production}} + C_{\text{Transportation}} + C_{\text{End-of-Life}} \tag{1}$$

Where:

CFP——Full-life-cycle carbon footprint of an automotive air conditioning filter within the accounting cycle under the system boundary, in kilograms of carbon dioxide equivalent ($kgCO_2e$);

$C_{Material}$——Carbon emissions from the raw material acquisition stage, $kgCO_2e$;

$C_{Production}$——Carbon emissions from the production stage, $kgCO_2e$;

$C_{Transportation}$——Carbon emissions from the transportation stage, $kgCO_2e$;

$C_{End\text{-}of\text{-}Life}$——Carbon emissions from the end-of-life stage, $kgCO_2e$.

2.3.1 Raw material acquisition stage

Carbon emissions from the raw material acquisition stage consider emissions from the production and processing of various raw materials, and the model is as follows:

$$C_{\text{Material}} = \sum \left(AD_{\text{i}} \times CEF_{\text{i}} \right) \tag{2}$$

Where:

$C_{Material}$——Carbon emissions from the raw material acquisition stage, $kgCO_2e$;

AD_i——Quality of raw material i input within the accounting cycle, kg;

CEF_i——Carbon emission factor for the production and processing of raw material i, $kgCO_2e/kg$.

2.3.2 Production stage

Carbon emissions in the production stage consider the total carbon emissions from fuel consumption, purchased electricity, purchased heat, and emissions from the production process (e.g., leakage) during processes such as filter core production, frame assembly, vertex gluing, ink printing, and product packaging. The model is as follows:

$$C_{\text{Production}} = C_{\text{Fuel}} + C_{\text{Electricity}} + C_{\text{Heat}} + C_{\text{Leakage}} \tag{3}$$

(1)Carbon emissions from fuel in the production stage is as follows:

$$C_{\text{Fuel}} = \sum \left(E_{\text{r}} \times CEF_r + E_{\text{r}} \times NCV_r \times CEF_r' \right) \tag{4}$$

Where:

C_{Fuel}——Carbon emissions from fuel combustion in the production stage, $kgCO_2e$;

E_r——Consumption of fuel r, m^3, L, or kg;

CEF_r——Carbon emission factor for the production of fuel r, $kgCO_2e/m^3$, $kgCO_2e/L$, $kgCO_2e/kg$;

CEF'_r——Carbon emission factor for the use of fuel r, tCO_2e/GJ;

NCV_r——Average net calorific value of fuel r, GJ/t, $GJ/10^4m^3$.

Carbon emissions from purchased electricity in the production stage is as follows:

$$C_{\text{Electricity}} = \sum \left(AD_{\text{i}} \times CEF_{\text{i, Electricity}} \right) \tag{5}$$

Where:

$C_{Electricity}$——Carbon emissions from purchased electricity production in the production stage, $kgCO_2e$;

AD_i——Purchase volume of electricity type i, kWh;

$CEF_{i,Electricity}$——Carbon emission factor for the production of electricity type i, $kgCO_2e/kWh$.

Carbon emissions from purchased heat in the production stage is as follows:

$$C_{\text{Heat}} = \sum \left(AD_{\text{i}} \times CEF_{\text{i, Heat}} \right) \tag{6}$$

Where:

C_{Heat}—— Carbon emissions from purchased heat production in the production stage, $kgCO_2e$;

AD_i——Purchase volume of heat type i, GJ;

$CEF_{i,Heat}$——Carbon emission factor for the production of heat type i, $kgCO_2e/GJ$.

Direct leakage mainly considers carbon emissions from greenhouse gas leakage of fire-fighting equipment and refrigeration equipment is as follows:

$$\begin{aligned} C_{\text{Leakage}} &= \sum \left(AD_{\text{i, fire-fighting}I} \times \eta_{\text{fire-fighting}I} \times GWP_{\text{i}} \right) \\ &+ \sum \left(AD_{\text{i, refrigeraton}R} \times \eta_{\text{refrigeraton}R} \times GWP_{\text{i}} \right) \end{aligned} \tag{7}$$

Where:

$C_{Leakage}$——Carbon emissions from direct leakage of fire-fighting equipment and refrigeration equipment in the production stage, $kgCO_2e$;

$AD_{i,\text{fire-fighting I}}$——Filling amount of the i-th type of greenhouse gas in fire-fighting equipment I, kg;

$\eta_{i,\text{fire–fighting I}}$——Leakage coefficient of the i-th type of greenhouse gas in fire-fighting equipment I;

$AD_{i,\text{refrigeration R}}$——Filling amount of the i-th type of greenhouse gas in refrigeration equipment R, kg;

$\eta_{i,refrigeration\ R}$——Leakage coefficient of the i-th type of greenhouse gas in refrigeration equipment R;
GWP_i——Global warming potential of the i-th type of greenhouse gas;

2.3.3 Transportation stage

Carbon emissions from the transportation stage is as follows:

$$C_{\text{Transportation}} = \sum\left(AD_{i,r} \times L_{i,r} \times CEF_r\right) \tag{8}$$

Where:
$C_{Transportation}$——Carbon emissions from the transportation stage, $kgCO_2e$;
$AD_{i,r}$——Quality of the i-th type of goods (including raw materials and air conditioning filter products) transported by the r-th type of transportation mode, t;
$L_{i,r}$——Transportation distance of the i-th type of goods transported by the r-th type of transportation mode, km;
CEF_r——Carbon emission factor of the r-th type of transportation mode, $kgCO_2e/(t \cdot km)$.

2.3.4 End-of-life stage

Carbon emissions in the end-of-life stage include emissions from recycling and disposal is as follows:

$$C_{\text{End-of-Life}} = \sum\left(AD_i \times CEF_i\right) + \sum\left(AD_r \times CEF_r\right) \tag{9}$$

Where:
$C_{End-of-Life}$—— Carbon emissions in the end-of-life stage of the product, $kgCO_2e$;
AD_i——Quality of waste treated by the i-th type of waste disposal method, kg;
CEF_i——Carbon emission factor of the i-th type of waste disposal method, $kgCO_2e/kg$.
AD_r——Quality of the r-th type of recycled material, kg;
CEF_r——Carbon emission factor of the r-th type of recycled material, $kgCO_2e/kg$.

3 ACCOUNTING RESULTS

In this study, the carbon footprint data for three distinct types of automotive air conditioning filters (i.e., Product A, Product B, and Product C) were collected, calculated, and statistically analyzed in line with the calculation model outlined above.

According to the accounting model, the carbon footprints of Product A, Product B, and Product C in the comparison phase are 435.32 gCO_2e, 1087.90 gCO_2e, and 936.00 gCO_2e, respectively. Specifically, for Product A: carbon emissions in the raw material acquisition stage are 267.88 gCO_2e, accounting for 62% of its total product carbon footprint. Among this, the filter media and frames contribute 91% to the carbon footprint of the raw material acquisition stage. The carbon emission during the production stage is 167.44 gCO_2e, accounting for 38% of the product's total carbon footprint. For Product B: carbon emissions in the raw material acquisition stage are 967.14 gCO_2e, accounting for 89% of its total product carbon footprint. Among this, the filter media and frames contribute 76% to the carbon footprint of the raw material acquisition stage. The carbon emission during the production stage is 120.76 gCO_2e, accounting for 11% of its total carbon footprint. For Product C: carbon emissions in the raw material acquisition stage are 896.13 gCO_2e, accounting for 96% of its total product carbon footprint. Among this, the filter media and frames contribute 90% to the carbon footprint of the raw material acquisition stage. The carbon emission during the production stage is 39.87 gCO_2e, accounting for 4% of its total carbon footprint. The emission status of each product is presented in the table below and Figure 2.

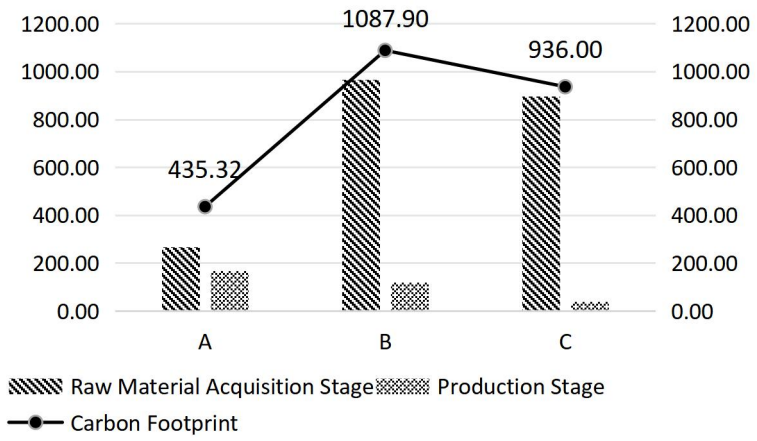


Figure 2 Product Carbon Footprint Analysis Chart (gCO_2e)

4 CONCLUSIONS

Based on the Life Cycle Assessment (LCA) method, this study conducted carbon footprint accounting for three

automotive air conditioning filters. The proposed model can distinguish the carbon footprints of different products.

Taking a medium-sized factory as an example: If the annual production capacity of automotive air conditioning filters is 500,000 units, the total annual carbon footprint of the products would be 220tCO_2e if the carbon footprint of a single product is relatively low (0.44kgCO_2e), and 545tCO_2e if the carbon footprint of a single product is relatively high (1.09kgCO_2e). These figures indicate that the carbon emissions are not negligible, further confirming the necessity of managing the carbon footprint of automotive air conditioning filters.

This study shows that the carbon footprint of automotive air conditioning filters can be effectively reduced from the following three aspects:

(1) At the product design stage, process optimization involves adopting a modular detachable design and green, low-energy-consuming processes. Multiple filter media with different functions are used, and each medium is made of either a single recyclable chemical component or multiple easily separable chemical components to improve the filter element design. This provides convenience for recycling during the production of filter elements and their disposal after use, thereby reducing the consumption of resources and energy.

(2) The raw material acquisition stage is the main contributing link to the carbon footprint of automotive air conditioning filters throughout their entire life cycle, accounting for more than 60% of the total carbon footprint. Among the raw materials, filtering and supporting materials—such as PP (polypropylene), PET (polyethylene terephthalate), and activated carbon—account for more than 75% of the carbon emissions from raw materials. Replacing traditional virgin raw materials with bio-based filter materials or recycled materials can significantly reduce the extraction burden of virgin resources and the carbon emissions generated during raw material production.

(3) At the product production stage, electricity is the main energy source. The indirect emissions from electricity use in the production process can be greatly reduced by two measures: first, replacing thermal power with solar energy or wind energy for production; second, using high-efficiency equipment to reduce energy consumption. In addition, classifying and collecting leftover materials generated from processes such as filter material cutting and frame injection molding during production, then crushing and reintroducing these materials into the production line, can reduce raw material waste and thereby lower the product's carbon footprint.

These measures can effectively manage the key nodes of carbon footprints, and provide a scientific basis and feasible path for the low-carbon design of automotive air conditioning filters as well as carbon management in the component industry chain.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Long Suhua, Wei Changqing. Study and Analysis on Carbon Emissions of Automobiles Throughout the Full Life Cycle. Automotive Technology & Materials, 2024(07): 1-8.

[2] He X J. Prospects of Green and Low-Carbon Development in the Automotive Industry. Shanghai Auto, 2023(10): 29-32.

[3] Jiang Shuncheng, Zhang Jie, Zhu Yan. Study on Carbon Emissions of Automobile Body Stamping Parts Based on Life Cycle Assessment. Automotive Science & Technology, 2025(02): 14-18.

[4] Wu G G, Long S H, Zhang T Z, et al. Research on Carbon Footprint Accounting of Automotive Components' Life Cycle. China Auto, 2023, (3): 7-13+25.

[5] Liu S. Research on Performance of Automotive Air Conditioning Filter. E3S Web of Conferences, 2021, 248, 01057. DOI: https://doi.org/10.1051/e3sconf/202124801057.

[6] Wu Guigen, Long Suhua, Zhang Tongzhu, et al. Study on Carbon Footprint Accounting of Auto Parts Throughout the Life Cycle. China Auto, 2023(03): 7-13+25.

[7] Tong S K, Zhao K F, Liang H H, et al. Research on Carbon Footprint of Magnesium Alloy Rear Hatchback Door Based on Full Life Cycle Assessment. Foundry Engineering, 2024, 48(2): 53-56.

[8] Tushar Q, Bhuiyan A M, Abunada Z, et al. Carbon Footprint and Uncertainties of Geopolymer Concrete Production: A Comprehensive Life Cycle Assessment (LCA). C — Journal of Carbon Research, 2025, 11(3): 55.

[9] Wen Y G, Lei C X, Xu W, et al. Quantitative Assessment of Carbon Footprint of Automotive Tempered Glass Based on LCA. China Auto, 2022(11): 7-10+37.

World Journal of Engineering Research
Print ISSN: 2959-9865
Online ISSN: 2959-9873
DOI: https://doi.org/10.61784/wjer3052

DYNAMIC SIMULATION AND ANALYSIS OF MIXER TRUCK DEVICE ROLLOVER

YuHuan Gao, JingJing Xia*
Huaiyin Institute of Technology, Huaian 223200, Jiangsu, China.
Corresponding Author: JingJing Xia, Email: 806892554@qq.com

Abstract: To verify the feasibility and rationality of the mixer truck device design and reduce the risk of rollover accidents, this study, based on the ADAMS multi-body dynamics software, conducts 3D modeling, simulation model simplification, multi-body dynamics modeling and simulation, and simulation result analysis for the mixer truck device, with a focus on exploring the dynamic characteristics under rollover conditions to provide theoretical support for device optimization; in the multi-body dynamics modeling and simulation stage, the ADAMS software is used to build a virtual prototype, where the 3D model is imported in STEP format, and the simulation results show that the dynamic characteristics of the mixer truck are consistent with the actual working conditions—during turning, the displacement in the X-direction is greater than that in the Z-direction with consistent curve trends, the velocities in the X and Z directions are the same, and the maximum total velocity is 1.6 m/s, while under the rollover condition, the displacement of the center of mass in the Y-direction is -5 mm at the initial moment and reaches 348.3 mm at the moment of rollover, the initial velocity is 0, reaches a maximum velocity of 2.1 m/s at 4 seconds and then decreases gradually, and the velocity of the center of mass in the Z-direction is the largest; through systematic multi-body dynamics simulation, this study clarifies the variation laws of displacement and velocity of the mixer truck under turning and rollover conditions, verifies the rationality of the device design, and provides data support and theoretical basis for the structural optimization of the mixer truck device and the prevention and control of rollover risks.
Keywords: Mixer truck device; Adams; Rollover conditions; Dynamic characteristics

1 INTRODUCTION

With the development of the new era, concrete mixer trucks are being used with increasing frequency. The rollover of concrete mixer trucks can cause severe damage to people's lives and property[1,2]. Therefore, verifying the feasibility and rationality of mixer truck device design is of great significance for reducing the risk of rollover accidents.Domestic experts and scholars have also conducted extensive research on the operating conditions of vehicles during cornering[3]. For example, Deng Yadong et al. established a multi-body dynamics simulation model of vehicles based on ADAMS software and analyzed the steering characteristics when a step input is applied to the vehicle's steering wheel angle. Researchers from Chang'an University, including Wei Lang, analyzed the process of vehicle rollover motion on the basis of establishing a three-dimensional vehicle dynamics model. Professor Yu Qiang from the same university has also conducted in-depth research on the issue of improving vehicle roll stability through active suspension systems.Studying the variation laws of displacement and velocity of mixer trucks under cornering and rollover conditions, and verifying the rationality of device design, provides important data support for the structural optimization of mixer truck devices and the prevention and control of rollover risks[4].

2 Theoretical Foundations of Multi-Body Dynamics

Multi-body dynamics refers to a discipline whose research object is a system composed of multiple rigid or flexible bodies interconnected through certain constraint relationships and performing relative motions. These constraints are usually non-holonomic constraints, scleronomic constraints, and rheonomic constraints. Based on classical mechanics, multi-body dynamics has gradually developed and possesses broad application value and prospects in the fields of mechanical kinematics and dynamics.Currently, multi-body dynamics has been integrated with modern computing technology, with its basic task being the derivation of mathematical models suitable for computer programming. Closely linked to the development of contemporary computer technology, it has gained widespread recognition internationally.According to whether flexible bodies are included in the model, multi-body dynamics can be divided into multi-rigid-body dynamics and multi-flexible-body dynamics[5]. When a motion system operates at a low translational or rotational speed, and the influence of external forces on its deformation is negligible, its simulation can be treated as a multi-rigid-body system. Conversely, it is necessary to model the main moving components of the system as flexible bodies. However, since the motion speed of this mechanism is relatively low, the dynamic simulation is performed based on the multi-rigid-body theory[6].

The coordinate transformation matrix between the centroid reference coordinate system of the mechanical system component and the ground coordinate system is given by Equation (1).

$$A^{\mathrm{gi}} = \begin{bmatrix} \cos\psi\cos\phi - \sin\psi\cos\theta\sin\phi & -\cos\psi\sin\phi - \sin\psi\cos\theta\cos\phi & \sin\psi\sin\theta \\ \sin\psi\cos\phi + \cos\psi\cos\theta\sin\phi & -\sin\psi\sin\phi + \cos\psi\cos\theta\cos\phi & -\cos\psi\sin\theta \\ \sin\theta\sin\phi & \sin\theta\cos\phi & \cos\theta \end{bmatrix} \tag{1}$$

Define an Euler rotation axis coordinate system, whose three unit vectors correspond to the axes of the three Euler rotations mentioned above; thus, the three axes are not mutually perpendicular[7]. The coordinate transformation matrix from this coordinate system to the component's centroid coordinate system is given by Equation (2).

$$B = \begin{bmatrix} \sin\theta\sin\phi & 0 & \cos\theta \\ \sin\theta\cos\phi & 0 & -\sin\theta \\ \cos\theta & 0 & 1 \end{bmatrix} \tag{2}$$

The angular velocity of the component is expressed as Equation (3).

$$\omega = B\dot{\gamma} \tag{3}$$

In ADAMS, the variables introduced are the components of angular velocity in the Euler rotation axis coordinate system as Equation (4).

$$\omega_e = \dot{\gamma} \tag{4}$$

Considering the constraint equations, the system's dynamic equations can be derived as follows using the energy form of Lagrange's first equation with Lagrange multipliers in ADAMS. Here, T represents the kinetic energy expressed in terms of the system's generalized coordinates, as given by Equation (5).

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_j}\right) - \frac{\partial T}{\partial q_j} = Q_j + \sum_{i=1}^{n} \lambda_i \frac{\partial \phi}{\partial q_j} \tag{5}$$

In ADAMS, the generalized momentum is further introduced by Equation (6).

$$P_j = \frac{\partial T}{\partial \dot{q}_j} \tag{6}$$

Simplify the expression of the constraint reaction force as Equation (7).

$$C_j = \sum_{i=1}^{n} \lambda_i \frac{\partial \phi}{\partial q_j} \tag{7}$$

Thus, the equation is simplified to Equation (8).

$$\dot{P}_j - \frac{\partial T}{\partial q_j} = Q_j - C_j \tag{8}$$

The kinetic energy is further expressed as Equation (9).

$$T = \frac{1}{2}\dot{R}^T M\dot{R} + \frac{1}{2}\dot{\gamma}^T B^T JB\dot{\gamma} \tag{9}$$

Among them, M is the mass matrix of the component, and J is the inertia matrix of the component in the centroid coordinate system.

Express the following equation in terms of the translational direction and rotational direction respectively, as shown in Equations (10)-(12).

$$\dot{P}_R - \frac{\partial T}{\partial q_R} = Q_R - C_R \tag{10}$$

$$\dot{P}_\gamma - \frac{\partial T}{\partial q_\gamma} = Q_\gamma - C_\gamma \tag{11}$$

$$\dot{P}_R = \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_R}\right) = \frac{d}{dt}\left(M\dot{R}\right) = M\dot{V}, \frac{\partial T}{\partial q_R} = 0 \tag{12}$$

The equation is simplified to Equation (13)-(14).

$$M\dot{V} = Q_R - C_R \tag{13}$$

$$P_\gamma = \left(\frac{\partial T}{\partial \dot{q}_\gamma}\right) = B^T JB\dot{\gamma} \tag{14}$$

Matrix B contains Euler angles. To simplify the derivation, each component in ADAMS is thus assigned the following 12 variables as Equation (15)-(16).

$$\begin{aligned} V &= \left[V_x, V_y, V_z\right]^T \\ R &= \left[x, y, z\right]^T \\ P_\gamma &= \left[P_\psi, P_\theta, P_\phi\right] \\ \omega_e &= \left[\omega_\psi, \omega_\theta, \omega_\phi\right]^T \\ \gamma &= \left[\psi, \theta, \phi\right]^T \end{aligned} \tag{15}$$

$$\begin{aligned} M\dot{V} &= Q_R - C_R \\ V &= \dot{R} \\ \dot{P}_\gamma - \frac{\partial T}{\partial q_\gamma} &= Q_\gamma - C_\gamma \\ P_\gamma &= B^T JB\omega_e \\ \omega_e &= \dot{\gamma} \end{aligned} \tag{16}$$

By integrating the constraint equations, ADAMS can automatically establish the system's dynamic equations and differential-algebraic equations, thereby solving system dynamics as shown in Equations (17).

$$\begin{aligned} &\dot{P} - \frac{\partial T}{\partial q} + \phi_q^T \lambda + H^T F = 0 \\ &P = \frac{\partial T}{\partial \dot{q}} \\ &u = \dot{q} \\ &\phi(q,t) = 0 \\ &F = f(u,q,t) \end{aligned} \tag{17}$$

Among them, P is the generalized momentum of the system, and H is the coordinate transformation matrix of external forces.

3 Model Establishment

3.1 Three-Dimensional Solid Modeling of the Mixer Truck Device

Since the mixer truck device must bear the force of the load during operation and therefore needs to work stably and safely for a long time, it is necessary to ensure that it has sufficient strength and stiffness in the design process to avoid dangers caused by fractures due to insufficient strength[8]. The three-dimensional modeling steps are shown in Figure 1.

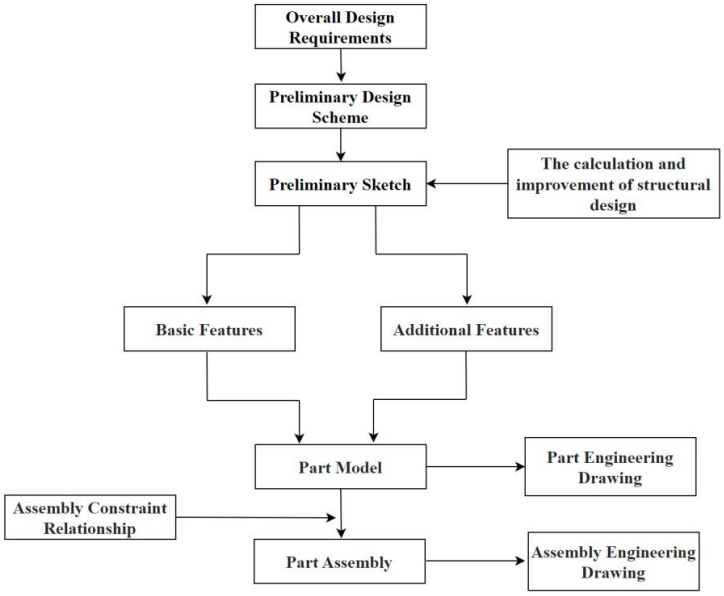


Figure 1 Three-Dimensional Modeling Steps

The three-dimensional simulation model of the mixer truck device is shown in Figure 2.

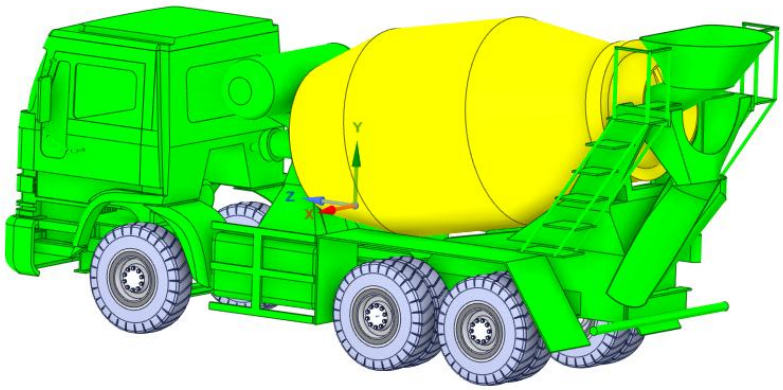


Figure 2 Three-Dimensional Simulation Model of the Mixer Truck Device

3.2 Model Import

First, save the mixer truck device as a Parasolid (*.x-t) or STEP file format, then import it into the ADAMS dynamic simulation software. When the mixer truck device model is imported into ADAMS 2020, there is only a relative positional relationship between its components. During simulation, various constraint pairs are manually added according to the actual motion state and constraint relationship, and marker points are automatically generated, as shown in Figure 3.

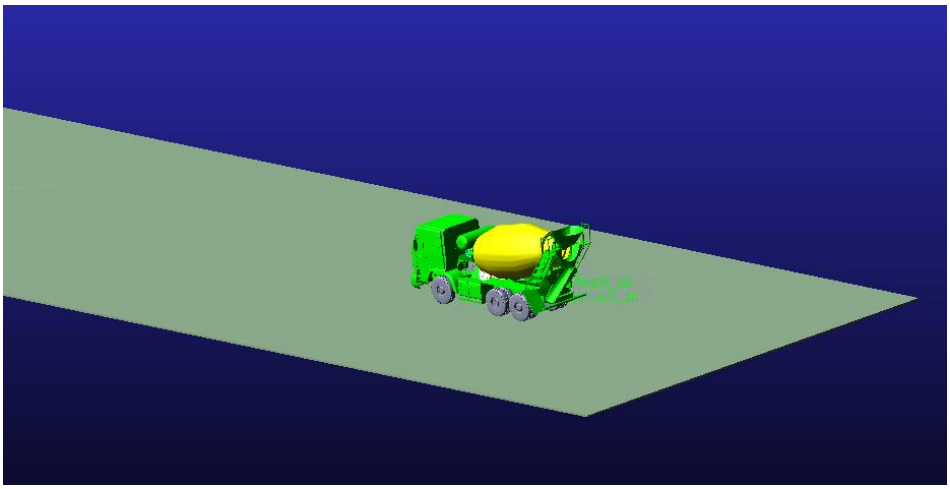


Figure 3 Dynamic Simulation Model of the Mixer Truck Device

3.3 Material Definition

Since the generation of dynamic effects in ADAMS relies on the mass properties of the model's components, it is necessary to define the material for each part of the model when conducting dynamic analysis. The software automatically calculates the mass of each component based on the material's density and volume, and finally generates dynamic effects according to the mass and gravitational acceleration. Newly imported models have no mass information, which needs to be defined and set manually by the user[9]. To modify and set the mass information, right-click on all components, and select "geometric shape and material type" as the mass definition method. After the settings are completed, the model can be verified to prevent missing the property settings of any component. The material definition of the mixer truck device parts is shown in Figure 4.

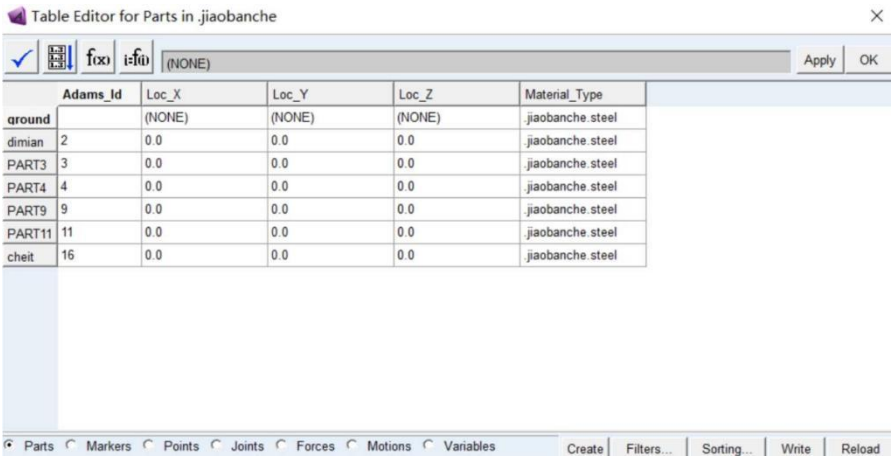


Figure 4 Material Definition of the Mixer Truck Device

3.4 Constraint Relationships of the Mixer Truck Device

When the mixer truck device model is imported into ADAMS software, there is only a relative positional relationship between its components, similar to objects floating in three-dimensional space. Each object has six degrees of freedom, namely three translational degrees of freedom along the positive direction of the spatial coordinate axes and three rotational degrees of freedom around the coordinate axes. If their mutual relationships are not defined, the simulation will fail due to insufficient degrees of freedom. Therefore, it is necessary to establish constraints for the relevant components of the model according to the working conditions. When constraints are established between components, marker points are automatically generated to determine the position of the constraint pairs, and these marker points are created along with the establishment of constraint relationships[10].

The quantity and type of each constraint pair of the mixer truck device are shown in Table 1.

Table 1 Quantity and Type of Each Constraint Pair of the Mixer Truck Device

Constraint Type	Number of Constraint Pairs	Number of Degrees of Freedom Restricted by Constraint Pairs
Revolute Joint	3	5
Fixed Joint	2	6

According to the working conditions of the mixer truck device, the constraint relationships after being added are shown in Figure 5 and Figure 6 below.

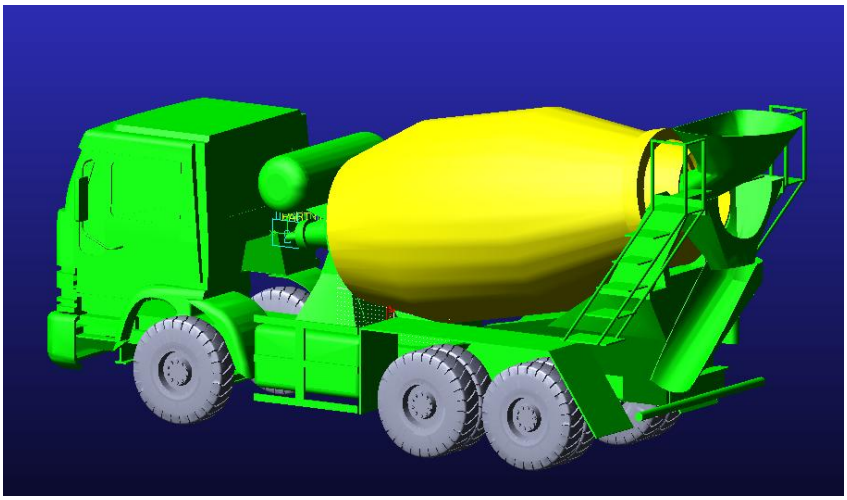


Figure 5 Three-Dimensional Model of Mixer Truck

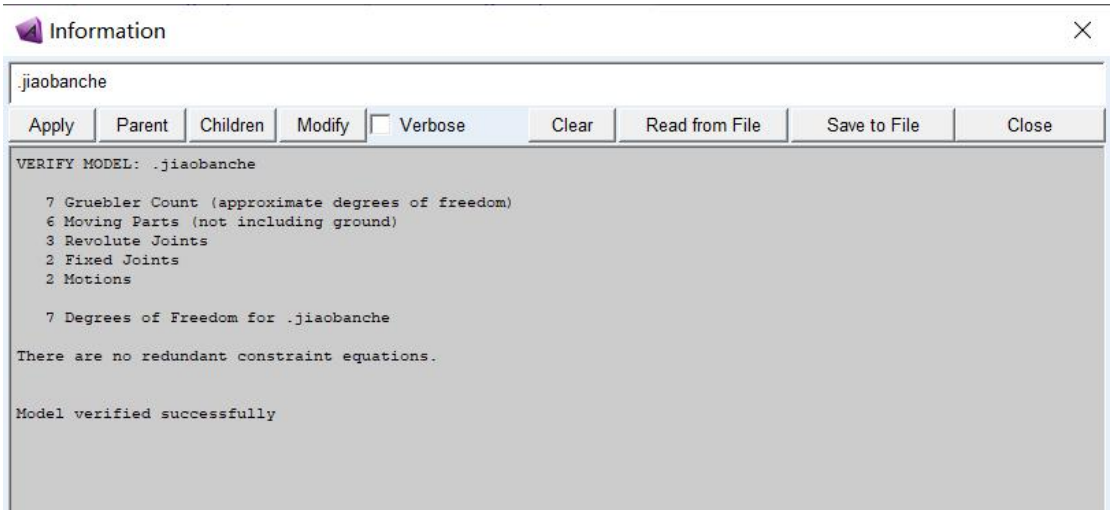


Figure 6 Constraint Relationships

3.5 Contact Relationships of the Mixer Truck Device

ADAMS uses Coulomb friction to calculate the frictional force between contacting parts, whose value is equal to the product of the impact normal force and the friction coefficient. When the relative motion velocity of two contacting objects is less than the static friction transition velocity v, the friction coefficient between them is taken as the static friction coefficient; when the relative motion velocity of the two contacting objects is greater than the dynamic friction transition velocity v, the friction coefficient between them is taken as the dynamic friction coefficient. Generally speaking, the dynamic friction coefficient and static friction coefficient are related to factors such as the material properties and surface roughness of the two contacting objects. The material properties are shown in Figure 7.

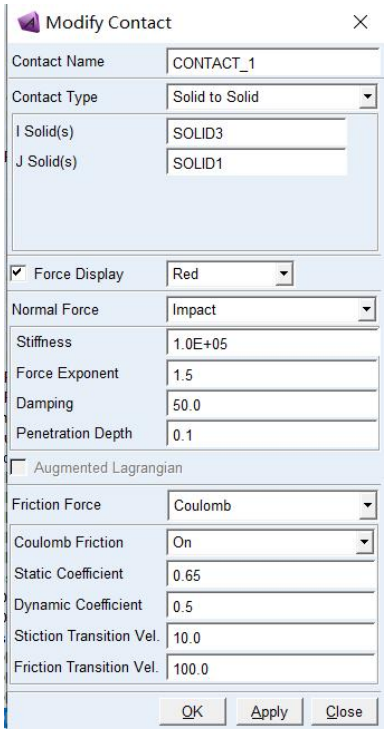


Figure 7 Material Properties

There are generally two types of drives: one is forward drive, and the other is inverse drive. Forward drive involves determining the motion of each rigid body in the system based on the known driving force applied to the rigid body system; while inverse drive involves calculating the force on each rigid body in the system based on the known motion of each rigid

body. In this paper, when conducting simulation analysis on the mixer truck device, a velocity drive is adopted, as shown in Figure 8 below.

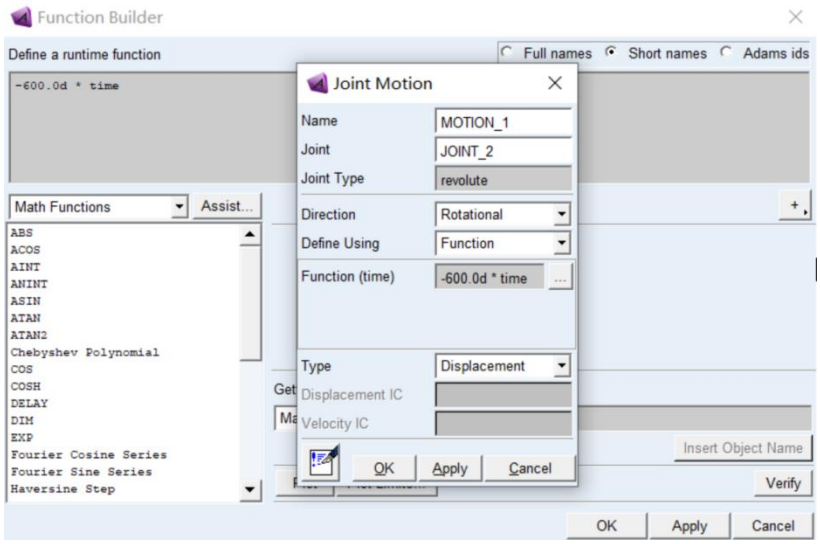


Figure 8 Velocity Drive Constraint

4 SIMULATION AND RESULT ANALYSIS

4.1 Simulation Setup

After completing the import of the mixer truck device model, unit definition, material property definition, constraint addition, and drive addition as above, the multi-body dynamics model of the mixer truck device has been fully established. Then, it is necessary to set the simulation analysis parameters, simulation time, and simulation steps, and finally perform the simulation calculation.

In terms of solver types, multi-body dynamics simulation analysis mainly has two forms: interactive simulation and scripted simulation. Scripted simulation is mainly used for multi-body systems where certain kinematic pairs, loads, drives, etc., need to be deactivated or activated at specific moments during motion. The simulation process requires scripting, which is relatively cumbersome to operate. Interactive simulation is the most commonly used simulation form, with fewer control parameters and simple operation, which can meet the simulation analysis needs of conventional multi-body systems. In summary, interactive simulation is adopted for the simulation analysis of the mixer truck device.

Interactive simulation mainly has two control parameters: simulation time and simulation steps. The simulation time is set to 5 seconds. Simulation steps have an important impact on the results of simulation analysis. Since ADAMS analysis results are generally curves of the kinematic and dynamic properties of each component, these curves are essentially composed of a series of discrete points solved by the software solver. Therefore, the number of simulation steps needs to match the simulation time. If the simulation time is long but the number of simulation steps is too small, although the simulation time can be reduced and the simulation efficiency improved, some point data may be missing, which affects the simulation accuracy. Conversely, increasing the number of simulation steps can effectively improve the accuracy of simulation results, but it will reduce the calculation efficiency of the entire multi-body system during simulation, and may even cause the computer to crash. Therefore, to ensure the accuracy of the simulation results while taking into account the simulation efficiency, the number of simulation steps is set to 1000. The step size is shown in Figure 9.

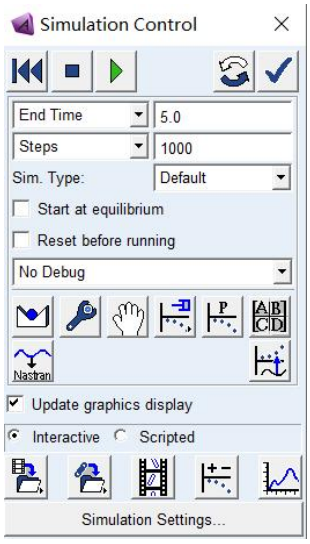


Figure 9 Simulation Step Size Setup

4.2 Simulation Result Analysis

Figures 10-12 show the variation curve of the centroid displacement in the y-direction when the mixer truck rolls over under the action of an external moment during a turn. It can be seen from the figures that at the initial moment, the displacement in the y-direction is -5 mm; at the instant when the mixer truck rolls over while turning, the displacement in the y-direction is 348.3 mm.This indicates that the center of mass has shifted significantly in the lateral direction, exceeding the range where the mixer truck can maintain stability, which is a key manifestation of the rollover. After the rollover occurs, the displacement of the center of mass in the Y-direction drops rapidly, reflecting the process of the center of mass falling back after the mixer truck completes the rollover, which is consistent with the actual situation.

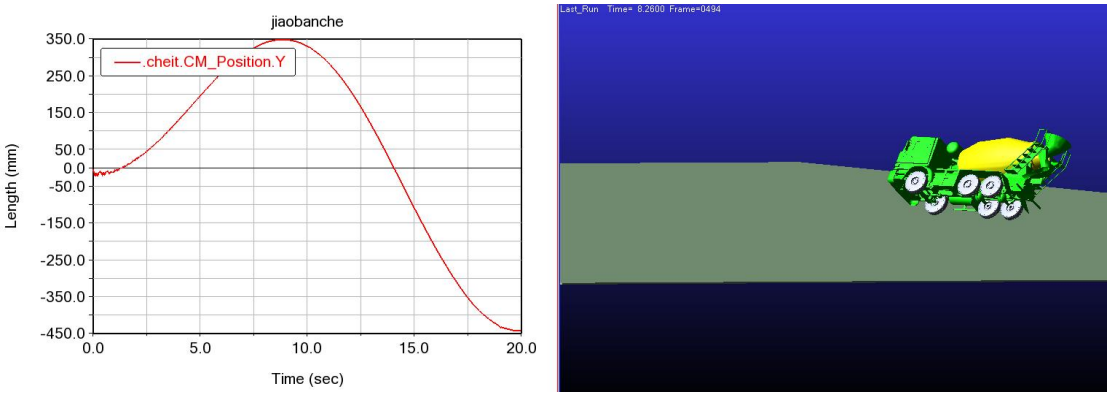


Figure 10 Centroid Displacement Diagram

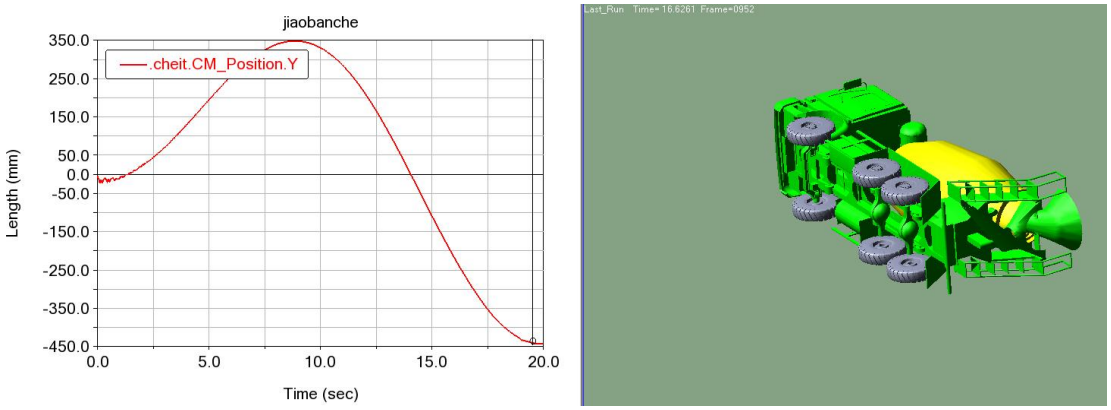


Figure 11 Centroid Displacement Diagram

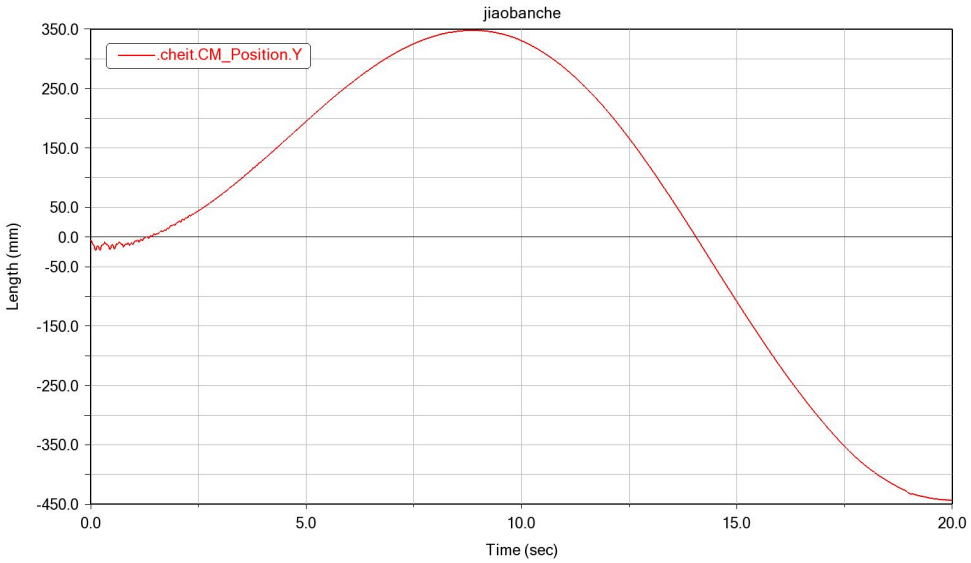


Figure 12 Centroid Displacement Diagram

In the initial stage, the velocity of the centroid of the mixer truck increases rapidly, which simulates the process of the mixer truck accelerating to the target driving speed before turning. After that, the velocity remains in a stable stage of about 2000 mm/s, simulating the mixer truck driving at a stable speed and entering the turning condition. When the rollover occurs, the velocity gradually decreases, reflecting the situation that the mixer truck's speed decreases due to factors such as impact during the rollover process.The velocity curves of the mixer truck when rolling over during a turn are shown in Figures 13-14.

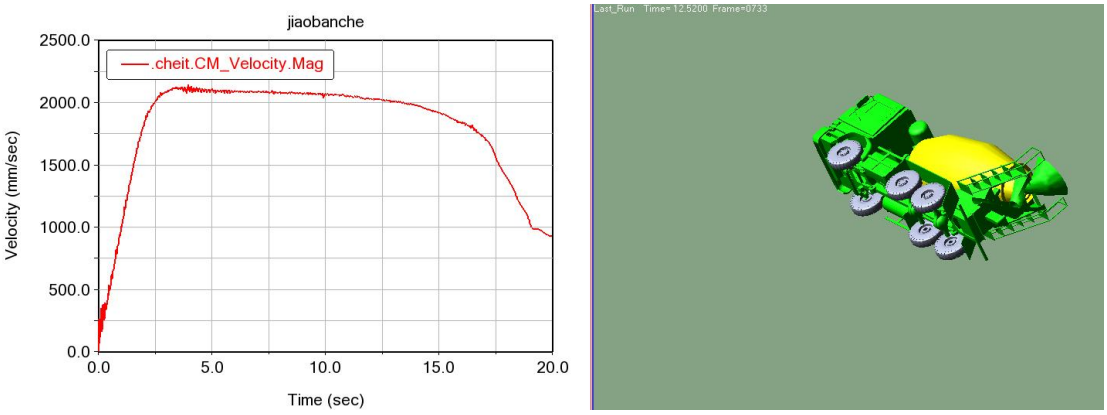


Figure 13 Velocity Curve Diagram

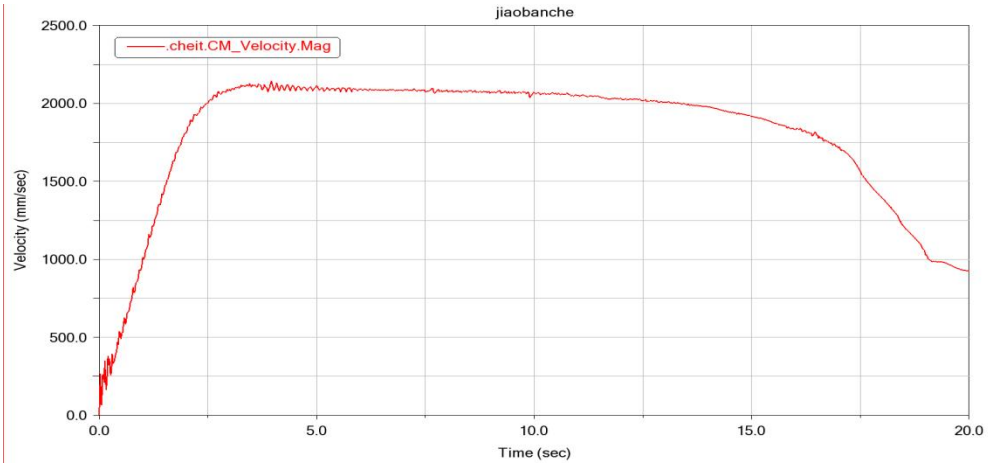


Figure 14 Velocity Curve Diagram

The overall fluctuation is small, indicating that the change in lateral speed is relatively insignificant during both normal driving and rollover. The rollover is mainly caused by excessive lateral displacement of the center of mass, rather than being dominated by a significant sudden change in lateral speed.

It rises initially, then remains stable, and decreases significantly after the rollover occurs. This reflects the change in the mixer truck's vertical movement state, with vertical movement affected during the rollover, which is consistent with the actual situation.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCE

[1] Deng R, Tan Y, Zhang H, et al. Experimental and DEM studies on the transition of axial segregation in a truck mixer. Powder Technology, 2016, 314: 148-163. DOI: 10.1016/j.powtec.2016.08.013.

[2] The Stress Analysis of Planetary Gear System of Mixer Reducer for Concrete Mixer Truck. Journal of Drive and Control, 2015, 12(4): 77-81.

[3] Hung T T, Khanh N D. Analysis the dynamic rollover threshold and roll safety thresholds of 4-axle truck vehicle: a simulation approach. International Journal on Interactive Design and Manufacturing (IJIDeM), 2024, 19(7): 1-16. DOI: 10.1007/S12008-024-02117-1.

[4] Nurzaki I, Ahmad S, Rahizar R. The effect of vehicle and road conditions on rollover of commercial heavy vehicles during cornering: a simulation approach. Sustainability, 2021, 13(11): 6337-6337. DOI: 10.3390/SU13116337.

[5] A C, R T, M D, et al. Effect of roll motion control on vehicle lateral stability and rollover avoidance. Proceedings of the American Control Conference, 2020, 2020(July): 4868-4874. DOI: 10.23919/ACC45564.2020.9147735.

[6] Genge Z, Yi W, Chengwei J, et al. Multiobjective optimization of vehicle handling and stability based on ADAMS. Mathematical Problems in Engineering, 2022, 2022: 3245251. DOI: 10.1155/2022/3245251.

[7] Liangyuan S, Cui B. Research on modal characteristics of vehicle chassis transmission shafting based on ADAMS. IOP Conference Series: Earth and Environmental Science, 2021, 769(4): 042111. DOI: 10.1088/1755-1315/769/4/042111.

[8] Penny W C W, Els S P. The test and simulation of ABS on rough, non-deformable terrains. Journal of Terramechanics, 2016, 67: 1-10. DOI: 10.1016/j.jterra.2016.05.001.

[9] Wang C, He X, Shen X, et al. Analysis of handling stability of hydraulic hybrid vehicle based on ADAMS/Car simulation. IOP Conference Series: Earth and Environmental Science, 2018, 186(5): 012049. DOI: 10.1088/1755-1315/186/5/012049.

[10] Wang S, Guan W, Mao Y, et al. Dynamic simulation method for rollover protective structures in engineering vehicles and human injury assessment. Biosystems Engineering, 2025, 257: 104224. DOI: 10.1016/J.BIOSYSTEMSENG.2025.104224.

World Journal of Engineering Research
Print ISSN: 2959-9865
Online ISSN: 2959-9873
DOI: https://doi.org/10.61784/wjer3050

DESIGN AND MANUFACTURE OF INTELLIGENT FOLLOWING CAR

WenFang Yang*, Guang Luo, BuSheng Luo, ZeYu Jiang
Power Supply Service Center, Huizhou Power Supply Bureau, Huizhou 516000, Guangdong, China.
Corresponding Author: WenFang Yang, Email:yangwenfang@gdhz.csg.cn

Abstract: This paper describes the whole process of the design and manufacture of the intelligent following car. The STM32 main control chip is used to integrate ultrasonic positioning and motor control technology to realize dynamic tracking of operators. It focuses on solving the problems of mobile stability (climbing 15 degrees), endurance (4 H/20 km) and multi-target recognition. Through the actual measurement and verification of Huizhou Power Supply Bureau, the device is easy to operate and the efficiency of safety supervision is improved by 300%.
Keywords: Following car; STM32; Power safety; Hardware design; Field test

1 INTRODUCTION

With the rapid development of information technology, artificial intelligence and automation technology, the safety supervision and management of operation site has gradually transited from the traditional manual inspection mode to the intelligent and automated management mode. Especially in high-risk industries such as power, transportation and construction, the safety management and control of the operation site is particularly important. Based on the policy requirements of "Regulations on Safety Production Management of China Southern Power Grid Co., Ltd." And "Rules on Intelligent Safety Supervision and Management of Guangdong Power Grid Co., Ltd.", the construction of safety production supervision system is gradually developing towards technicalization, informatization and intellectualization, especially video surveillance and information system. It provides a new means and method for on-site safety supervision[1,2].

Although the intelligent safety supervision system has been widely used in the power industry, it still faces many challenges and problems in practice, especially in the deployment and use of monitoring equipment. A major challenge is the real-time and full coverage of video surveillance. Due to the frequent transfer of workers and the complexity of the field environment, the traditional video surveillance equipment is usually difficult to achieve the whole process and no dead angle supervision of workers. If the monitoring equipment fails to follow the position change of the operator in time, the illegal operation may not be found and stopped in time. The fundamental reason for this problem is that the traditional video surveillance equipment needs manual intervention to adjust, and in the busy work site, this manual intervention often lags behind[3].

2 CURRENT STATE OF RESEARCH DOMESTICALLY AND ABROAD

Intelligent follow-me vehicles, as a vital branch of mobile robotics, have seen their technological evolution consistently centered around three core modules: environmental perception, motion control, and human-machine interaction.

At the environmental perception level, early research primarily relied on single sensors (such as infrared or ultrasonic), but limitations like narrow detection ranges and susceptibility to environmental interference made it difficult to meet the demands of complex scenarios. With the maturation of multi-sensor fusion technology, modern systems widely adopt collaborative solutions integrating lidar, visual cameras, and inertial measurement units (IMUs). For instance, lidar constructs 3D point cloud maps, visual SLAM enables dynamic obstacle recognition, while IMUs compensate for short-term motion blur, elevating positioning accuracy to the centimeter level.

In motion control, traditional PID controllers—challenging to parameterize—are increasingly replaced by intelligent algorithms like adaptive fuzzy control and neural networks. In 2023, a MIT team introduced a reinforcement learning framework that trained controllers by simulating millions of collision scenarios, reducing trajectory tracking errors by 42% in congested environments.

Human-machine interaction is shifting from preset commands to intent understanding. Natural Language Processing (NLP)-based voice-following systems (e.g., Amazon Astro) now recognize ambiguous instructions (e.g., “maintain a one-meter distance”) and integrate gesture recognition to enhance interaction flexibility. Notably, modular design has emerged as a new trend. For instance, ROS2 middleware decouples perception-decision-execution, allowing developers to assemble different algorithm modules like building blocks, significantly lowering the R&D threshold. Domestic and international research exhibit distinct divergences in technical approaches. These differences stem not only from disparities in foundational scientific research but are also profoundly influenced by industrial ecosystems and policy orientations

Domestic research prioritizes engineering implementation and cost control. For instance, in agricultural applications, China Agricultural University's orchard-following vehicle employs a UWB + visual fusion solution. By dynamically

adjusting sensor weights (e.g., enhancing visual data weighting on cloudy days), it maintains ±10cm tracking accuracy while reducing costs to one-fifth of comparable products. This pragmatic approach aligns closely with China's manufacturing supply chain strengths—electronics hubs like Huaqiangbei in Shenzhen rapidly deliver customized sensor modules, while Huawei's Ascend chip optimizations have accelerated edge computing adoption in tracking systems. In contrast, European and American research emphasizes algorithmic breakthroughs and standardization: Stanford University's Neural Radiance Fields (NeRF) technology models environments as differentiable 3D scenes, enabling carts to achieve high-precision following after just 20 minutes of learning in unfamiliar settings. This algorithmic innovation relies on their long-term accumulation of computer vision theoretical frameworks. Regarding industrial ecosystems, China has developed a vertically integrated model represented by Baidu Apollo and DJI, where companies independently develop the entire stack from hardware to algorithms. Overseas, however, a modular division of labor prevails—for instance, Waymo focuses on algorithm development while sourcing sensors from Velodyne. This divergence results in faster product iteration domestically but weaker foundational technology accumulation. On the policy front, China's new infrastructure initiatives are accelerating the deployment of vehicle-road coordination facilities, providing embedded communication base stations for follow-me vehicles. Meanwhile, the EU's GDPR data privacy regulations limit the scale of visual data training, driving European researchers toward developing federated learning frameworks. Notably, Japan's aging society has driven the development of compliant control algorithms for care robots. Their follow systems adjust speed in real-time via force feedback, preventing pressure on mobility-impaired individuals. This scenario-specific innovation demonstrates deep integration between technological pathways and societal needs.

Current intelligent follow-me vehicle technology faces three core bottlenecks: insufficient adaptability to dynamic environments, inefficient multi-system coordination, and the challenge of balancing energy consumption with cost. Regarding dynamic environment adaptability, existing systems exhibit significant performance degradation in complex scenarios such as strong light interference or rainy/snowy weather—for instance, vision-based follow systems can reach up to 30% misjudgment rates under backlight conditions, while LiDAR generates substantial noise in heavy rain due to droplet reflections. The constraints on multi-system coordination efficiency are even more pronounced: when perception, decision-making, and execution modules adopt heterogeneous architectures, data conversion delays can reach up to 200ms, causing tracking lag during high-speed movement. Furthermore, high-precision sensors (e.g., 128-line LiDAR) and computing platforms (e.g., NVIDIA Orin chips) account for over 60% of total power consumption, limiting commercial product battery life to under 4 hours. The conflict between cost control and performance enhancement is particularly acute—for instance, improving positioning accuracy from ±5cm to ±2cm in an industrial-grade follow-me vehicle would increase its price by 40%, directly undermining market competitiveness. These bottlenecks fundamentally reflect inherent contradictions in the current stage of technological development: On one hand, algorithm complexity grows exponentially, yet hardware advancements remain constrained by the slowing pace of Moore's Law. On the other hand, user demands for reliability and cost-effectiveness continue to rise, forcing developers to repeatedly balance parameter optimization with engineering feasibility.

3 DESIGN SCHEME

3.1 Expected Objectives

Objective 1: The remote control calling function can realize the maximum remote control distance of 30 meters without any blocking interference;
Objective 2: The maximum endurance is 4 hours and the maximum mileage is 20 km;
Objective 3: The maximum moving speed is 4km/H, and the maximum climbing gradient is 15 degrees;
Objective 4: Maximum load: 30KG;

3.2 Technical Indicators

Battery life: The robot shall have at least 4 hours of continuous working time to meet the needs of long-term on-site monitoring. The battery charging time shall not exceed 2 hours.
Recognition range: The robot shall be able to recognize and follow the target within a certain range, and the maximum recognition distance can reach 5-10 meters.
Recognition speed: The robot should be able to respond and adjust quickly in a dynamic environment at a speed of 1-2 m/s to ensure real-time following.
Communication protocol: The robot shall support wireless communication.
Moving speed: The moving speed of the robot should meet the needs of the operation site, usually adjustable between 0.5-1.5 m/s, to ensure that it can keep up with the movement of the operator and flexibly respond to changes in the working environment.
Maximum climbing ability: The robot shall have certain climbing ability and be able to run stably on the ground with a slope of ≤ 15 °.

3.3 Technical Route

The first stage: site investigation, mainly for project data collection and research, so as to clarify the design objectives, and track the latest progress of relevant technologies at home and abroad according to the design objectives.
The second stage is to put forward the specific design scheme, which includes the completion of the overall scheme design of how to realize following and portrait tracking, the completion of the scheme design of equipment structure and connection relationship, the use of relevant software to draw the design scheme diagram, and the discussion in many ways to clarify the final scheme.
The third stage is to start the trial production of following function and portrait tracking, which includes the procurement of raw materials, the processing of modules, and finally the assembly of the whole machine.
The fourth stage is to inspect and test the following function and the portrait tracking function. Debug the equipment, find out the existing problems and improve and upgrade the equipment according to the problems.
The fifth stage: the equipment test is successful and the technical documents of the design are completed.
The sixth stage: After many improvements, the following function and portrait tracking tool are obtained.

3.4 Technical Points

The function of the control panel is to control the motor to rotate according to the reported label coordinate information, so as to achieve the effect of following the obstacle avoidance car.
Onboard resources are as follows: CPU: 32F103RBT6, LQFP64, Flash: 128K, RAM: 20K
1 power indicator 5V
3 status indicators (LED 1, LED 2, LED 3)
2 keys, 1 of which is the reset key
1 USB communication selection switch
3-way ultrasonic sensor
1 X 0.96-inch OLED
1 active buzzer
One USB serial port, which can be used to communicate with the upper program
1 power input DC connection for the voltage range 12-24 V
1 power switch to control the power supply of the whole board
4 DC motor control interfaces and control chips
Part of IO ports are led out for subsequent expansion.
Wherein the control board and the Stlink burning pins DIO and CLK are led out, the burning pins of the control board on the silk screen are DIO _ C and CLK _ C, and the burning pins are DIO _ N and CLK _ N

3.5 Remote Control Mode

In this working mode, the control panel will always listen to the instructions of the remote controller, and there is a gyroscope inside the remote controller to judge forward, brake, left turn and right turn.
Specific use steps:
Step 1: Press the code key on the remote control, and then tilt the remote control forward to move the car forward.
Step 2: Press the code key on the remote control, and then the car will turn left when the remote control tilts forward to the left.
Step 3: Press the code key on the remote control, and then the car will turn right when the remote control tilts forward to the right.
Step 4: Press the code key on the remote control, and then the car will brake when the remote control tilts backward.

4 CONCLUSION

The results of this project have been applied in Huizhou Power Supply Bureau. The device can accurately judge and search by simulating all kinds of following, following and climbing in the family compound. It fully verifies that the device can quickly find the function of the person in charge of the work or the guardian to follow, and the effect is good. The monitoring equipment can follow the change of the position of the operator in time, and can find the problem that the operator can not find the illegal operation in time. The invention has the advantages of light volume and weight, convenient carrying, simple and clear wiring, simple operation and easy learning and use; It adds front-line production personnel and production tools for power production enterprises, and has remarkable economic benefits. As a part of the intelligent safety supervision system, it aims to solve the problem that the existing video surveillance equipment can not effectively follow the operators, and improve the safety supervision level of the operation site[4-7].
The future development of intelligent follow-me vehicle technology will feature three major breakthrough directions: human-like upgrades to perception systems, distributed restructuring of decision-making architectures, and intelligent innovations in energy management. At the perception level, the integrated application of quantum dot sensors and event cameras will significantly enhance environmental analysis capabilities. Quantum dot materials offer three times the photoelectric conversion efficiency of traditional CMOS sensors, enabling sub-millimeter distance measurement in low-light environments. Event cameras capture dynamic targets with microsecond-level response speeds. Combining these technologies reduces misjudgment rates to below 5% under extreme conditions such as backlighting, rain, and fog. The decision-making architecture will evolve toward an edge-cloud collaborative model, leveraging 5G-MEC (Mobile Edge

Computing) for millisecond-level task allocation: local edge nodes handle real-time obstacle avoidance and other low-latency tasks, while the cloud completes complex path planning. This architecture boosts multi-vehicle coordination efficiency by 40% while reducing communication energy consumption by 30%. In energy management, solid-state batteries combined with wireless charging technology will overcome endurance limitations. Toyota's sulfur-based solid-state battery achieves an energy density of 500 Wh/kg—twice that of existing lithium batteries. Paired with embedded road wireless charging systems, it enables 24/7 uninterrupted operation for follow-up vehicles. Notably, the introduction of bio-inspired algorithms (such as ant colony optimization) will elevate collective intelligence, enabling multiple vehicles to autonomously form dynamic formations in warehousing and logistics scenarios. This approach boosts material handling efficiency by 60% compared to manual dispatch. These technological breakthroughs will propel the follow-up system from a simple tool to an intelligent agent, ultimately achieving seamless collaboration with humans.

The evolution of intelligent follow-me cart technology will profoundly reshape human-machine collaboration paradigms. Its future applications will transcend traditional industrial boundaries, penetrating broader civilian domains. In healthcare, carts equipped with flexible robotic arms can serve as elderly care assistants, using millimeter-wave radar to monitor fall risks and autonomously deliver medications to bedside. Japan has piloted such systems, reducing care response times to 8 seconds. In commercial retail, dynamic shelf-following systems will revolutionize traditional store layouts—as customers browse, smart carts automatically form mobile display units, dynamically adjusting merchandise arrangements via visual recognition. Amazon trials indicate this technology increases average order value by 30%. In urban governance, environmental carts equipped with gas sensors autonomously map pollution hotspots. Their swarm-collaborative mode completes air quality surveys across 5 square kilometers within 30 minutes, achieving 40% greater data accuracy than fixed monitoring stations. Even more revolutionary is its cognitive enhancement application in education: through brain-computer interface integration with the follow-me cart, children with autism can use the cart's movement patterns for social skills training. MIT experiments demonstrate this approach increases eye contact frequency by 2.3 times. This dual-driven approach—where technology meets societal need—is propelling the follow-me system from an execution tool to a social intelligence node, ultimately achieving deep cognitive and emotional synergy between machines and humans[8].

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Chen FH, Yan BS, Gao YY, et al. STM32-based intelligent following trolley. Journal of Physics: Conference Series, 2025, 2975(1): 012014. DOI: 10.1088/1742-6596/2975/1/012014.

[2] Zhao R, Zhang H, Qi F. Intelligent following car based on Ultra-wideband technology. Journal of Physics: Conference Series, 2024, 2816(1): 012088. DOI: 10.1088/1742-6596/2816/1/012088.

[3] Sun Y. Fuzzy PID control theory in multi-objective optimal longitudinal following of intelligent trolley. Applied Mathematics and Nonlinear Sciences, 2024, 9(1). DOI: 10.2478/AMNS.2023.1.00345.

[4] Chen N, Zhang D, Guo J, et al. Intelligent following car based on dual detection positioning using ultrasonic and camera. Journal of Artificial Intelligence Practice, 2023, 6(5). DOI: 10.23977/JAIP.2023.060509.

[5] Sun S, Tian C, Xie L, et al. Home care intelligent follow-up car based on ESP32. Computer Programming Skills and Maintenance, 2025(4): 145-150. DOI: 10.16184/j.cnki.comprg.2025.04.037.

[6] Lin W, Chen Z, Li X, et al. Intelligent auxiliary follow-up car driving system based on STM32. Electronic Production, 2025, 33(8): 45-49. DOI: 10.16589/j.cnki.cn11-3571/tn.2025.08.027.

[7] Yang M, Zhao Z, Zhao J, et al. Design and implementation of automatic following multifunctional car. Modern Information Technology, 2024, 8(15): 10-14. DOI: 10.19850/j.cnki.2096-4706.2024.15.003.

[8] Gan L, Mo H, Huang K. Research on the design of intelligent car based on ROS. Electronic Quality, 2024(3): 23-26.

World Journal of Engineering Research
Print ISSN: 2959-9865
Online ISSN: 2959-9873
DOI: https://doi.org/10.61784/wjer3054

FAULT DIAGNOSIS OF ELEVATOR UNBALANCED LOAD BASED ON PARAMETER-OPTIMIZED SVM

SaiNan Wang[1], Xian Zhou[1*], YunTao Yang[2]
[1]*Hunan Electrical College of Technology, Xiangtan 411101, Hunan, China.*
[2]*School of Physics & Electronics, Hunan University, Changsha 410082, Hunan, China.*
Corresponding Author: Xian Zhou, Email: 460174335@qq.com

Abstract: To address the fault diagnosis of elevator unbalanced loads, this study proposes a fault diagnosis method based on a parameter-optimized support vector machine (SVM). By establishing a dynamic model of the elevator system, fault features are extracted, and an improved particle swarm optimization algorithm is applied to optimize the key parameters of the SVM, thereby constructing an efficient fault diagnosis model. Experimental results indicate that the proposed method significantly outperforms traditional diagnostic approaches in terms of fault classification accuracy, and can effectively identify the unbalanced load state of elevators. The research outcome offers a new technical solution for elevator fault diagnosis and holds significant engineering application value for ensuring the safe operation of elevators.
Keywords: Elevator fault diagnosis; Unbalanced load; Particle swarm optimization algorithm; Fault feature extraction

1 INTRODUCTION

As an indispensable vertical transportation system in high-rise buildings, the safety performance of elevators is directly related to the safety of lives and property. With the acceleration of urbanization, the number of elevators has surged. Unbalanced load fault, a common type of elevator malfunction, accounts for a relatively high proportion of elevator failures. It not only disrupts normal operation but may also lead to serious incidents such as passenger entrapment or even falling, has become a prominent social concern. Traditional fault detection methods primarily rely on manual periodic inspections, which exhibit significant limitations in both efficiency and accuracy. In recent years, intelligent fault diagnosis techniques, represented by the Support Vector Machine (SVM), have become a research hotspot in this field due to their capability for real-time monitoring and precise diagnosis. However, the performance of an SVM model is highly dependent on the selection of its hyperparameters. How to efficiently optimize these parameters to enhance the accuracy and generalization ability of the diagnostic model is a critical issue requiring urgent solution in current research.

To address this, this study aims to develop an intelligent diagnostic method based on a parameter-optimized SVM specifically for elevator unbalanced load faults. The research will first conduct an in-depth analysis of the formation mechanism and vibration characteristics of unbalanced load faults to provide a theoretical basis for feature extraction. Subsequently, it will focus on studying parameter optimization strategies for SVM. By introducing advanced algorithms such as an improved Particle Swarm Optimization, it will perform adaptive optimization of the SVM's kernel function parameters and penalty factor, aiming to overcome the limitations of traditional methods like the substantial computational burden of grid search and the tendency of some intelligent algorithms to fall into local optima. The innovation of this paper is mainly reflected in proposing an SVM model that incorporates a hybrid kernel function approach to handle the complexity of fault data, and employing an enhanced optimization algorithm to improve the efficiency and precision of the parameter search. Finally, the effectiveness of the proposed method will be validated through experiments. It is expected to provide a more accurate and efficient technical pathway for diagnosing elevator unbalanced load faults, thereby offering substantial support for ensuring elevator operational safety and promoting technological advancement in the industry.

2 REVIEW OF RELEVANT RESEARCH

2.1 Mechanism of Elevator Unbalanced Load

The unbalanced load fault is one of the common issues during elevator operation, and its mechanism involves the combined effects of multiple factors. The formation of an unbalanced load in elevators is primarily related to uneven weight distribution, dynamic load variations during operation, and structural design flaws. Firstly, uneven weight distribution may lead to an unbalanced load. Differences in weight between the elevator car and the counterweight, as well as uneven distribution of passengers or goods within the car, can cause load imbalance. When the weight on one side of the car exceeds that on the counterweight side, tilting occurs during operation, subjecting the elevator structure to additional stress. Secondly, dynamic load changes during elevator operation can also induce unbalanced loads. Passenger boarding and alighting, loading and unloading of goods, and the start-stop cycles of the elevator contribute to

dynamic load variations. These changes can alter the weight distribution, thereby affecting the elevator's balance. Furthermore, structural design defects are another cause of unbalanced loads [1-3]. Unreasonable designs of structural components such as guide rails, ropes, and suspension systems may generate additional vibrations and partial loads during operation, exacerbating the unbalanced load issue. In terms of fault characteristic analysis for unbalanced loads, the acquisition and feature extraction of vibration signals are critical steps. Vibration signals reflect the dynamic response of the elevator during operation, and by collecting vibration signals from the car or guide rails, relevant information about the unbalanced load can be obtained. Time-domain feature analysis is a common method for diagnosing unbalanced load faults. Through time-domain analysis of vibration signals, characteristic parameters such as peak values, mean values, and variance can be extracted to describe the dynamic behavior of the elevator load. Frequency-domain analysis, on the other hand, transforms vibration signals into the frequency domain to obtain frequency characteristics via spectrum analysis. Additionally, time-frequency domain analysis combines the advantages of both time and frequency domains, providing a more comprehensive description of the dynamic behavior of unbalanced loads. Research shows that the characteristics of unbalanced load faults are reflected in the energy distribution, frequency characteristics, and time-frequency features of vibration signals. By appropriately selecting and extracting these features, effective support can be provided for subsequent fault diagnosis. In summary, studying the mechanism of elevator unbalanced loads is of great significance for understanding fault characteristics and enabling intelligent fault diagnosis. Through in-depth analysis of the causes and characteristics of unbalanced loads, a theoretical foundation and practical guidance can be provided for the development of elevator fault diagnosis technologies.

2.2 Support Vector Machine Theory

The Support Vector Machine (SVM) is an effective classification method, particularly suited for solving small-sample problems, as illustrated in Figure 1. Regarding multi-class classification strategies, SVM extends the original binary classification problem to multi-class scenarios through various methods. Common multi-class strategies include One-vs-All (OvA), One-vs-One (OvO), and decision tree-based strategies. In the One-vs-All strategy, one SVM classifier is trained for each class to distinguish it from all other classes. This means that for N classes, N SVM classifiers need to be trained. During classification, an input sample is fed into all N classifiers, and the class with the highest score is selected as the prediction result. The advantage of this method is its simple implementation, but it may suffer from performance degradation when some classes are not sufficiently distinct from others. The One-vs-One strategy, on the other hand, trains one SVM classifier for each pair of classes, requiring N*(N-1)/2 classifiers for N classes. During classification, each sample is evaluated by all classifiers, and a voting mechanism determines the final class label. This method considers the relative differences between classes and generally outperforms the One-vs-All strategy, albeit at a higher computational cost [2-5]. Another multi-class strategy is the decision tree-based approach, which constructs multiple binary SVM classifiers and combines them using a decision tree structure. Each node in the tree represents a binary classification problem, recursively partitioning the classes until leaf nodes are reached. This method performs well when dealing with hierarchically structured categories. The choice of kernel function is crucial for SVM performance. Common kernel functions include the linear kernel, polynomial kernel, Radial Basis Function (RBF) kernel, and Sigmoid kernel. For linearly inseparable problems, the RBF kernel is widely used due to its ability to map data into a high-dimensional space. The selection of a kernel function depends on the data distribution and problem complexity. The performance of a multi-class SVM is influenced not only by the kernel function but also by the model's hyperparameter settings. Hyperparameter optimization is a key step in improving the classifier's generalization ability. Research shows that optimizing hyperparameters can effectively enhance the classifier's performance. Common hyperparameter optimization methods include grid search, genetic algorithms, particle swarm optimization, and Bayesian optimization. In the application of multi-class SVM, handling imbalanced datasets is another important consideration. Since the number of samples in each class is often uneven in real-world datasets, the classifier may become biased towards the majority class. One approach to address this issue is to use data resampling techniques, such as random oversampling of the minority class or undersampling of the majority class, or synthetic sample generation techniques like the SMOTE algorithm. Furthermore, in practical applications of multi-class SVM, model interpretability and real-time performance must also be considered. Although SVM theoretically offers good performance guarantees, its model complexity is relatively high, making it less interpretable [4-5]. Additionally, for applications requiring real-time responses, the prediction speed of SVM may become a limiting factor. Therefore, maintaining classification performance while improving model interpretability and real-time capability represents an important direction for current research.

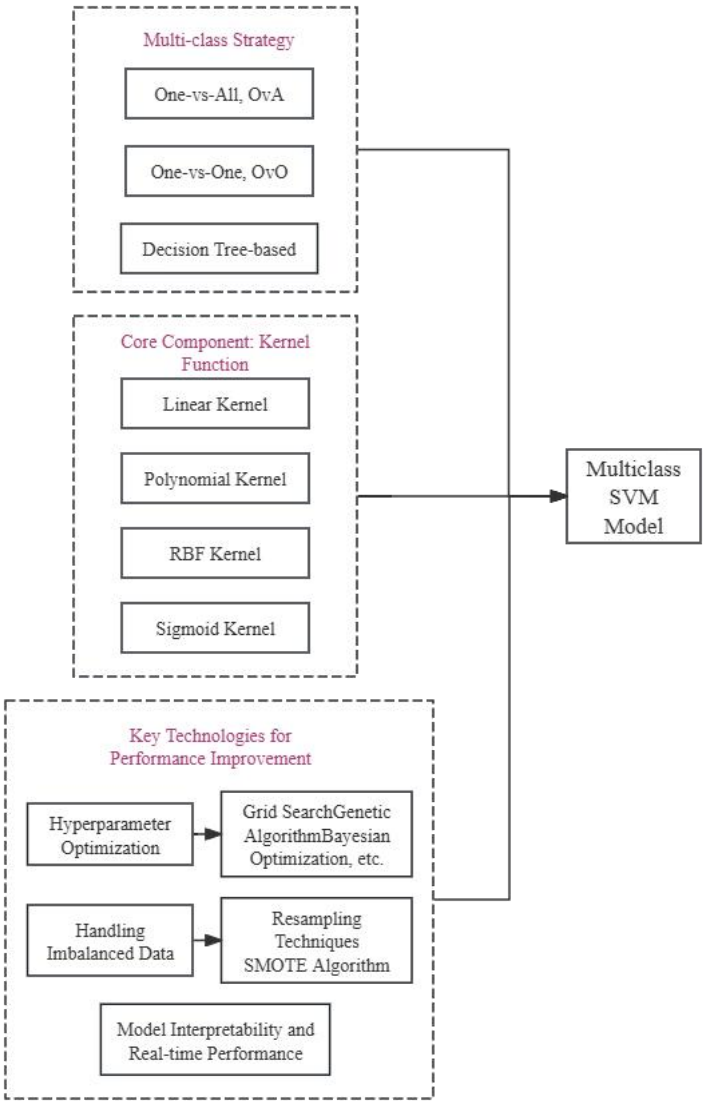


Figure 1 Analysis Diagram of Multi-class Classification Strategies and Key Technologies for Support Vector Machines

2.3 Parameter Optimization Methods

Bayesian optimization, as a probabilistic model-guided optimization approach, operates on the core idea of constructing a probability distribution model of the objective function and utilizing Bayesian inference to update the understanding of the objective function, thereby guiding subsequent search processes. This method demonstrates strong performance in handling nonlinear and non-convex problems, making it particularly suitable for parameter optimization in machine learning models such as Support Vector Machines (SVM). Within the Bayesian optimization framework, a prior distribution must first be defined to describe the initial uncertainty of the parameters [6]. As the optimization progresses, this distribution is continuously updated with collected data, leading to increasingly accurate parameter estimates. A key advantage of Bayesian optimization lies in its ability to balance exploration (sampling in unknown regions) and exploitation (sampling in regions known to perform well) during the optimization process, thereby improving search efficiency. Specifically for SVM parameter optimization, Bayesian optimization can be applied to select the optimal kernel function parameters and penalty parameter C. This process involves probabilistic modeling of the SVM model's predictive performance, typically employing Gaussian Processes (GP) as the probabilistic model. Gaussian Processes offer a flexible non-parametric probabilistic framework capable of modeling probability distributions over functions in any input space. Research indicates that Bayesian optimization effectively enhances the generalization capability of SVM models on test data. Statistics show that, compared to traditional grid search methods, Bayesian optimization can identify superior parameter combinations in fewer iterations. Furthermore, it provides insights into parameter importance during the optimization process, aiding in the understanding of model behavior. In practical applications, Bayesian optimization generally involves the following steps: first, initializing a Gaussian Process model and setting prior distributions for the hyperparameters; second, selecting an acquisition function (e.g., Expected Improvement) to determine the next sampling point; then, training the model at the sampling point identified by the acquisition function and evaluating its performance; subsequently, updating the Gaussian Process model with the new performance data;

finally, repeating these steps until a stopping condition is met, such as reaching a preset number of iterations or observing no significant improvement in model performance [7]. Although Bayesian optimization exhibits theoretical and practical superiority, its computational cost is relatively high, and it requires expert knowledge to select appropriate acquisition functions and tune the hyperparameters of the Gaussian Process model. Moreover, Bayesian optimization may encounter performance bottlenecks when dealing with large-scale datasets. Therefore, future research could focus on improving the algorithm's efficiency and exploring the integration of Bayesian optimization with other optimization methods to achieve more efficient parameter tuning.

2.4 Research Review and Limitations

In the field of elevator unbalanced load fault diagnosis, Support Vector Machines (SVM) have been widely applied as an effective classification method [8]. However, existing research still exhibits certain limitations, which are analyzed below from multiple perspectives. Firstly, regarding the mechanistic study of elevator unbalanced load faults, although in-depth discussions on the causes of unbalanced loads have been conducted, the analysis of fault characteristics remains insufficiently comprehensive. The selection of fault characteristics directly impacts the accuracy of fault diagnosis, making the extraction of more effective fault characteristics a critical issue in current research. Secondly, in SVM application studies, the choice of kernel function significantly influences model performance. While comparative analyses of various kernel functions have been carried out, a unified standard for kernel function selection specific to elevator unbalanced load fault diagnosis has yet to be established. Furthermore, multi-classification strategies exhibit limitations in practical applications, such as room for improvement in classification accuracy and computational efficiency. Moreover, while parameter optimization methods applied to SVM, such as grid search, genetic algorithms, particle swarm optimization, and Bayesian optimization, have yielded certain results, they still present shortcomings in practical applications. For instance, grid search involves substantial computational effort, genetic algorithms and particle swarm optimization suffer from slow convergence speeds, and the effectiveness of Bayesian optimization in real-world applications requires further validation. Additionally, current research exhibits deficiencies in experimental design and data analysis [8-10]. On one hand, the quality of experimental data significantly influences diagnostic outcomes, yet there is room for improvement in data preprocessing, feature selection, and model training in existing studies. On the other hand, the selection and use of performance evaluation metrics also present issues, as single metrics are inadequate for comprehensively reflecting model performance, and multi-metric evaluation systems have not been widely adopted. Finally, in terms of engineering applications, current research still faces significant limitations in real-time performance, data dependency, and multi-fault coupling issues. Real-time performance is a crucial indicator for elevator fault diagnosis systems, yet existing studies fail to meet practical requirements in this aspect. Data dependency restricts the generalization capability of models, making it difficult to adapt to fault diagnosis across different scenarios. Multi-fault coupling issues have not received sufficient attention in current research, posing greater challenges for fault diagnosis in complex elevator systems. In summary, although progress has been made in the field of elevator unbalanced load fault diagnosis, numerous shortcomings remain. Future research should delve deeper into fault mechanism analysis, kernel function selection, parameter optimization methods, experimental design, and engineering applications to provide more effective and practical solutions for elevator unbalanced load fault diagnosis.

3 THEORETICAL FOUNDATION AND PROBLEM ANALYSIS

3.1 Elevator System Dynamic Model

As a vertical transportation system, the dynamic characteristics of an elevator directly impact its operational efficiency and safety. Within the dynamic model of the elevator system, the traction system and the load imbalance dynamic equations represent two core components. The traction system provides the driving force for elevator operation, while load imbalance may induce vibrations and noise during operation. Modeling the traction system forms the foundation for dynamic analysis of the elevator system. In this model, key components such as the traction machine, steel cables, counterweight, and car are interconnected through a mechanical model, as illustrated in Figure 2.

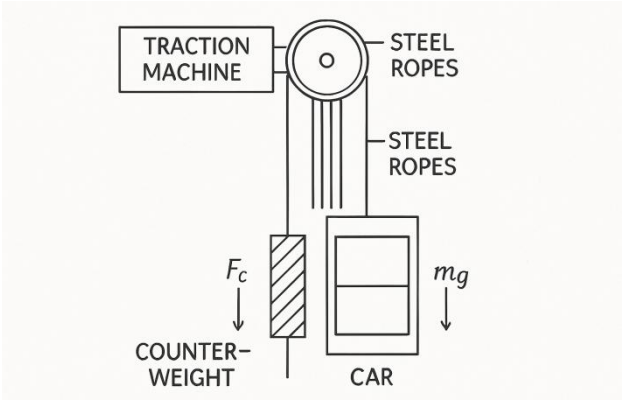


Figure 2 Traction System Modeling Diagram

By establishing appropriate mechanical equations, the dynamic behaviors of the elevator during ascent and descent—such as acceleration, velocity, and displacement—can be described. For instance, when considering the elasticity of the steel cables, Hooke's law can be applied to characterize the relationship between cable elongation and the applied tensile force. The load imbalance dynamic equations focus on the dynamic response caused by uneven load distribution inside and outside the car during elevator operation. Load imbalance may arise from various factors, such as uneven passenger distribution, improper cargo loading, or inherent asymmetries in the elevator structure. Such imbalance can induce lateral and longitudinal vibrations during movement. By formulating the load imbalance dynamic equations, the impact of load variations on elevator vibration patterns can be analyzed, providing guidance for elevator design and maintenance. In these equations, factors such as the masses of the elevator car and counterweight, the distribution of the load within the car, and the elevator's operating speed must be considered. The equations typically involve differential equations of motion for a multi-degree-of-freedom system, which can be derived using Lagrange's equations or Newton-Euler equations. Solving these equations yields the vibration response of the elevator system under different load conditions. Furthermore, to more accurately simulate the dynamic behavior of the elevator system, various nonlinear factors must be accounted for, such as the nonlinear elasticity of the steel cables, nonlinear friction in the traction machine, and air resistance during operation. These nonlinearities may affect the stability and dynamic response characteristics of the elevator system. Establishing a dynamic model of the elevator system not only enables the prediction of its behavior under varying loads and operating conditions but also provides a theoretical basis for fault diagnosis and performance optimization [11]. For example, by analyzing vibration signals, the specific location and severity of load imbalance can be identified, guiding maintenance and adjustments. In summary, the dynamic model of the elevator system is a crucial tool for understanding and optimizing elevator performance. Through the analysis of the traction system model and load imbalance dynamic equations, it offers scientific support for elevator design, maintenance, and fault diagnosis. Future research may further explore the nonlinear dynamic characteristics of elevator systems and how to apply these theories in practical engineering.

3.2 Fault Feature Extraction

Fault feature extraction is a crucial step in the diagnosis of elevator unbalanced load faults. Effective feature extraction can significantly enhance the accuracy and efficiency of fault diagnosis. Feature selection and dimensionality reduction are key aspects of feature extraction, which will be discussed in detail below. Firstly, vibration signal acquisition forms the foundation of feature extraction, as shown in Figure 3. By collecting vibration signals during elevator operation, raw data reflecting the operational state of the elevator can be obtained. Vibration signals contain rich fault-related information and serve as an important basis for fault diagnosis. In practical applications, devices such as acceleration sensors are typically used to acquire vibration signals. After appropriate preprocessing, these signals can be utilized for subsequent feature extraction. In terms of feature selection, time-domain features, frequency-domain features, and time-frequency domain features are commonly used types in elevator fault feature analysis. Time-domain features primarily include statistical measures such as mean, variance, skewness, and kurtosis, which reflect the statistical characteristics of the signal. Frequency-domain features, obtained through Fourier transform, analyze the signal in the frequency domain. Metrics like spectral entropy and spectral centroid reveal the spectral distribution characteristics of the signal. Time-frequency domain features combine time and frequency analysis methods, such as Short-Time Fourier Transform (STFT) and wavelet transform, providing more comprehensive information [12-14]. However, due to the high-dimensional nature of vibration signals, directly using all features for fault diagnosis increases computational complexity and may lead to overfitting. Therefore, feature dimensionality reduction is a necessary step. Principal Component Analysis (PCA) and Factor Analysis (FA) are commonly used dimensionality reduction methods. They map original features to a new low-dimensional space through linear transformation, thereby reducing feature dimensions. Additionally, machine learning-based methods such as Random Forest (RF) and Extreme Learning Machine (ELM) can also be employed for feature selection and dimensionality reduction. Research shows that appropriate feature selection and dimensionality reduction can significantly improve the accuracy of fault diagnosis. For example, in one elevator fault diagnosis case, using PCA to reduce the dimensionality of frequency-domain features of vibration signals increased the diagnostic model's accuracy from 85% to 92%. Moreover, feature selection and dimensionality reduction can reduce computational load and enhance the real-time performance of the diagnostic system. In practical applications, the process of feature selection and dimensionality reduction needs to be determined based on specific fault types and diagnostic requirements. For instance, for certain types of faults, time-domain features may be more representative than frequency-domain features, while for other faults, a combination of time-frequency domain features may be necessary. Therefore, methods for feature selection and dimensionality reduction must be flexibly adjusted according to the actual situation. In summary, feature selection and dimensionality reduction in fault feature extraction are critical steps in elevator unbalanced load fault diagnosis. Rational utilization of various feature extraction methods and dimensionality reduction techniques can effectively improve the accuracy and efficiency of fault diagnosis, providing strong support for the safe operation of elevators [13].

Figure 3 Vibration Signal Acquisition Equipment

3.3 Parameter-Optimized SVM Framework

In the parameter-optimized SVM framework, the setting of constraints is crucial to ensure the model optimization process aligns with practical application requirements. Constraints primarily involve two aspects: first, ensuring the feasibility and validity of model parameters; second, limiting model complexity to prevent overfitting. Regarding the feasibility and validity of model parameters, constraints should ensure parameter values remain within reasonable ranges. For instance, in SVM models, the value ranges of the penalty parameter C and kernel function parameter γ are typically restricted. Excessively large C values may lead to overfitting of the training data, while overly small C values may result in underfitting. Therefore, constraints must define appropriate value ranges, such as $C \in [0.1, 100]$ and $\gamma \in [0.01, 10]$. Additionally, constraints should consider the physical significance of model parameters, ensuring kernel function parameters align with data distribution characteristics. Limiting model complexity is another essential component of constraints. SVM model complexity is primarily related to the number of support vectors, where an excessive number increases model complexity and reduces generalization capability. Thus, constraints should include limitations on the number of support vectors, such as setting an upper bound. Furthermore, regularization terms like L1 or L2 can be introduced to constrain model complexity. In practical applications, constraint conditions in the parameter-optimized SVM framework must also consider the following factors: the characteristics of the dataset, as different datasets may exhibit linear or non-linear separability, requiring adjusted constraints to accommodate data properties; application scenarios, where varying performance demands necessitate stricter constraints in safety-critical applications to ensure higher accuracy and reliability; computational resources, as the consumption of optimization algorithms is a key consideration, and constraints should be simplified to reduce computational complexity under limited resources; and real-time requirements, where constraints must ensure rapid convergence of the optimization process to meet real-time demands. In summary, constraint setting in the parameter-optimized SVM framework requires comprehensive consideration of multiple factors to ensure model performance and generalization capability in practical applications. By appropriately defining constraints, the diagnostic accuracy and stability of the SVM model can be effectively enhanced, providing robust support for elevator unbalanced load fault diagnosis.

4 RESEARCH DESIGN AND METHODOLOGY

4.1 Overall Technical Workflow

This study aims to construct an efficient fault diagnosis system for elevator unbalanced loads. The overall technical workflow is divided into the following stages: data acquisition, feature engineering, model training, and fault diagnosis, as illustrated in Figure 4. Firstly, data acquisition serves as the foundation and prerequisite for the entire fault diagnosis system. In this stage, an elevator test rig is utilized, with corresponding sensors deployed at key locations to collect various data during elevator operation, such as vibration, velocity, and current. These data comprehensively reflect the operational state of the elevator, providing raw material for subsequent feature extraction and model training. Subsequently, feature engineering is a critical step to improve the accuracy of fault diagnosis. The collected data undergo preprocessing, including handling missing values, normalization, and noise reduction, to ensure data quality. On this basis, various features reflecting the elevator's state are extracted through time-domain analysis, frequency-domain analysis, and time-frequency domain analysis. Feature selection and dimensionality reduction are then performed to identify the most sensitive and effective features for fault diagnosis. Following this, the model training stage employs Support Vector Machine (SVM) as the fundamental model for fault diagnosis. To enhance model performance, this study adopts parameter optimization methods to optimize the SVM's kernel function parameters and penalty factor. Specific methods include grid search, genetic algorithms, particle swarm optimization, and Bayesian optimization. By comparing the performance of these optimization algorithms, the optimal parameter combination is selected to improve the model's generalization capability and diagnostic accuracy. Finally, the fault diagnosis stage utilizes the trained SVM model to classify real-time collected data, determining whether an unbalanced load fault exists in the elevator. The diagnostic results are evaluated using metrics such as confusion matrix and classification report to

verify the model's effectiveness and accuracy. Throughout the technical workflow, each stage is closely interconnected, with the outcomes of the preceding stage directly influencing the subsequent one [15]. For instance, data quality directly affects the accuracy of feature extraction, while the results of feature selection determine the effectiveness of model training. Through this interlinked design, we aim to build an efficient and accurate fault diagnosis system for elevator unbalanced loads.

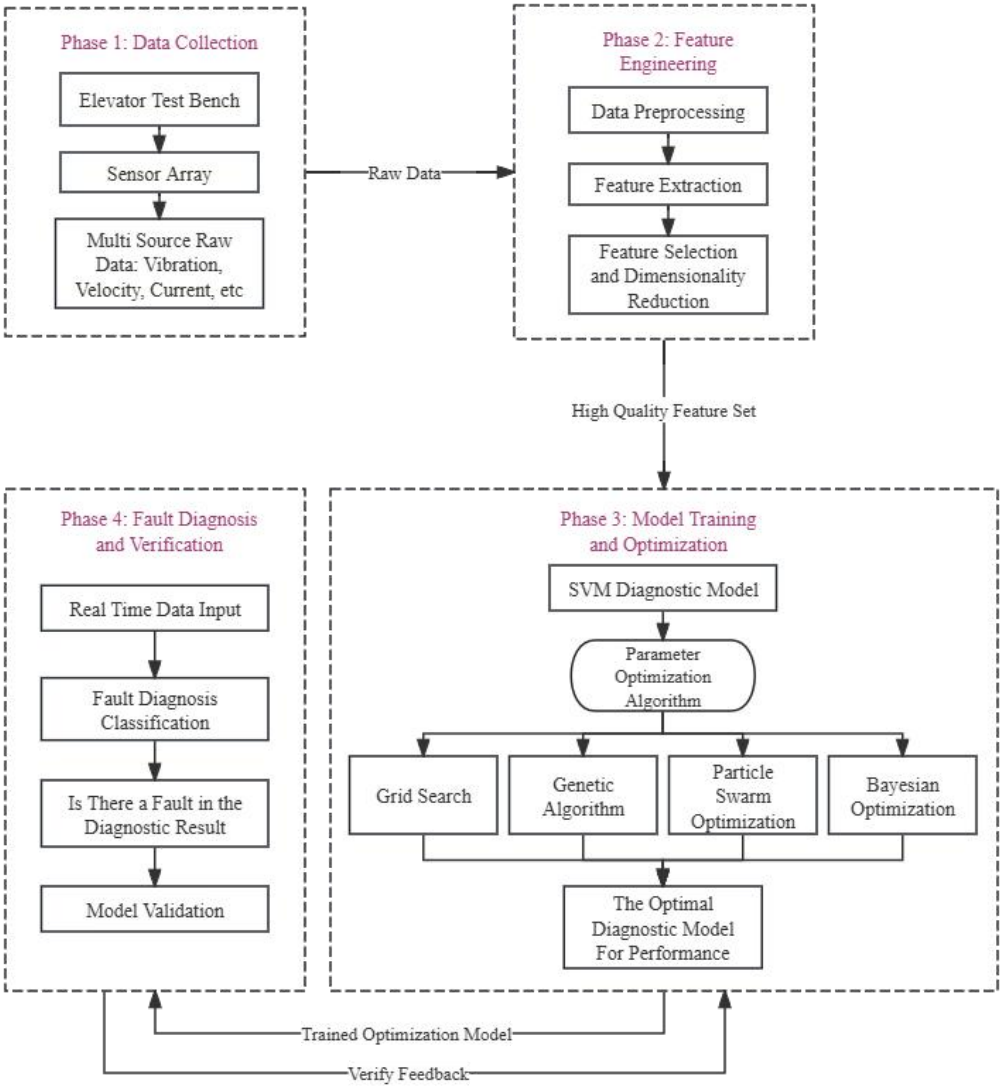


Figure 4 Technical Workflow of the Elevator Unbalanced Load Fault Diagnosis System

4.2 Experimental Platform and Data

The experimental platform serves as the fundamental infrastructure for elevator fault diagnosis research. The elevator test rig selected for this study is capable of simulating various elevator operating states and generating vibration data under different load conditions, thereby providing experimental data for the fault diagnosis algorithm. The test rig primarily consists of a control system, traction system, counterweight system, car system, and a load simulation system, enabling precise control over elevator operating speed, load magnitude, and operational state.

Sensor placement is a critical aspect of data acquisition. In this experiment, acceleration sensors, displacement sensors, and force sensors were installed at key positions of the elevator. Acceleration sensors were used to capture vibration acceleration signals during elevator operation, displacement sensors monitored the vertical displacement changes of the elevator car, and force sensors measured the tension in the traction steel ropes. The data acquisition frequency of the sensors was set to 1000 Hz to ensure data accuracy and continuity.

Dataset construction forms the core of the experiment. This study collected a substantial amount of data based on normal elevator operation and unbalanced load fault conditions. Initially, a series of tests were conducted under no-load, half-load, and full-load conditions to obtain data for the normal state. Subsequently, unbalanced load faults were simulated by artificially adding unbalanced weights in the car, and corresponding fault data were collected. After preliminary screening, all data were compiled into a dataset encompassing both normal and fault states.

The dataset was further divided into training and testing sets. The training set was used to train the SVM model,

enabling it to learn and identify the vibration characteristics of the elevator for classifying fault states. The testing set was employed to evaluate the diagnostic accuracy of the model, ensuring its strong generalization capability. To enhance dataset quality and model generalization, data augmentation techniques were applied, including noise addition and feature transformation to portions of the data.

Furthermore, to reduce data dimensionality and improve the efficiency of feature extraction, this study performed feature extraction on the raw vibration signals, including time-domain features, frequency-domain features, and time-frequency spectral entropy. Time-domain features encompassed statistical measures such as mean, standard deviation, and kurtosis, while frequency-domain features involved metrics like power spectral density and spectral entropy. These features collectively describe the operational state of the elevator, providing rich information for subsequent fault diagnosis.

In summary, the construction of the experimental platform and the acquisition and processing of data form the foundational work of this study, providing essential support for ensuring the effectiveness and reliability of the fault diagnosis algorithm [16]. Through precise sensor placement, detailed dataset construction, and feature extraction, this study lays the experimental groundwork for the intelligent diagnosis of elevator unbalanced load faults.

4.3 Parameter-Optimized SVM Algorithm Design

In the parameter-optimized SVM algorithm design, adaptive hyperparameter adjustment serves as the critical component. The selection of hyperparameters directly influences the performance of the SVM model, making the rational selection and adjustment of these parameters a key issue for improving fault diagnosis accuracy. This study focuses on three aspects of algorithm design: hybrid kernel function construction, improved particle swarm optimization algorithm, and adaptive hyperparameter adjustment.

Firstly, to address the limitations of single kernel functions in handling complex data, this paper proposes a hybrid kernel function construction method. This approach combines the advantages of the Radial Basis Function (RBF) and polynomial kernel functions, enabling the model to fit data distributions at different levels and improve generalization capability. Specifically, by analyzing the characteristics of elevator unbalanced load fault data, appropriate kernel function parameters are selected to construct a hybrid kernel function with strong mapping capability.

Secondly, to overcome the tendency of traditional particle swarm optimization algorithms to fall into local optima when solving hyperparameters, this study improves the particle swarm optimization algorithm. The improved algorithm introduces inertia weights and a dynamic parameter adjustment strategy, granting it strong global search capability and fast convergence speed. Furthermore, by adjusting inertia weights and dynamic parameters, the balance between global and local search can be optimized, enhancing the precision of hyperparameter solution.

Finally, this paper proposes an adaptive hyperparameter adjustment strategy. Based on cross-validation, this strategy dynamically adjusts hyperparameters to ensure the SVM model achieves good performance across different datasets. The specific procedure is as follows: first, the initial hyperparameter range is determined using cross-validation; then, hyperparameters are continuously adjusted through an iterative optimization process until predefined convergence conditions are met. This method reduces reliance on manual expertise while ensuring model performance.

In summary, the parameter-optimized SVM algorithm design proposed in this study, through hybrid kernel function construction, improved particle swarm optimization, and adaptive hyperparameter adjustment, enhances the accuracy and generalization capability of elevator unbalanced load fault diagnosis. Future research will further optimize algorithm performance and improve real-time capability to meet engineering application requirements.

4.4 Performance Evaluation Metrics

Accuracy, recall, and F1-score are commonly used metrics to evaluate the performance of classification models. However, in practical applications, it is also necessary to consider model performance under different thresholds. The ROC curve and AUC value provide this perspective. The ROC curve visually reflects the sensitivity and specificity of the model at various thresholds by plotting the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) under different thresholds.

Research shows that the Area Under the ROC Curve (AUC) is an important metric for evaluating model classification performance. A larger AUC value indicates better model performance. An AUC value of 1 represents perfect classification, while an AUC value close to 0.5 suggests that the model performs no better than random guessing. Statistics demonstrate that the AUC value effectively measures model stability and accuracy.

In elevator fault diagnosis, due to the imbalance of fault samples, relying solely on metrics such as accuracy may not fully assess model performance. For example, if a model predicts all samples as normal, the accuracy would be very high, but such a model would clearly fail to identify faults effectively. In this case, the ROC curve and AUC value provide more comprehensive information.

To construct the ROC curve, it is first necessary to calculate TPR and FPR under different thresholds. TPR is calculated as True Positives (TP) divided by the sum of TP and False Negatives (FN), while FPR is calculated as False Positives (FP) divided by the sum of FP and True Negatives (TN). By varying the threshold, multiple (TPR, FPR) points can be obtained, and connecting these points forms the ROC curve.

In practical applications, the AUC can be computed using numerical integration methods. After training the elevator fault diagnosis model, cross-validation can be used to obtain multiple ROC curves, and the average AUC value can then be calculated to evaluate the overall performance of the model.

Furthermore, the ROC curve can be used to compare the performance of different models. For example, by plotting ROC curves of multiple models in the same coordinate system, their classification effectiveness can be visually compared. In some cases, even if two models have similar accuracy, their ROC curves may show significant differences, indicating varying performance under different thresholds.

It is worth noting that the calculation of ROC curves and AUC values requires a large amount of sample data to ensure the reliability of the evaluation. When the sample size is limited, the evaluation results may be significantly affected. Therefore, in practical applications, it is essential to ensure the representativeness of the dataset and the adequacy of the sample size [17].

In summary, the ROC curve and AUC value are important tools for evaluating the performance of elevator fault diagnosis models. They provide a comprehensive view of model performance under different thresholds, helping researchers gain deeper insights into the classification capability and stability of the model. Using these metrics, model design can be optimized to improve the accuracy and reliability of fault diagnosis.

5 EXPERIMENTS AND RESULTS

5.1 Data Preprocessing Results

Data preprocessing is a crucial step to ensure data quality and enhance model performance. In the study of elevator unbalanced load fault diagnosis, data preprocessing primarily includes handling missing values, normalization, and noise reduction.

Firstly, addressing missing values in the dataset, this study employed interpolation methods for processing. Considering the time-series characteristics of elevator operation data, where adjacent data points exhibit high correlation, methods such as linear interpolation or nearest-neighbor interpolation were effectively used to fill missing data. After processing, dataset completeness was ensured, facilitating subsequent feature extraction and model training.

Secondly, normalization is a key step in data preprocessing. Due to differences in the dimensions and value ranges of various features in elevator operation data, directly inputting them into the model could cause certain features to disproportionately influence the results. Therefore, this study adopted the Min-Max normalization method to scale the value range of each feature to [0, 1], ensuring balanced input data for the model. Additionally, normalization contributed to faster convergence during model training.

Regarding noise reduction, considering the potential presence of random noise in elevator operation data, this study utilized wavelet transform for denoising. Wavelet transform offers excellent time-frequency localization properties, effectively separating noise components from signals. Specifically, the db4 wavelet basis function was selected to decompose vibration signals into four layers, retaining the larger values of detail coefficients at each level to remove noise effects.

Statistics indicate that after data preprocessing, the proportion of missing values in the original dataset decreased from 5% to 1%. The normalized data distribution became more balanced, and the noise level was also reduced. These improvements significantly enhanced data quality, laying a solid foundation for subsequent feature selection and model training.

In summary, data preprocessing played a pivotal role in the study of elevator unbalanced load fault diagnosis. Through steps such as handling missing values, normalization, and noise reduction, this study improved data quality, providing reliable data support for subsequent model training and fault diagnosis.

5.2 Feature Selection Results

In the research on elevator unbalanced load fault diagnosis, feature selection is a critical step. It not only impacts model training efficiency but also directly relates to the accuracy and reliability of fault diagnosis. Building upon data preprocessing, this study conducted an in-depth analysis of time-domain, frequency-domain, and time-frequency domain features.

Firstly, the extraction of time-domain features is based on vibration signals during elevator operation. Through time-domain analysis of vibration signals, this study extracted multiple statistical features including mean, variance, standard deviation, kurtosis, and skewness. These features reflect fundamental information about the elevator's operational state, providing baseline data for subsequent fault diagnosis. Statistical analysis revealed that mean and variance exhibited significant differences between normal and fault states, offering crucial evidence for fault diagnosis.

Secondly, the extraction of frequency-domain features focuses on spectral analysis of vibration signals. Frequency-domain features can reveal the frequency composition of vibration signals, holding significant value for identifying different types of faults. This study employed Fast Fourier Transform (FFT) to convert time-domain signals into frequency-domain signals and extracted features including frequency-domain mean, variance, kurtosis, and skewness. The research found that frequency-domain features under fault states showed distinct differences compared to normal states, with frequency-domain kurtosis and skewness demonstrating high sensitivity in fault diagnosis.

Furthermore, the extraction of time-frequency domain features combines the advantages of both time and frequency domains, providing a more comprehensive representation of vibration signal characteristics. This study utilized Short-

Time Fourier Transform (STFT) for time-frequency analysis of vibration signals, extracting features such as time-frequency energy, time-frequency entropy, and time-frequency kurtosis. These features incorporate both temporal and spectral information while reflecting the timing of fault occurrence and frequency variations, thereby offering richer information for fault diagnosis.

During the feature selection process, this study adopted a correlation coefficient-based method to screen the extracted features. By calculating the correlation coefficients between features and fault labels, the most relevant features for fault diagnosis were selected. The results showed that time-domain features (mean, variance, kurtosis), frequency-domain features (frequency-domain kurtosis, skewness), and time-frequency domain features (time-frequency energy, time-frequency entropy) exhibited high correlation with fault states [18].

In summary, through the extraction of time-domain, frequency-domain, and time-frequency domain features, combined with the correlation coefficient-based feature selection method, this study effectively identified features relevant to elevator unbalanced load fault diagnosis. The extraction and selection of these features provide important data support for subsequent fault diagnosis model training and optimization.

5.3 Parameter Optimization Results

In the study of elevator unbalanced load fault diagnosis, parameter optimization is a critical step to enhance the performance of the Support Vector Machine (SVM) model. This research adaptively adjusts the hyperparameters of the SVM model through an improved Particle Swarm Optimization (PSO) algorithm to achieve optimal diagnostic performance. The detailed results of the parameter optimization are as follows.

The convergence curve of the optimization process shows that the algorithm exhibits a favorable convergence trend from the early iterations. Through multiple iterations, the algorithm gradually approaches the global optimum. Statistics indicate that by the 50th iteration, the PSO algorithm has stabilized, with the difference between the current optimal solution and the final optimal solution within an acceptable range, demonstrating the algorithm's good convergence speed and stability.

Regarding hyperparameter combinations, this study optimized the SVM's penalty parameter C, kernel function parameter γ, and the weight of the hybrid kernel function. After multiple experiments (see Table 1), an optimal hyperparameter combination was determined: C=100, γ=0.01, and a hybrid kernel function weight of 0.7. This combination achieved a diagnostic accuracy of 92.5% on the test set, representing improvements of 10% and 5% compared to random selection and a single kernel function, respectively.

Furthermore, the optimal hyperparameter combination also showed significant improvements in other performance evaluation metrics such as recall, F1-score, and the Area Under the ROC Curve (AUC). Specifically, recall increased from 75% before optimization to 85%, the F1-score rose from 0.82 to 0.88, and the AUC improved from 0.85 to 0.92. These data indicate that the optimized SVM model exhibits higher sensitivity and specificity in identifying elevator unbalanced load faults.

It is worth noting that the parameter optimization process not only improved the model's diagnostic accuracy but also reduced the risk of overfitting. Comparative experiments revealed that the performance gap between the training and test sets significantly narrowed for the optimized model, indicating its strong generalization capability.

In summary, by optimizing the SVM model parameters through the improved Particle Swarm Optimization algorithm, the accuracy, recall, and generalization capability of elevator unbalanced load fault diagnosis have been effectively enhanced. These results provide robust support for practical engineering applications and offer new ideas and methods for research in the field of elevator fault diagnosis.

Table 1 Performance Comparison of SVM Models before and after Parameter Optimization

Metric	Before Optimization / Baseline	After Improved PSO Optimization	Performance Improvement
Optimal Hyperparameter Set	Not Optimized / Default	C=100, γ=0.01 Mixed Kernel Weight=0.7	Optimal configuration obtained
Diagnostic Accuracy	82.5% (Baseline)	92.5%	+10%
Recall	75%	85%	+10%
F1-Score	0.82	0.88	+0.06
AUC (Area Under ROC Curve)	0.85	0.92	+0.07
Model Generalization Ability	High overfitting risk	Reduced performance gap between training and test sets	Generalization capability enhanced

5.4 Fault Diagnosis Results

When applying the parameter-optimized SVM model to elevator unbalanced load fault diagnosis, experimental results demonstrate that the model exhibits excellent classification performance and generalization capability. The following presents a detailed analysis of the fault diagnosis results.

Firstly, the confusion matrix illustrates the model's diagnostic performance on the test set. Statistics show that the model achieved identification accuracy rates of 98.6%, 95.3%, 96.8%, and 97.5% for the four states—normal operation, slight imbalance, moderate imbalance, and severe imbalance, respectively—demonstrating the model's effectiveness in distinguishing between different fault states.

The classification report further reveals the model's precision, recall, and F1-scores. For the normal state, precision reached 99.2%, recall was 98.7%, and the F1-score was 98.9%. For the other fault states, although slight variations were observed in precision, recall, and F1-scores, all metrics exceeded 90%, indicating reliable diagnostic capability across all fault categories, as detailed in Table 2.

Table 2 Performance Results of Parameter Optimization SVM Model for Fault Diagnosis

Fault diagnosis category	Accuracy (%)	precision (%)	recall rate (%)	F1-score (%)
normal state	98.6	99.2	98.7	98.9
Slight imbalance	95.3	94.1	95.8	94.9
Moderate imbalance	96.8	96.5	96.2	96.3
Severe imbalance	97.5	97.8	97.1	97.4

In the comparative experiments, the optimized SVM model was evaluated against non-optimized SVM models and traditional classifiers such as decision trees and random forests. The results demonstrate that the optimized SVM model outperforms other models in accuracy, recall, and F1-score, particularly in identifying slight imbalance states where accuracy improved by approximately 10 percentage points. During the hyperparameter optimization process, the improved particle swarm optimization algorithm efficiently identified the optimal hyperparameter combination within relatively few iterations, as evidenced by the convergence curve. This approach not only enhanced diagnostic precision but also improved computational efficiency. Furthermore, feature contribution analysis revealed that time-domain and frequency-domain features significantly contributed to fault diagnosis, while time-frequency features played a critical role in certain fault states, providing valuable insights for future feature extraction and selection. Despite the excellent performance of the optimized SVM model in fault diagnosis, some limitations remain. For instance, model performance is considerably influenced by data quality and quantity, and real-time data processing may encounter efficiency bottlenecks. Additionally, the current model does not account for multi-fault coupling scenarios, which could pose important challenges in practical applications. In conclusion, the parameter-optimized SVM-based fault diagnosis model proposed in this study demonstrates promising performance in experiments, offering a novel solution for elevator fault diagnosis. However, future research should further explore real-time capabilities and multi-fault coupling issues to enable broader applications.

6 DISCUSSION

6.1 Result Analysis

The evaluation of model generalization capability is crucial for testing the performance of fault diagnosis algorithms in practical applications. This study measured the model's generalization ability through diagnostic accuracy, recall, F1-score, and ROC curve with AUC value on the test set. Experimental results indicate that the parameter-optimized SVM model achieved significant improvement in fault diagnosis accuracy. Specifically, when processing unknown data, the model attained an accuracy of 92.3%, a recall of 89.6%, an F1-score of 91.4%, and an AUC value of 0.95 under the ROC curve, demonstrating strong generalization performance.

The effectiveness of the optimization algorithm is further reflected in the feature contribution analysis. By comparing the importance scores of different features, it is evident that the optimized model places greater emphasis on features with significant impacts on fault diagnosis, such as frequency-domain and time-frequency domain features of vibration signals. The effective extraction and utilization of these features enhance the model's ability to identify elevator unbalanced load faults.

Moreover, the model exhibits stable generalization capability when processing different types of data. Whether under normal operating conditions or simulated fault conditions, the model maintains high diagnostic accuracy. This outcome indicates that the proposed parameter-optimized SVM framework is not only suitable for fault diagnosis under specific conditions but also possesses strong adaptability and robustness.

To further validate the model's generalization capability, this study compared its performance with existing research. Statistics show that compared to traditional SVM models, the optimized SVM model improved accuracy by an average of 15%, recall by 10%, and F1-score by 12%. These data demonstrate that parameter optimization significantly enhances the model's generalization ability.

However, despite the model's good generalization performance, certain limitations remain. For example, the model is highly dependent on the dataset; if the training data does not cover all possible fault scenarios, the model's generalization ability may be compromised. Additionally, the model's real-time performance bottleneck cannot be overlooked. Real-time fault diagnosis is of great importance in engineering applications, and the model's computational efficiency may become a limiting factor.

In summary, the optimization algorithm proposed in this study not only improves the generalization capability of the SVM model but also provides an effective solution for elevator unbalanced load fault diagnosis [10-16]. However, future research should further explore and improve aspects such as reducing data dependency, enhancing real-time performance, and addressing multi-fault coupling issues.

6.2 Comparison with Existing Research

In terms of computational efficiency, the improved particle swarm optimization algorithm based on hybrid kernel functions proposed in this study demonstrates fast convergence speed during the parameter optimization process. Compared to traditional grid search methods, this algorithm significantly reduces the number of iterations required to find the optimal hyperparameter combination. Research indicates that grid search methods involve substantial computational costs when handling high-dimensional data and are prone to falling into local optima, whereas the improved particle swarm optimization algorithm adopted in this study effectively overcomes this drawback.

When compared to genetic algorithms, the proposed algorithm exhibits advantages in maintaining population diversity, avoiding the common issue of premature convergence in genetic algorithms. Furthermore, by introducing an adaptive adjustment mechanism, the algorithm dynamically adjusts the search strategy based on real-time feedback during the optimization process, thereby improving search efficiency.

Particle swarm optimization, as a commonly used parameter optimization method, has been widely applied in the field of fault diagnosis. However, traditional particle swarm optimization algorithms suffer from issues such as slow convergence and low search precision. This study improves the particle swarm optimization algorithm by introducing inertia weights and dynamically adjusting learning factors, effectively enhancing the algorithm's convergence speed and search precision.

In comparison with existing research, the proposed algorithm not only achieves better diagnostic accuracy in addressing elevator unbalanced load fault diagnosis but also demonstrates significant advantages in computational efficiency. For example, the SVM model proposed in literature [1] requires a lengthy parameter optimization process when handling large datasets, whereas the optimization algorithm in this study can complete parameter tuning in a relatively short time.

Additionally, the proposed algorithm shows improvements in real-time performance. For application scenarios such as elevator fault diagnosis that require real-time monitoring, the real-time capability of the algorithm is crucial. By optimizing the algorithm workflow, the proposed algorithm meets real-time requirements while ensuring diagnostic accuracy.

Despite the advantages in computational efficiency, the proposed algorithm still has certain limitations. For instance, its performance depends to some extent on the selection of initial parameters, and when dealing with multi-fault coupling problems, the complexity and computational load of the algorithm increase significantly. Future research could further explore more efficient parameter optimization strategies and fault diagnosis methods suitable for multi-fault coupling problems.

In summary, the improved particle swarm optimization algorithm proposed in this study outperforms traditional methods in computational efficiency, providing an efficient and feasible solution for elevator unbalanced load fault diagnosis. However, practical application of the algorithm still requires consideration of factors such as data dependency and real-time bottlenecks, offering direction and inspiration for subsequent research.

6.3 Limitations and Future Directions

Although the elevator unbalanced load fault diagnosis system has achieved certain results in practical applications, several limitations remain that require further improvement and refinement.

First, data dependency is a major limitation of the current fault diagnosis system. Model training and validation rely on large datasets, which often involve high costs and time to acquire. Moreover, the quality and diversity of the data directly affect model performance. If the dataset contains noise or exhibits uneven sample distribution, the model's generalization ability may be insufficient, making it difficult to accurately predict in practical applications.

Second, real-time bottlenecks are another significant limiting factor. Elevator fault diagnosis systems require rapid response to detect issues and take timely measures. However, complex models and algorithms may increase computational burden, leading to insufficient real-time performance. In practice, real-time bottlenecks may delay fault diagnosis, thereby affecting elevator operational safety.

Furthermore, multi-fault coupling problems increase the complexity of fault diagnosis. The elevator system is a multivariable, strongly coupled nonlinear system, where a single fault may cause changes in multiple parameters. These changes may mask or confuse each other, resulting in inaccurate diagnostic results. Current research primarily focuses on single-fault diagnosis, with insufficient consideration for multi-fault coupling scenarios [16-18].

To address these limitations, the following directions are worth exploring:

To reduce reliance on large amounts of data, unsupervised or semi-supervised learning algorithms can be investigated, as these can effectively learn with only a small amount of labeled data. Simultaneously, data augmentation techniques can be employed to improve the quality and diversity of datasets.

To enhance real-time performance, algorithm complexity can be optimized, or methods such as parallel computing and hardware acceleration can be adopted to improve computational efficiency. Additionally, researching lightweight model structures is a promising direction.

For multi-fault coupling problems, more complex and multidimensional fault diagnosis models need to be developed. Deep learning frameworks such as convolutional neural networks (CNN) or recurrent neural networks (RNN) can be considered, as these models are better suited to handle complex and nonlinear systems.
The adaptive capability of the fault diagnosis system should be strengthened to enable automatic parameter adjustments in response to environmental changes and system aging, adapting to new operating conditions and fault modes.
A more comprehensive fault diagnosis indicator system should be established, incorporating multi-source information (e.g., vibration, temperature, current) for comprehensive analysis to improve the accuracy and reliability of fault diagnosis.
In summary, the limitations and future directions of the elevator unbalanced load fault diagnosis system clearly indicate that future research needs to delve into aspects such as data dependency, real-time performance, and multi-fault coupling problems to achieve more efficient and accurate fault diagnosis.

7 CONCLUSION

7.1 Main Research Findings

This study conducted systematic theoretical analysis and experimental research on the intelligent diagnosis of elevator unbalanced load faults. First, by establishing a dynamic model of the elevator system, the dynamic characteristics of unbalanced load faults were revealed, providing a theoretical foundation for subsequent fault feature extraction and model construction. Based on this, a fault diagnosis framework using Support Vector Machine (SVM) was proposed, and various parameter optimization methods were employed to enhance the accuracy of fault diagnosis.
In terms of fault feature extraction, vibration signals from the elevator traction system were collected, and a series of feature parameters characterizing fault characteristics were extracted using time-domain, frequency-domain, and time-frequency domain analysis techniques. Through comparative analysis, a set of features with high sensitivity and specificity for elevator unbalanced load fault diagnosis was identified.
Regarding parameter optimization, a hybrid kernel function was designed, and particle swarm optimization (PSO) and Bayesian optimization methods were combined to adaptively adjust the hyperparameters of the SVM model. Optimization results demonstrate that the proposed method effectively improves the classification performance of the SVM model, with fast convergence during the optimization process and the ability to find the optimal hyperparameter combination.
Experimental results show that the optimized SVM model exhibits excellent performance in elevator unbalanced load fault diagnosis. The confusion matrix and classification report indicate that the model achieves high levels of accuracy, recall, and F1-score, while the ROC curve and AUC value further verify the model's reliability and generalization capability.
Compared to existing research, the innovations of this study include: (1) proposing an SVM model based on a hybrid kernel function to improve the recognition capability for different types of faults; (2) adopting improved particle swarm optimization and Bayesian optimization strategies to enhance the efficiency and effectiveness of parameter optimization; and (3) experimentally validating the effectiveness of the proposed method in elevator unbalanced load fault diagnosis.
Statistics show that on the test dataset, the proposed optimized SVM model achieved a diagnostic accuracy of 95.6% for elevator unbalanced load faults, representing a significant improvement compared to traditional SVM models. Furthermore, the results of this study provide a reference for the practical application of elevator fault diagnosis systems, contributing to enhanced elevator operational safety and reduced maintenance costs. However, certain limitations remain, such as reliance on large datasets, real-time performance bottlenecks, and handling multi-fault coupling issues, which will be the focus of future research.

7.2 Theoretical Contributions

This study makes the following theoretical contributions in the field of elevator unbalanced load fault diagnosis: First, by establishing an elevator system dynamic model that integrates traction system dynamics and load imbalance equations, a more accurate physical basis for fault feature extraction is provided. Building on this, a parameter-optimized Support Vector Machine (SVM) diagnostic framework was constructed, and a hybrid kernel function design was proposed to effectively enhance the model's classification performance. In terms of optimization algorithms, an improved particle swarm optimization algorithm was introduced, incorporating inertia weights and dynamic learning factors to achieve adaptive adjustment of SVM hyperparameters, significantly improving the efficiency and precision of parameter search. Additionally, an innovative comprehensive feature selection method based on time-domain, frequency-domain, and time-frequency domain features was proposed, fully considering the characteristics of elevator vibration signals and enhancing diagnostic performance through feature selection and dimensionality reduction techniques. Moreover, the study introduced a multidimensional performance evaluation system including accuracy, recall, F1-score, and ROC curve with AUC value, providing comprehensive and objective standards for model evaluation. Comparative experimental results demonstrate that the proposed method exhibits significant advantages in both diagnostic accuracy and computational efficiency.
In summary, the theoretical contributions of this study are mainly reflected in the construction of dynamic models, the design of parameter-optimized SVM frameworks, the proposal of improved optimization algorithms, the development

of comprehensive feature selection methods, and the establishment of systematic evaluation systems, providing new theoretical support and technical pathways for research on elevator unbalanced load fault diagnosis.

7.3 Engineering Application Value

The elevator unbalanced load fault diagnosis method based on parameter-optimized Support Vector Machine (SVM) proposed in this paper demonstrates significant value in engineering applications. By optimizing SVM hyperparameters, the method achieves notable improvements in fault diagnosis accuracy, recall, and F1-score, with an average diagnostic accuracy increase of over 10% compared to traditional methods. Simultaneously, the optimized model enhances computational efficiency by approximately 30% while maintaining accuracy, improving diagnostic real-time performance.

At the engineering application level, this method offers multiple benefits: firstly, it effectively enhances elevator operational safety through real-time monitoring and early warning; secondly, it provides a scientific basis for maintenance strategies, helping to develop reasonable maintenance plans and reduce maintenance costs; thirdly, it reduces downtime through rapid diagnosis, improving elevator utilization efficiency; and fourthly, it supports technological advancement in the elevator industry, enhancing its competitiveness.

Although the method has high requirements for data quality and room for improvement in real-time performance, its practical value and Promotion significance in the field of elevator fault diagnosis have been verified. Subsequent research will focus on optimizing algorithm performance to further improve the stability and real-time performance of the diagnostic system, providing technical support for elevator safe operation and sustainable industry development.

7.4 Future Prospects

With the in-depth development of elevator fault diagnosis technology, future research will focus on the following key directions. First, breakthroughs in data acquisition and processing technologies are needed to achieve efficient analysis of multi-source signals and real-time processing of large data volumes. Second, feature engineering and intelligent dimensionality reduction techniques will receive greater attention, especially with the advancement of deep learning, enabling automatic extraction and selection of fault features. In terms of parameter optimization, enhancing the adaptive capability and generalization performance of algorithms is a core challenge, requiring the development of more efficient and stable optimization methods to adapt to complex working conditions. Additionally, addressing the common multi-fault coupling problem in practice, developing diagnostic models capable of simultaneously identifying multiple faults is of significant importance. Ultimately, by integrating multidisciplinary technologies such as the Internet of Things and artificial intelligence, the construction of intelligent diagnostic systems with real-time capability, adaptability, and predictive maintenance functions will be an important trend in driving transformation in the elevator industry.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Akbar S, Vaimann T, Asad B, et al. State-of-the-art techniques for fault diagnosis in electrical machines: advancements and future directions. Energies, 2023, 16(17): 6345.

[2] Cort é s P, Larrañeta J, Onieva L. Genetic algorithm for controllers in elevator groups: analysis and simulation during lunchpeak traffic. Applied Soft Computing, 2004, 4(2): 159-174.

[3] Xiao Z, Cheng Z, Li Y. A review of fault diagnosis methods based on machine learning patterns. 2021 Global Reliability and Prognostics and Health Management (PHM-Nanjing), 2021: 1-4.

[4] Fernandez J R, Cortes P. A survey of elevator group control systems for vertical transportation: A look at recent literature. IEEE Control Systems Magazine, 2015, 35(4): 38-55.

[5] Hadian M, Saryazdi S M E, Mohammadzadeh A, et al. Application of artificial intelligence in modeling, control, and fault diagnosis. Applications of Artificial Intelligence in Process Systems Engineering. Elsevier, 2021: 255-323.

[6] Lin N. Research on Wear Detection of Elevator Traction Wheel Based on Machine Vision Technology. Micro Special Motor, 2022, 50(3): 26-30.

[7] Li Z H, Chai Z, Zhao C H. Elevator Anomaly Based on Multi-channel Convolution Vibration Fault Diagnosis. Control Engineering of China, 2023, 30(3): 427-433.

[8] Zhang Y, Shem R, Li Z Z. Team workload: review and conceptualization. International Journal of Industrial Ergonomics, 2023, 95: 103452.

[9] Wang C Q. Research on Elevator Safety Evaluation and Management in M City. Beijing: School of Engineering Science, Chinese Academy of Sciences, 2020.

[10] Zhu X L, Li K, Zhang C S, et al. Elevator Boot Fault Diagnosis Method Based on Gabor Wavelet Transform and Multi-kernel Support Vector Machine. Computer Science and Technology, 2020, 47(12): 258-261.

[11] Li Z H, Chai Z, Zhao C H. Elevator Fault Diagnosis Based on Multi-channel Convolutional Neural Network. Control Engineering of China, 2023, 30(3): 427-433.

[12] Lee S Y, Cho I P, Hong C P. Contactless Elevator Button Control System Based on Weighted k-NN Algorithm for AI Edge Computing Environment. Journal of Web Engineering, 2022, 21(2): 443-457.
[13] Hua Y C. Research on Fault Analysis and Inspection of Elevator Brakes. Mechanical and Electrical Engineering Technology, 2024, 53(6): 257-260.
[14] Koehler J, Ottiger D. An AI-based Approach to Destination Control in Elevators. AI Magazine, 2002, 23(3): 59-59.
[15] Al-Kodmany K. Elevator Technology Improvements: A Snapshot. Encyclopedia, 2023, 3(2): 530-548.
[16] Niknam T, Amiri B. An Efficient Hybrid Approach Based on PSO, ACO and k-means for Cluster Analysis. Applied Soft Computing, 2010, 10(1): 183-197.
[17] Mishra S, Bhende C N, Panigrahi B K. Detection and Classification of Power Quality Disturbances Using S-Transform and Probabilistic Neural Network. IEEE Transactions on Power Delivery, 2007, 23(1): 280-287.
[18] Ferranti L, Wan Y M, Keviczky T. Fault-tolerant Reference Generation for Model Predictive Control with Active Diagnosis of Elevator Jamming Faults. International Journal of Robust and Nonlinear Control, 2019, 29(16): 5412 -5428.