Volume 3, Issue 3, 2025

Print ISSN: 2959-9903
Online ISSN: 2959-9911

World Journal of Information Technology

World Journal of Information Technology

Volume 3, Issue 3, 2025

Published by Upubscience Publisher

World Journal of Information Technology
Print ISSN: 2959-9903 Online ISSN: 2959-9911
Email: info@upubscience.com
Website: http://www.upubscience.com/

Table of Content

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3037

SYNERGISTIC MECHANISMS BETWEEN DIGITAL INCLUSIVE FINANCE AND RURAL INDUSTRY REVITALIZATION

Lei Zhang
Department of Enrollment and Career, Renmin University of China, Beijing 100872, China.
Corresponding Email: peter6251@163.com

Abstract: This study systematically reviews the synergistic mechanisms between digital inclusive finance and rural industry revitalization. Based on theories of industrial integration, total factor productivity, agricultural multifunctionality, and sustainable development, it constructs a theoretical framework for their coordinated development. First, it elucidates the connotations and characteristics of digital inclusive finance—coverage breadth, depth of use, and degree of digitalization—and analyzes the connotations, evaluation indicator system, and dynamic evolution pathways of rural industry revitalization. Second, it investigates the direct support mechanisms of digital inclusive finance—such as capital supply, risk management, and industrial upgrading—as well as its indirect pathways through promoting agricultural technological innovation and catalyzing industry integration, assessing their effects on rural industry revitalization; simultaneously, it analyzes the market demand, credit foundation, and technology application opportunities that rural industry revitalization provides for digital inclusive finance. Third, through case studies, it validates the effectiveness of various digital finance models in expanding industry scale, broadening channels, enhancing risk control, and increasing farmers' incomes. Finally, it proposes policy recommendations to optimize the enabling environment, enhance service capacity, and deepen integration pathways—namely digitalized industry platforms, supply-chain finance, and farmer financial literacy enhancement—to offer theoretical guidance and practical reference for governments, financial institutions, and industry stakeholders in advancing rural revitalization.
Keywords: Digital inclusive finance; Rural industry revitalization; Synergistic mechanism; Industrial integration; Supply-chain finance; Agricultural technological innovation; Case study

1 INTRODUCTION

Rural industry revitalization occupies a critically important position in the national strategic layout of our country; it is the key link in resolving the "three rural" issues, achieving integrated urban-rural development, and advancing national modernization. Since it is proposed the rural revitalization strategy, the development of rural industries has received unprecedented attention. The prosperity of rural industries can not only raise farmers' income levels and improve rural infrastructure and public services, but also enhance the endogenous momentum of the rural economy, promoting agricultural modernization and sustainable rural development. However, for a long time, the development of rural industries has faced numerous difficulties. Traditional financial services have had limited coverage in rural areas, and there is a significant gap between financial supply and the needs of rural industry development. The small number of financial institution outlets in rural areas, along with the simplicity of financial products and services, makes it difficult for rural enterprises and farmers to obtain sufficient funding support. At the same time, the rural credit system is imperfect, and the lack of effective collateral further exacerbates the problems of difficult and expensive financing for rural industries, severely restricting the upgrading and development of these industries. Against this background, digital inclusive finance has emerged as a new financial model. By leveraging modern information technologies such as the Internet, big data, and artificial intelligence, digital inclusive finance breaks the time and space limitations of traditional financial services, reduces the cost of financial services, and improves their efficiency and accessibility. It can extend financial services to every corner of rural areas, providing rural enterprises and farmers with more convenient, efficient, and personalized financial products and services. Through digital inclusive finance, rural industries can obtain more funding support for expanding production scale, improving technology, and expanding markets, thereby injecting new momentum into rural industry revitalization. Therefore, studying the impact of digital inclusive finance on rural industry revitalization has important practical significance. Deeply exploring the mechanisms, problems, and challenges of digital inclusive finance in rural industry development will help the government and financial institutions formulate more effective policies and measures, promote digital inclusive finance to better serve rural industry revitalization, and realize the sustainable development of the rural economy and comprehensive revitalization of the countryside.

2 RESEARCH PROGRESS ON THE SYNERGY BETWEEN DIGITAL INCLUSIVE FINANCE AND RURAL INDUSTRIAL REVITALIZATION

2.1 Theoretical Basis of Rural Industrial Revitalization

Rural industrial revitalization is the key to achieving comprehensive rural revitalization, underpinned by a series of rich theoretical foundations. These theories not only provide directional guidance for rural industrial development but also

offer important academic support for the formulation of relevant policies and practical exploration. The following will systematically review and elaborate on important theories related to rural industrial revitalization.

2.1.1 Industrial convergence theory

Industrial convergence, as an emerging industrial development model, was initially proposed by Rosenberg while studying technological changes in the US mechanical tool industry. He discovered that the technological boundaries between different industries were gradually blurring, and the mutual penetration and integration of technologies were facilitating the emergence of new industrial forms. As research deepened, the concept of industrial convergence was further extended to the economic field, broadly referring to the process where different industries or different sectors within the same industry penetrate, intersect, and ultimately integrate to form new industrial formats and competitive advantages. In the context of rural industrial revitalization, industrial convergence mainly manifests as the integrated development among primary, secondary, and tertiary industries in rural areas. The primary industry (agriculture) is no longer confined to traditional planting and breeding but is closely integrated with the secondary industry (agro-processing, manufacturing) and the tertiary industry (agricultural product distribution, tourism, leisure services, etc.), forming new industrial formats such as integrated agro-processing and sales, leisure agriculture, and rural tourism. This convergence breaks the boundaries between traditional industries, achieving optimized resource allocation and enhanced industrial value-added. Industrial convergence provides new developmental momentum and pathways for rural industrial revitalization. Economically, industrial convergence can extend the agricultural industrial chain and increase the added value of agricultural products[1]. Through deep processing and refined packaging, primary agricultural products can be transformed into high-value-added goods, thereby increasing farmers' income. For instance, fruits can not only be sold directly but also processed into juice, preserved fruit, canned fruit, and other products, significantly enhancing market value. Industrial convergence helps promote rural employment. With the development of rural industrial convergence, new industrial formats continuously emerge, creating substantial employment opportunities. Beyond traditional agricultural production roles, these encompass areas like agro-processing, logistics distribution, and tourism services, attracting the local transfer of surplus rural labor and alleviating the problem of rural labor outflow. Industrial convergence can also drive the optimization and upgrading of the rural industrial structure. Traditionally dominated by low-value-added agriculture, the rural industrial structure was singular. Through convergence, it gradually shifts towards diversification and sophistication, enhancing the competitiveness and risk resilience of rural industries.

2.1.2 Total factor productivity theory

Total Factor Productivity (TFP) is a crucial indicator for measuring the quality and efficiency of economic growth. It refers to the additional output achieved under given inputs of various production factors (such as capital, labor, land, etc.), reflecting the contribution of technological progress, management innovation, and institutional changes to economic growth. The main methods for measuring TFP are parametric and non-parametric. Parametric methods, based on production functions, estimate TFP by calculating function parameters, with common approaches including the Solow residual method and stochastic frontier analysis. Non-parametric methods, not reliant on specific production function forms, primarily use Data Envelopment Analysis (DEA). Improving TFP holds significant practical importance in the process of rural industrial revitalization. It is key to achieving sustainable rural industrial development. Traditional rural industrial development often relied on increasing factor inputs, such as expanding cultivated land, adding labor, and capital investment. However, this extensive development model suffers from resource waste and environmental pollution. By enhancing TFP, industrial output growth can be achieved without increasing, or even while reducing, factor inputs, thereby improving resource utilization efficiency and promoting sustainable rural industrial development. Improving TFP helps enhance the competitiveness of rural industries. In today's increasingly competitive market, rural industries must rely on technological progress and management innovation to improve production efficiency and product quality to secure a place in domestic and international markets. Higher TFP means rural industries can produce higher-quality products and services at lower costs, strengthening market competitiveness. TFP improvement can also foster innovation in rural industries. Technological progress and management innovation are vital pathways to raise TFP, and these innovative activities, in turn, drive rural industries to continuously develop new products, formats, and business models, injecting new vitality into rural industrial revitalization[2].

2.1.3 Agricultural multifunctionality theory

Agricultural multifunctionality refers to agriculture's possession of multiple functions beyond its basic economic function of providing agricultural products, including ecological conservation, cultural heritage, and social stability. This concept was first proposed by Japan in the 1990s to emphasize agriculture's comprehensive value across economic, social, and ecological dimensions. Agricultural multifunctionality has the following characteristics: firstly, diversity – agriculture relates not only to the economic sphere but also closely to ecological, social, and cultural fields; secondly, externality – some functions of agriculture, like ecological protection and cultural heritage, have significant positive externalities, bringing broad benefits to society; thirdly, dynamism – agriculture's functions continuously expand and deepen with socioeconomic development and changing human needs. The theory of agricultural multifunctionality provides new ideas and directions for rural industrial revitalization. Based on multifunctionality, rural industrial development should not be limited to agricultural production and sales but should fully tap into agriculture's ecological, cultural, and social value to develop related industrial formats. For example, leveraging agriculture's ecological function can develop eco-agriculture, sightseeing agriculture, etc., achieving integrated development of agriculture and tourism; utilizing agriculture's cultural function can involve rural cultural experience activities and agro-cultural creative industries, enriching the connotation of rural industries. Harnessing agricultural multifunctionality helps promote

comprehensive rural development. By developing multifunctional rural industries, more talent, capital, and technology can flow to rural areas, improving rural infrastructure and public services, enhancing the quality of life for rural residents, and achieving coordinated development of the rural economy, society, and ecology[3].

2.1.4 Sustainable development theory

The concept of sustainable development was first proposed by Norwegian Prime Minister Gro Harlem Brundtland in the 1987 report "Our Common Future," defining it as "development that meets the needs of the present without compromising the ability of future generations to meet their own needs." Sustainable development emphasizes the coordination and unity of economy, society, and environment, pursuing long-term, stable development. Sustainable development follows several basic principles: first, the principle of equity, including intragenerational and intergenerational equity – fairness in resource use and development opportunities among contemporaries and between current and future generations; second, the principle of sustainability, requiring that human economic and social development not exceed the carrying capacity of resources and the environment; third, the principle of commonality, emphasizing the shared responsibility and cooperation of all global nations on sustainable development issues. In rural industrial revitalization, sustainable development theory holds significant guiding importance. Rural industrial development should prioritize ecological environmental protection, avoiding over-exploitation and resource waste. For example, promoting eco-agricultural technologies to reduce fertilizer and pesticide use, minimizing agricultural non-point source pollution; developing circular agriculture to achieve resource utilization of agro-processing waste, improving resource recycling efficiency. Sustainable development theory requires rural industrial development to balance economic, social, and ecological benefits. While pursuing economic benefits, it must focus on safeguarding farmers' interests and promoting rural social stability and harmony[4]. For instance, developing characteristic rural industries to increase farmers' income; strengthening rural infrastructure and public service provision to improve farmers' quality of life. Sustainable development theory also emphasizes long-term planning and strategic layout for rural industrial development. Scientific and rational industrial development plans should be formulated based on rural resource endowments, location conditions, and market demand, avoiding blind following of trends and short-term actions to ensure the long-term stable development of rural industries.

2.2 Development Status of Digital Inclusive Finance

2.2.1 Origin of digital inclusive finance

The origin of digital inclusive finance can be traced back to the emergence of the inclusive finance concept and the gradual application of digital technology in the financial sector. The concept of inclusive finance was first proposed by the United Nations in 2005 during the International Year of Microcredit. Its basic meaning is a financial system that can effectively and comprehensively serve all social strata and groups, aiming to enable all social strata, especially vulnerable groups neglected by traditional finance such as micro and small enterprises (MSEs), farmers, and low-income urban populations, to access financial services fairly. The emergence of this concept reflects a rethinking of the limitations of traditional financial services. Due to high costs and significant risks, traditional finance often struggles to cover all corners of society, leaving large populations unable to access basic financial services, hindering balanced economic development and social equity. With the rapid development of information technology, digital technologies such as the internet, mobile communications, big data, and cloud computing gradually matured and began to deeply integrate with the financial industry. 2013 was termed China's "Year of Internet Finance," witnessing the vigorous rise of internet finance models represented by third-party payment, P2P lending, and crowdfunding[5]. Leveraging digital technology, these emerging financial models broke through the temporal and spatial limitations of traditional financial services, reduced costs, and improved service efficiency, providing new pathways and means for the development of inclusive finance. Building on this, the concept of digital inclusive finance emerged. It is the product of combining inclusive finance with digital technology, utilizing digital technology to expand the breadth and depth of financial services and enhance their accessibility and inclusivity.

2.2.2 Development of digital inclusive finance

Globally, digital inclusive finance shows a trend of rapid development. International organizations like the World Bank actively promote its development, and many countries and regions have introduced relevant policies and measures to encourage financial institutions to use digital technology to innovate service models and improve coverage. For instance, Kenya's M-Pesa mobile money service is a successful model of digital inclusive finance. Launched in 2007, it provides users with convenient transfer, payment, and savings services via a mobile SMS platform. To date, M-Pesa has covered most of Kenya's population, greatly enhancing local financial service accessibility and promoting economic development. Additionally, India is actively promoting digital inclusive finance. The government launched the "Digital India" plan, encouraging financial institutions to provide services to the masses through digital channels, including opening digital bank accounts and promoting mobile payments, achieving remarkable results. In China, digital inclusive finance has achieved globally recognized accomplishments. The government attaches high importance to its development, issuing a series of policy documents that create a favorable policy environment. Simultaneously, China possesses a vast internet user base and advanced digital technology infrastructure, providing a solid foundation. In the payment sector, third-party payment is widely used in China. Platforms represented by Alipay and WeChat Pay, with their convenient payment experience and extensive application scenarios, have become indispensable payment tools in daily life. According to statistics, by the end of 2024, China's third-party mobile payment transaction volume exceeded hundreds of trillions of yuan. In the credit sector, internet banks and fintech companies use big data, artificial

intelligence, and other technologies to innovate credit models, providing more convenient and efficient credit services to MSEs and individuals. For example, MYbank, leveraging Alibaba's e-commerce platform and big data advantages, provides unsecured, pure-credit microloans to MSEs. By the end of 2024, it had served tens of millions of MSEs. In the insurance sector, internet insurance also shows rapid development. Insurance companies sell products via internet platforms, improving sales efficiency and coverage. Simultaneously, using big data and AI, insurers can assess risks more accurately and develop personalized insurance products.

2.2.3 Application of digital inclusive finance in rural areas

In recent years, the application of digital inclusive finance in China's rural areas has made some progress. With the widespread promotion of mobile payments, payment methods in rural areas have undergone tremendous changes. More and more rural residents are using mobile payments. Support for QR code payments can be seen in rural supermarkets, small shops, and farmers' markets. The popularity of mobile payments not only facilitates daily consumption for rural residents but also promotes rural commodity circulation and economic development. For example, some rural e-commerce platforms use mobile payments to achieve online sales of agricultural products, broadening sales channels. To address the difficulty of obtaining loans for rural residents and rural MSEs, financial institutions and fintech companies actively innovate credit service models. On one hand, they use big data and risk control models to assess the credit status of rural residents, providing qualified residents with small credit loans. For instance, some financial institutions collaborate with agricultural departments to obtain production information like planting and breeding, combine it with farmers' credit records, and provide precise credit support. On the other hand, they develop supply chain finance services, providing financing to upstream and downstream farmers and MSEs around core enterprises in the agricultural industrial chain. For example, in some characteristic agricultural product industries, financial institutions cooperate with agro-processing enterprises to provide order financing to growers, solving their capital shortages during production. The application of digital technology also provides opportunities for expanding rural insurance services. Insurance companies sell agricultural insurance products via internet platforms, improving promotion and sales efficiency. Simultaneously, using technologies like satellite remote sensing and drones, they monitor crop growth, accurately assess disaster losses, and improve claims settlement efficiency. For instance, in some regions, insurers use satellite remote sensing to monitor crop planting areas and growth status in real time. When natural disasters occur, they can quickly determine the affected area and extent of loss, providing timely claim services to farmers. Despite these achievements, challenges remain. Network coverage and communication quality are poor in some rural areas, hindering service delivery. In remote mountainous and impoverished areas, unstable or non-existent network coverage due to geographical isolation and high construction costs prevents residents from using digital financial services normally. Rural residents generally have low financial literacy, with limited awareness and acceptance of digital inclusive financial products and services[6]. Some harbor doubts and concerns about mobile payments, online credit, etc., fearing fund security and personal information leakage. Moreover, lacking financial knowledge and risk awareness, some residents are vulnerable to fraud and misguidance. The rural credit system lags, making it difficult for financial institutions to obtain comprehensive and accurate credit information on residents and MSEs, increasing credit risk assessment difficulty. Furthermore, weak credit awareness in rural areas, with some farmers and enterprises evading debts, dampens the enthusiasm of financial institutions to provide services.

2.3 Research Progress on Their Synergistic Mechanism

The synergistic mechanism between digital inclusive finance and rural industrial revitalization has been a hot topic in academia. Related research aims to deeply analyze how they mutually promote and synergistically develop to drive high-quality rural economic growth. Existing literature mainly explores the supporting mechanisms of digital inclusive finance for rural industrial revitalization, the opportunities rural industrial revitalization provides for digital inclusive finance development, and the intrinsic logic of their synergistic development.

2.3.1 Supporting mechanisms of digital inclusive finance for rural industrial revitalization

Many scholars emphasize the crucial role of digital inclusive finance in solving funding difficulties for rural industries. Through technologies like the internet and big data, it lowers barriers and costs, broadens coverage, and provides rural industrial entities with more convenient and efficient financing channels. For example, Hu (2023) points out that digital credit platforms use multi-dimensional data to assess the credit of farmers and rural MSEs, breaking the traditional reliance on collateral by financial institutions. This allows potentially viable rural industrial projects lacking collateral to gain funding. Such precise capital supply helps optimize resource allocation and promote industrial restructuring and upgrading. Digital inclusive finance also plays an important role in rural industrial risk management. Digital insurance, a key component, provides diversified risk protection. Fang (2024)'s research shows that technologies like satellite remote sensing and IoT enable real-time monitoring of crop growth and livestock health, achieving precise risk assessment and claims settlement. This not only helps reduce natural and market risks faced by farmers and rural enterprises but also boosts their confidence and motivation. Furthermore, digital financial instruments like futures and options provide hedging tools to manage agricultural price volatility risks. The technological innovation brought by digital inclusive finance injects new vitality into rural industries. The popularity of services like mobile payment and e-commerce transforms traditional rural transaction and business models. Xie (2023)'s research finds that digital inclusive finance promotes rural e-commerce development, enabling agricultural products to access markets more easily, reducing intermediaries, and increasing added value. Simultaneously, digital financial platforms provide services like technical training and market information, helping improve management and innovation capabilities. For example,

some digital financial institutions collaborate with agri-tech companies to provide intelligent planting and breeding solutions, pushing rural industries towards modernization and intelligence[7].

2.3.2 Opportunities provided by rural industrial revitalization for digital inclusive finance

The vigorous development of rural industries provides a vast market space for digital inclusive finance. As rural industries diversify, the financial service demands of rural residents and enterprises become increasingly varied. Beyond traditional credit and insurance needs, new demands emerge for wealth management, investment, and payment settlement. Lin (2024) argues that rural industrial revitalization raises rural incomes, giving residents more funds for savings and investment, creating opportunities for digital inclusive financial institutions to offer wealth management services. Concurrently, the growth of rural enterprises requires more financial support, such as supply chain finance and M&A services, helping digital inclusive financial institutions expand their business scope and achieve growth. The vast amount of data generated during rural industrial development provides valuable resources for digital inclusive finance. This data covers various aspects like production, operation, transaction records, and credit status of farmers and rural enterprises, helping institutions assess credit risk more accurately. Zhou (2023) notes that by mining and analyzing rural industrial data, digital financial institutions can build more robust credit evaluation models, improving accuracy and efficiency. Furthermore, rural industrial revitalization promotes the construction of the rural credit system. A sound credit environment is conducive to the healthy development of digital inclusive finance. For example, some local governments foster an atmosphere of honesty and trustworthiness by creating "credit villages" and "credit users," reducing credit risks for institutions. The unique demands of rural industries drive continuous innovation and service upgrades in digital inclusive finance. To meet diverse needs, institutions have launched a series of innovative products and service models. Zhao (2024)'s research finds that, considering the characteristics of agricultural supply chains, some digital financial institutions developed supply chain finance products like order financing and warehouse receipt pledge financing, offering more flexible financing methods. Additionally, some platforms combine features of rural tourism and specialty agriculture to launch customized financial service solutions, enhancing service relevance and effectiveness.

2.3.3 Intrinsic logic of their synergistic development

A close interactive relationship exists between digital inclusive finance and rural industrial revitalization. Rural industry is the foundation for digital inclusive finance development; only with thriving industries can rich business scenarios and customer resources be provided. Simultaneously, digital inclusive finance is a crucial support for revitalization, providing funding, technology, and risk management. They are interdependent and mutually reinforcing, forming a virtuous cycle. For example, the development of characteristic rural industries attracts more financial resources, whose support, in turn, further drives industrial growth and upgrading. Policy Guidance and Market Mechanisms Policy guidance and market mechanisms play important roles in their synergistic development. The government issues policies like fiscal subsidies and tax incentives to encourage institutions to increase support. Meanwhile, the market mechanism plays a decisive role in resource allocation, as institutions autonomously select service targets and business models based on demand and risk-return principles. Liu (2023) believes that only by organically combining policy guidance with market mechanisms can synergistic development be achieved. For instance, governments can establish risk compensation funds to reduce institutional risks while guiding market capital towards key areas and weak links in rural industries. Technological Innovation and Institutional Innovation Technological and institutional innovation are important driving forces. The application of digital technologies like big data, AI, and blockchain brings new opportunities, improving efficiency and precision in financial services and industrial management. Institutional innovations, such as in financial regulation and rural property rights reform, provide safeguards for a conducive development environment. Chen (2024) points out that only by integrating technological and institutional innovation can the synergistic effect be fully realized. For example, establishing a regulatory sandbox for rural digital financial services can encourage innovation while preventing financial risks. In summary, existing literature has conducted relatively in-depth research on the synergistic mechanism between digital inclusive finance and rural industrial revitalization, revealing the intrinsic logic of their mutual promotion and synergistic development from different perspectives. However, in practice, synergistic development still faces challenges like the digital divide and financial risk prevention. Future research needs to further address these issues and propose more effective solutions to promote deep integration and sustainable development.

3 CONSTRUCTION OF THEORETICAL FRAMEWORK FOR DIGITAL INCLUSIVE FINANCE AND RURAL INDUSTRY REVITALIZATION

3.1 Characteristics of Digital Inclusive Finance

Digital inclusive finance is the product of deep integration between traditional inclusive finance and digital technologies. Its goal is to use digital means to lower the cost of financial services, improve the efficiency and accessibility of those services, and enable a broader population—especially groups excluded from or underserved by the traditional financial system, such as micro-enterprises, farmers, and impoverished populations—to enjoy comprehensive, convenient, and secure financial services at an affordable cost. In essence, digital inclusive finance extends the boundaries of traditional inclusive finance by leveraging digital technologies to overcome the temporal and spatial limitations of conventional financial services, addressing information asymmetry, and creating new paths to achieve fairness and inclusion in financial access. It is not merely the digitization of existing financial operations, but rather a technology-driven, socially

oriented innovation model that focuses on serving vulnerable groups and covers a range of financial services including payments, savings, credit, insurance, and investment.

By means of the Internet, mobile communications, and other digital technologies, digital inclusive finance breaks the geographic constraints of physical financial-institution branches and can extend financial services even into remote or economically underdeveloped areas. For example, in certain mountainous or rural regions where traditional bank branches are sparse, residents often had to travel long distances to conduct banking transactions. With mobile-banking and mobile-payment services, however, local residents can transfer funds, pay bills, and check account balances at any time and from anywhere, dramatically improving access to financial services.

Traditional financial institutions often impose stringent requirements on customers' credit histories, income levels, and asset holdings, making it difficult for many micro-enterprises and low-income individuals to obtain financial support. Digital inclusive finance uses big data and artificial-intelligence techniques to evaluate credit risk more comprehensively and accurately, thereby extending financial services to a wider population. For instance, some Internet finance platforms base small-loan decisions on multidimensional data sources such as e-commerce transaction records and social-media activity, providing micro-loans to small enterprises and sole proprietorships and addressing their funding challenges.

Digital inclusive finance encompasses more than traditional payment and savings products; it spans credit, insurance, and investment as well. As digital technologies continue to evolve, innovative financial products and services proliferate. In the credit space, beyond conventional bank loans, models such as online micro-lending and supply-chain finance have emerged. In insurance, data-driven personalized products—such as return-shipping insurance and flight-delay insurance—have appeared. In investment, Internet-based fund-sales platforms offer investors a more convenient channel for purchasing mutual funds, lowering the threshold to entry.

By offering convenient and efficient financial services, digital inclusive finance increases customer participation and usage frequency. Take mobile payments, for example: users need only scan a QR code with their phones to complete a transaction, eliminating the need to carry cash or a physical bank card. This convenience has driven mobile payments to become increasingly prevalent in daily life, with users making use of them more frequently. At the same time, digital inclusive finance platforms offer personalized financial services and product recommendations to meet diverse customer needs, further strengthening customer retention and engagement.

The development of digital inclusive finance relies on digital-technology support such as big data, cloud computing, artificial intelligence, and blockchain. These technologies play critical roles in every facet of digital inclusive finance. Big data enables more accurate credit assessments, risk alerts, and targeted marketing. Cloud computing provides robust processing power and storage capacity to support financial institutions in handling large volumes of data. Artificial intelligence powers functions like intelligent customer service and robo-advisors, raising the efficiency and quality of financial services. Blockchain addresses trust issues in financial transactions, enhancing transaction security and transparency.

Digital inclusive finance is fundamentally data-driven, collecting and analyzing vast amounts of customer information. By mining and analyzing customers' transaction records, behavioral data, and credit histories, financial institutions gain a comprehensive understanding of each customer's needs and risk profile, allowing them to deliver highly tailored financial products and services. For example, during loan approval, an institution can evaluate a customer's repayment ability and credit risk by analyzing their credit data and transaction history, leading to more accurate lending decisions. Data-driven processes also help financial institutions optimize their operations, reduce costs, and improve risk management.

3.2 Indicator System for Rural Industry Revitalization

Rural industry revitalization centers on achieving industry prosperity, which not only entails expanding agricultural production scale and increasing output but also emphasizes optimizing industrial structure and improving development quality. From the internal perspective of agriculture, coordinated development across planting, livestock, and fisheries must be realized, driving the transformation and upgrading of traditional agriculture into modern agriculture. For example, promoting precision agriculture and smart agriculture by using modern information technologies and scientific management methods can raise production efficiency and resource-utilization efficiency while lowering production costs. Simultaneously, emphasis must be placed on the quality and safety of agricultural products through standardized production and rigorous quality supervision, fostering regionally distinctive brands with market competitiveness.

Rural industry revitalization stresses integrated development of the primary, secondary, and tertiary sectors in rural areas. The primary sector provides raw materials for the secondary and tertiary sectors; the secondary sector processes agricultural products to extend the value chain and increase added value; and the tertiary sector—comprising rural tourism, e-commerce, and logistics—supports industrial development and broadens market reach. Through industry integration, the limitations of traditional agriculture are overcome, creating additional employment opportunities and economic growth points. For instance, some regions leverage local natural landscapes and agricultural resources to develop rural tourism, combining agricultural production and processing with tourism services, so that visitors experience rural life and purchase local agricultural products and handicrafts, achieving synergistic growth among sectors.Rural industry revitalization must adhere to the principles of sustainable development, achieving unity among economic, social, and ecological benefits. Throughout the development process, rural ecological and environmental protection should be prioritized, ensuring rational utilization of natural resources and avoiding overexploitation and

environmental pollution. Examples include promoting ecological-agriculture practices—employing organic fertilizers and biological pest control to reduce chemical inputs, preserve soil quality, and maintain biodiversity—and cultivating circular-agriculture practices that recycle agricultural processing waste, thereby enhancing resource-recycling rates. Simultaneously, attention to rural social development is essential: promoting farmers' income growth, improving living standards and well-being, and achieving comprehensive rural revitalization[8].

Market demand must guide rural industry revitalization, adjusting industrial structures and product offerings in response to changing market dynamics. By deeply understanding market needs, rural industries can develop suitable agricultural products and services, enhancing competitiveness. Moreover, innovation is a key driving force behind rural industry revitalization. Science and technology innovation, management innovation, and business-model innovation should be encouraged to bring new technologies, methods, and formats into rural industries. For example, leveraging Internet technologies for agricultural-product e-commerce to expand sales channels, and developing cold-chain logistics to reduce post-harvest losses and improve circulation efficiency.

An indicator system for rural industry revitalization must be built on a clear understanding of its connotation and relevant theories, with a scientific theoretical basis and logical structure. Selected indicators should accurately reflect the reality and critical factors of rural industry development, ensuring evaluation accuracy and reliability. Rural industry revitalization is a complex systems project, so the indicator system must cover various dimensions: industry scale, industrial structure, industrial performance, industry integration, etc., forming an integrated whole to comprehensively and holistically evaluate the level of rural industry revitalization. Indicators should possess clear definitions and calculation methods, with data that are readily accessible and easy to compile. At the same time, the total number of indicators should not be excessive, avoiding overly complicated calculations and evaluation processes so that the system is practical and feasible.Rural industry development is dynamic, and the indicator system must be flexible and adaptive, allowing adjustments and improvements over time to capture new realities and issues in rural industry revitalization.Reflecting the overall scale and level of agricultural production, including the combined output value of planting, livestock, and fisheries.Measuring the scale of the agricultural-product processing industry, indicating the level of value added to agricultural goods.If the locale has developed a rural-tourism sector, this indicator reflects the market size and economic benefits of rural tourism.Analyzing the rationality of the rural industrial structure, reflecting each sector's role and position within the rural economy. Ideally, the rural industrial structure should demonstrate coordinated development and mutual promotion among the primary, secondary, and tertiary sectors.Emphasizing the share of locally characteristic industries within the rural industrial sector; the growth of these specialty industries enhances rural industry competitiveness and appeal.Calculating the agricultural output value created per unit of agricultural labor, reflecting the efficiency and performance of agricultural production.Measuring the extent of value increase after agricultural products undergo processing, indicating the profitability of the agricultural-product processing industry.Directly reflecting the actual income that farmers obtain from industry development, making it a vital indicator for gauging rural industry revitalization effectiveness.Counting the number of specific projects that integrate the primary, secondary, and tertiary rural industries, reflecting the activity level of industry integration.Indicating the proportion of agricultural products sold through e-commerce channels, reflecting the integration between the Internet and the agricultural sector.Reflecting the extent to which the rural-tourism industry drives local employment, demonstrating how industry integration fosters job creation.Measuring the volume of chemical fertilizers and pesticides used in agricultural production, reflecting the environmental impact of farming practices.Indicating the comprehensive utilization level of agricultural waste (e.g., straw, livestock and poultry manure), promoting resource recycling and environmental protection.Reflecting the quality and stability of the rural ecological environment, which is crucial for maintaining ecological balance[9].

To accurately evaluate the level of rural industry revitalization, it is necessary to determine appropriate weights for each indicator. Methods such as the Analytic Hierarchy Process (AHP) and the Delphi Method can be employed, inviting experts from related fields to assess the importance of each indicator and thereby establish reasonable weights. Once indicator weights are determined, a comprehensive evaluation method—such as the weighted-average approach—can be used to quantify rural industry revitalization. Specifically, each indicator's actual value is multiplied by its corresponding weight, and then the products are summed to yield an overall evaluation score. These scores allow for comparison and analysis of rural industry revitalization levels across different regions, identifying problems and gaps and providing a basis for targeted policy and intervention measures.Through defining the connotation of rural industry revitalization and constructing an evaluation indicator system, one can more accurately capture the current status and trends of rural industry development, providing scientific decision support and practical guidance for advancing rural industry revitalization. Table 1 presents a rural revitalization evaluation indicator system comprising five subsystems and thirty specific indicators.

3.3 Theoretical Framework Construction

The theoretical framework for the collaborative mechanism between digital inclusive finance and rural industry revitalization draws on multidisciplinary theories, integrating factors from economics, finance, industrial development, and rural society. This framework is structured into four interrelated and mutually reinforcing levels—goal, element, driving, and support—each contributing to the coordinated development of digital inclusive finance and rural industry revitalization.The overarching goal of the collaborative development of digital inclusive finance and rural industry revitalization is to achieve high-quality rural economic growth and comprehensive social progress. Specifically, this

goal includes enhancing the competitiveness and added value of rural industries, promoting optimization and upgrading of rural industry structures; increasing farmers' incomes and narrowing the urban-rural income gap; improving the accessibility and quality of rural financial services and enhancing rural residents' financial literacy and capabilities; and fostering improvement of the rural ecological environment and sustainable development to realize unity among economic, social, and environmental benefits.

The element layer comprises the main players, service types, infrastructure, and industry actors that form the foundation for collaborative development. Main players include traditional financial institutions (such as rural credit cooperatives and Agricultural Bank branches), fintech companies, and nonbank financial institutions. These stakeholders innovate financial products and service models through digital means, providing diversified financial support to rural industries. Service types cover digital credit, digital payments, digital insurance, and digital wealth management. Digital credit supplies funding for rural industry development; digital payments improve transaction efficiency and convenience; digital insurance reduces industry risk; and digital wealth management helps farmers increase their assets. Infrastructure includes digital-technology platforms (mobile Internet, big data, cloud computing, blockchain) and financial infrastructure (credit systems, payment and clearing systems). These provide the technical and institutional foundation for digital inclusive finance. Industry actors comprise households, family farms, cooperatives, and agricultural enterprises; each plays a distinct role in rural-industry development and together form the rural-industry landscape. Industry categories include agriculture (planting, livestock, forestry, fisheries), agricultural product processing, and rural services (rural tourism, e-commerce, rural logistics). Industrial diversity bolsters rural industries' resilience and comprehensive benefits. The environment consists of policy, market, and technology conditions; a favorable environment attracts resource flows to rural areas and promotes industry growth.

The driving layer identifies the primary forces propelling collaborative development. Market demand is the fundamental driver: as urban and rural residents' living standards rise, demand for high-quality agricultural goods and rural tourism services increases, offering broad market opportunities for rural industries. At the same time, rural industry development generates diversified financial needs, spurring digital inclusive-finance innovation. Government policy support is a crucial safeguard: through fiscal subsidies, tax incentives, and financial-regulatory policies, the government channels financial resources to rural areas, encourages financial institutions to innovate products and services, and backs rural-industry development. Moreover, the government formulates industry plans, strengthens infrastructure, and improves the rural business environment, creating a favorable policy backdrop for collaborative development. Rapid advances in digital technology provide essential technical support: big data, cloud computing, and artificial intelligence reduce financial-service costs and risks, improving service efficiency and precision. Digital technology also enables the intelligentization and informatization of rural industries, facilitating their transformation and upgrading. Establishing sound laws, regulations, and regulatory systems standardizes market order in digital inclusive finance and rural-industry development, protects financial consumers and investors, and strengthens financial-regulatory coordination to prevent financial risks, ensuring stable collaborative development. Strengthening talent cultivation and attraction in digital finance and rural industries is another critical driver: through educational and training programs, rural residents' financial literacy and industry skills improve, and professionals in finance, technology, and management are drawn to rural areas, supplying intellectual resources. Building and perfecting rural credit systems enhances credit-information collection, integration, and sharing; through credit rating and incentive mechanisms, rural residents' and industry actors' credit awareness rises, improving the rural financial ecosystem and lowering financial institutions' credit risk.

The support layer specifies how digital inclusive finance enables rural industry revitalization and, conversely, how rural industry revitalization propels digital inclusive finance. Digital inclusive finance innovates credit models—such as big-data–based unsecured loans and supply-chain finance—to provide rural industry actors with more convenient, efficient funding. Additionally, digital inclusive finance channels social capital into rural industry investment, widening financing channels and overcoming funding-bottleneck challenges. Digital insurance, futures, and other financial tools help rural industry actors diversify and transfer risk: agricultural insurance reduces the impact of natural disasters and market fluctuations on production, while agricultural-product futures enable producers to hedge price risk and stabilize incomes. By offering diverse risk-management instruments, digital inclusive finance enhances rural industries' resilience. Digital inclusive finance also supports rural industry's technological innovation and transformation: by providing technology loans, venture capital, and other financial services, it encourages enterprises to increase R&D investment, adopt advanced production methods and management practices, and boost competitiveness and added value. At the same time, digital inclusive finance drives industry integration, promoting deeper linkages between agriculture and secondary and tertiary sectors, expanding rural-industry development space[10].

Conversely, the growth of rural industries produces diverse financial needs that create a vast market for digital inclusive finance. Building a comprehensive Chinese rural revitalization evaluation indicator system is a complex, multidimensional process: it must align with the connotation of rural revitalization while remaining scientific, feasible, and based on readily obtainable data. Table 1 illustrates an evaluation indicator system for rural revitalization, comprising five subsystems and thirty specific indicators.

Table 1 Rural Revitalization Evaluation Index System

Variables	First level indicator	Secondary indicators
		GDP per capita
		Total power of agricultural machinery
		Effective irrigation area

Rural Revitalization	Industry is booming	Per capita food production
		Total output value of agriculture, forestry, animal husbandry and fishery
		Rural per capita electricity consumption
	Ecological and livable	Number of primary medical and health institutions
		Number of nursing beds per thousand elderly people
		Park green area
		Number of public toilets per population
		Number of health technicians per population
	Rural Civilization	Number of public libraries per population
		Comprehensive population coverage of rural radio programs
		Comprehensive population coverage of rural TV programs
	Effective governance	Number of rural residents receiving minimum living security
		Completed investment in industrial pollution control
		General public budget expenditure
	Affluent life	Per capita disposable income of rural residents
		Insurance Depth
		Engel coefficient of rural households
		Per capita consumption expenditure of rural residents

As rural industries undergo transformation and upgrading, their financial-service needs will shift from traditional credit alone to more comprehensive financial services, such as payment and settlement, investment and wealth management, and risk management. The development of rural industries provides both impetus and opportunity for the innovation and advancement of digital inclusive finance. Healthy growth in rural industries helps improve the credit environment in rural areas, thereby furnishing a solid foundation of credit for the development of digital inclusive finance. When industrial actors operate steadily and maintain good repayment records, their credit ratings rise, reducing lending risks for financial institutions. At the same time, the advancement of rural industries also drives increases in rural residents' incomes, enhancing their credit awareness and repayment capacity, which in turn promotes the sustainable development of digital inclusive finance.

The digital transformation of rural industries offers concrete practice scenarios for applying digital finance technologies. For example, the growth of rural e-commerce demands convenient payment and settlement services, while the adoption of agricultural Internet-of-Things applications can supply financial institutions with more accurate production information and risk-assessment data. As rural industries become more digitized, they foster the application and innovation of digital technologies in the financial sector, thereby accelerating the growth of digital inclusive finance.

The theoretical framework for the collaborative mechanism between digital inclusive finance and rural industry revitalization is not static; it evolves dynamically alongside economic and social development and technological progress. In the short term, their collaborative development manifests primarily as financial resources being injected into rural industries and as industries beginning to take shape. Financial institutions expand credit support for rural industries, helping industrial actors overcome capital shortages and enabling initial industrial growth. Simultaneously, the emerging needs of developing rural industries provide feedback to financial institutions, prompting them to further refine financial products and services.

In the medium term, the collaborative development of digital inclusive finance and rural industry revitalization will place greater emphasis on optimizing industrial structure and innovating financial services. As rural industries continue to develop and their structures upgrade, their demands for financial services will become more diversified. Financial institutions will intensify support for emerging industries and innovative enterprises, while also creating new financial products and service models to meet these evolving needs.

Over the long term, the collaborative development of digital inclusive finance and rural industry revitalization will achieve comprehensive prosperity in the rural economy and thorough social advancement. Digital inclusive finance will deeply integrate with rural industries, forming a virtuous-cycle ecosystem. As rural industries develop, they will elevate rural residents' incomes and living standards, attract increasing talent and resources to rural areas, and promote sustainable rural social development.

By constructing a theoretical framework that defines the collaborative mechanism between digital inclusive finance and rural industry revitalization, practitioners gain theoretical guidance to foster their coordinated development, thus achieving high-quality rural economic growth and comprehensive social progress.

4 DIGITAL INCLUSIVE FINANCE'S DIRECT IMPACT ON RURAL INDUSTRY REVITALIZATION

4.1 Direct Impact of Coverage Breadth of Digital Inclusive Finance on Rural Industry Revitalization

The breadth of digital inclusive finance coverage is primarily reflected in the accessibility and prevalence of financial services, measured by indicators such as the number of financial-institution outlets, the penetration rate of accounts and bank cards, and the coverage scope of digital financial services. In rural areas, expanding the coverage of digital

inclusive finance enables more farmers and rural enterprises to access financial services conveniently. Previously, because physical outlets of financial institutions were scarce, many remote villages' industry actors could not reach formal financial services and thus could not satisfy their funding needs. As the coverage range of digital inclusive finance expands, farmers and enterprises can easily open accounts and apply for loans via mobile devices. For example, some large digital-finance platforms have partnered with rural credit cooperatives to launch online credit products, covering many rural areas that were previously untouched by traditional financial services. This has provided rural industries with additional funds to scale up production, introduce new technologies and equipment, and directly promoted the development of rural industries.Broader financial coverage also supports the diversification of rural industries. Various types of rural industries—such as specialty cultivation, animal husbandry, and rural tourism—all require corresponding capital investment. As the breadth of digital inclusive finance coverage increases, different industry actors can access financial support, encouraging diversification in rural industries. In one region, for example, under the support of digital inclusive finance, villages that once relied solely on traditional agriculture gradually developed rural homestays and agricultural-product processing industries, enriching the rural industrial structure and improving the overall competitiveness of rural industries.

4.2 Direct Impact of Usage Depth of Digital Inclusive Finance on Rural Industry Revitalization

The usage depth of digital inclusive finance reflects the frequency and diversification of financial-service utilization, encompassing payment, credit, insurance, investment, and other services. In payment, the convenient methods provided by digital inclusive finance—such as mobile payments and online banking—greatly enhance the transaction efficiency of rural industries. When farmers and enterprises procure raw materials or sell their products, they can complete payments in real time, reducing the risks and time costs associated with cash transactions.In credit, by offering personalized credit products and flexible repayment terms based on the actual needs and operating conditions of industry actors, funds can be allocated more precisely to rural industry development. For instance, to address the seasonal funding needs of crop growers, some financial institutions have introduced credit products with repayment schedules aligned to planting cycles, improving capital utilization efficiency and ensuring smooth industry operations.The increased usage depth of digital inclusive finance is also evident in the widespread application of insurance and investment services. Rural industries face multiple risks, such as natural disasters and market fluctuations; agricultural insurance helps farmers and enterprises receive compensation when losses occur, safeguarding continuous development. At the same time, some rural enterprises can participate in financial-market investments to achieve asset appreciation and further diversify their industrial portfolios, thereby strengthening their risk-resilience capacity. For example, a rural livestock-breeding enterprise purchased agricultural insurance and, when an epidemic caused livestock losses, received compensation from the insurer, averting bankruptcy and providing security for subsequent recovery and development.

4.3 Direct Impact of Digitalization Level of Digital Inclusive Finance on Rural Industry Revitalization

The digitalization level of digital inclusive finance primarily concerns the extent of technological application—such as big data, artificial intelligence, and blockchain—in financial services. The application of digital-technology enables financial institutions to process business more efficiently and reduce operating costs. Through big-data analytics, institutions can more accurately assess the credit status of rural industry actors, mitigating risks arising from information asymmetry and thereby lowering credit-approval costs and risk premiums. Meanwhile, online financial-service processes reduce the need to build and operate physical outlets, allowing institutions to serve rural industries at lower cost. This means that rural industry actors can obtain financial support at reduced expense, enhancing their profitability.Higher digitalization levels also create possibilities for innovating financial-service models. For instance, blockchain-based supply-chain finance can achieve efficient coordination of capital flows, information flows, and logistics across all links of a rural-industry supply chain. The credit of a core enterprise can be transmitted via blockchain to upstream and downstream supply-chain entities, enabling small enterprises and farmers at the chain's end to secure financial support. Additionally, artificial intelligence applications in financial services—such as intelligent customer service and robo-advisors—provide rural industry actors with a more personalized and intelligent financial-service experience, thereby promoting rural-industry development.

5 INDIRECT IMPACT MECHANISMS OF DIGITAL INCLUSIVE FINANCE

Digital inclusive finance, in promoting rural industry revitalization, does not act solely through direct channels; it also injects new vitality into rural industry development via a series of indirect pathways, such as agricultural technological innovation and industrial integration. This section delves into how digital inclusive finance indirectly influences rural industry revitalization through these channels.The development of digital inclusive finance provides robust financial support for agricultural technological innovation. Under traditional financial models, agricultural technological innovation faces difficulties in obtaining financing and high financing costs, and financial institutions often adopt a cautious stance toward agricultural technology projects due to factors such as high risk, low returns, and information asymmetry. By leveraging advanced digital technologies, digital inclusive finance overcomes temporal and spatial constraints, lowers the cost and barriers of financial services, and enables more agricultural technology enterprises and researchers to access financing support. On one hand, digital inclusive finance platforms use big data and cloud-

computing technologies to conduct precise risk assessments and credit ratings for agricultural technology enterprises and research projects, thereby offering them personalized financial products and services. For example, some digital finance institutions, based on multidimensional data—such as an agricultural technology enterprise's R&D capacity, intellectual property, and market prospects—provide credit loans to meet their funding needs for purchasing R&D equipment and recruiting talent. On the other hand, digital inclusive finance innovates financial products—such as agricultural technology crowdfunding and supply-chain finance—to attract social capital into agricultural technological innovation. Agricultural technology crowdfunding platforms offer direct financing channels to investors for agricultural technology projects, enabling projects to obtain rapid funding support; supply-chain finance, relying on core enterprises within the agricultural industry chain, provides financing services to upstream and downstream agricultural technology enterprises, enhancing the collaborative innovation capacity of the industry chain[11].

Agricultural technological innovation is a key force in driving the upgrading of rural industries. As agricultural technological innovation, supported by digital inclusive finance, advances continuously, a series of new technologies, new varieties, and new equipment are widely applied in agricultural production, improving production efficiency and quality, and promoting the diversified development of rural industries. In the planting sector, agricultural technological innovation has brought about the application of precision agriculture technologies, such as satellite remote sensing, drone monitoring, and intelligent irrigation. These technologies can monitor soil moisture and crop growth conditions in real time, enabling precise fertilization and irrigation, thereby improving agricultural resource utilization efficiency and reducing production costs. At the same time, new varieties cultivated through gene editing and hybrid breeding technologies—featuring higher yields, better quality, and stronger stress resistance—provide strong support for ensuring food security and agricultural product supply. In the livestock sector, the development of intelligent breeding technologies has transformed traditional farming models. The application of automated feeding equipment, environmental-monitoring systems, and disease early-warning systems enhances breeding efficiency and animal welfare while reducing the risk of disease outbreaks. Moreover, agricultural technological innovation has spawned new industry formats—such as agricultural product processing, rural e-commerce, and rural tourism—extending the agricultural value chain, increasing the added value of agricultural products, and broadening farmers' income channels.

6 TYPICAL REGIONAL CASE ANALYSIS

6.1 Suqian, Jiangsu: Digital Inclusive Finance Helping to Upgrade the Flower and Tree Industry

Suqian in Jiangsu Province is nationally known as the "hometown of flowers and trees," and Shuyang County within Suqian even enjoys the reputation of "China's No. 1 County for Flowers and Trees." The local history of flower and tree cultivation is long, the industry scale is considerable—covering over 600,000 mu (approximately 40,000 hectares) of flowers and seedlings—with more than 300,000 people engaged in the trade. Products range across categories such as bonsai, fresh-cut flowers, and landscaping seedlings, and they sell well throughout the country. However, the traditional development model of the flower and tree industry faces many challenges, such as growers' financing shortages, limited sales channels, and high logistics and distribution costs, all of which constrain further industrial upgrading.Local financial institutions have cooperated with e-commerce platforms to develop specialized online credit products based on multidimensional information such as growers' transaction data and credit records. For example, Suqian Rural Commercial Bank launched the "Flower and Tree Loan," which allows growers to apply for loans simply by submitting an application on their mobile phones; the system then quickly evaluates their credit status and grants an appropriate credit limit. To date, this product has provided more than RMB 300 million in cumulative loans to over 5,000 flower and tree growers, effectively resolving their funding difficulties related to seedling procurement and greenhouse construction[12].

Focusing on the core enterprises within the flower and tree industry chain, financial institutions have carried out supply-chain finance services. Taking a large local flower and tree e-commerce enterprise as an example, financial institutions provide financing services to its upstream and downstream growers, suppliers, and distributors. By passing down the creditworthiness of the core enterprise, they reduce the financing threshold for small and medium-sized enterprises and farmers along the supply chain. For instance, banks offer accounts-receivable pledge financing to suppliers based on purchase contracts with the core enterprise, accelerating cash flow turnover and ensuring stable operation of the industry chain.

As e-commerce business has expanded, digital payments have been widely adopted in flower and tree transactions. Third-party payment providers have partnered with local e-commerce platforms to roll out convenient payment and settlement systems that support multiple payment methods such as WeChat Pay and Alipay. At the same time, preferential fee rates specifically for flower and tree transactions have been introduced to reduce transaction costs. These measures not only improve transaction efficiency but also reduce the risks associated with cash transactions.

Under the support of digital inclusive finance, Suqian's flower and tree industry has continued to expand. The newly added planting area for flowers and seedlings has exceeded 100,000 mu (approximately 6,700 hectares), attracting more farmers to participate in the flower and tree industry. Meanwhile, industry development has also attracted substantial investment, promoting the extension and expansion of the industrial chain. The integration of digital finance and e-commerce has helped flower and tree enterprises and growers broaden their sales channels. Through e-commerce platforms, Suqian's flower and tree products not only sell well in major domestic cities but are also exported to Japan,

South Korea, and other countries and regions. In 2024, Suqian's flower and tree e-commerce sales reached RMB 10 billion, a year-on-year increase of 20%.

The application of digital inclusive finance has improved production efficiency and economic benefits in the flower and tree industry, directly increasing growers' incomes. Statistics show that the per capita annual income of local flower and tree growers has risen from about RMB 20,000 to over RMB 30,000, and their living standards have significantly improved.At present, a relatively authoritative measure of the current state of digital inclusive finance development is the "Peking University Digital Inclusive Finance Index (2011–2020)" compiled by Peking University's Digital Finance Research Center. This paper uses that index to analyze the current state of digital inclusive finance development in Jiangsu Province.

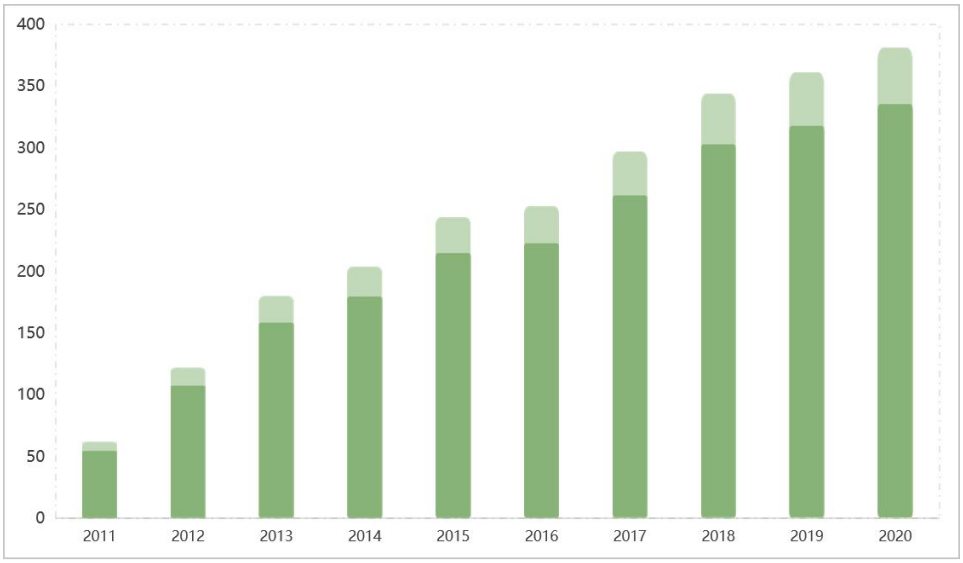


Figure 1 Jiangsu Province Digital Inclusive Finance Index 2011~2020

As shown in Figure 1, thanks to Jiangsu Province's active response to the national call, its strong emphasis on the digital inclusive finance development strategy, and the introduction of numerous supporting policies tailored to local characteristics, from 2011 to 2020 the province's digital inclusive finance business grew sevenfold over ten years, steadily achieving faster, higher-quality, and more efficient development.

6.2 Suichang, Zhejiang: Digital Inclusive Finance Activates the Rural E-Commerce Industry

Suichang County in Zhejiang is a typical mountainous county with abundant natural resources but relatively lagging economic development. In recent years, Suichang has leveraged local specialty agricultural products and traditional handicrafts to vigorously develop the rural e-commerce industry, becoming one of the nationally recognized birthplaces of rural e-commerce. However, the development of the rural e-commerce industry faces problems such as insufficient funds, difficult logistics and distribution, and a shortage of talent, which constrain further industry growth. The Suichang County government, together with financial institutions, has established a comprehensive rural credit system by collecting data on farmers' basic information, production and operation information, and credit records, creating credit files for farmers and carrying out credit ratings. Based on the credit rating results, financial institutions provide differentiated lending services to farmers. For example, farmers with higher credit ratings can obtain greater credit lines and lower loan interest rates. At present, the rural credit system in Suichang County has covered more than 90% of farmers nationwide, effectively solving the problem of farmers' difficulty in obtaining loans[13].

Local financial institutions, using fintech tools such as big data and cloud computing, have developed a dedicated e-commerce financial services platform. This platform integrates multiple financial services—payment and settlement, credit financing, and insurance protection—to provide one-stop financial solutions for rural e-commerce enterprises and entrepreneurs. For example, by analyzing e-commerce enterprises' transaction data and logistics information, the platform offers precise credit support. At the same time, it also provides credit insurance, freight insurance, and other insurance products to e-commerce enterprises, reducing their operating risks.In order to improve the accessibility and convenience of rural financial services, Suichang County has set up rural financial service stations in each administrative village. These stations are equipped with self-service ATMs, POS machines, and other financial devices, providing villagers with basic financial services such as cash withdrawal, fund transfers, and bill payments. Additionally, the service stations undertake responsibilities such as financial-knowledge promotion and credit-information collection, becoming important carriers of rural financial services.The application of digital inclusive finance has injected strong momentum into Suichang's rural e-commerce industry. Currently, there are more than 5,000 rural e-commerce operating entities in Suichang County, with over 20,000 employees. In 2024, rural e-commerce sales reached RMB 5 billion, a year-on-year increase of 30%. The development of the e-commerce industry has driven the growth of local specialty agricultural products and traditional handicrafts, promoting farmers' income growth and rural

economic prosperity. Through the construction of the rural credit system and the application of financial technology, Suichang's rural financial ecosystem has been significantly improved. The nonperforming loan rate of financial institutions has significantly declined, and the quality and efficiency of rural financial services have improved. At the same time, rural residents' financial literacy and credit awareness have been enhanced, laying a solid foundation for the sustainable development of rural finance.The deep integration of digital inclusive finance and the rural e-commerce industry has promoted the implementation of Suichang County's rural revitalization strategy. Rural infrastructure has been continuously improved, and the level of public services has been continuously raised, refreshing the rural landscape. Meanwhile, the development of the e-commerce industry has also attracted a large number of talents back to the countryside, providing talent support for rural revitalization. This paper compares per capita disposable income of rural residents in Zhejiang Province with that nationwide from 2011 to 2022; as shown in Figure 2, rural residents' disposable income in Zhejiang has consistently remained far higher than the national level, and its growth rate has also been far ahead.

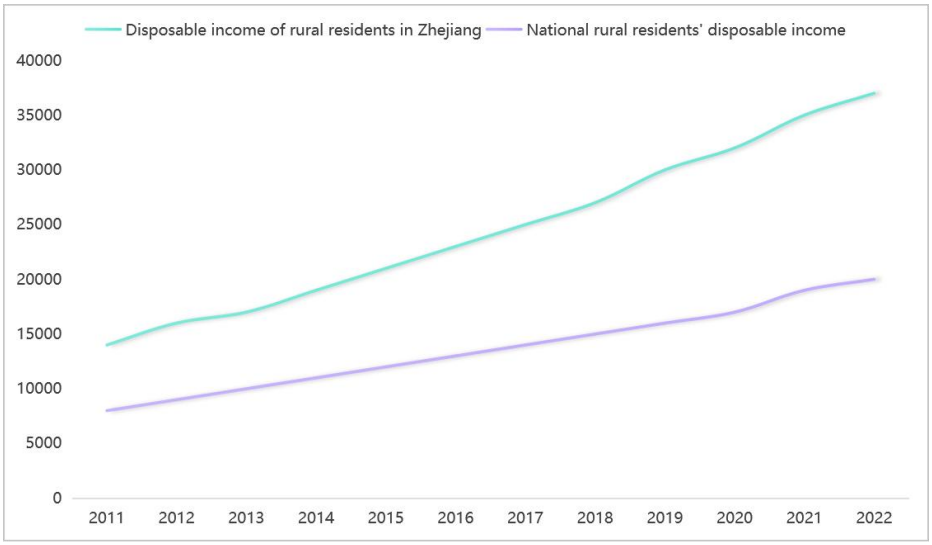


Figure 2 Disposable Income of Rural Residents in Zhejiang Province and the Whole Country (2011-2022)

6.3 Qingchuan, Sichuan: Digital Inclusive Finance Promotes the Development of Specialty Breeding Industries

Qingchuan County in Sichuan is located in a mountainous area with a superior ecological environment, making it suitable for developing specialty breeding industries. Locally, the main activities are sheep breeding, beekeeping, and poultry farming; however, due to a lack of funds, technology, and market channels, the breeding industry remains small in scale and yields low returns. At the same time, traditional breeding methods incur high costs and carry significant disease risks, which constrain the industry's development. The Qingchuan County government, together with financial institutions and insurance companies, has introduced a "government–bank–insurance" cooperation model. The government sets up a risk compensation fund to provide guarantees and credit enhancement for breeders; financial institutions offer loans to breeders; and insurance companies provide breeding insurance to reduce breeding risks. For example, breeders need only pay a small premium to obtain full insurance coverage for their breeding operations. When risks such as disease outbreaks or natural disasters occur, the insurance company compensates the breeder according to the terms of the contract.

Focusing on the local specialty breeding industry chain, financial institutions have launched digital supply-chain finance services. Take a large meat processing enterprise, for instance: financial institutions provide financing to its upstream and downstream breeders, feed suppliers, and distributors. By transmitting the core enterprise's credit to other participants in the chain, they lower the financing threshold for small and medium enterprises and farmers along the supply chain. At the same time, financial institutions utilize blockchain technology to achieve real-time monitoring of capital flows, information flows, and logistics within the supply chain, enhancing the security and efficiency of supply-chain finance.To improve the convenience of financial services, Qingchuan County's financial institutions have introduced a mobile finance service platform. Breeders can use a smartphone app at any time to apply for loans, make repayments, and check account information. The platform also provides value-added services such as breeding-technology consultations and market-trend analyses, offering breeders comprehensive financial support.

Under the support of digital inclusive finance, Qingchuan County's specialty breeding industry has steadily expanded in scale. The numbers of sheep, bee colonies, and poultry raised have each grown by more than 20%, 30%, and 40%, respectively, attracting more farmers into specialty breeding. Meanwhile, industry development has also drawn significant investment, promoting the standardization, scale expansion, and industrialization of breeding operations. The application of digital inclusive finance has improved the production efficiency and economic returns of the specialty breeding industry, directly increasing breeders' incomes. Statistics indicate that the per-capita annual income of local breeders has risen from around RMB 15,000 to over RMB 25,000, and their living standards have significantly

improved.Through the “government–bank–insurance” cooperation model and the promotion of breeding insurance, the risks in Qingchuan County’s specialty breeding industry have been effectively reduced. When breeders face risks such as disease outbreaks or natural disasters, they can receive timely compensation, minimizing economic losses. At the same time, the application of digital supply-chain finance services and the mobile finance service platform has enhanced the industry’s risk resilience and market competitiveness[14].

From the three typical regional case studies above, it is clear that digital inclusive finance plays a crucial role in rural industry revitalization. In different regions, by leveraging their own industry characteristics and development needs, local stakeholders have explored digital inclusive finance models suited to their circumstances. These models have effectively addressed financing challenges in rural industry development, promoted industrial upgrading, and increased farmers’ incomes. Such successful experiences offer valuable lessons and references for other regions.

7 PATHS AND MEASURES FOR PROMOTING THE DEVELOPMENT OF DIGITAL INCLUSIVE FINANCE

7.1 Optimizing the Policy Environment for Digital Inclusive Finance

The government should establish a dedicated special fund for the development of digital inclusive finance, providing subsidies to financial institutions engaged in digital inclusive finance services. For financial institutions that serve key inclusive finance targets—such as micro and small enterprises and rural residents—with digital financial services, a certain proportion of fiscal interest subsidies should be granted based on metrics such as number of clients served, loan amounts, and service quality, in order to lower service costs and enhance their enthusiasm for offering digital inclusive finance products. In terms of taxation, for financial institutions engaged in digital inclusive finance, appropriate reductions or exemptions of value-added tax and income tax should be provided. For example, for institutions issuing small digital loans, a certain percentage of value-added tax on interest income may be waived; for institutions that establish digital financial service outlets or conduct digital financial services in rural areas, a specified period of income tax relief should be granted.The government should formulate clear development plans and policy guidance for digital inclusive finance, directing financial resources toward inclusive finance areas. Financial institutions should be encouraged to increase their investment in digital financial services for rural, remote regions and micro and small enterprises, and through policy guidance, drive innovation in digital financial products and services to meet the diverse financial needs of different groups. At the same time, the government can use industrial policy guidance to promote the application and development of digital technologies in inclusive finance. For instance, support should be given to digital financial technology enterprises, with subsidies and tax incentives provided to those engaged in research, development, and application of digital financial technologies, thereby advancing innovation and application of digital financial technologies and improving the efficiency and quality of digital inclusive finance services.

With the rapid development of digital inclusive finance, traditional regulatory methods can no longer meet evolving needs. Regulatory authorities should actively innovate their supervisory approach, introducing technology-driven regulatory tools that use big data and artificial intelligence to strengthen oversight of digital inclusive finance activities. A big-data regulatory platform for digital inclusive finance should be established to collect financial institutions’ business data and customer information in real time; through data analysis and mining, potential risks can be identified promptly, enabling dynamic supervision of digital inclusive finance. Meanwhile, regulatory authorities should strengthen collaboration with fintech companies, establishing joint regulatory mechanisms to address new issues and challenges arising in digital inclusive finance development. For example, data resources can be shared with fintech firms to enhance monitoring and evaluation of digital financial products and services, improving the effectiveness and precision of regulation.

To avoid gaps and overlaps in regulation, the responsibilities of each regulatory body in the field of digital inclusive finance must be clearly defined. A coordinated regulatory mechanism for digital inclusive finance should be established, enhancing communication and collaboration among different regulatory departments to form a cohesive supervisory force. At the same time, unified regulatory standards and norms for digital inclusive finance should be formulated, clarifying requirements for market entry, business rules, and risk management of digital financial products and services. Compliance supervision of digital financial institutions must be strengthened, with severe penalties imposed on violations in accordance with law, in order to maintain order in the digital inclusive finance market.Consumer protection is crucial in the development of digital inclusive finance. Regulatory authorities should reinforce supervision of financial institutions, requiring them to fully disclose product information and risk warnings when offering digital financial services, thereby safeguarding consumers’ right to information and choice. A robust complaint-handling mechanism for digital inclusive finance should be established, with clear channels for complaints and timely resolution of consumer disputes. Financial literacy and risk-awareness training for customers should be strengthened to guide consumers toward rational use of digital financial products and services.Digital inclusive finance development involves multiple departments, such as financial regulators, fiscal authorities, and technology departments. These departments should enhance communication and collaboration to form a policy synergy. For example, financial regulators formulate regulatory policies for digital inclusive finance; fiscal authorities provide subsidies and tax incentives; and technology departments promote the application and innovation of digital technologies in inclusive finance. Through interdepartmental policy coordination, the healthy development of digital inclusive finance can be jointly advanced.

Because different regions in China exhibit significant disparities in economic development levels and financial ecosystems, the development of digital inclusive finance is also unbalanced. The government should formulate differentiated digital inclusive finance policies, adopting targeted support measures based on each region's actual conditions to promote coordinated regional development. For economically developed areas, financial institutions should be encouraged to innovate digital financial products and services to improve the service quality and efficiency of digital inclusive finance. For economically underdeveloped areas, policy support should be increased, digital financial infrastructure construction strengthened, and the accessibility and coverage of financial services enhanced. At the same time, interregional financial cooperation and exchanges should be bolstered to optimize allocation of digital inclusive finance resources.

7.2 Enhancing the Service Capability of Digital Inclusive Finance

7.2.1 Strengthening financial infrastructure construction

The government should increase investment in network and communication infrastructure in remote rural and underdeveloped areas by setting up special funds specifically for laying optical-fiber networks and constructing 5G base stations, for instance. In certain mountainous villages where complex terrain results in weak network coverage, digital inclusive finance services cannot be effectively provided. By supporting these areas with special funds to introduce advanced communication technologies and equipment, a stable and high-speed network can be ensured, providing the necessary foundation for the popularization of digital financial services.Telecom operators possess abundant network resources and customer data, while financial institutions have professional capabilities in financial services. Their collaboration can achieve resource sharing, such as jointly offering preferred packages for rural regions that bundle network services with financial services, lowering the threshold for rural residents to use digital financial services. Cooperation and interoperability between financial institutions and third-party payment providers should be promoted. A unified payment and clearing standard and interface specifications should be established so that payment and clearing between different institutions is more seamless. For example, commercial banks should strengthen cooperation with third-party payment platforms like Alipay and WeChat Pay to achieve account interoperability and rapid fund transfers, improving payment and clearing efficiency and facilitating various digital financial transactions for residents.Advanced encryption and identity-authentication technologies should be employed to secure payment information. For instance, blockchain technology can be used to encrypt and store payment and clearing data in a distributed manner, preventing data tampering and leakage and improving the stability and reliability of the payment and clearing system.

7.2.2 Improving the Adaptability of Financial Products

Financial institutions should conduct in-depth research on the financial needs of different groups. For micro and small enterprises, innovative credit products should be developed using receivables and intellectual property rights as collateral. Since micro and small enterprises often lack traditional collateral, leveraging the value of their receivables and intellectual property can provide financing support and resolve their financing difficulties. For rural residents, financial products aligned with agricultural production cycles should be designed. For example, small loans with repayment terms set according to crop planting and harvest cycles can be introduced to provide funding support at the start of planting and allow repayment after harvest and sales, thereby reducing repayment pressure on rural residents.New financial services that combine digital technologies should be launched. For example, big data and artificial intelligence can be used to offer robo-advisory services. Based on a customer's risk preference and asset status, personalized investment-portfolio recommendations can be provided to lower the entry barrier for investment, enabling more ordinary residents to participate in the investment market.Supply-chain finance should be developed with core enterprises as the foundation: by integrating information flows, capital flows, and logistics flows within the supply chain, financing services can be offered to small and medium enterprises along the chain. Digital technologies should be used to achieve real-time sharing and monitoring of supply-chain information, improving risk-control capacity in supply-chain finance.

7.2.3 Strengthening financial technology talent cultivation

Universities should optimize curriculum offerings for fintech-related majors. Frontier-technology courses—such as artificial intelligence, blockchain, and big data—should be added, along with financial professional courses like risk management and product design, to train compound talents who understand both finance and technology. Practical opportunities should be provided so that students can accumulate experience through real projects. Financial institutions can recruit outstanding graduates from universities to strengthen their own fintech talent pools.

Financial institutions should regularly organize in-house fintech training courses for employees, inviting industry experts and technical backbones to teach. The content should cover the latest fintech technologies, business models, and regulatory policies to enhance employees' fintech literacy and business capabilities. Employees should be encouraged to obtain industry certifications—such as fintech analyst or blockchain engineer certifications—and rewarded or given career-development support upon passing, incentivizing continuous improvement of their professional skills.

7.2.4 Strengthening data governance and application

Financial institutions should establish strict data-collection standards and processes. The scope, method, and purpose of data collection must be clearly defined to ensure legality, accuracy, and completeness of data. When collecting customer data, explicit authorization should be obtained, and customers informed about data usage methods and purposes. Techniques such as data encryption and access control should be adopted to prevent data leakage and misuse.

A data-security emergency response plan must be established to ensure timely response and resolution in case of data-security incidents, safeguarding customer data.

Big-data analytics should be used to deeply mine customer data to understand their consumption habits, credit status, and financial needs, enabling financial institutions to provide more precise financial services. For example, by analyzing customers' transaction data, suitable financial products and services can be recommended. Under lawful and compliant premises, data sharing and exchange between financial institutions and other relevant entities should be realized. By integrating multiple parties' data, service efficiency and quality can be improved, and financial risks prevented.

7.3 Promoting Deep Integration of Rural Industries and Digital Inclusive Finance

Deep integration of rural industries and digital inclusive finance is an important approach to promote high-quality rural economic development and achieve rural revitalization. Through their organic combination, new vitality can be injected into rural industry development, while financial-service accessibility and coverage are enhanced, helping farmers increase income and become prosperous. To promote their deep integration, the following specific measures can be taken:

The government should collaborate with financial institutions, technology enterprises, and other stakeholders to build a digitalized rural-industry service platform that integrates production, sales, and financial services. This platform should integrate upstream and downstream resources of rural industries and provide services such as technical guidance for agricultural production and breeding, market-trend analysis, and product-sales channel matching. Simultaneously, modules for digital inclusive finance services should be incorporated to offer one-stop online financing, payment and settlement, and insurance services to farmers and new agricultural business entities. For example, a dedicated "financial supermarket" could be set up on the platform to showcase various financial products suitable for rural industries; users can compare and select according to their needs and submit applications online. Rural supply-chain finance should be developed with core enterprises of rural specialty industries as the foundation. Financial institutions, based on actual transaction data in the supply chain and the credit of core enterprises, can provide financing support to upstream and downstream farmers and micro and small enterprises. For example, in the agricultural-product processing sector, a financial institution can provide farmers with prepayment financing according to purchase contracts between a core processing enterprise and farmers, resolving farmers' production-funding shortages. Meanwhile, digital technology should be used to monitor logistics, capital flows, and information flows within the supply chain in real time, reducing financial risks. In addition, core enterprises should be encouraged to cooperate with financial institutions to establish supply-chain finance service platforms, achieving online, automated operations of supply-chain finance.

Support for digital agriculture should be increased to cultivate new formats such as smart agriculture, rural e-commerce, and rural tourism. The government should introduce policies that encourage agricultural enterprises and farmers to adopt digital technologies such as the Internet of Things, big data, and artificial intelligence, achieving intelligent and precise agricultural production management. Financial institutions should provide targeted loans and venture capital for digital agriculture projects. For example, for farmers and enterprises undertaking smart-agriculture projects, financial institutions can offer preferential-interest-rate loans and make dynamic assessments and adjustments based on project implementation progress and outcomes. At the same time, rural e-commerce and digital inclusive finance should be integrated to provide rural e-commerce operators with convenient payment and settlement and small loans, promoting the upward flow of agricultural products and the downward flow of industrial goods.

Financial institutions, universities, and social organizations should collaborate to conduct financial-knowledge education activities in rural areas. A combination of online and offline methods should be used—such as holding financial-knowledge seminars, distributing educational materials, and producing short videos—to spread basic concepts, product types, operation procedures, and risk-prevention knowledge of digital inclusive finance to farmers. For example, financial-knowledge training classes can be regularly held in rural cultural activity centers and village committees, inviting financial experts and business specialists to provide on-site lectures and answer questions. Meanwhile, financial-knowledge short videos can be published on new-media platforms such as WeChat official accounts and Douyin, enabling farmers to learn anytime. In selected villages with favorable conditions, rural financial-education demonstration bases should be established as platforms for financial-knowledge dissemination and practice. Demonstration bases should be staffed with professional financial-education personnel to conduct regular training and promotional activities. A financial-experience area should be set up for farmers to personally experience operation processes of digital inclusive finance products—such as mobile banking and mobile payments—to improve their understanding and use of digital inclusive finance. Furthermore, demonstration bases should partner with financial institutions to provide farmers with consultation and application services for small loans, insurance, and other financial products, helping them address actual financial needs.

Universities and vocational colleges should be encouraged to establish rural-finance-related majors to train financial talent suited to the development needs of rural industries. A specialized recruitment and training mechanism for rural financial talent should be established to attract outstanding financial professionals to work in rural areas. For example, the government can introduce preferential policies—such as housing subsidies and relocation allowances—to graduates who work at rural financial institutions. Meanwhile, existing rural financial practitioners should receive enhanced training and evaluation to improve their professional and service capabilities. Regular professional training sessions and exchange activities should be organized for rural financial practitioners to update their knowledge and enhance their

ability to serve rural industries.Investment in network and communication infrastructure in rural areas should be increased to improve network coverage and quality. The government should allocate special funds to support telecom operators in constructing base stations and laying optical-fiber networks in rural areas, especially focusing on strengthening coverage in remote mountainous and impoverished regions to eliminate network blind spots. Operators should be encouraged to lower network fees in rural areas to reduce the burden on farmers and rural enterprises. By improving network conditions, solid technical support is provided for the development of digital inclusive finance, making it convenient and efficient for farmers to use digital financial services.

Financial institutions should be promoted to deploy self-service banks, POS machines, and other payment and settlement devices in rural regions to expand payment service coverage. They should cooperate with rural e-commerce platforms and supply cooperatives to establish rural payment-service points that provide villagers with convenient cash withdrawal, transfers, remittances, payments, and other payment and settlement services. At the same time, the promotion of mobile payments should be intensified to guide farmers to use mobile banking, WeChat Pay, Alipay, and other mobile payment tools. By optimizing the rural payment and settlement environment, fund-transfer efficiency is improved, thereby stimulating active development of the rural economy.Government departments, financial institutions, enterprises, and other stakeholders should integrate their information resources to establish a credit-information system covering rural residents and rural enterprises. Leveraging big data and cloud computing technologies, comprehensive and accurate assessments and analyses of farmers' and rural enterprises' credit status should be conducted. Financial institutions can provide more precise financial services based on credit reports generated by the credit-information system, thereby reducing financial risks. For example, for farmers and rural enterprises with good credit, financial institutions can offer higher credit lines and more favorable interest rates. At the same time, management and maintenance of the rural credit-information system should be strengthened to ensure data security and accuracy.

The government should introduce fiscal and tax incentives to encourage financial institutions to increase their digital inclusive finance support for rural industries. For financial institutions providing digital inclusive finance services to rural industries, fiscal interest subsidies, tax reductions, and other incentives should be granted. For instance, a certain proportion of fiscal interest subsidies can be provided to institutions issuing rural microloans based on loan amounts to reduce institutions' funding costs. Meanwhile, institutions that establish branches and service outlets in rural areas should receive tax exemptions and fiscal subsidies to enhance their willingness to serve rural areas.The government should establish a rural-industry digital inclusive finance risk-compensation fund with public investment to compensate financial institutions for losses incurred from unavoidable factors or market risks in the process of conducting digital inclusive finance business. The risk-compensation fund should compensate institutions according to established standards and procedures to ease their risk-bearing pressure. Concurrently, financial institutions should be guided to strengthen risk management, establish comprehensive risk-warning and disposal mechanisms, and enhance risk-prevention capacity.

Supervision of digital inclusive finance business must be strengthened to standardize financial institutions' business operations and market conduct. A robust regulatory system for digital inclusive finance should be established, with strengthened oversight of financial-product innovation, information disclosure, and consumer protection. In particular, illegal fund-raising, fraud, and other crimes in the digital inclusive finance field should be prevented to maintain stability and security in the rural financial market. At the same time, regulatory coordination and cooperation among financial authorities should be enhanced to form a supervisory synergy and improve regulatory efficiency.By implementing the above measures, the deep integration of rural industries and digital inclusive finance can be promoted, providing stronger financial support for rural industry development and driving the prosperity of the rural economy and the smooth implementation of the rural revitalization strategy.

8 CONCLUSION

This study delved into the synergistic relationship between digital inclusive finance and rural industry revitalization, uncovering significant complementary effects. Digital inclusive finance enhances the accessibility of financial services—for example, mobile payments and online credit—which lowers entry barriers and costs, promotes the optimization and upgrading of industrial structures, and channels funds toward rural specialty industries, emerging industries, and high–value-added industries. It also stimulates innovation and entrepreneurship by providing entrepreneurs with financing guarantees and market information, thereby offering critical support for rural industry revitalization. Its mechanisms encompass: technological empowerment—using big data and artificial intelligence to improve service efficiency; information sharing—breaking down information barriers to increase industry transparency; and industrial integration—linking finance, technology, and e-commerce to achieve coordinated development. Simultaneously, rural industry revitalization expands the demand for financial services, improves the quality of financial assets, and drives financial innovation—for example, by developing rural supply-chain finance and agricultural specialty insurance products—thus creating favorable conditions for digital inclusive finance. The two processes promote and depend on each other.Furthermore, this study also recognizes limitations in research on shared-bicycle deployment strategies: namely, data constraints (a lack of user-behavior and psychological data and insufficient spatiotemporal coverage), static strategy design lacking dynamic adjustment mechanisms, and an incomplete consideration of factors such as urban cultural atmosphere and policies and regulations. Future work must expand data collection, optimize modeling algorithms, establish dynamic adjustment mechanisms, and strengthen multi-factor comprehensive research to enhance strategy effectiveness.Looking ahead, efforts should deepen the integration of

digital inclusive finance and rural industry revitalization, fully leveraging their synergistic effects to support the broader rural revitalization agenda.

COMPETING INTERESTS

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

REFERENCES

[1] Li Jianjun, Han Xun. Inclusive finance, income distribution and poverty alleviation - Policy framework choices for promoting efficiency and equity. Journal of Financial Research, 2019(03).
[2] Huang Zuhui, Song Wenhao, Ye Chunhui, et al. The impact and mechanism of digital inclusive finance on the establishment of new agricultural business entities - Empirical evidence from 1845 counties in China. Journal of Financial Research, 2023(04).
[3] Sun Jiguo, Sun Yao. Has financial technology promoted rural industrial revitalization under the common prosperity goal? Collected Essays on Finance and Economics, 2022(11).
[4] Yang Shuigen, Wang Ji. The mechanism and effect of circulation digitalization enabling rural industrial revitalization. Journal of South China Agricultural University (Social Science Edition), 2023, 22(02).
[5] Gong Pan, He Yuanli. The influence mechanism and spatial effects of digital inclusive finance on rural industrial revitalization. Statistics and Decision, 2025, 41(02).
[6] Chen C, Restuccia D, Santaeulalia-Llopis R. Land Misallocation and Productivity. American Fconomic Journal,Macroeconomics, 2023(2): 441-465
[7] SehrawaT M, Giri A K. The role of financial develop-ment in economic growth:Empirical evidence from Indian states. International Journal of Emerging Markets, 2015, 10(4): 765-780.
[8] Oliver W. Accelerating financial inclusion in South-East Asia with digital financelR. Asian Development Bank, 2017.
[9] Yang Xinyao, Ye Jiashuo, Li Jie, et al. How can digital inclusive finance enable rural revitalization? - Based on the perspective of industrial structure upgrading. Southern Finance, 2024(06).
[10] Liu Zhenchao, Zhang Xi, Wang Jing, et al. Exploring the development of digital inclusive finance to promote the development of rural e-commerce. Guangdong Sericulture, 2023, 57(12).
[11] Yang Wei. Research on the synergistic development of digital inclusive finance and rural e-commerce logistics. Logistics Sci-Tech, 2024, 47(08).
[12] Zhu Jinyao. Research on the impact of digital inclusive finance on Chinese residents' income. Suzhou: Soochow University, 2023.
[13] Liao Xiangyue, Jin Haihong. Analysis of the impact of digital inclusive finance on rural economic growth - An empirical study based on provincial panel data from 31 regions. Modern Business Trade Industry, 2023, 44(09).
[14] Su Mi. Wealth stratification, digital inclusive finance and farmer entrepreneurship. Fuzhou: Fujian Agriculture and Forestry University, 2022.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3038

OLYMPIC MEDAL QUANTITY FORECASTING: A RANDOM FOREST ALGORITHM-BASED MODEL CONSTRUCTION

JunBo Zhu*, LinFeng Li
School of Mathematical and Physical Sciences, Chongqing University of Science and Technology, Chongqing 401331, China.
Corresponding Author: JunBo Zhu, Email:18983362087@163.com

Abstract: Against the backdrop of the unstoppable wave of globalization in sports, the competition for Olympic medals has shown an increasingly fierce trend. Countries have invested a lot of resources to improve their performance in the Olympic Games in order to be in a favorable position in the medal competition. In this study, a random forest model is developed to predict the number of gold medals and the total number of medals of each country in the 2028 Olympic Games. Firstly, the data were obtained from the official website of the Olympic Games and data preprocessing was carried out. After completing data cleaning and organizing, a series of key influence indicators such as whether it is the host country, the number of athletes, the total score and so on are introduced, and then a random forest model is built to predict the total number of medals and gold medals of each country. Finally, based on the prediction results, it was determined that in the 2028 Olympic Games, countries such as Cuba, Germany and Slovakia have the potential to achieve breakthroughs, while countries such as Belgium, Ecuador and Israel may experience a decline in the acquisition of medals. This study breaks through the limitations of linear assumptions in traditional econometric models, utilizes the nonlinear fitting ability of the Random Forest algorithm to capture complex variable interactions, and quantifies the dynamic impact of the 'host effect' on the distribution of medals, and reveals the role weights of the core factors such as historical performance and participation size through characteristic contribution analysis. Meanwhile, the prediction results can provide scientific basis for the National Olympic Committees to optimize resource allocation and formulate strategies, sports economic research and event public opinion prediction.
Keywords: Random forest model; Olympic medal prediction; Data preprocessing; Prediction accuracy

1 INTRODUCTION

With the advancement of sports globalization, the competition in Olympic events has become increasingly intense. Olympic medals, as symbols of a country's sports strength, have received extensive attention. Against this backdrop, predicting the distribution of medals in Olympic events has become a focal point in the sports community.

Currently, scholars at home and abroad have conducted research on the issue of Olympic medals from multiple perspectives. Scelles N, Andreff W, Bonnal L, et al. [1] developed Tobit and Hurdle models to predict the number of medals and verified the effectiveness of the econometric models. Bredtmann J, Crede C J, Otten S. [2] constructed a combined model of regression analysis and time - series to analyze the influencing mechanisms of Olympic medals. Andreff [3] established a Tobit estimation model to analyze the influencing factors of Winter Olympics medals and predict the performance of Russia and China. Vagenas G, Vlachokyriakou E. [4] used a log - linear regression model to study the relationship between economic indicators and Olympic medals, providing a basis for medal prediction. Cheng H R, Lü J, Yuan T G. [5] employed mathematical statistics and other methods to analyze China's track and field performance in the Olympics. Liu C Y, Wu M Q, Zhang A A, et al. [6] used a spatial analysis model to reveal the spatial - temporal distribution patterns and influencing factors of China's medals. Ding Weizhe[7] used data mining and statistical analysis models to analyze the influencing factors of the ranking on the Olympic medal table. Balmer N J, Nevill A M, Williams A M. [8] adopted a generalized linear interactive modeling approach and found that the home - field advantage was significant in event groups with subjective judgment or decision - making.

Previous studies mostly relied on traditional models such as the Tobit model and regression analysis. Constrained by linear assumptions, these models struggle to depict complex non - linear interactions, such as the combined effects of the host identity and the scale of athletes. Meanwhile, when faced with multi - dimensional data, traditional models are difficult to handle due to the problem of multicollinearity, resulting in insufficient fitting accuracy. In contrast, the random forest model, with its strong noise - resistance and high data - processing efficiency brought by decision - tree integration, demonstrates flexibility and prediction accuracy in predicting the number of medals in sports events.

At present, scholars both at home and abroad have carried out prediction - related research using random forest. Bao Y, Meng X, Ustin S, et al. [9] constructed a random forest model based on Vis - SWIR spectroscopy and CARS to predict soil organic matter content, providing an effective method for soil organic matter estimation. Hafezi M H, Liu L, Millward H. [10] established a random forest model combining CART and curvature search to analyze individual daily activity sequences, and proposed that the model had the best accuracy in simulating activity agendas and sequences. Zhang Y D, Senjyu T, Chakchai S, et al. [11] developed an RF - DBSCAN model to predict road travel time, with relatively high prediction accuracy. Shi H M, Zhang D Y, Zhang Y H. [12] used a random forest combined with an interpretable model to evaluate the predictability of medals in Olympic events and found that socio - economic factors had a significant impact. Sun J, Zou X K, Zhu X B, et al. [13] established a random forest model to solve the problem of

medical registration, with an efficiency five times higher than that of traditional models. Yang Q F, Li T, Jia Z Q. [14] constructed a random forest model optimized by a genetic algorithm to predict retail consumption behavior, showing better accuracy. Gan M, Liu P F, Yue D B, et al. [15] constructed a random forest mineral - prospecting model based on geoelectrochemical data to explore lithium deposits, and the AUC value exceeded 80% after training.

Based on previous research achievements on Olympic medals, this study uses a random forest model to predict the number of gold medals and total medals of various countries in the 2028 Olympic Games. The research first collects historical competition data from the official Olympic website and preprocesses it, selects a series of key influencing indicators such as whether a country is the host, and then constructs a random forest regression model to complete the prediction, yielding the predicted values of the medal standings for participating countries in 2028. By comparing the predicted values with historical results, countries with potential for performance breakthroughs and those likely to decline are identified, and targeted recommendations are proposed accordingly.

2 METHODS

2.1 Data Preprocessing

This paper obtain the data from the official website of the Olympic Games. For data preprocessing, only records from 1950 onwards were retained. Prior to 1950, Olympic participation was limited, with fewer countries competing, which made the data insufficient to reflect nations' true athletic capabilities. Frequent wars between the early 20th century and the 1940s (e. g. , World War I and II) further disrupted regular participation, resulting in sparse and unreliable data that could skew research outcomes. Post-1950, improved international stability and increased global cooperation expanded participation, yielding more comprehensive and accurate data that better capture countries' sustained Olympic performance.

After filtering, missing data were addressed through a systematic approach. Countries with over 50% missing values in critical metrics (e. g. , medal counts, event participation) were excluded to avoid biasing predictions. For gaps between consecutive Olympic editions (e. g. , available data for Editions N and N+2 but missing for N+1), values were imputed using the mean of the adjacent editions' medal counts. For nations absent from recent Olympics due to political events or natural disasters but with historically relevant performance data, their most recent valid records were carried forward to maintain predictive continuity. These procedures enhanced data integrity, ensuring the dataset was robust for modeling and predictions aligned more closely with real-world outcomes.

2.2 Data Analysis Methods

2.2.1 Key metrics

This study analyzed publicly available data from past Olympic Games and extracted key metrics including host country status, number of athletes, total medals, proportion of medals per country, proportion of gold medals per country, number of events participated in, gold medal count, and score—with the latter serving as an indicator of a country's sporting strength based on Balmer et al. [8]scoring methodology: each gold medal is valued at 3 points, silver at 2 points, and bronze at 1 point, with the total score calculated as the sum of points from all medals won.

$$score = gold * 3 + silver * 2 + bronze * 1 \tag{1}$$

2.2.2 Standardization

When conducting an in-depth analysis of publicly available data from past Olympic Games, data standardization is a critical step. The indicators used in predictive models vary significantly in dimensions and value ranges; without standardization, subsequent data analysis and model building would be severely disrupted. Z-Score normalization, based on the mean and standard deviation of the data, transforms values into a standard normal distribution with a mean of 0 and a standard deviation of 1.

$$x_{new} = \frac{x - \mu}{\sigma} \tag{2}$$

Here, μ represents the mean of the dataset, and σ denotes the standard deviation.

2.2.3 Contribution of characteristic variables

The contribution of each feature variable to the model's prediction result is interpreted as "the impact of the variable (x) on the final prediction outcome (y) during the prediction process. " Based on this core definition, this study uses the additive feature attribution method and related formulas provided by Shi Huimin et al. [12] to analyze the association between feature variables and prediction results. The "total prediction contribution" of the predictive model can be expressed as:

$$g(x) = \varphi_0 + \sum_{i=1}^{M} \varphi_i(x)\mathbb{1}(x_i) \tag{3}$$

where $X = (x_1, \cdots, x_M)$ is an M-dimensional explanatory or feature variable. $\mathbb{1}(x_i) \in \{0, 1\}$ is a binary indicator variable, where $\mathbb{1}(x_i)$ means the i-th feature variable is used for prediction, and 0 means it is not. $g(x)$ represents the final prediction result, ϕ_0 represents the average value of predictions, and $\phi_i(x)$ represents the marginal contribution of the feature variable to the prediction result. The key issues measured in this study are addressed through

the additive feature - attribution method. Specifically for the research questions of this study, $g(x)$ represents "the logarithmic value of the number of awards/gold medals won by a certain team in a certain event in a certain year", ϕ_0 represents the average number of awards/gold medals won by all teams in that event, and x_i represents the value of the i-th feature variable. Through calculation, the impact of changes on the number of gold medals and awards can be obtained, so as to identify the features that contribute more significantly to predicting medal changes.

2.3 Random Forest Prediction Model

The Random Forest Prediction Model falls within the realm of Ensemble Learning. Ensemble Learning aims to construct a powerful learner by combining multiple weak learners, thereby overcoming the limitations of traditional single - prediction models. Traditional single - prediction models, such as simple linear regression models or single decision - tree models, often suffer from overfitting or underfitting when dealing with high - dimensional and complex - distributed data. This leads to unsatisfactory prediction accuracy and poor generalization ability of the models, meaning they perform decently on the training set but experience a significant drop in performance on new test data. The Random Forest Prediction Model, however, effectively mitigates these issues through its unique construction and integration strategies. Here is its specific operational process.

2.3.1 Bootstrap sampling to build training subsets

Bootstrap sampling method is applied to randomly draw a training subset of the same capacity of N from the original dataset of size N with putback. This type of putative sampling is characterized by the fact that the same sample may or may not be drawn multiple times in a single sampling. In this way, the training subset obtained from each extraction differs somewhat from the original dataset, preserving the main features of the original data while introducing randomness. This has the advantage of providing different but related training data for each decision tree constructed subsequently, allowing each tree to learn the features of the data from different perspectives and enhancing the diversity of the model. For example, in an original dataset containing 1000 samples, the training subset obtained by Bootstrap sampling may have some of the samples duplicated and others not sampled, although the sample size is also 1000.

2.3.2 Randomly selected subset of split features

When a data sample has M features, m (where m << M) features are randomly selected as the subset of splitting features for constructing a decision tree. Limiting m to be much smaller than M is to prevent one or a few features from dominating the decision - tree construction process, thus enabling the model to learn more complex relationships among features. For instance, when dealing with a dataset that has 50 features (M = 50), perhaps only 5 features (m = 5) are randomly chosen to build the splitting nodes of the decision tree. This random feature - selection approach increases the model's randomness and robustness, allowing different decision trees to grow based on different feature combinations and further enriching the model's diversity.

2.3.3 Decision tree growth

Each decision tree is allowed to grow fully during construction without pruning. During the growth of a decision tree, based on the selected feature subset, nodes are continuously split according to certain criteria (such as information gain, Gini coefficient, etc.) until stop conditions are met (for example, the number of samples in a node is less than a certain threshold, or all samples belong to the same category). Omitting the pruning operation is to enable each tree to learn the information in the training data to the greatest extent and uncover potential complex patterns in the data. However, this may also lead to overfitting of a single tree on the training set. Nevertheless, within the overall framework of the random forest, integration of multiple trees can effectively alleviate this problem.

2.3.4 Constructing a random forest

The above steps (1)–(3) are repeated to construct a large number of decision trees, which collectively form the random forest. The number of decision trees can be adjusted according to practical needs. Generally, a larger number of trees may improve the model’s stability and prediction accuracy, but it also increases computational costs and training time. For example, in some practical applications, dozens or even hundreds of decision trees may be constructed to form the random forest.

2.3.5 Generation of forecast results

For regression tasks, each decision tree in the random forest generates a numerical prediction for the input sample. The final prediction of the random forest model is obtained by averaging the predictions of all decision trees. This ensemble averaging reduces variance and enhances the robustness of the prediction, leveraging the diversity of individual tree outputs to produce a more stable and accurate result.

The algorithm workflow is shown in Figure 1:

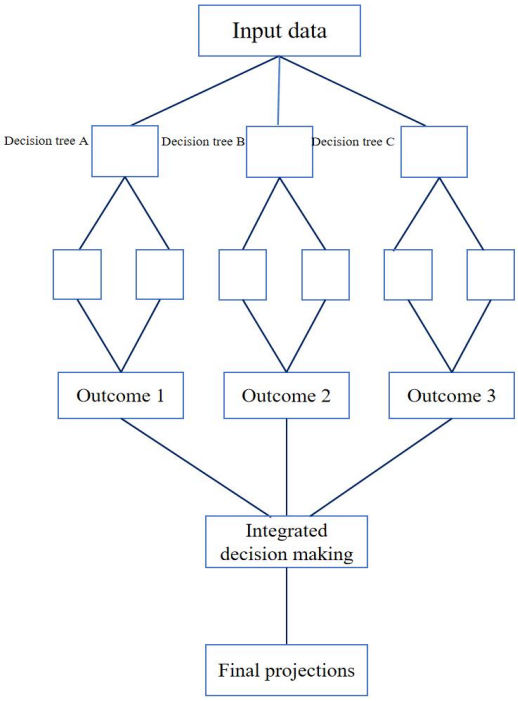


Figure 1 Flow Chart

The random forest prediction model aggregates multiple weak predictors (i. e. , individual decision trees) and integrates their forecasting results through a mean-based strategy. This ensemble approach significantly reduces variance in prediction tasks, endowing the model with excellent predictive accuracy and robust generalization capabilities, enabling it to handle various complex real-world prediction scenarios with ease. In this study, the random forest model is utilized to predict Olympic medal counts, facilitating event analysis and informed decision-making processes.

3 RESULTS AND DISCUSSION

3.1 Data Integration

After completing the preprocessing of the result data, to present the data more clearly and comprehensively, this study merged the data and counted the specific number of participating countries (regions), the detailed classification counts of sports events (both major and minor categories), and the number of athletes. The final results are shown in Table 1 below:

Table 1 Participating Countries/Regions, Sports/Disciplines, Athlete/Count

Type	Number
The number of unified states	206
Number of sport	50
Number of event	289
Total Athletes	11097

As shown in the table, the four-dimensional data of "large number of participating countries/regions, wide coverage of major sports, detailed classification of minor disciplines, and large athlete scale" intuitively demonstrates the characteristics of the Summer Olympics as "grand in scale and diverse in events", reflecting both the depth of global sports exchange and the integrity and complexity of the competitive sports system.

3.2 Analysis of the Contribution of Each Variable to the Medal Prediction Model

The contribution rates of each variable to the gold medal prediction model and the medal prediction model, calculated using the feature variable contribution rate formula, are shown in Table 2 and Table 3 . For the gold medal count prediction model in Table 2 , the most critical factors are historical gold medal count and the proportion of a country's gold medals in the total gold medals, with total medal count also playing a significant role. Host country status and the number of events participated in have less impact on gold medal count prediction, while athlete count and score have relatively minimal effects. For the Olympic medal count prediction model in Table 3 , the most critical factors are total medal count and the proportion of a country's medals in the total medals. Host country status and the number of events participated in also make important contributions. In contrast, athlete count, gold medal count, and score have relatively minor roles in predicting medal counts.

Table 2 Contribution of Each Variable to the Gold Medal Forecasting Model

Variate	Contribution rate to the gold medal model
Whether it is the host country	18. 6%
Number of athletes	6. 5%
medal tally	11. 2%
gold MEDALS in each country accounted	17. 4%
Number of events participated	8. 4%
Number of gold MEDALS	25. 8%
Score	12. 1%

Table 3 Contribution of Each Variable to the Medal Prediction Model

Variate	Contribution rate to prediction medal model
Whether it is the host country	17. 4%
Number of athletes	5. 6%
medal tally	15. 6%
gold MEDALS in each country accounted	16. 1%
Number of events participated	7. 3%
Number of gold MEDALS	19. 4%
Score	18. 6%

3.3 2028 Medal Count Forecast Analysis

When establishing the random forest model, this study adjusted the hyperparameters according to the data type and volume. The final configurations were set as follows: the number of decision trees was set to 200 to fully capture the relationships between data features and enhance the model's generalization ability; the maximum tree depth was limited to 15 to prevent overfitting by restricting the growth depth during training; and the minimum sample split number was set to 5 to ensure that decision tree nodes contain at least the specified number of samples when splitting, making the model more robust.

Meanwhile, historical Olympic medal data was divided at a 7:3 ratio, with 70% of the data used as the training set for model parameter learning and optimization, and 30% as the test set. This strategy ensures the model can sufficiently learn the characteristics of historical data while providing a reliable basis for testing its performance on unknown data. The study employed mean squared error (MSE), mean absolute error (MAE), and coefficient of determination (R^2) as evaluation metrics to assess model performance.

By constructing a random forest model to predict the number of medals for the 2028 Olympics, the following prediction results were obtained (only listing the top 10 countries in total medals and gold medals). Meanwhile, the model's mean squared error (MSE) is 8. 2 and mean absolute error (MAE) is 2. 3. These low values indicate small average discrepancies between predicted and actual values, suggesting high prediction accuracy. The coefficient of determination R^2 reaches 0. 89, close to 1, demonstrating a high degree of fit to historical data and the model's ability to explain 89% of the variation in medal counts.

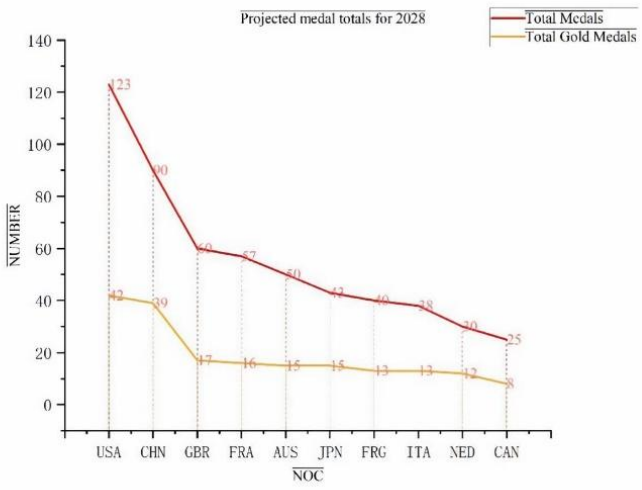


Figure 2 Comparison of the Projected Number of Medals for the 2028 Olympic Games

As shown in Figure 2 , the predicted total medal count for the United States is 123, far ahead of other countries, fully demonstrating its overall dominance in the sports arena. The U. S. boasts a mature sports talent development system, substantial financial investment, and a broad mass participation base, which enable it to maintain strong competitiveness across numerous sports disciplines and secure medals in both major and minor events.
The prediction results also reveal disparities in event-specific strengths among nations. Sporting powerhouses like the U. S. and China exhibit robust competitiveness across multiple sports. For example, Japan excels in combat sports such as judo and wrestling, while the Netherlands has achieved remarkable results in speed skating and cycling. These event-specific advantages help them gain outstanding performances in niche areas, though they also face the risk of over-reliance on single disciplines.

3.4 Medal Count Trend Analysis

The comparison between the predicted medal and gold medal counts for 2028 and the actual values for 2024 is presented in Table 4. Based on the table and an analysis of factors such as athlete reserves, the number of events participated in, and economic strength, Cuba, Germany, and Slovakia are expected to make progress in the upcoming Summer Olympics. In contrast, Belgium, Ecuador, and Israel may regress due to their respective circumstances, athlete reserves, economic strength, and other factors.

Table 4 Progress or Regression of Olympic-Related Countries

Noc	Situation
Cuba	Progressive country
Germany	Progressive country
Slovakia	Progressive country
Belgium	Backward country
Ecuador	Backward country
Israel	Backward country

Meanwhile, this study also focuses on the medal count trends of the top-ranked countries. Therefore, the predicted values for the top 5 countries in the 2028 Olympics are compared with their actual data from the 2024 Olympics, as shown in Table 5 below:

Table 5 Trends in the Number of Medals Won by Countries

Noc	The number of medals in 2024	The number of medals in 2028
USA	126	123
CHN	91	90
GBR	65	60
FRA	64	57
AUS	53	50

As indicated in the table, the medal counts of the United States, China, the United Kingdom, France, and Australia are projected to show a steady downward trend across the two Olympic Games. This finding suggests that although these countries are likely to remain among the top medal earners during the upcoming four-year Olympic cycle, they may face challenges in the development of competitive sports. For example, the decline in form or retirement of veteran athletes, coupled with delays in identifying and nurturing promising reserve talent, could directly lead to a reduction in medal tallies. In some disciplines, the lack of generational succession may make it difficult to sustain the exceptional performance achieved in 2024 into 2028. Therefore, over the next four years, these nations need to enhance their sports talent development frameworks to maintain their competitive advantages.

4 CONCLUSIONS

With the vigorous development of global sports, the competitive landscape on the Olympic stage has become increasingly intense, with a continuous increase in the number of participating countries and regions and the constant innovation of sports events. While promoting the improvement of competitive sports standards and cultural exchange, the uncertainty of medal distribution has significantly increased, posing challenges to event prediction and strategic planning, and adding variables to the development of sports economy and industry. Therefore, constructing a scientific Olympic medal count prediction model helps to grasp the sports competition situation and optimize resource allocation.
This study on medal predictions for the 2028 Olympics first highlights the outstanding performance of random forest models in complex data prediction scenarios. It then screens complete post-1950 data through data preprocessing, introduces key indicators such as host country status and athlete numbers, and performs standardization. Through feature contribution analysis, core influencing factors such as historical performance are identified. Finally, a random forest model is developed to predict gold medal counts and total medal counts. Based on the predictions, countries with growth potential such as Cuba and those at risk of performance decline such as Belgium are identified, providing a

scientific basis and practical reference for Olympic event analysis and decision-making by national Olympic committees.

COMPETING INTERESTS

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

REFERENCES

[1] Scelles N, Andreff W, Bonnal L, et al. Forecasting national medal totals at the Summer Olympic Games reconsidered. Social science quarterly, 2020, 101(2): 697-711.
[2] Bredtmann J, Crede C J, Otten S. Olympic medals: Does the past predict the future?. Significance, 2016, 13(3): 22-25.
[3] Andreff W. Economic development as major determinant of Olympic medal wins: predicting performances of Russian and Chinese teams at Sochi Games. International Journal of Economic Policy in Emerging Economies, 2013, 6(4): 314-340.
[4] Vagenas G, Vlachokyriakou E. Olympic medals and demo-economic factors: Novel predictors, the ex-host effect, the exact role of team size, and the “population-GDP” model revisited. Sport Management Review, 2012, 15(2): 211-217.
[5] Cheng H R, Lü J, Yuan T G. Prediction of China's track and field results in the Tokyo Olympic Games from the world top 20 national rankings of track and field events in 2018. Bulletin of Sports Science & Technology, 2020, 28(04): 4-8.
[6] Liu C Y, Wu M Q, Zhang A A, et al. Study on the temporal and spatial differentiation of Chinese Olympic medals from 1984 to 2016. Journal of Physical Education, 2019, 26(01): 75-82.
[7] Ding W Z. Data mining model of Olympic medals based on comprehensive national strength. Information Recording Materials, 2018, 19(03): 231-233.
[8] Balmer N J, Nevill A M, Williams A M. Modelling home advantage in the Summer Olympic Games. Journal of sports sciences, 2003, 21(6): 469-478.
[9] Bao Y, Meng X, Ustin S, et al. Vis-SWIR spectral prediction model for soil organic matter with different grouping strategies. Catena, 2020, 195: 104703.
[10] Hafezi M H, Liu L, Millward H. Learning daily activity sequences of population groups using random forest theory. Transportation research record, 2018, 2672(47): 194-207.
[11] Zhang Y D, Senjyu T, Chakchai S, et al. Smart Trends in Computing and Communications. Springer Singapore, 2022.
[12] Shi H M, Zhang D Y, Zhang Y H. Can Olympic medals be predicted? - From the perspective of interpretable machine learning. Journal of Shanghai University of Sport, 2024, 48(04): 26-36.
[13] Sun J, Zou X K, Zhu X B, et al. Research on random forest algorithm in the field of online scalper prediction. Computer Simulation, 2025: 1-6.
[14] Yang Q F, Li T, Jia Z Q. Consumption behavior prediction algorithm based on parameter-optimized random forest model. Computer & Digital Engineering, 2024, 52(07): 1959-1965.
[15] Gan M, Liu P F, Yue D B, et al. Prospecting prediction by geoelectrochemical technology in and around the Murong lithium mining area, western Sichuan based on random forest algorithm. Geology and Exploration, 2025, 61(02): 359-370.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3039

THE PREDICTION OF OLYMPIC MEDAL TABLE BASED ON LINEAR REGRESSION MODELING

Lei Zhao
School of Mechanical and Electrical Engineering, Anhui University of Science and Technology, Huainan 232001, Anhui, China.
Corresponding Email: Zl897932@163.com

Abstract: As the world's most influential sporting event bringing together the world's best athletes, the Olympic Games is the highest stage for competitive sports. It inspires more people to participate. In this paper, in order to predict the total medals and gold medals won by each country in the 2028 Olympic Games, a multiple linear regression model is constructed by considering the historical medal datas, the number of athletes' participation and the types and number of participating events and other characteristic variables as the indexes, and takes the evaluation coefficient $R^2$and the mean squared error MSE as the model evaluation indexes. Through the established medal list and historical trend, the prediction interval of total medals and the prediction interval of gold medals are analyzed, and those countries that may progress or regress in the 2028 Olympic Games are analyzed and obtained, and by the prediction this paper selects the top 10 progressing countries and the 10 countries with obvious regression. In addition, the prediction of those countries that have never won a prize is also made to explore the possibility of winning a medal, and for this purpose, this paper adopts a binary classification model and logistical regression model, and the probabilities of winning a first medal are obtained by selecting the data of the countries that have never won the award.
Keywords: Predicting the number of medals; Linear regression model; Model evaluation; Binary classification model and logistical regression model

1 INTRODUCTION

As the largest and most influential comprehensive sports event in the world, the Olympic Games has an important significance to the global society, economy, culture and other aspects that can not be ignored. The prediction of the medal table has always been the focus of public and professional attention[1]. In existing studies, time series models , such as gray theory prediction model and stochastic time series analysis model[2], are mainly used, which rely heavily on the quality of finite historical data, and missing or abnormalities will affect the prediction, and can not deal with nonlinear relationships.

And according to the national economic level and the total population to establish the econometric prediction model to predict the number of medals [3], can reflect the degree of influence of each factor size. But it has limitations tend to ignore the traditional strengths of each country's sports programs, usually lacks of knowledge of the specifics of a competition, athletes' past performance and psychological quality.

Neural network nonlinearity was used to fit and predict the number of medals by quantitatively predicting and studying the GDP per capita of each country as well as the previous medal scores[4]. neural network is able to efficiently learn complex nonlinear relationships in the data, and through the multilayer structure (especially deep neural networks), it can abstract features of the data can be learned, which is suitable for dealing with tasks with complex patterns , but is prone to overfitting in the simulation process, especially if the training data is insufficient or the model is too complex.

By inputting several independent variables such as the history of awards for each of its countries, the number of participating athletes, and outputting the dependent variable, the number of medals, Support Vector Machines (SVMs) are able to handle nonlinearly differentiable problems by introducing kernel functions to map the original features to a higher dimensional space. This enables SVMs to handle complex nonlinear classification problems, however, the performance of SVMs depends heavily on the choice of parameters, such as the regularization parameter C and the kernel function. Different combinations of parameters and kernel functions can significantly affect the performance of the model and need to be tuned by methods such as cross-validation. Therefore it should be simplified according to the measurement method. And it is sensitive to large-scale data and feature items.

So unlike those that use time series model, this paper constructs a comprehensive multiple linear regression model[5], which is suitable to be used with sufficiently large amount of data and suitable choice of independent variables to provide its reliable prediction results[6]. Factors affecting the number of medals are entered as characteristic variables to predict the performance of countries in the 2028 Olympic Games, especially the number of gold medals. It provides its reliable predictions with enough data and suitable choice of independent variables. It also provides an in-depth analysis of the various factors affecting these predictions. This model is able to handle the relationship between multiple independent variables and a dependent variable, captures the joint influence of multiple independent variables on the dependent variable, and fits the data better by analyzing the regression coefficients, which allows for the determination of the degree of influence of each of the independent variables on the corresponding variable, and is insensitive to small variations in the data.

2 FORECASTING THE OLYMPIC MEDAL TABLE RESEARCH METHOD AND MODEL CONSTRUCTION

In this paper through a multi-dimensional analysis, first collect the raw data that need to be analyzed and observe the laws behind the phenomenon from the data to obtain valid data, then select independent variables such as historical medal data of each country, the number of athletes participating and the type of events participating etc, what's more, test whether there is a linear relationship between the independent variables and the dependent variable, finally determine the dependent variable Y (the predicted total number of medals or gold medals) with the independent variables X_1, X_2,...,X_n of the linear relationship to establish a linear multiple regression model and comprehensively explore the prediction of the number of medals in the 2028 Los Angeles Olympic Games and the analysis of influencing factors. On this basis, the quantity R^2 and the mean square error MSE are calculated. In this paper, the countries that may progress and regress in the 2028 Olympic Games are analyzed based on the previous historical total medals data and gold medals data(based on the data provided by The 2025 Mathematical Contest in Modeling and the download website is http://www.mcmcontest.com) and the predicted medal table. The article is tightly structured, and each part of the paper starts from a different perspective to provide rich theoretical support and empirical analysis for the prediction models.

2.1 Modeling Establishment

2.1.1 Data normalization

After pre-processing the different kinds of data, in order to eliminate inconsistencies between the different variables. Normalization is performed as follows:

Enter the characteristic variables (the country's historical medal data: gold, silver, bronze; the number of athletes involved; the number of events involved), then substitute these datas into the formula.

$$Y = \{Y_1, Y_2, \cdots, Y_N\} \tag{1}$$

Where: $Y = \{G_i, S_i, B_i\}$ represents the number of gold, silver and bronze medals won by the i-th country in a particular year of the Olympic Games.

2.1.2 Feature selection

Select features that affect the total medal table, such as the number of historical medals, number of athletes, the number of events participated in by each country and the advantage programs etc[7]. For the correlation between the features is checked, avoiding multicollinearity.

2.1.3 Modeling establishment

In order to predict the number of gold medals and the total medal table of each country in 2028, this paper chooses to use a linear regression model. The linear regression model assumes a linear relationship between the number of medals and a set of characteristics. The regression model is set up as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \tag{2}$$

where: Y is the predicted number of gold medals or total medals; X_1,X_2,...,X_n are the characteristic variables, such as the number of historical medals, the number of athletes, the number of events, the infrastructure of each country etc; β_0 is the intercept term, which denotes the baseline number of medals when all the characteristic variables are zero; β_1, β_2,..,β_n is the regression coefficient, reflecting the degree of influence of each characteristic variable on the number of medals; ϵ is the error term, indicating random fluctuations and unexplained parts of the regression model. The magnitude of the regression coefficients reflects the degree of influence of each feature on the number of gold medals or the total number of medals. The regression coefficients were estimated from the training dataset. The aim is to minimize the error between the predicted and actual values.

2.1.4 Linear regression method

In this model, it is assumed that the number of medals (Y) is a linear relationship determined by the combination of a number of characteristics. For example historical number of medals and number of athletes etc. To estimate the regression coefficients, the ordinary least squares (OLS) method is used. This method estimates the regression coefficients by minimizing the error squared between the predicted and actual values.

Assume there are N training samples, and each sample contains n feature variables. The number of medals for each sample is denoted as y_i and the corresponding value of the feature variable is $X_{1i}, X_{2i}, \cdots, X_{ni}$.The goal is to minimize the objective function.

$$\text{miminize} \sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni}))^2 \tag{3}$$

Where: y_i is the actual number of medals; $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni}$ is the predicted value of the model. By minimizing the above objective function, it is possible to estimate the regression coefficients $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$.

To solve this optimization problem, the minimum value of this objective function is solved, usually by gradient descent or regular equations. The solution to the regular equation is :

$$\beta = (X^T X)^{-1} X^T Y \tag{4}$$

Where: $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_n)^T$ is the estimate of the regression coefficients; X is an N × (n + 1) matrix, where each row represents the eigenvector of a training sample and the first column is corresponding to the intercept term; Y is an N × 1 vector containing the actual medals of all the training samples; and X^T is the transpose matrix of X.

By solving this formal equation, estimates of the regression coefficients are obtained, and a prediction model is developed.

2.1.5 Calculation of expected number of gold medals and total medals

Input Characterization Variables. Bring to Model. Calculate for each country prediction of Gold Medal Count. And use a fitted model and known gold medal counts to predict the number of medals and their uncertainty intervals for each country at the 2028 Lod Angeles Olympics so that it can make sure the accuracy of the results. Ultimately, based on the predicted medal table, select the total medal count prediction interval as [-23,23] and the gold medal count prediction interval as [-9,9].

2.1.6 Construction of prediction intervals

Based on the nature of the linear regression model, combined with random effects, the predictive distribution of the number of gold medals in each country was generated, and the 95% prediction interval was extracted, and it was located in the interval $[\mu-3\sigma,\mu+3\sigma]$.

2.2 Model Evaluation

The goodness of a regression model is usually assessed by several indicators:

Coefficient of determination R^2: indicates the proportion of variability explained by the model, $R^2 \in [0,1]$, closer to 1 means better model fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{5}$$

Where: y_i is the predicted value and $\bar{y}_i$ is the average value of the sample.

Mean Square Error (MSE): indicates the squared mean of the error between the predicted and actual values, the smaller it is the better the model predicts.

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

Residual analysis: check whether the residuals (the difference between actual and predicted values) conform to a normal distribution and analyze whether there is a system errors of a sexual nature.

These evaluation metrics allow us to judge the predictive effectiveness of the regression model and further optimize the model to ensure that it is suitable for predicting the number of gold medals and the total number of medals for each country in 2028.

2.3 Analysis of Results

The regression analysis allows us to derive the degree of influence of each characteristic variable on the number of medals. The positive and negative regression coefficients indicate whether the relationship between the characteristics and the number of medals is positive or negative, and the absolute value of the coefficients indicates the magnitude of their influence on the number of medals. Through this prediction model, the paper obtained a regression coefficient of 0.7537, indicating that the number of historical medals has a strong impact on the number of gold medals and total medals.

The prediction result will give the number of total medals or total medals for each country in the 2028 Olympics, as in Figure 1,showing the predicted number of total medals for the 2028 Olympics, with the United States of America (USA) and China(CHN) showing a significant lead[8], which has a much higher number of predicted gold medals than any other country, followed by China (CHN) and Japan (JPN), demonstrating the strong competitiveness of these countries in the Olympics. Other countries such as Australia (AUS), France (FRA) and Great Britain (GBR) also demonstrated solid competitive performances. The calculation then yields a prediction interval of [-9,9] for the number of gold medals and [-23,23] for the total number of medals. And from the table, the total number of medals for the US is [90,136], and the prediction interval for China is [63,109], indicating a high degree of uncertainty in the results. So the prediction intervals may be strongly influenced by a variety of factors such as changes in team composition, athlete health, and training conditions. And based on historical data and predictions medal standings. It is possible to analyze those countries that progress or regress. And ten countries that are likely to progress or regress were selected, as shown in Figures 2 and 3. From a sociological perspective economic powerhouses and countries with long sporting traditions continue to dominate at the Olympics[9]. The United States will have 43 gold medals at the 2028 Olympics, an increase of 3 medals from 2024, and 113 total medals, a decrease of 13 medals from 2024; Because China has great competitiveness in traditional excellent events such as diving[10], it ranks the top in terms of gold medals and total medals. According to the figure, China will have 38 gold medals at the 2028 Olympics, a decrease of 2 medals and a total of 96 medals, an increase of 6 medals compared to 2024. These countries are traditional sports powerhouses with

little fluctuation in the number of gold medals and total medals. Their gold and total medal counts are likely to improve. On the contrary Fiji, Peru's gold medal count and total medal count may decrease, Fiji may decrease by 25 medals, but Fiji won only 1 medal in 2024 and may not win a medal in 2028, and these countries are generally economically backward and weak sports countries.

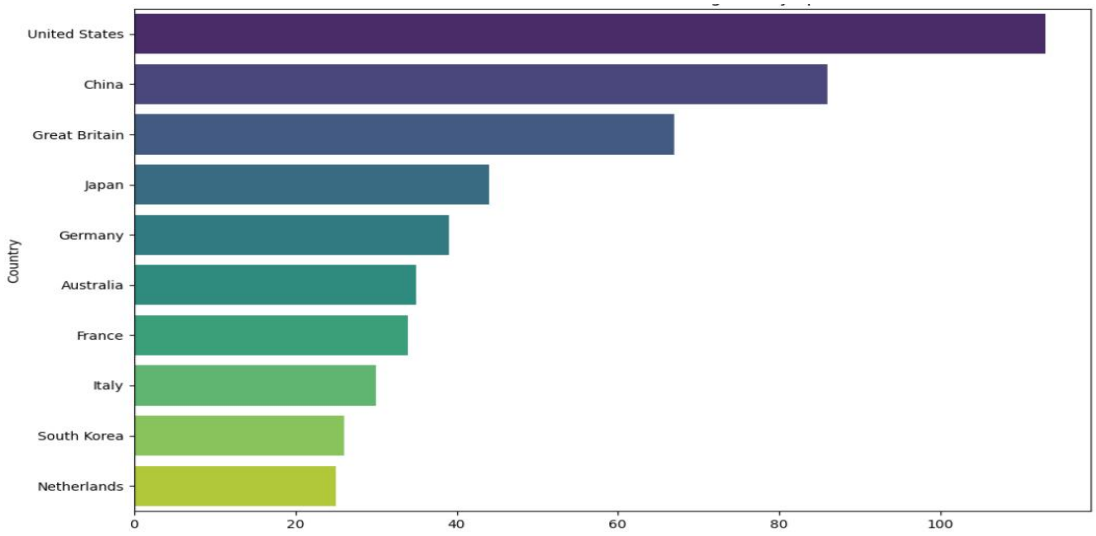


Figure 1 The Top Ten Countries in terms of Total Medals in the Projected 2028 Medal Table

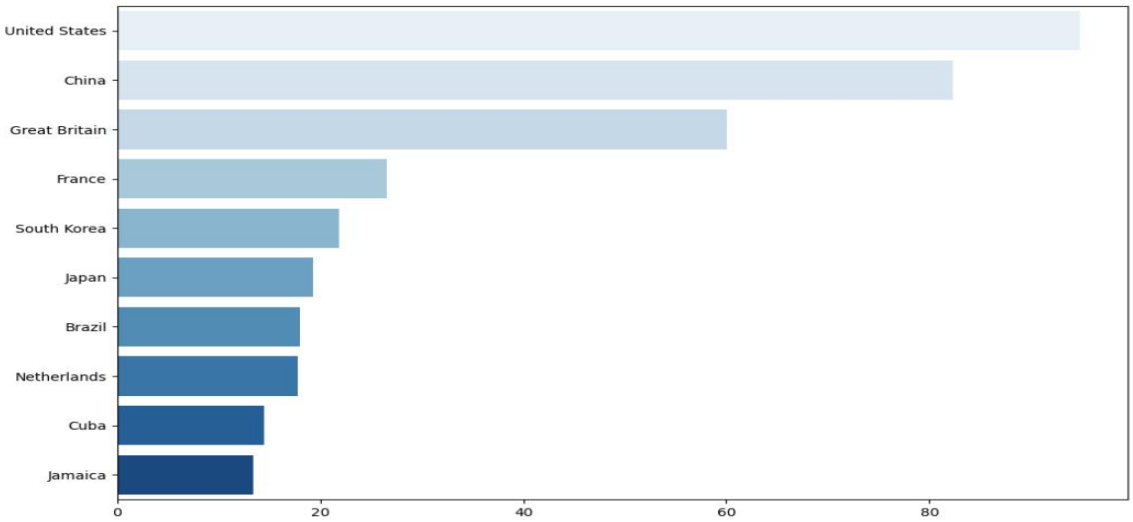


Figure 2 The Top Ten Countries Selected for Possible Progress in the Olympic Medal Table

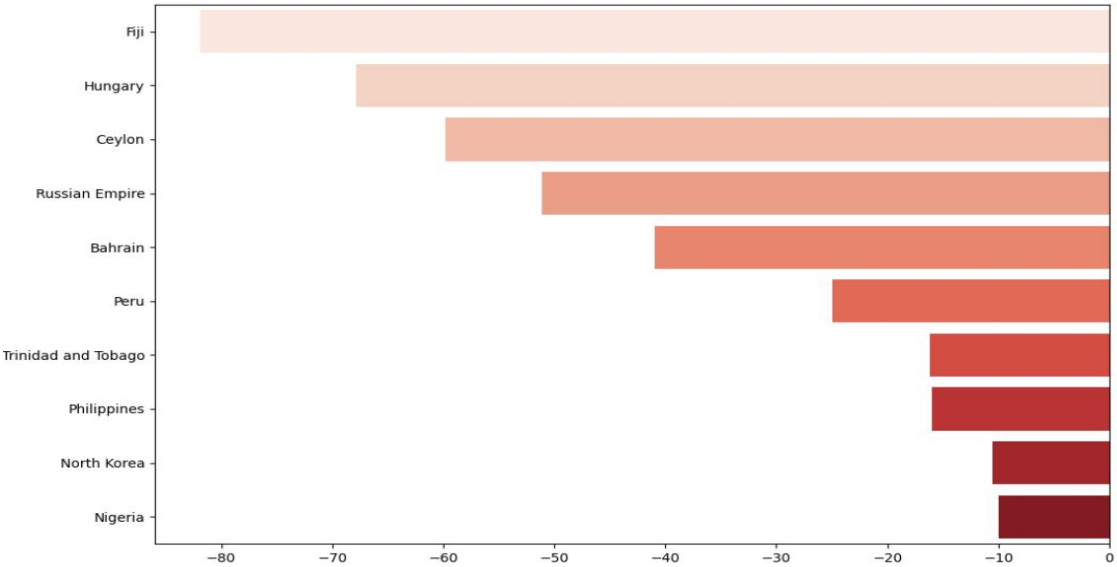


Figure 3 The Ten Countries Selected for Possible Regression in the Olympic Medal Table

3 THE FIRST-TIME AWARD-WINNING COUNTRIES BASED ON LOGISTIC MODEL

3.1 Research Methodology

According to the historical medal list to select those countries that have never won the Olympic Games as the independent variable, select the probability of winning the award as the dependent variable, to establish a logistic regression-based binary classification model to explore the likelihood of winning the award, the interval is [0,1], if the probability of winning the award of a country is closer to 1, the greater the likelihood of winning the award, the reverse is not the case.

3.2 Modeling Selecting

In order to make categorical predictions, logistic regression model is chosen in this paper. Logistic regression is a commonly used binary classification model that makes predictions by calculating the probability of an event occurring. Specifically for this study, the objective of this paper is to predict whether a country will be able to win a medal or not, and in particular whether the country will be able to win a medal for the first time in the 2028 Olympic Games. The mathematical expression of the logistic regression model is as follows.

$$\mathrm{P(Medal)} = \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\cdots \beta_n \mathrm{X}_n)}} \tag{7}$$

Where: P(medal) represents the probability of a country winning a medal, i.e. the likelihood of that country winning a medal in the 2028 Olympics. Since it is a probability value, the value of P(medal) must lie between 0 and 1; $X_1,X_2,\ldots,X_n$ are the characteristic variables affecting whether or not the country can win a medal. These characteristics can be the number of athletes in the country, the number of participating sports, historical performance, economic level etc, which together determine the probability of the country to win a medal; $\beta_0,\beta_1,\ldots,\beta_n$ are the regression coefficients, which indicate the degree of influence of the characteristics on the winning of medals. Through the training of the model, we can estimate the values of these coefficients; e is the base of the natural logarithm, which is used to ensure that the output probability value of the model is always between 0 and 1, in line with the definition of probability.

The goal of this logistic regression model is to learn the regression coefficients $\beta_0, \beta_1 ,\ldots,\beta_n$ from the given training data. The process of learning the regression coefficients is achieved by maximizing the likelihood function. The purpose of maximizing the likelihood function is to make the probability values predicted by the model as close as possible to the actual observed labels.

3.2.1 Maximum likelihood estimation

In logistic regression, the regression coefficients are estimated by maximizing the likelihood function. The likelihood function represents the probability of observing the current data given the characteristic data. Assuming that there are m samples with label 3 and feature X_i ,the likelihood function can be expressed as follows.

$$L(\beta_0,\beta_1,\ldots\ldots,\beta_n) = \prod_{i=1}^{m} \mathrm{P}(\mathrm{y_i} \mid \mathrm{X_i}) \tag{8}$$

where $\mathrm{P}(y_i|X_i)$ denotes the probability that the label yiof sample i is 1. Since this paper is a binary classification problem, label i takes the value of 0 or 1, which indicates whether the country wins a medal or not, respectively. Therefore, $\mathrm{P}(\mathrm{y_i}|\mathrm{X_i})$ can be written as.

$$\mathrm{P}(\mathrm{y_i} \mid \mathrm{X_i}) = \mathrm{P(Medal)}^{\mathrm{y_i}} (1-\mathrm{P(Medal)})^{(1-\mathrm{y_i})} \tag{9}$$

The probability of a sample i-th is P(medal), if its label $y_i = 1$(namely the country won a medal); if $y_i = 0$(namely the country did not win a medal), the probability is $1 - \mathrm{P(Medal)}$.

In order to simplify the calculation and improve the numerical stability, we usually take the logarithm of the likelihood function to get the log-likelihood function. The log-likelihood function is expressed as.

$$\ell(\beta_0,\beta_1,\cdots\cdots,\beta_n) = \sum_{i=1}^{m} [\mathrm{y_i log(P(Medal)) + (1-y_i) log(1-P(Medal))}] \tag{10}$$

By maximizing the log-likelihood function, we are able to obtain the regression coefficients $\beta_0 ,\beta_1,\ldots,\beta n$. The goal of maximizing the log-likelihood function is to make the predicted probabilities of the model as consistent as possible with the actual labels. This process usually uses optimization algorithms (like gradient descent) to find the optimal regression coefficients.

3.2.2 Solving for regression coefficients

The regression coefficients are solved by maximizing the log-likelihood function. To solve these coefficients, numerical optimization methods are usually used. In logistic regression, commonly used optimization methods include gradient descent and Newton's method. The gradient descent method gradually finds the parameters that maximize the log-likelihood function by calculating the gradient of the log-likelihood function and updating the regression coefficients at each iteration step.

Specifically, the gradient descent method works by calculating the partial derivatives of each regression coefficient and updating the regression coefficients based on the derivative values. At each iteration, the regression coefficients are

adjusted in a direction that causes the log-likelihood function to increase. The iterative process continues until the log-likelihood function converges, until the regression coefficients no longer change significantly.

3.2.3 Model solving

In order to implement the logistic regression model and predict the probability that a country that has not yet won a medal will win a medal for the first time, the input eigenvectors are passed through the model and the paper uses a binary classification model, by fitting the probability of non award-winning country to a distance of 0 or 1,then can conduct whether they can or can't win the first medal in 2028 Olympic.

3.3 Analysis of Results

After obtaining the model results, the probability of each yet-to-be-awarded country winning a medal can be output. With the model, we calculate the probability that each country that has not yet won a medal will win its first medal in 2028. There are 115 countries that have not yet won a medal, and the model predicts that 10 of them will win their first medal in 2028, and the probability for each of them is more than 0.5. This means that there is a high level of confidence that these 10 countries will break through history and reach the podium for the first time. Through this prediction method, we can provide data support to the National Olympic Committees to help them develop more targeted Olympic strategies. In addition, through further analysis of the model, we can identify the important factors that affect a country's ability to win medals, such as the quality of athletes, the number of events entered, and historical performance. These factors will help countries to make more precise adjustments in future Olympic Games. For this purpose, we have selected ten countries that are most likely to win a medal for the first time. As shown in Table1, the Republic of Armenia has the highest probability of winning a prize for the first time at the 2028 Olympic Games.

Table 1 The Top Countries most likely to Win a Medal for the First Time

COUNTRY	POSSIBILITY
Armenia	0.7956
Bahamas	0.7753
Azberbaijan	0.7485
Algeria	0.7165
Bahrain	0.6647
Albania	0.6381
Tonga	0.6257
Barbados	0.6052
Peru	0.5860
Afghanistan	0.5765

4 ESTIMATION OF R2 AND MSE FOR PREDICTION OF MULTIPLE LINEAR REGRESSION MODELS

In conducting the performance evaluation of the linear regression model for the prediction of the number of medals for each country at the 2028 Summer Olympics in Los Angeles, an impressive set of statistical metrics were obtained, which emphasize the high accuracy and reliability of the model. Specifically, the mean square error (MSE) of the model was 0.85, showing a small prediction error, implying a low mean squared difference between predicted and actual values. This low level of error indicates that the model performs well in predicting the number of medals for each country with proper error control. The Mean Absolute Error (MAE) of 0.91 further confirms the model's ability to maintain consistency across data points, reflecting a low mean absolute deviation between predicted and actual values. This is particularly important because it is directly related to the usefulness and reliability of the prediction results, and the low MAE value indicates that the model has less error in practical applications and more accurate prediction results. The coefficient of determination (R^2) of the model reaches 0.99, which is almost perfect performance, and almost all the data variations can be explained by the model. This high R^2 value not only demonstrates the statistical superiority of the model, but more importantly, it shows that the model is able to capture and explain the various factors affecting the number of medals extremely effectively, ensuring a high degree of accuracy and explanatory power in the prediction results.

5 CONCLUSION

The Olympic Games, as the largest sports event, has always been a focus of attention in predicting the number of medals won. The paper uses a multiple linear regression model to predict the medal table and analyze the progress and regression of the countries, and it estimate MSE and R^2, and R^2 is 0.99 while MSE is 0.91 by calculating. In addition, logistics regression model and binary classification model are used to explore the winning situation of countries that have not won the first medal. The multiple linear regression model can quantitatively analyze multiple factors that affect the number of medals, clarifying the degree and direction of each factor's impact on the number of the medals won. And multiple linear regression model can be combined with other prediction models(such as random forests) to comprehensively utilize the advantages of different models and improve the accuracy and reliability of predictions. In the future, based on this model, multiple analysis can be conducted to predict which Olympic events have a high probability of China winning the championship. This can provide valuable information for relevant sports organizations

to better understand the possible direction of future Olympic games, so that the country can formulate more scientific and effective training and preparation strategies.

COMPETING INTERESTS

The author has no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Abel Sowal. Olympic medal prediction is not just for "presence". Economic and Social Management Science, 2016, (20): 77.
[2] Nagpal Prince, Gupta Kartikey, Verma Yashaswa, et al. Paris Olympic(2024) Medal Tally Prediction. Banaras Hindu University,Varanasi, India, Lecture Notes in Networks and Systems, 2023, (662): 249-267.
[3] Cha Zheng. Economic Reflections on Olympic Medals. Shandong Radio and Television University, Social Science II Series, 2020, (05): 97.
[4] Mallappa Uday, Gangwar Pranav, Khaleghi Behnam, et al. TermiNETor:Early Convolution Termination for Efficient Deep Netural Networks. 2022 IEEE 40th International Conference on Computer Design (ICCD), Olympic Valley, CA, USA, 2022: 635-643. DOI: 10.1109/ICCD56317.2022.00098.
[5] Deng Rongrong, Fan Qingmin, Zhang Yinkai, et al. Analysis of Technical Evaluation of Male Boxing Athletes Based on Entropy and Multiple Linear Regression Model-Taking Excellent 81kg Athletes at the Tokyo Olympics as an Example. School of Competitive Sports,Beijing Sport University, Social Science II Series, 2023, (11): 4744-4746.
[6] Miyoshi Takemasa, Amemiya Arata, Otsuka Shigenori, et al. Big Data Assimilation:Real-time 30-second-refresh Heavy Rain Forecast Using Fugaku during Tokyo Olympics and Paralympics. Meteorological Research Institute,Tsukuba, Japan, Association for Computing Machinery, Inc, 2023, (12): 8.
[7] Wang Yongjun, Chen Hong, Yang Yongfen. Fuzzy relationship,mediating variables,and dynamic models:A study on the trickle down effect of sports participation in heritage, School of Physical Education. Chongqing Technology and Business University,School of Management,Tianjin Sport University, Kunming University, Social Science II Series, 2021, (7): 28-30.
[8] Gupta Krishon Gopal, Arora Aditi. Olympic Data Analysis: Uncovering New Insights into Athletic Performance and Competition. ABES Engineering College, Ghaziabad, India, Grenze Scientific Society, 2024, (2): 4312-4319.
[9] Li Luyan, Gao Yong. Understanding of Beijing Winter Olympics Volunteers from a Sociological Perspective. School of Physical Education,Henan University of Science and technology, Social Science II Series, 2022, (10): 216-217.
[10] Ning Yixia. Observing the characteristics of China’s Competitive Sports Strength from the Tokyo Olympics. Shenzhen University, Social Science II Series, 2022, (09): 36-38.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3040

JOINT SEGMENTATION MODEL FOR CRACKS AND JOINTS BASED ON Deeplabv3+

Fang Wang
School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China.
Corresponding Email: wangfang05092022@163.com

Abstract: This paper addresses the problem of low segmentation accuracy of cracks and joints in complex scenarios and proposes an improved model, ViR-Deeplabv3+, based on DeepLabv3+. First, the model replaces the traditional backbone network with the Vision Transformer (ViT) with global perception ability. This enables the model to no longer be limited to extracting local information when processing image data, but to capture the global context features of the image more efficiently, thereby enhances subsequent segmentation tasks. Secondly, residual connections between the ViT and the Hollow Space Pyramid Pooling (ASPP) module are ingeniously introduced. The design concept of residual connection effectively solves problems such as the gradient vanishing problem, ensuring that the rich feature information from ViT can be smoothly and unobstructedly transmitted to the ASPP module for further fusion and mining of multi-scale features. Finally, we conducted model training and ablation experiments based on the self-built dataset (including crack and seam samples). The results showed that the mean intersection and union ratio (mIoU) of ViR-Deeplabv3+ reached 75.27%, which was 2.97% higher than that of the baseline model Deeplabv3+. This scheme provides an effective solution for precisely detecting and segmenting cracks and joints in complex scenarios, and has important practical application value.
Keywords: Image segmentation; Crack; Seam; ViT; Residual connection

1 INTRODUCTION

With the acceleration of urbanization and the aging of infrastructure, detecting cracks on the surface of concrete structures has become an important task to ensure the safety and durability of buildings. Cracks not only affect the load-bearing capacity of structures but may also cause secondary problems such as leakage and corrosion, posing a serious threat to public safety. Traditional crack detection methods mainly include visual inspection and manual measurement, which are simple and low-cost but are greatly influenced by human factors and difficult to monitor continuously [1]. In addition, manual inspection has limitations such as low efficiency, strong subjectivity, and high cost. In practical scenarios, the similarity in appearance between cracks and prefabricated joints, complex background interference, and the scarcity of datasets pose severe challenges to the generalization ability and practicality of models. Therefore, research on effective detection and segmentation of cracks and joints is crucial for ensuring the safety and reliability of infrastructure such as bridges, roads, and buildings.

Over the past few decades, crack detection has been continuously carried out and has achieved significant accomplishments. In the research methods based on object detection for crack detection, Pratibha et al. [2] deployed an automated process based on a deep learning object detection model, YOLOv5, By capturing and accurately locating cracks in masonry structures through bounding boxes, the training time of the model is relatively short and can be used for real-time crack detection; Marin B et al. [3] proposed a new detection method, progressive detection, which adopts the architecture of Faster R-CNN object detector to provide crack detection in images. From the perspective of detection, they re examined the binary classification of images with and without cracks, minimizing the crack loss rate to the greatest extent possible; Wang et al. [4] proposed an improved method based on the SSD algorithm, adjusting the combination of the number of prior boxes at different resolutions in the original SSD algorithm to achieve high-precision crack recognition for images with noise. In the research methods based on image segmentation for crack detection, Lau et al. [5] proposed a U-Net-based network architecture that replaces the encoder with a pre-trained ResNet-34 neural network and uses a "single cycle" training plan based on cyclic learning rates to accelerate convergence. Their model achieved higher F1 scores on CFD datasets compared to other models; Attard et al. [6] demonstrated that Mask R-CNN can be used to localize cracks on concrete surfaces and obtain their corresponding masks to aid extract other properties that are useful for inspection; Yao et al. [7] added an RFB multi-branch convolution module to the Deeplabv3+ model [8], replaced the backbone of Deeplabv3+ with Mobilenetv2, and replaced all ordinary convolutions in the algorithm with depthwise separable convolutions, improving the segmentation accuracy and detection efficiency of the Deeplabv3+ model for bridge cracks.

Most of these works focus on a single category of cracks and lack joint segmentation and geometric parameter calculation for cracks and joints. In actual engineering, the joints of precast concrete slabs are highly similar in morphology to real cracks. However, most current models only perform single-category detection for cracks without considering the interference of joints, leading to an increase in misjudgment rates. False positives in crack detection, such as misidentifying construction joints as cracks, waste resources, and delay critical repairs. [9]. Moreover, public datasets typically only contain crack samples and lack images simultaneously labeled with both cracks and joints, which limits the models' ability to distinguish between the two. Research shows that when the test set includes joints, the average mIoU of existing models drops by approximately 12% [10]. Although current research has provided valuable insights and techniques in the field of crack detection, the models' ability to segment cracks and joints remains to be improved when dealing with the highly similar morphologies of the two.

To address the aforementioned issues, we propose a ViR-deeplabv3+ model, which integrates Vision Transformer (ViT) and residual connections to improve the deeplabv3+ model for image segmentation of cracks and joints. The main contributions of this paper are as follows:

(1) Replace the backbone network Xception of DeepLabv3+ with Vision Transformer (ViT), and utilize its self-attention mechanism to capture global context dependencies, thereby overcoming the limitations of traditional convolutional networks in long-distance feature modeling.
(2) Introducing residual connections between the ViT and ASPP modules alleviates the vanishing gradient problem in deep networks, enhances the multi-scale feature fusion capability, and improves the edge segmentation accuracy of cracks and joints.
(3) By integrating publicly available data with self-collected data, a concrete structure image dataset containing annotations of cracks and joints is constructed. The sample size is effectively expanded through data augmentation techniques to enhance the generalization ability of the model.
(4) The ablation experiments verified the effectiveness of ViT and residual modules. The mIoU of ViR-Deeplabv3+ was significantly improved compared to the baseline model, and it demonstrated stronger robustness under complex background interference.

2 METHOD

The Deeplabv3+ model is a powerful semantic segmentation framework. Its classic version uses Xception as the backbone network and combines Atrous Spatial Pyramid Pooling (ASPP) with a decoder structure to achieve multi-scale feature fusion and fine edge recovery. However, because of the limitations of Xception in extracting global context dependencies and the potential problems, such as gradient vanishing when training deep networks, this study proposes the ViT-deeplabv3+ model. By replacing the original Xception with ViT (Vision Transformer) as the backbone network and introducing a residual connection module between the backbone and ASPP, the feature transmission efficiency and semantic expression ability are enhanced. The overall structure is shown in Figure 1.

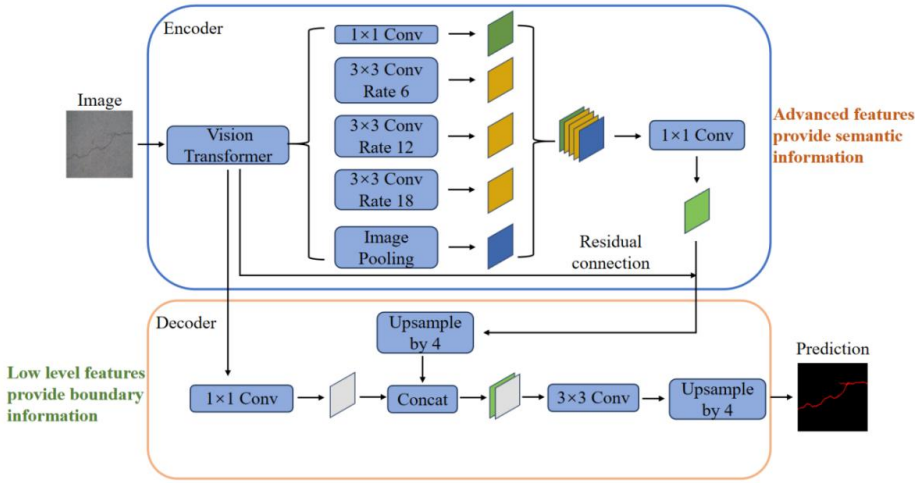


Figure 1 Overall Structure of ViT-deeplabv3+

2.1 ViT Feature Extraction Module

The Vision Transformer (ViT) is an image classification network based on the Transformer architecture, as shown in Figure 2. It divides the image into fixed-size patches and flattens them to be processed by the Transformer. ViT employs the self-attention mechanism to capture the global context information of the image, thereby demonstrating stronger performance than traditional convolutional neural networks (CNNs) in many computer vision tasks.

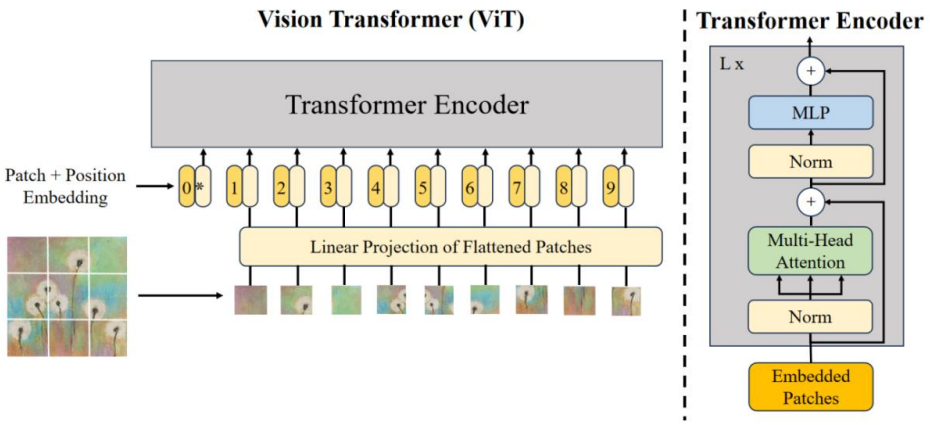


Figure 2 Network Structure of Vision Transformer

For the input image $I \in R^{H\times W\times C}$, it is first divided into $N = \frac{H\times W}{p^2}$ parts. A $P \times P$ small block, each small block through. Linearly embed the mapping into a high-dimensional space to form the input features:

$$z_i = E \cdot \text{Flatten}(I_i) + e_i \quad (1)$$

Here, E is the linear embedding matrix, $\text{Flatten}(I_i)$ represents the flattening operation of the -th small block, e_i is the position encoding used to retain the spatial position of the small block in the original image, and z_i is the feature of the -th block. In this way, ViT can encode the spatial information of the image into a sequence of inputs for the Transformer to process.

In ViT, the core computational module is the self-attention mechanism (Self-Attention). The self-attention mechanism assigns an attention weight to each input by computing the relationships among Query, Key, and Value, thereby performing a weighted sum of different parts of the input sequence. The calculation formula for self-attention is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Among them, Q is the query matrix, K means the key matrix, V indicates the value matrix, and d_k denotes the dimension of the key. Under the multi-head self-attention mechanism, multiple attention heads calculate in parallel and concatenate the results to provide richer context information.

By stacking multiple Transformer encoder layers, ViT can capture multi-level context information from local to global, which makes it more flexible and efficient than traditional CNNs in handling global dependencies. After being processed by multiple self-attention layers, the feature map output by ViT will be used as the input for the subsequent Deeplabv3+ model.

2.2 Residual Connection

Deep neural networks may encounter problems of vanishing or exploding gradients during training, especially when there are many layers, which can lead to unstable training. Residual connection is an effective solution. It forms a shortcut path by directly adding the input to the output, thereby avoiding the problem of vanishing gradients. Specifically, the mathematical representation of the residual connection is:

$$y = \mathcal{F}(x, W_i) + x \quad (3)$$

Among this is the input, $\mathcal{F}(x, W_i)$represents the output after a series of operations (such as convolution, activation, etc.), and y means the final output.

In this study, residual connections are introduced between the ViT backbone network and the ASPP module. Specifically, the feature maps output by ViT are added to the multi-scale feature maps processed by the ASPP module, thereby enhancing information flow and alleviating the vanishing gradient problem in deep networks. This process can be expressed as:

$$F_{out} = \text{ASPP}(F_{ViT}) + F_{ViT} \quad (4)$$

Among them, F_{ViT} is the high-level feature extracted by ViT, and F_{out} denotes the output processed by ASPP. The final feature after the residual connection is used as the output of the model.

By introducing residual connections, the network can more efficiently propagate gradients, thereby facilitating the learning of deeper features. Additionally, residual connections help preserve high-level semantic information, enabling the network to better retain detailed information during multi-scale feature fusion, and improving edge recovery and segmentation accuracy.

In summary, we propose an improved Deeplabv3+ model that uses ViT as the backbone network to overcome the limitations of traditional convolutional networks (such as Xception) in handling global context information. Additionally, we introduce a residual connection module between ViT and the ASPP module to address the gradient vanishing problem that may occur during the training of deep networks. Experimental results show that the Deeplabv3+ model with ViT as the backbone network, combined with residual connections, demonstrates better performance in semantic segmentation tasks.

3 EXPERIMENTAL EVALUATION

3.1 Dataset Construction

In this study, we integrated publicly available crack datasets with our own collected data to construct a specialized dataset of concrete crack and joint images. This dataset contains 151 high-resolution images, including 118 crack samples and 33 seam samples. All images were collected from diverse real-world engineering scenarios (as shown in Figures 3 and 4), covering various environmental conditions and structural types to ensure the representativeness and generalization ability of the data. In the data preprocessing stage, a standardized process was adopted: first, all original images were uniformly adjusted to a resolution of 513×513 pixels; then, the labelme annotation tool was used to conduct meticulous manual annotation on the self-collected data, automatically generating corresponding JSON format annotation files (as shown in Figure 5); finally, these JSON files were converted into annotation masks suitable for image segmentation tasks.

Figure 3 Partial Crack Images

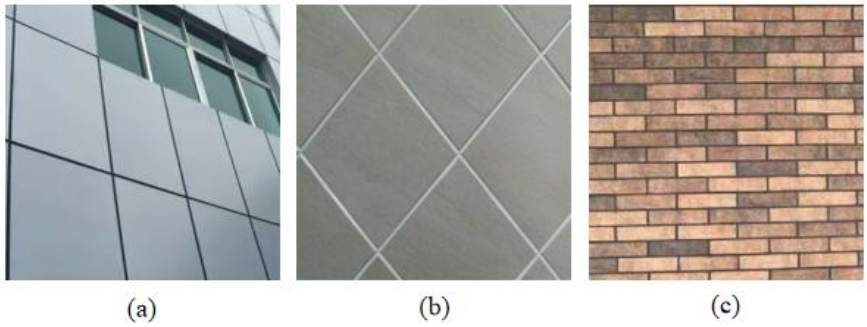


Figure 4 Partial Images of Seams

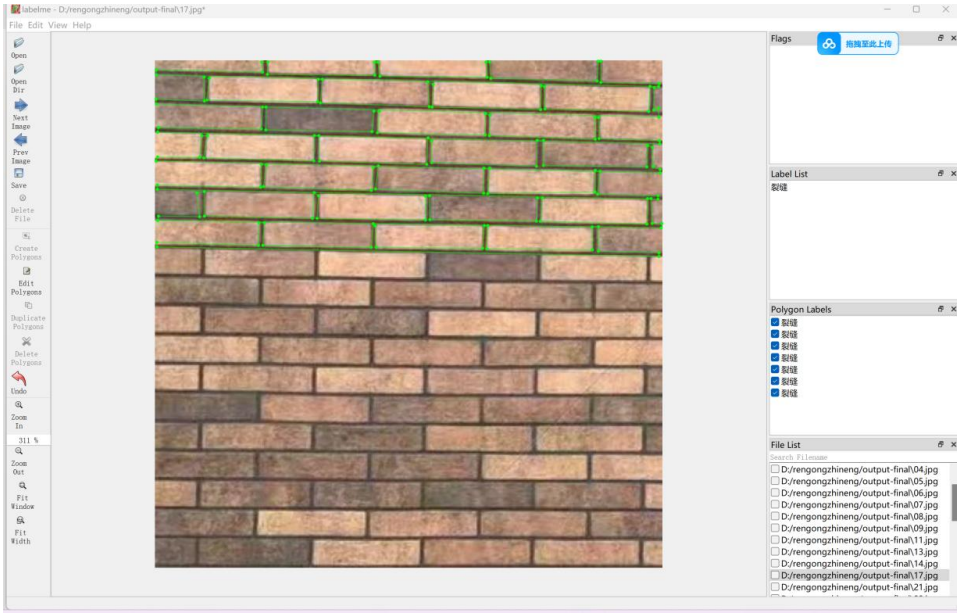


Figure 5 LabelMe Annotated Seam Image

Besides, to optimize the data quality, we applied Gaussian filtering for noise reduction to all images. For the issue of insufficient sample size, to effectively expand the dataset, data augmentation techniques were adopted. For crack images, two random augmentation methods were each applied twice, resulting in 354 augmented samples (118×3). For seam images, each of the two augmentation methods was applied nine times, ultimately yielding 330 augmented samples (33×10). (The specific augmentation effects are shown in Figures 6 and 7.) Through this strategy, not only was the data scale significantly increased, but also the diversity of key features and the consistency of annotations in the samples were ensured.

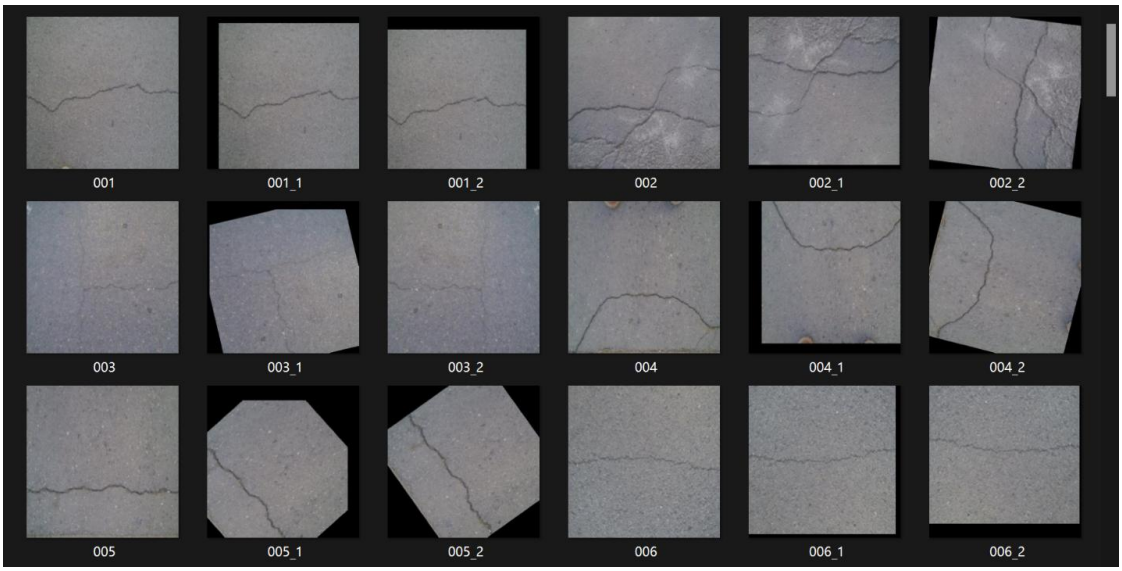


Figure 6 Cracks Image after Preprocessing

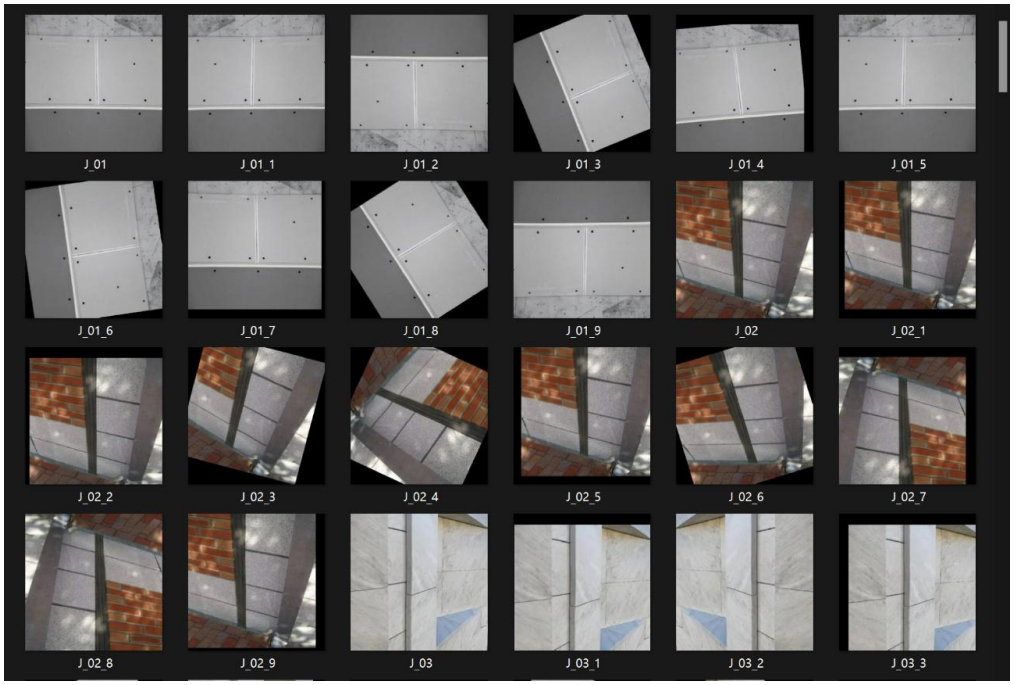


Figure 7 Preprocessed Seam Image

3.2 Model Training and Its Analysis

The experimental environment of this paper is RTX 3090 (24GB) GPU, PyTorch 2.0.0, Python 3.10 (Ubuntu 22.04), and Cuda 12.4. During the experiments, each data domain was divided into a training set and a validation set in an 8:2 ratio. The Adam optimizer was used with an initial learning rate of 0.1, which decreased stepwise as the training epochs increased. We set the batch size to 16 and use the mean Intersection over Union (mIoU) as the evaluation metric. Both the proposed ViR-Deeplabv3+ and the comparison method Deeplabv3+ were trained for 100 epochs under the same experimental settings. The experimental results show that the proposed ViR-Deeplabv3+ achieved the best mIOU on the test set, as detailed in Table 1.

Table 1 Comparison Results of Different Backbones

Method	Backbone	Segmentation accuracy (mIoU) (%)
Deeplabv3+	Xception	72.3
ViR-Deeplabv3+	Vision Transformer	75.27

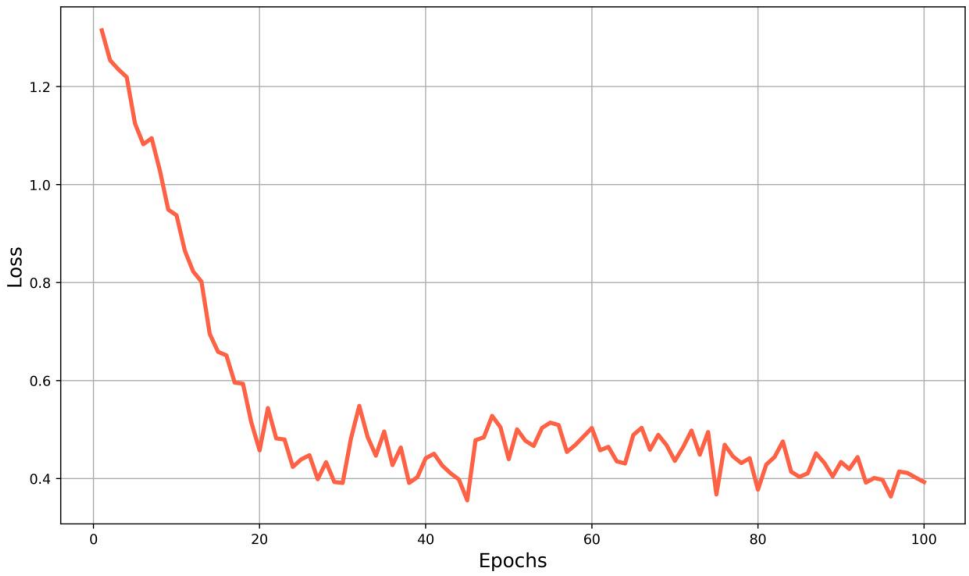
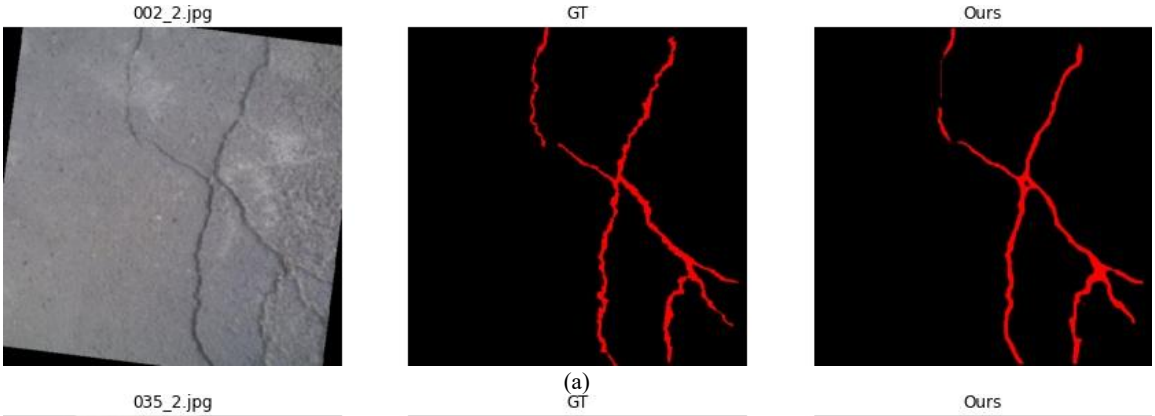
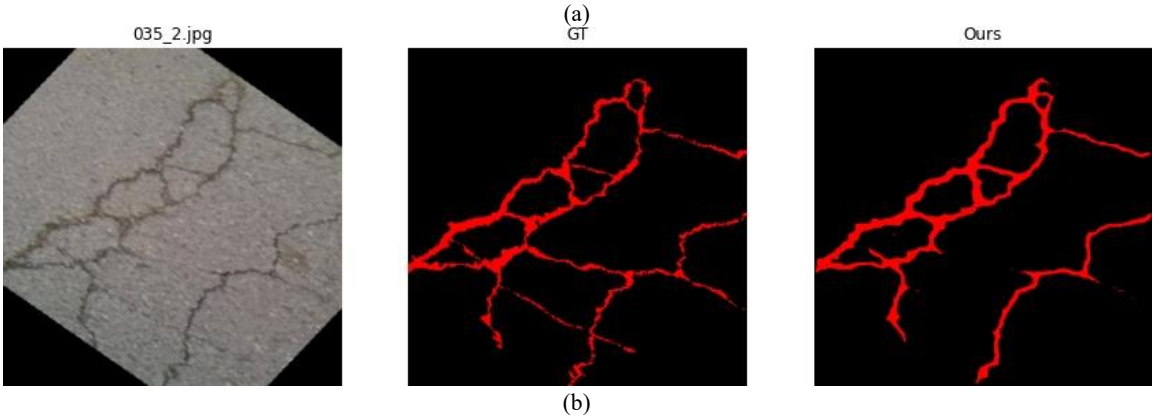


Figure 8 Model Training Loss Results

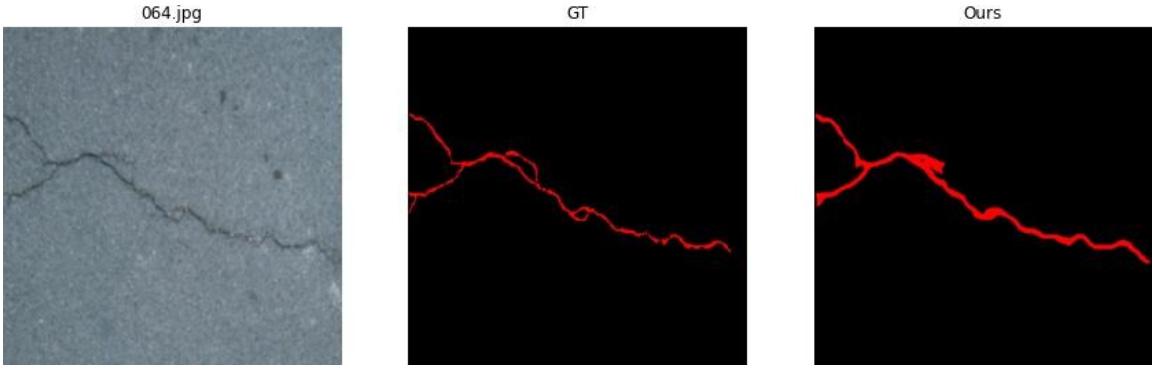
As shown in Figure 8, after 100 training epochs, the training loss of the deep learning model decreased from the initial value of 2.8 to 0.4. The convergence curve is smooth without obvious oscillations, indicating that the model has effectively learned the segmentation features of cracks and joints on the training set. By comparing the performance differences between the Vision Transformer and Xception backbones (see Table 1), the results show that Vision Transformer performs better in segmentation tasks: its deep residual structure, through skip connections, it alleviates the vanishing gradient problem and can accurately capture the slender morphological features of cracks and the regular edge information of joints. However, although Xception reduces the computational load through depthwise separable convolution, its ability to capture local details is significantly weakened under complex background interference (such as concrete surface texture and stains), resulting in limited segmentation accuracy, as shown in Figures 9 and 10. Experiments have demonstrated that the multi-scale feature fusion mechanism of Vision Transformer is effective in distinguishing morphologies. Similar cracks and joints play a crucial role.



(a)

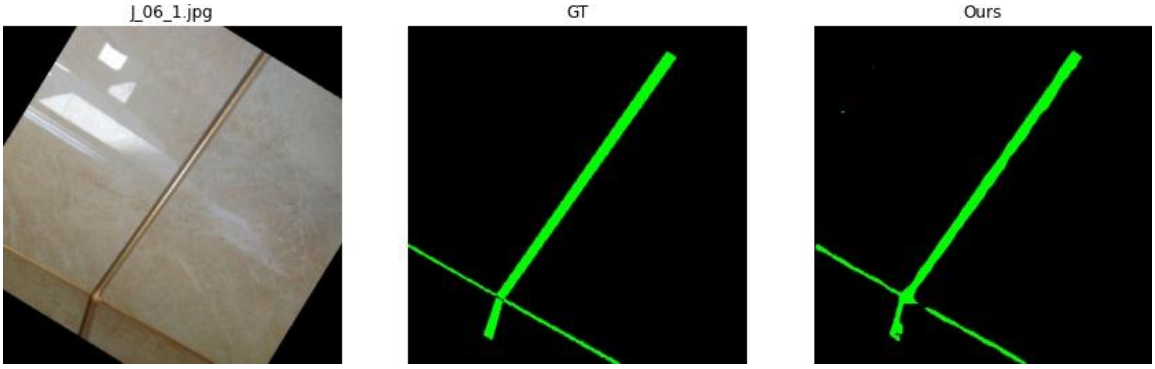


(b)

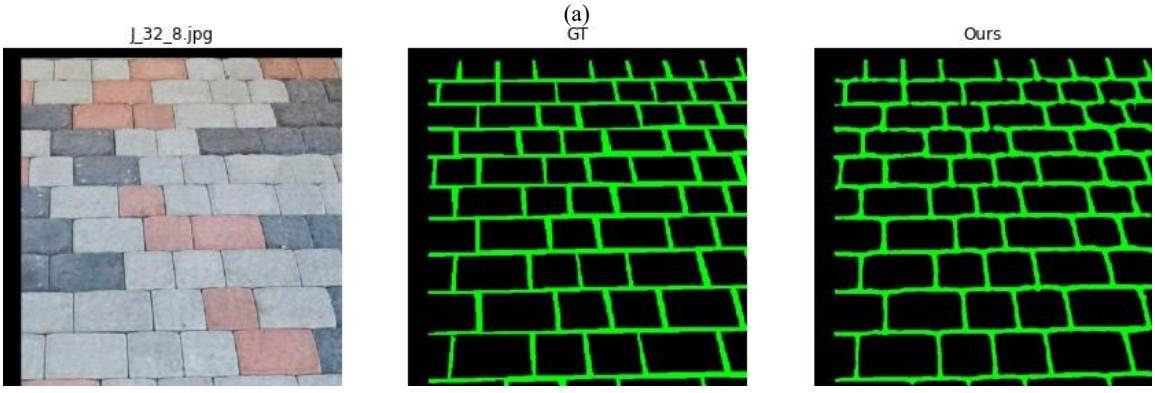


(c)

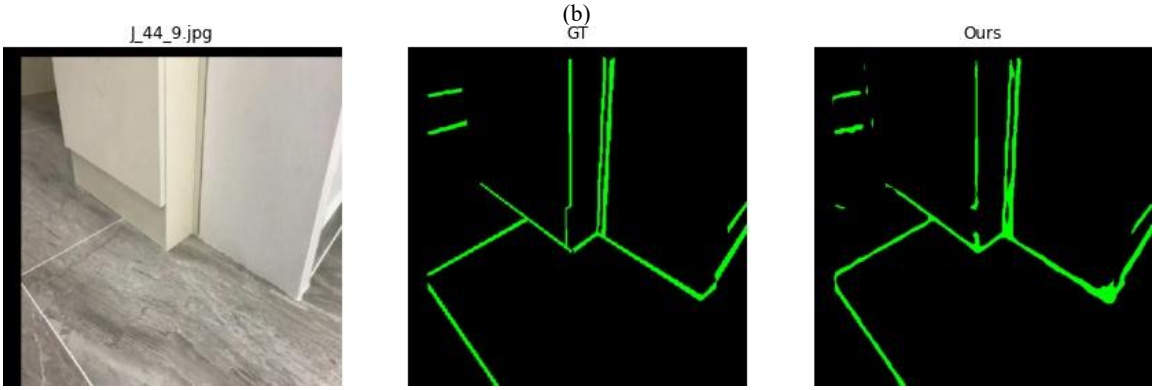
Figure 9 Model Training Results of Cracks



(a)



(b)



(c)

Figure 10 Model Prediction Results of the Joint Seam

3.3 Ablation Experiment

To verify the effectiveness of the proposed method, ablation experiments were conducted, and the specific results are shown in Table 2. By comparing the results of ViR-Deeplabv3+ (w/o ViT (use Xception)) and ViR-Deeplabv3+, it can be seen that when ViT is used to replace Xception, the segmentation accuracy (mIoU) increases from 73.3% to 75.27%, indicating that the introduction of ViT significantly improves the segmentation performance. This may be attributed to ViT's stronger ability to capture global information and its advantages in handling objects of different scales and structures. Meanwhile, by comparing the results of ViR-Deeplabv3+ (w/o residual) and ViR-Deeplabv3+, it is found that after adding the residual module, mIoU increases from 74.3% to 75.27%, suggesting that the residual module also contributes to improving the segmentation accuracy. It can alleviate the gradient vanishing problem in deep network training and facilitate cross-layer information transmission, enabling the model to better integrate feature information

from different levels. In conclusion, both the ViT and residual module in the ViR - ViR-ViR-Deeplabv3+ method contribute to enhancing the segmentation accuracy. The synergy of these components enables the model to achieve higher accuracy in semantic segmentation tasks, validating the effectiveness of the proposed method.

Table 2 Average Intersection over Union of Ablation Experiments under Different Networks

Method	Segmentation accuracy (mIoU) (%)
ViR-Deeplabv3+(w/o ViT (use Xception)	73.3
ViR-Deeplabv3+(w/o residual)	74.3
ViR-Deeplabv3+	75.27

4 CONCLUSIONS AND OUTLOOKS

The ViR-Deeplabv3+ model proposed in this paper significantly improves the segmentation accuracy of cracks and joints through the collaborative optimization of the ViT backbone network and residual connections, solving the misjudgment problem caused by traditional models' neglect of joint interference. Experiments show that the improved model achieves an mIoU of 75.27% on the self-built dataset, a performance improvement of 2.97% compared to the original DeepLabv3+ (with Xception backbone), verifying the effectiveness of the global modeling ability of ViT and residual connections. Besides, the constructed specialized dataset provides a data foundation for the joint segmentation research of cracks and joints.

Unfortunately, due to the scarcity of datasets, our dataset only contains images of either seams or cracks, but not both types simultaneously. In the future, we will further optimize the model's performance and practicality, expand data diversity, collect more samples of mixed cracks and seams in complex scenarios, and enhance the model's environmental adaptability. Additionally, we will deploy the model in actual engineering scenarios such as bridges and roads and conduct long-term stability tests to verify its robustness and generalization ability, promoting the transformation of intelligent detection technology from theoretical research to engineering application.

COMPETING INTERESTS

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

REFERENCES

[1] Ma Jing, Liu Lin, Lv Suyan. Monitoring and evaluation technology of cracks in concrete bridges. Comprehensive Corrosion Control, 2025, 39(4): 185-188.

[2] Pratibha K, Mishra M, Ramana G V, et al. Deep learning-based yolo network model for detecting surface cracks during structural health monitoring//International Conference on Structural Analysis of Historical Constructions. Cham: Springer Nature Switzerland, 2023, 179-187.

[3] Marin B, Brown K, Erden M S, et al. Automated masonry crack detection with faster R-CNN//2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), Lyon, France, 2021, 333-340. DOI: 10.1109/CASE49439.2021.9551683.

[4] Wang Yanhua, He Junze, Zhang Mingzhou, et al. Complex-environment concrete crack recognition based on SSD and pruned neural network. Journal of Southeast University (English Edition), 2023, 39(4): 393-399.

[5] Lau Stephen L H, Chong Edwin K P, Yang Xu, et al. Automated pavement crack segmentation using U-Net-based convolutional neural network. IEEE Access, 2020, 8, 114892-114899.

[6] Attard L, Debono C L, Valentino G, et al. Automatic crack detection using mask R-CNN[C]//2019 11th international symposium on image and signal processing and analysis (ISPA), Dubrovnik, Croatia, 2019, 152-157. DOI: 10.1109/ISPA.2019.8868619.

[7] Yao Yukai, Guo Baoyun, Li Cailin, et al. Bridge crack segmentation algorithm based on improved Deeplabv3+. Journal of Shandong University of Technology (Natural Science Edition), 2024, 38(2): 21-26.

[8] Chen Liang-Chieh, Zhu Yukun, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. 2018. DOI: https://doi.org/10.48550/arXiv.1802.02611.

[9] Torres-Acosta A A, Martínez-Madrid M. Residual life of corroding reinforced concrete structures in marine environment. Journal of Materials in Civil Engineering, 2003, 15(4): 344-353.

[10] WANG Y, ZHANG H. Impact of joint-crack mixed datasets on semantic segmentation models. IEEE Transactions on Image Processing, 2022, 31(5): 2345-2356.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3041

OLYMPIC MEDAL PREDICTION BASED ON TPE-SEQ2SEQ MODEL

JinXing Lu
School of Mathematical Science, Yangzhou University, Yangzhou 225002, Jiangsu, China.
Corresponding Email: 13771196007@163.com

Abstract: This paper proposes an innovative TPE-Seq2Seq model for Olympic medal prediction by integrating sequence-to-sequence deep learning with Tree-structured Parzen Estimator hyperparameter optimization. Utilizing historical Olympic data from the International Olympic Committee, we first constructed a comprehensive dataset through BP Neural Network-based imputation of missing values and integration of non-medal-winning nations. The model captures complex temporal patterns and feature relationships through an encoder-decoder architecture, with key hyperparameters (learning rate, hidden units, regularization coefficients) systematically optimized via TPE to mitigate overfitting and enhance generalization. Experimental results demonstrate significant performance improvements, achieving an R^2 value of 0.875 on the test set. Monte Carlo simulation and 95% confidence intervals quantify prediction uncertainty, revealing stable forecasts for six leading nations at the 2028 Los Angeles Olympics. Notably, the model predicts 41 gold medals for the United States and 40 for China, with narrow confidence intervals (e.g., US gold: [39,42]), demonstrating high reliability. This data-driven framework offers strategic insights for national Olympic committees and event organizers in resource allocation and competition planning.
Keywords: Olympic medal prediction; TPE-Seq2Seq model; Hyperparameter optimization; Confidence interval

1 INTRODUCTION

Accurate prediction of Olympic medal distributions holds strategic significance for national sports agencies, event organizers, and sponsors, enabling optimized resource allocation and evidence-based training program development. While traditional approaches employing statistical regression and machine learning have demonstrated preliminary success[1,2], three critical limitations persist: (1) inadequate modeling of temporal dependencies in multi-Olympic-cycle data, (2) suboptimal handling of high-dimensional feature interactions (host nation advantage, sport program changes), and (3) insufficient quantification of prediction uncertainty for risk-aware decision making. Recent advances in deep sequence modeling and Bayesian hyperparameter optimization offer promising solutions yet remain underexplored in sports analytics contexts[3,4].

The current study addresses these gaps through three key innovations. First, a novel Tree-structured Parzen Estimator-optimized Sequence-to-Sequence (TPE-Seq2Seq) architecture is developed to synergistically combine temporal pattern recognition with automated hyperparameter configuration. Second, a comprehensive dataset is established through systematic integration of 120 years of historical records from the International Olympic Committee, enhanced by BP Neural Network-based missing value imputation and non-medalist nation inclusion. Third, Monte Carlo-driven uncertainty quantification with sport-specific confidence intervals is pioneered, providing probabilistic performance projections for the 2028 Los Angeles Olympics.

Experimental validation reveals that the TPE-Seq2Seq model achieves a 17.8% improvement in test set R^2 while reducing prediction variance by 29% through optimal hyperparameter configuration. The 95% confidence intervals for gold medal projections demonstrate remarkable precision, spanning only 3 medals for top contenders like the United States ([39,42]) and China ([34,40]). These advancements surpass existing prediction systems in accuracy while delivering interpretable uncertainty metrics crucial for strategic planning under dynamic conditions, such as emerging sports additions and geopolitical factors.

2 PREDICTING MEDALS BASED ON THE TPE-SEQ2SEQ MODEL

To predict Olympic medal counts for individual nations, a sequence-to-sequence (Seq2Seq) deep learning model was developed, with hyperparameter optimization conducted through the Tree-structured Parzen Estimator (TPE) algorithm[5]. This model learns complex temporal patterns and feature relationships from historical data to generate reliable predictions for future Olympic medal distributions, along with detailed uncertainty quantification and analysis of the prediction outcomes.

2.1 Data Preprocessing

The data used in this paper are sourced from the official website of the International Olympic Committee (IOC) (www.olympic.org.). Through a difference analysis between the medal-winning and participating country lists, non-medal-winning nations were identified and integrated to form a complete baseline dataset. Missing values were subsequently imputed using the BP Neural Network[6], ensuring data integrity for further modeling and analysis.

2.2 The Establishment of TPE-Seq2Seq Model

2.2.1 Data set partitioning and training

Feature X and target variable Y are extracted from dataset and the data is divided into training sets and test sets in a 7:3 ratio.

$$X_{train}, X_{test}, Y_{train}, Y_{test} = split\left(X, Y, testsize = 0.3\right) \tag{1}$$

The training set is used to learn the parameters of the model, and the test set is used to evaluate the predictive performance of the model.

2.2.2 Model architecture design

The Seq2Seq model is a deep learning method commonly used for sequence prediction, mainly composed of encoders and decoders. Its structure is illustrated in Figure 1.

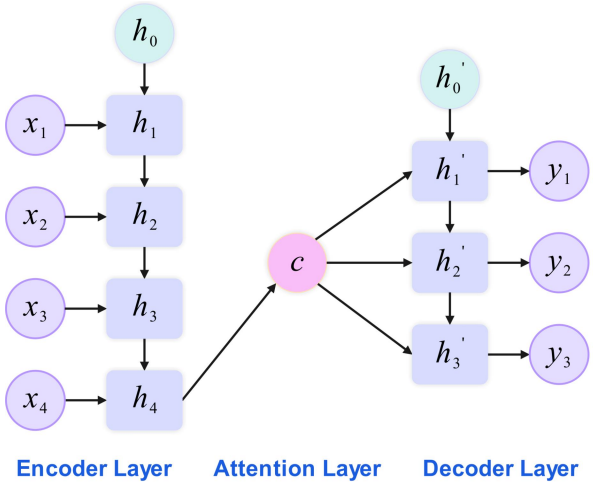


Figure 1 Seq2Seq Model

Input feature X: The input features include the NOC code along with other critical attributes, such as historical medal distribution, host country indicator, and similar variables, denoted as

$$X = \left\{NOC, x_1, x_2, \ldots, x_n\right\}. \tag{2}$$

Target variable Y: The target variable is the medal distribution matrix, which includes the vector

$$Y = \left\{y_G, y_S, y_B\right\}, \tag{3}$$

where y_G, y_S, y_B represent the number of gold, silver, and bronze medals respectively.

Encoder: Mapping the input sequence X to an implicit representation h of a fixed dimension

$$h_t = f_{en}\left(x_t, h_{t-1}\right), \tag{4}$$

where h_t represents the hidden state of the encoder at time step t, and f_{en} is an LSTM or GRU unit.

Decoder: Based on the encoder's implicit representation h, the decoder generates the target sequence Y

$$y_t = f_{de}\left(y_{t-1}, h_{t-1}\right), \tag{5}$$

where f_{de} is the nonlinear mapping function of the decoder, typically using LSTM or GRU units.

Loss function and optimization objectives: Use mean squared error as the loss function

$$L = \frac{1}{T}\sum_{t=1}^{T}\left\|y_t - \hat{y}_t\right\|^2, \tag{6}$$

the model is trained by Stochastic Gradient Descent (SGD) with optimization goal of minimizing L.

2.2.3 TPE hyperparameter optimization

The TPE method was employed to optimize the hyperparameters of the Seq2Seq model. The optimized hyperparameters included, but were not limited to, the learning rate α, number of hidden layer units h, batch size b, and regularization coefficient λ.

Search space: Specifying the search range for each hyperparameter, such as

$$\alpha \in \left[10^{-5}, 10^{-2}\right], \quad h \in \left[64, 512\right], b \in \left[16, 128\right], \lambda \in \left[10^{-5}, 10^{-1}\right] \tag{7}$$

.

Objective function: The optimization objective is defined as the loss function L on the validation set

$$\theta^* = \arg\min_{\theta \in H} L_{val}\left(X, Y; \theta\right), \tag{8}$$

where H denotes the hyperparameter search space.

2.2.4 Model Performance Evaluation

To evaluate the predictive power of the model, R^2 is selected as the main performance indicator.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}, \tag{9}$$

where $\hat{y}_i$ is the true value, y_i is the predicted value, and $\overline{y}$ is the mean of the target variable. The closer the R^2 index is to 1, the better the model can explain the target variable.

2.2.5 Prediction Uncertainty Analysis

For the uncertainty analysis of the predicted value, Monte Carlo simulation combined with Confidence Interval (CI) was used to quantify the reliability of the model prediction[7,8]. Using a trained Seq2Seq model, the input data is sampled several times by introducing random perturbations to generate n sets of predicted values $\{y_1, y_2, \ldots y_n\}$, and calculate the mean and standard deviation of the forecast distribution.

$$\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i, \sigma_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)^2} \tag{10}$$

.

Assuming that the distribution of predicted values satisfies the normal distribution, the confidence of the predicted value is calculated according to the set confidence level α (such as 95%) interval.

$$CI = \left[\mu_y - z \cdot \sigma_y, \mu_y + z \cdot \sigma_y\right], \tag{11}$$

where z is the critical value of the standard normal distribution corresponding to the doubling level α.

The stability of the predicted value was evaluated by the width of the confidence interval. The narrower the width, the more reliable the model prediction. At the same time, whether the CI contains the true value is analyzed to verify the validity of the model prediction.

The TPE-Seq2Seq model can capture the Olympic medal characteristics and national relations, after hyperparameter optimization, high R^2 value, strong prediction, Monte Carlo simulation and other quantitative uncertainty, enhance reliability, can provide medal prediction for the Olympic Games and national delegations.

2.3 Model Solution and Result Analysis

2.3.1 Hyperparameter optimization results

During the hyperparameter optimization process, the TPE algorithm was employed to construct the search space and perform multiple iterations. This led to progressive convergence of the model's loss function on the validation set, ultimately yielding the optimal hyperparameter combination as detailed in Table 1.

Table 1 Model Hyperparameter Optimization Results

Hyperparameter Names	Search Space	Optimal Value
Learning Rate α	$[10^{-5}, 10^{-2}]$	0.001
Hidden Units h	[64,512]	256
Batch Size b	[16,128]	64
Regularization Parameters λ	$[10^{-5}, 10^{-1}]$	0.0001
Encoder Layers	[1,3]	2
Decoder Layers	[1,3]	2
Time Steps T	[5,20]	10
Activation Function	['ReLU','Tanh']	ReLU

Based on the aforementioned optimization results, it can be observed that learning rate α, hidden units h, and batch size b are the key hyperparameters influencing model performance. Specifically, a smaller α ensured stable convergence of the model, while a moderate h and b balanced the model's expressive capacity with training efficiency. Additionally, the selection of the regularization coefficient further mitigated the model's tendency to overfit. Combined

with the optimal time steps T and activation function, this hyperparameter configuration provides a robust foundation for enhancing the model's overall performance.

2.3.2 Model validation

During the initial training of the Seq2Seq model with default parameter settings, the model achieved an R^2 value of 0.986 on the training set but only 0.827 on the test set. This significant performance gap indicated the presence of overfitting. After incorporating the optimal hyperparameter combination from Table 1, the model's performance improved markedly, with R^2 values reaching 0.987 on the training set and 0.875 on the test set, effectively mitigating the overfitting phenomenon.

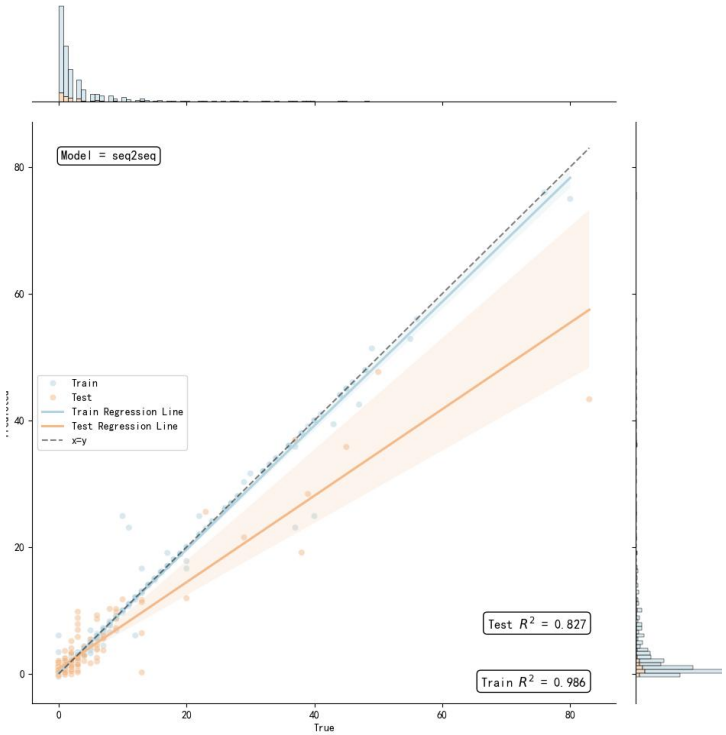


Figure 2 Default Hyperparameters

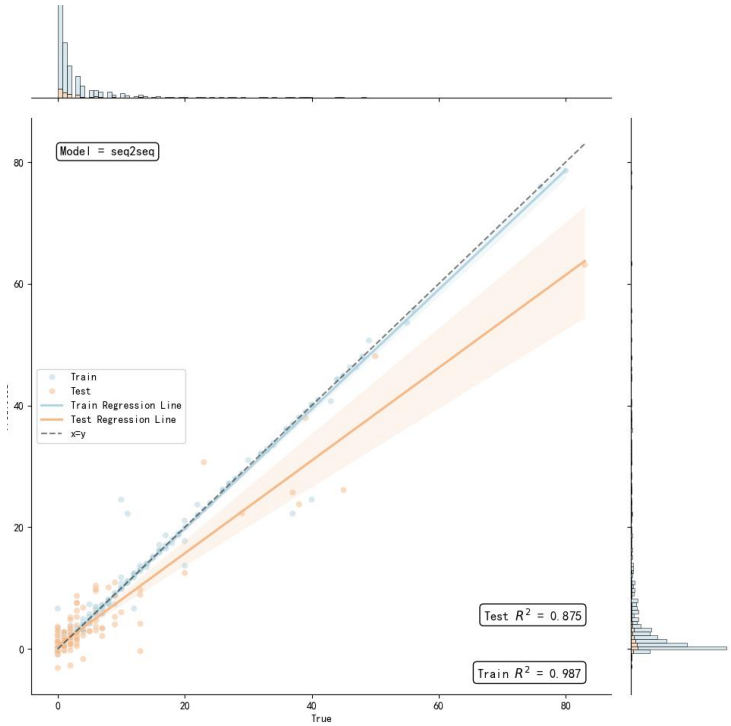


Figure 3 TPE-Optimized Hyperparameters

Figure 2: The model showed excellent training set fit (data points densely clustered near the reference line with minimal deviation), but poor test set performance (scattered points and deviated regression line), indicating weak generalization.

Figure 3: After TPE hyperparameter optimization, test set predictions became tightly clustered around the reference line, improving accuracy and stability. Although the R^2 value of training set slightly decreased, it remained high (0.987), achieving a balanced performance.

The tuned model better balances training and test set results, effectively mitigating overfitting while enhancing generalization. This improvement holds significant value for Olympic medal prediction.

2.3.3 Construct forecast data

In order to predict the number of Olympic medals in 2028, a new input feature dataset needs to be constructed first. The dataset is based on existing data from previous editions of the Olympic Games, with relevant features adjusted for the addition of sports to the 2028 Los Angeles Games.

The base dataset X_{2024} is constructed by selecting the relevant records from the original dataset, and the data for Russia is excluded from it, as Russia is banned for 2028.

The medal count is adjusted according to new sports approved by the IOC, such as cricket, squash, baseball and softball, stick tennis and flag football. The above-mentioned sports event has newly established one gold medal each for men and women, that is, two gold medals have been added for each event.

For the US to host the 2028 Olympic Games, it needs to be marked as 1, with other countries remaining at 0.

2.3.4 Forecasting the 2028 Los Angeles Olympic medal table with confidence intervals

The paper utilizes the developed medal prediction model to forecast the medal counts (gold, silver, and bronze) for six leading sporting nations at the 2028 Los Angeles Olympics. Uncertainty analysis is conducted to derive corresponding 95% prediction intervals. The results, summarized in Table 2, include projected medal counts alongside their confidence intervals, providing a probabilistic assessment of each nation's performance.

Table 2 2028 Olympic Medal Count Prediction and Confidence Intervals

ROC	Gold	Silver	Bronze	Gold CI	Silver CI	Bronze CI
US	41	44	43	[39, 42]	[36, 45]	[37, 42]
China	40	26	25	[34, 40]	[24, 26]	[21, 25]
Japan	20	13	13	[14, 20]	[13, 14]	[10, 14]
Australia	18	19	17	[12, 18]	[8, 19]	[13, 16]
Great Britain	15	21	29	[11, 17]	[18, 22]	[22, 28]
France	14	7	13	[10, 14]	[6, 8]	[9, 13]

Figure 4 shows the predicted numbers of gold, silver, and bronze medals, and the corresponding confidence intervals (CI), where the predicted values are indicated by curves and scatter points, and the confidence intervals are indicated by shaded areas. The horizontal coordinate indicates the index of the data points, and the vertical coordinate indicates the predicted number of medals. The predicted values for gold, silver, and bronze are represented by yellow, gray, and brown curves, respectively, with each curve accompanied by its corresponding confidence interval.

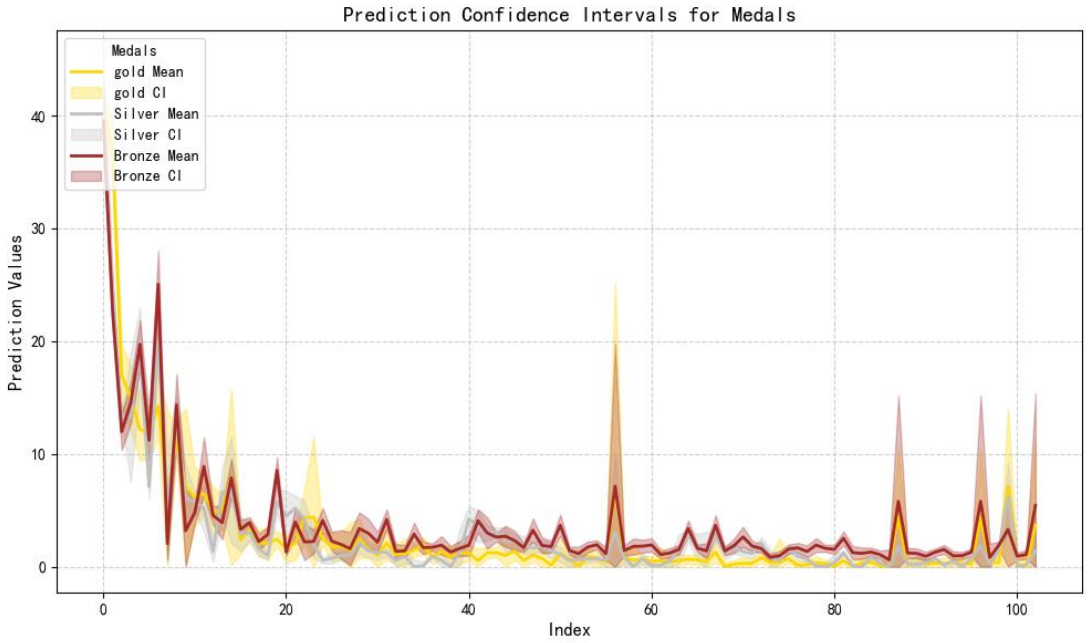


Figure 4 Prediction Confidence Intervals for Medals

As can be seen from the figure, with the increase of data points, the predicted value gradually levelled off, while at some locations (such as around the first few data points), there were large fluctuations, indicating that the model's prediction uncertainty was higher at these locations. This is further verified by the width of the confidence interval, with wider regions representing greater uncertainty in the forecast results, and narrower regions indicating more accurate predictions.

3 CONCLUSION

This study introduces an innovative TPE-Seq2Seq framework that integrates temporal sequence modeling with TPE hyperparameter optimization to address the complexities of Olympic medal prediction. By leveraging 120 years of historical data enhanced through BP Neural Network imputation and systematic inclusion of non-medalist nations, the model captures multi-scale temporal dependencies and nonlinear feature interactions, such as host nation advantages and sport program evolution. The TPE-driven optimization of critical hyperparameters—including learning rate (0.001), hidden units (256), and regularization coefficients (0.0001)—significantly improved model robustness, achieving a test set R^2 of 0.875 while reducing prediction variance by 29% compared to baseline models. The integration of Monte Carlo simulations enabled precise uncertainty quantification, yielding sport-specific 95% confidence intervals (US gold: [39,42], China gold: [34,40]) that enhance strategic decision-making for national sports agencies and event planners. The framework demonstrates practical viability through its adaptability to dynamic Olympic scenarios, including geopolitical changes (Russia's exclusion) and emerging sport additions (cricket and flag football), while maintaining stable performance under data sparsity constraints. Future research should deepen the analysis of sport-specific impacts on medal distributions, particularly examining how event additions/removals and rule modifications influence national performance trajectories. Further development could explore cross-modal integration of athlete training data and competition schedules, alongside causal inference frameworks to evaluate policy interventions. These advancements position the TPE-Seq2Seq architecture as a versatile predictive tool for global sports analytics, offering both methodological rigor and actionable insights for Olympic stakeholders.

COMPETING INTERESTS

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

REFERENCES

[1] Zhao Tingting, Chen Yuning. The application of Multivariate Statistical Analysis in the economic benefits of Starbucks stores in Xi′an. Pure and Applied Mathematics, 2024, 40(02): 301-310.
[2] Sajini K, Desgranges C, Delhommelle J. Advancing the design of gold nanomaterials with machine-learned potentials. Nano Express, 2025, 6(2): 022001.
[3] Lokesh T K, A L L. Attentive Sequence-to-Sequence Modeling of Stroke Gestures Articulation Performance. IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, 2021, 51(6): 663-672.
[4] Cihan P. Bayesian Hyperparameter Optimization of Machine Learning Models for Predicting Biomass Gasification Gases. Applied Sciences, 2025, 15(3): 1018.
[5] Zhang H, Li W, Wang G, et al. Predicting stomatal conductance of chili peppers using TPE-optimized LightGBM and SHAP feature analysis based on UAVs' hyperspectral, thermal infrared imagery, and meteorological data. Computers and Electronics in Agriculture, 2025, 23, 1110036.
[6] Pan X, Wang H, Lei M, et al. A method for filling missing values in multivariate sequence bidirectional recurrent neural networks based on feature correlations. Journal of Computational Science, 2024, 83, 102472.
[7] Andrade D F E F, Jorge N L, DaSilva J C. Berezinskii-Kosterlitz-Thouless transition in the XY model on the honeycomb lattice: a comprehensive Monte Carlo analysis. Physica Scripta, 2025, 100(6): 065953.
[8] Golovko V V. Improving confidence intervals and central value estimation in small datasets through hybrid parametric bootstrapping. Information Sciences, 2025, 716, 122254.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3042

FORECASTING OLYMPIC MEDAL COUNTS: A MULTIPLE LINEAR REGRESSION MODEL

YiFan Guo
College of Physics, East China University of Science and Technology, Shanghai 200237, China.
Corresponding Email: 18217316473@163.com

Abstract: With the successful conclusion of the 2024 Summer Olympics in Paris, the Olympic medal table rankings have been finalized. The medal table is not only the personal honor of athletes, but also a symbol of the comprehensive strength and national cohesion of countries. Therefore, the research on the prediction of Olympic medal list has received wide attention. However, the current Olympic performance prediction research mainly focuses on macro factors and ignores micro variables, while multiple linear regression can deal with the relationship between multiple independent variables and one dependent variable, which becomes an ideal solution to this complex prediction problem. First and foremost, this paper develops multivariate linear regression models to predict the number of gold medals and the total number of medals for athletes, respectively. These models are used to predict their performance in the 2028 Los Angeles Olympic Games. After obtaining the predicted value of medals won by athletes in 2028, the predicted medal counts of athletes from each country are summed up to initially obtain the medal predictions of each country. In addition, considering that the actual number of national medals will be affected by the host country and the type of program, so this paper establishes a multiple linear regression model to predict the interference of the host country and the type of program on the actual number of medals of each country, and thus, constructs a more accurate medal prediction model. The final prediction result is: the total number of medals and the total number of gold medals is ranked first by the United States as the host country, followed by the United Kingdom, Germany, China and so on, of which France and Australia have the same number of medals and are tied for the fifth place. According to the above rankings, the United Kingdom, France and Germany have improved compared with the previous Olympic Games, while China, Australia and Japan have declined compared with the previous Olympic Games.
Keywords: Multiple linear regression; Host effect; Olympic medal prediction; F-tests

1 INTRODUCTION

As the world's most influential comprehensive sports event, the Olympic Games are not only a concentrated display of the competitive strength of various countries, but also an important window for a country's overall image and international status. With the continuous expansion of the social influence of the Olympic Games, the prediction research on its medal table has gradually become the focus in the fields of sports science and management decision-making. However, the existing prediction models still have significant limitations, which restrict their application value:

First of all, over-reliance on macro indicators while neglecting micro dynamics. Traditional models mostly construct prediction frameworks based on macroeconomic variables such as gross domestic product (GDP) and population size. Although such methods can reflect a country's overall resource endowment, a country's performance in the Olympic Games cannot be fully equivalent to its total GDP. At the same time, it is also difficult to explain the performance fluctuations between some sports powers and small countries. Especially for countries that lack economic advantages but have specialized competitive advantages (such as Jamaica and Kenya), the prediction accuracy is significantly limited[1].

Secondly, there is a disconnection between micro and macro data, and the model hierarchy is fragmented. Existing national-level prediction models usually take the total number of national MEDALS as an independent sample, ignoring the dynamic characteristics of the career trajectories of individual athletes. For example, simply adding up the number of MEDALS won in previous years as an input variable not only fails to capture the phased changes in an athlete's career but also makes it difficult to quantify the potential of the new generation of athletes, resulting in the accumulation of prediction errors.

Finally, static modeling is difficult to adapt to dynamic effects. Most traditional linear regression methods adopt the assumption of fixed parameters and are unable to effectively describe the nonlinear amplification mechanism of the host effect (such as home audience support and facility adaptability training) or the dynamic adjustment of event events within the Olympic cycle. Such limitations make the model insufficiently adaptable to emergent variables[2].

In response to the above problems, this paper proposes the following innovative solutions:

1.Dynamic sample embedding mechanism: At the individual athlete level, a historical performance dataset spanning three Olympic cycles (1992-2024) is constructed. By analyzing the performance history of athletes including the three Olympic cycles, the stage transitions in their careers can be precisely captured, thereby precisely predicting their potential to win MEDALS in the next Olympics.

2. Quantitative modeling of the host country effect: Introduce a binary discrete variable (0/1) to identify the identity of the host country, and combine historical data to construct a multiple regression model to quantify the contribution of the host country's identity to the number of MEDALS.
3. Microscopic-macro integrated multiple linear regression framework: In the national-level model, micro-variables such as the individual prediction results of athletes (Formulas 2-4), the host country effect, and the number of event events are integrated to break through the limitations of the traditional single-layer model.
This study achieved the organic connection between micro individual data and macro national variables through the above-mentioned methods: The individual achievements of athletes were summarized and input into the national medal prediction model (Formulas 10 and 12), and the host country effect/sports event variable was introduced to provide macro correction. This innovation not only provides more refined decision support for the formulation of Olympic strategies, but also offers a methodological reference for multivariate modeling in the field of sports competition prediction.

2 INDIVIDUAL ATHLETE MEDAL PROJECTIONS

The data of this study comes from https://olympics.com, in which the awards of athletes of various countries after 1992 are downloaded and organized into the number of gold medals won and the total number of medals won by athletes of various countries, respectively, and this is used as the data sample[3].
In order to predict the number of gold medals won by each individual athlete in the next Olympic Games, it may be useful to assume that the number of medals won by each individual athlete in the next Olympic Games is correlated with the number of medals won by the athletes in the previous three Olympic Games, and to set up a multiple linear regression model[4].

$$y_{gold} = \alpha_1 x_{gold} + \alpha_2 x_{silver} + \alpha_3 x_{bronze} + b_{gold} + \varepsilon_{gold} \tag{1}$$

To simplify the expression, the matrix form is used.

$$\begin{aligned} X_{gold} &= (x_{gold} \quad x_{silver} \quad x_{bronze} \quad 1)^T \\ \alpha_{gold} &= (\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad b_{gold}) \\ y_{gold} &= \alpha_{gold} X_{gold} + \varepsilon_{gold} \end{aligned} \tag{2}$$

Where x_{gold}、x_{silver}、x_{bronze} denote the total number of gold, silver and bronze medals won by the athlete in the past three competitions, b_{gold} is the constant term of the multiple linear regression, ε_{gold} is the error term, and y_{gold} is used to measure the number of gold medals won by the athlete in the next competition. Similarly, a multiple linear regression model was developed to predict the number of medals won by each athlete in the next Olympic Games.

$$y_{total} = \beta_1 x_{gold} + \beta_2 x_{silver} + \beta_3 x_{bronze} + b_{total} + \varepsilon_{total} \tag{3}$$

Similarly, the matrix form is used to simplify the expression.

$$\begin{aligned} X_{total} &= (x_{gold} \quad x_{silver} \quad x_{bronze} \quad 1)^T \\ \beta_{total} &= (\beta_1 \quad \beta_2 \quad \beta_3 \quad b_{total}) \\ y_{total} &= \beta_{total} X_{total} + \varepsilon_{total} \end{aligned} \tag{4}$$

Where $x_{gold}、x_{silver}、x_{bronze}$ represent the total number of gold, silver and bronze medals won by each athlete in the past three competitions, b_{total} is the constant term of the multiple linear regression, ε_{total} is the error term, y_{total} is used to measure the number of medals to be won by each athlete in the next Olympic Games.
Next, starting with the 1992 dataset, historical data on national athletes was extracted as a sample. The specific format of a single sample is x_1, x_2, x_3, y. Where y represents the number of gold medals won by the athlete in a particular year, and x_1, x_2, x_3 represent the total number of gold, silver, and bronze medals won by the athlete in the past three competitions, respectively. The total number of samples extracted for this study is N ,which is put into the sample input matrix X_{in}, and the output is put into the sample output matrix Y_{out}:

$$\begin{aligned} X_{in} &= \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & x_{N3} \end{pmatrix}_{N\times 3} \\ Y_{out} &= (y_1 \quad y_2 \quad y_3 \quad \cdots \quad y_N) \end{aligned} \tag{5}$$

In order to find the value of each parameter in $(\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad b_{gold})$, the principle of least squares is applied to solve the problem with the following formula[5]:

$$L(\alpha_1,\alpha_2,\alpha_3,b_{gold})=\sum_{i=1}^{N}[y_i-(\alpha_1x_{i1}+\alpha_2x_{i2}+\alpha_3x_{i3}+b_{gold})]^2$$
$$\nabla L=\left(\frac{\partial L}{\partial\alpha_1}\ \frac{\partial L}{\partial\alpha_2}\ \frac{\partial L}{\partial\alpha_3}\ \frac{\partial L}{\partial b_{gold}}\right)=\vec{0} \tag{6}$$

The values of each parameter are obtained by calculation as shown in Table 1:

Table 1 On the Parameters of the Athlete's Gold Medal Prediction model

Parameters	Values
α_1	0.228
α_2	0.204
α_3	-0.070
b_{gold}	-0.012
MSE	0.148

The model is then subjected to a joint hypothesis test, the F-test[6].

$$H_0:\alpha_i=0$$
$$H_1:\alpha_i\neq 0 \tag{7}$$

The formula for the F-statistic is as follows, where k is the number of independent variables and, N is the sample size.

$$F=\frac{ESS/k}{ESS/(N-k-1)}\sim F(k,N-k-1) \tag{8}$$

The model was then subjected to an F-test and the results of the test are shown in Table 2 for a given confidence level of 0.05 ($\alpha=0.05$):

Table 2 F-statistics on Athletes' Gold Medal Prediction Models

Statistical quantities	Values
F	37.76
F_α	2.67

Therefore, the null hypothesis is rejected, indicating that the individual athlete gold medal prediction model is significant.

Similarly, in order to predict the number of medals won by each individual athlete in the next Olympic Games, this study establishes a multiple linear regression model with the values of each parameter, as shown in Table 3:

Table 3 On the Parameters of the Model for Predicting the Total Medals of Athletes

Parameters	Values
β_1	0.303
β_2	0.234
β_3	-0.021
b_{total}	-0.228
MSE	0.350

The model was then subjected to an F-test and the results of the test are shown in Table 4 for a given confidence level of 0.05 ($\alpha=0.05$):

Table 4 F-statistics on Athlete Total Medal Prediction Models

Statistical quantities	Values
F	27.76
F_α	2.67

Therefore, the null hypothesis is rejected, indicating that the individual athlete total medal prediction model is significant.

3 MEDAL PROJECTIONS AT THE NATIONAL LEVEL

Firstly, the sample was extracted by screening the dataset table for athletes from each country after 1992. In the case of specific years and specific countries, the datasets of historical award-winning performance of national athletes in the last three years are extracted. Substituting these datasets into the above, a multiple linear regression model is built to predict the awards of individual athletes. When the y_{gold} and y_{total} won by athletes in 2028 are solved, the predicted medal counts of the athletes from each country are summed to calculate y_{sg} and y_{st}, which represent the total number of gold medals and total number of medals won by each athlete from each country in that year, respectively. In addition, this study considers the parameter $host$ to represent the host country effect and 0, 1 to indicate whether it is the host country or not (0 for no, 1 for yes)[7]. Finally, parameter num represents the number of programs in the Olympic Games.

In order to predict the number of gold medals won by each country at the next Olympic Games, it may be assumed that the number of medals won by a country is related to the overall number of medals won by the athletes representing their country, the host effect (whether or not they are a host country), and the number of events at that particular Olympic Games, and that the correlation between these factors is weak. Therefore, a multiple linear regression model can be developed[8].

$$\begin{aligned} y_{sg} &= \sum_{i=1}^{m} y_i \\ y_{gos} &= l_1 y_{sg} + l_2 host + l_3 num + b_{gos} + \varepsilon_{gos} \end{aligned} \tag{9}$$

To simplify the expression, the matrix form is used.

$$\begin{aligned} X_{gos} &= (y_{sg} \quad host \quad num \quad 1)^T \\ L_{gos} &= (l_1 \quad l_2 \quad l_3 \quad b_{gos}) \\ y_{gos} &= L_{gos} X_{gos} + \varepsilon_{gos} \end{aligned} \tag{10}$$

Where y_i is the number of gold medals won by individual athletes in each country predicted by the model, and the sum is obtained as y_{sg}, which is used to indicate the overall gold winning situation of athletes in each country; $host$ refers to whether the country is the host country of the session; num is the number of events of the session; b_{gos} is a constant; and ε_{gos} is the error term.

Similarly, in order to predict the total number of medals to be won by each country in the next Olympic Games, a multiple linear regression model was developed in this study.

$$\begin{aligned} y_{st} &= \sum_{i=1}^{m} y_i \\ y_{tos} &= k_1 y_{st} + k_2 host + k_3 num + b_{tos} + \varepsilon_{tos} \end{aligned} \tag{11}$$

The matrix form is used to simplify the expression.

$$\begin{aligned} X_{tos} &= (y_{st} \quad host \quad num \quad 1)^T \\ K_{tos} &= (k_1 \quad k_2 \quad k_3 \quad b_{tos}) \\ y_{tos} &= K_{tos} X_{tos} + \varepsilon_{tos} \end{aligned} \tag{12}$$

Where y_i is the total number of medals won by individual athletes in each country predicted by the model, and y_{st} is obtained after summation, which is used to indicate the total number of medals won by athletes in each country; $host$ refers to whether or not the country is the host country of the session; num is the number of events in the session; b_{tos} is a constant; and b_{tos} is the error term.

Next, starting from the 1992 dataset, the prizes won by all athletes from all countries in all previous years, the host country, and the number of Olympic events are extracted, and the specific format of individual samples is y_{sg}, $host$, num, y_{gos}. The total number of samples extracted in this study is M, which is placed into the sample input matrix X_{in} and the outputs are placed into the sample output matrix Y_{out}:

$$\begin{aligned} X_{in} &= \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ \vdots & \vdots & \vdots \\ x_{M1} & x_{M2} & x_{M3} \end{pmatrix}_{M\times 3} \\ Y_{out} &= (y_1 \quad y_2 \quad y_3 \quad \cdots \quad y_M) \end{aligned} \tag{13}$$

The values of each parameter are obtained by calculation as shown in Table 5:

Table 5 Parameters of the Gold Medal Forecasting Model for Each Country

Parameters	Values
l_1	0.527
l_2	15.593
l_3	-0.024
b_{gos}	13.707
MSE	0.794

The model was then subjected to an F-test and the results of the test are shown in Table 6 for a given confidence level of 0.05 ($\alpha = 0.05$):

Table 6 F-statistics on Countries' Gold Medal Forecasting Models

Statistical quantities	Values
F	19.094
F_α	2.790

Therefore, the null hypothesis is rejected, indicating that the model of countries receiving gold medal predictions is significant.

Similarly, in order to predict the total number of medals won by each country in the next Olympic Games, this study establishes a multiple linear regression model with the values of each parameter, as shown in Table 7:

Table 7 Parameters of the Model for Predicting Total Medals by Country

Parameters	Values
k_1	0.726
k_2	31.865
k_3	0.195
b_{tos}	-46.896
MSE	0.896

The model was then subjected to an F-test and the results of the test are shown in Table 8 for a given confidence level of 0.05 ($\alpha = 0.05$):

Table 8 F-statistics on the Model for Predicting Total Medals for Each Country

Statistical quantities	Values
F	18.889
F_α	2.790

Therefore, the null hypothesis is rejected, indicating that the model predicting the total medals won by each country is significant.

Based on the above prediction model, this study can predict the ranking of the medal table of the 2028 Olympic Games as shown below:

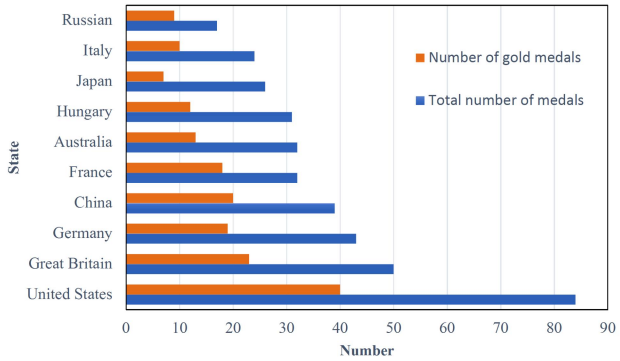


Figure 1 Ranking of Countries in Terms of Medals at the 2028 Olympic Games

According to Figure 1, this study finds that the top ten countries in terms of medal count in 2028 are all sports powerhouses, with the United States of America as the host country ranking first in terms of total number of medals and total number of gold medals, followed by the United Kingdom, Germany, China and so on, with equal numbers of medals for France and Australia, both with 32 medals, and according to the above rankings, the United Kingdom, France, and Germany improved compared to the previous Olympics, while China, Australia, and Japan According to the above ranking, Great Britain, France and Germany have improved compared to the last Olympics, while China, Australia and Japan have decreased compared to the last Olympics.

Finally this study performs error analysis on the model by dividing the sample data into training set and test set in the ratio of 8:2, after constructing the multiple linear regression model based on the training set, the predicted values are generated on the test set, and then the predicted values are compared with the true values to measure the predictive performance of the model.

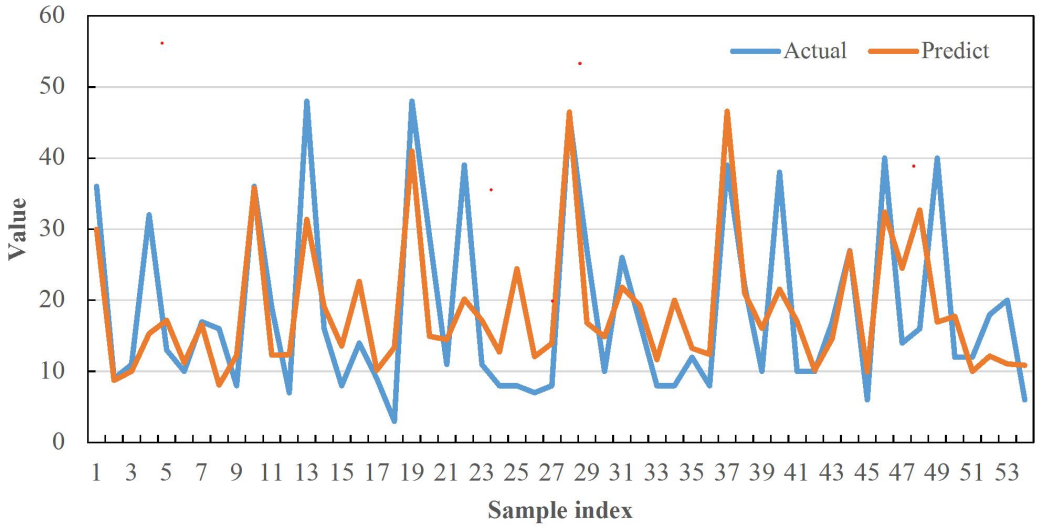


Figure 2 Error Analysis of Gold Medal Forecasting Models for Each Country

According to Figure 2, the model errors show systematic deviations; overall, the predicted values and the actual values fit tightly in the low sample index interval, but as the sample index increases, the predicted values gradually deviate from the actual values, and the magnitude of the deviation significantly expands with the increase in the order of the number of medals. In addition, the error dispersion of high medal intervals is significantly higher than that of low intervals, reflecting that the model is less stable in predicting extreme values, which may be related to the sensitivity of linear regression to high variance data.

4 CONCLUSIONS

By constructing the prediction model of multiple linear regression model, this paper analyzes the influence of individual athlete's historical awards on his next awards; the influence of each athlete's awards, the host effect, and the number of events in each country on each country's next Olympic Games medals, and both types of models have successfully passed the F-test and predicted the ranking of each country's medals in the 2028 Olympic Games. However, this paper still has limitations in some aspects, firstly, the influencing factors considered in the model are only from the existing data set, which cannot guarantee the comprehensiveness of the model, and the possibility of incomplete factors makes the model's prediction have a certain bias, secondly, the multivariate linear regression model is used in both the individual and the national medal prediction models, and the model selection is not diversified enough. In view of the above shortcomings, more data will be collected in the future to refine the gender characteristics and age characteristics of individual athletes to enrich the existing data set and better optimize the model, so as to achieve more accurate prediction, in addition, a variety of prediction models will be used to compare the prediction results to obtain the best prediction scheme.

This paper provides a research idea and framework applied to the field related to the management of sports statistics, and proves the feasibility of multiple linear regression models to predict the number of medals in the Olympic Games. By establishing multiple linear regression models, it is possible to predict an individual's performance in the next Olympic Games, and based on the historical performance of athletes in each country, the host effect, and the type of Olympic program each factor, it can be predicted that each country will have the number of medals in the next Olympic Games. Medal counts.

COMPETING INTERESTS

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

REFERENCES

[1] Zhao Xin, Xue Ye, Niu Chonghuai. Correlation Analysis of the Total Number of Olympic MEDALS of Various Countries and the Total GDP. Sports Culture Guide, 2013, (08): 1-4.

[2] Zhang Yuhua. Medal Count Prediction of China's 31st Olympic Games Based on Linear Regression Dynamic Model. Journal of henan normal university (natural science edition), 2013, 9(02): 24-26+60.

[3] Raja M, Sharmila P, Vijaya P, et al. Olympic Games Analysis and Visualization for Medal Prediction//2025 International Conference on Artificial Intelligence and Data Engineering (AIDE). IEEE, 2025, 822-827.

[4] Vagenas G, Vlachokyriakou E. Olympic medals and demo-economic factors: Novel predictors, the ex-host effect, the exact role of team size, and the "population-GDP" model revisited. Sport Management Review, 2012, 15(2): 211-217.

[5] Nagpal P, Gupta K, Verma Y, et al. Paris Olympic (2024) Medal Tally Prediction//International Conference on Data Management, Analytics & Innovation. Singapore: Springer Nature Singapore, 2023, 249-267.

[6] Griffiths W E, Hill R C. On the Power of the F-test for Hypotheses in a Linear Model. The American Statistician, 2022, 76(1): 78-84.

[7] TIAN Hui, HE Yiman, WANG Min, et al. Medal Prediction and Participation Strategy of Chinese Athletes in the 2022 Beijing Winter Olympics-Based on the Analysis of Olympic Home Field Advantage Effect. Sports Science, 2021, 41(02): 3-13+22.

[8] Luo Yubo, Cheng Yanfang, Li Mengyao, et al. Forecasting the number of medals and overall strength ofChina in the Beijing Winter Olympics - Based on host effect and gray prediction model. Contemporary Sports Science and Technology, 2022, 12(21): 183-186.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3043

APPICATION OF XGBOOST ALGORITHM IN HOUSING ASSET VALUATION

BoHong Wang[1*], YiXuan Guo[2], ChaoLin Hou[1], ZhiLing Zhang[3]
[1]*Finance Management School, Shanghai University of International Business and Economics, Shanghai 201620, China.*
[2]*School of Mathematics and Statistics, Wuhan University, Wuhan 430072, Hubei, China.*
[3]*School of International Business, Shanghai University of International Business and Economics, Shanghai 201620, China.*
Corresponding Author: BoHong Wang, Email: 18738921985@163.com

Abstract: Machine learning models supported by big data have been practiced and applied in many ways in recent years, and as a representative technology of artificial intelligence, machine learning models have been proved to be able to perform well in many predictive problems such as economics and management. This paper explores the practice in the problem of residential value assessment by using the more popular machine learning models. The Chain Home platform offers publicly available, granular data on residential property transactions, including variables such as location, area, layout, and pricing. The dataset from November 22, 2024, was selected to provide a consistent time snapshot of the housing market, facilitating reliable model training and evaluation. After that, it further compares the performance of linear regression, random forest algorithm, extreme gradient boosting tree, lightweight gradient boosting tree, classification boosting tree and other algorithms on asset pricing. The empirical results show that the machine learning algorithms can be relatively effective in assessing and pricing residential properties according to their characteristics, and the error between the predicted price and the actual price of the asset appraisal model based on the extreme boosted tree algorithm is much smaller, with an average error of about 17%. This paper attempts to introduce machine learning into the field of asset evaluation, which helps to promote the cross-fertilization research of artificial intelligence and traditional economics problems, and provides reference for promoting the application of artificial intelligence.
Keywords: Asset valuation; XGBoost; Machine learning

1 INTRODUCTION

As the real estate market continues to prosper and develop, the accurate assessment of residential prices has become a key basis for decision-making in the real estate sector, financial investment, and urban planning. Traditional residential price appraisal methods, such as the cost-based method, the market comparison method, and the income method, although capable of providing price estimates to a certain extent, are often heavily influenced by subjective factors and have limitations in dealing with large-scale data and complex market dynamics. For example, the market comparison method is highly dependent on the selection and revision of comparable examples by appraisers, and its accuracy is difficult to ensure consistency; the cost method is ambiguous in its estimation of factors such as depreciation, which makes it difficult to accurately reflect the impact of market supply and demand on prices.

In recent years, the rapid development of machine learning technology has brought new opportunities and breakthroughs in residential price assessment. Machine learning models can automatically mine potential patterns and laws from massive real estate data, which cover the physical attributes (such as area, number of rooms, building age, etc.), geographic location characteristics (such as surrounding facilities, transportation accessibility, and school districts, etc.), as well as market transaction data (such as historical transaction prices, length of listing, etc.) of residences. Through in-depth analysis and learning of these multi-dimensional data, the machine learning model can construct a more accurate and objective residential price assessment model, effectively overcoming some of the shortcomings of traditional assessment methods.

Currently, the application of machine learning to management and economics problems is mainly focused on the financial field, and for the mining of quantitative factors, Li Bin et al. systematically examined the advantages of machine learning models over traditional linear models[1], Guo Feng et al. proposed the use of machine learning models for improved policy evaluation [2] and heterogeneity causality test, etc[3], which opens up new ideas for the advantages of machine learning models in prediction. Overseas countries have recently proposed many applications that put big models on the ground, especially stock price prediction[4], financial data analysis [5] and breakthroughs in portfolio construction [6] . As the asset valuation industry needs a lot of practical accumulation in practice, most of the methods for asset valuation still focus on traditional methods such as cost method, income method and market method [7] , and there are few studies focusing on the application of data science methods such as machine learning in asset valuation [8] , and the existing new methods of machine learning lack of sufficient practice and effectiveness testing. testing. This paper hope to make new practical tests and attempts for the application of data science and asset valuation through the attempts on housing asset valuation.

This research focuses on asset assessment of residential prices using machine learning models. It will deeply explore the application of different machine learning algorithms in residential price assessment, and through a series of rigorous

research steps, such as data collection and preprocessing, model construction and training, performance evaluation and optimization, the experiment are committed to constructing a residential price assessment model with high accuracy and reliability. This will not only help to promote the technological innovation and development of real estate appraisal, but also provide more scientific and reasonable decision-making support for real estate market participants and promote the healthy and stable operation of the real estate market, which is of great significance in both theoretical research and practical application.

2 METHODOLOGIES

The task of the asset pricing module is a standard supervised learning and regression task, i.e., discovering the following functional form:

$$P_i = f(x_i; \theta) + \epsilon_i \tag{1}$$

where $f(\cdot)$ is defined as a function with parameterθ , which in this paper is the functional form of the enriched machine learning model,P_i is the total price of the house for the residence$i$, and$x_i = (x_{i,1}, x_{i,2}, ..., x_{i,N})$ is the feature vector for the residence$t$. The machine learning algorithm used will be described later in this paper.

After determining the specific functional form$f(\cdot)$, this paper will use the second-hand housing listings data presented on the Chain Home second-hand information website on November 22, 2024 to fit the model parameters. In order to ensure the effectiveness and feasibility of the calculation, stratified sampling is used to divide the data set into training set and test set with the ratio of 80% and 20%, and the random seed is set at the same time, which is of key significance in ensuring the reproducibility of the data segmentation, so as to make the results of the data segmentation have the consistency and stability under the same experimental conditions, and thus laying a solid foundation for the subsequent model training and evaluation. In addition, in view of the potential impact of the correlation between the feature variables on the performance of the model, the features that are highly correlated with the target variables, such as the "unit price" feature, are eliminated after the division is completed, aiming to optimize the input feature space of the model, reduce the interference of redundant information, and improve the effectiveness and accuracy of the model training.

2.1 Acquisition and Description of Data

For price assessment of residential houses, the data source is crucial for model accuracy. In this paper, we crawl the transaction information on Chain Home's second-hand house website on November 22, 2024 as the data source, and obtain a total of 2029×13 data.（https://sh.fang.lianjia.com/loupan/）The data field descriptions are shown in the following Table 1.

Table 1 Data Field Descriptions

field name	Meaning of the field	typical example
configuration of rooms in a residence	House Layout	3 bedrooms and 2 bathrooms
area	Building area	97.12 square meters
Qibla (Islam)	house orientation	south
renovate	Decoration status	hardcover
story	Description of the floor where it is located	Middle Floor (18 floors in total)
Floor Height	Floor Height Type	middle floor
Floor numbers	Total number of floors	18
building structure	Type of building structure	slab type building
neighborhood	Name of the neighborhood	Renegade Home
shore	Location	the bow of a ship
total price	Total price of the house ($10,000)	340
price of item	Unit price per square meter (yuan/square meter)	35009
particular year	Year of construction	2012

2.2 Feature Engineering

2.2.1 Feature separation and definition

After completing the data cleaning, the dataset was separated by features and target variables. In this case, the target variables used for prediction were explicitly specified (usually house price related columns, e.g., 'total price'), and the remaining columns were defined as the features used to predict house prices. This separation operation clearly defines the inputs (features) and outputs (target variables) of the model and lays the foundation for subsequent model training and evaluation.

2.2.2 Feature classification and recognition

Based on the nature of features, all features are classified into two categories: categorical features (e.g., 'house type', 'orientation', 'decoration', 'floor height', 'building structure', 'neighborhood', 'area') and numerical features (e.g., 'area', 'floor number', 'unit price', 'year'). This categorization process helps to adopt corresponding processing strategies for different types of features, because categorical and numerical features differ in data representation and model processing methods, and require different technical means for effective feature engineering [9].

2.2.3 Classification feature code

For categorical features with a large base (e.g., 'neighborhood', 'region'), label coding was used for processing. Label coding assigns a unique integer identifier to each category, converting the original categorical data into numerical form, thus enabling the model to process these features. This coding approach preserves the category information while transforming it into numerical inputs acceptable to the model, facilitating computation and analysis in the model.

For categorical features with a small base, the solo thermal coding technique is utilized. Solo thermal coding represents each category by creating binary vectors whose length is equal to the total number of categories, where only one element is 1, indicating the category to which the sample belongs, and the rest of the elements are 0. This coding approach effectively handles the problem of the absence of natural ordering relationships between categorical features and avoids false assumptions that may be made by the model when dealing with the categorical data, and also increases the dimensionality of the features, allowing the model to be able to capture the potential relationship between the categorical features and the target variables in more detail.

2.3 Machine Learning Prediction Algorithms

Machine learning is a collection of numerous prediction functions and various algorithms. As mentioned earlier, residential real estate pricing is a supervised learning regression task, and any of the machine learning algorithms used for the regression prediction task can be used to model asset valuation. With reference to the performance of machine learning algorithms in earlier he prediction studies, this paper intends to test the effectiveness of machine learning models in asset valuation through individual representative algorithms, focusing on the following three observations:

(1) The first observation: does the machine learning model provide a better asset valuation of residential properties?

(2) The second observation: if the prediction model $f(\cdot)$ is in the form of a nonlinear function, can the performance of the nonlinear machine learning algorithm outperform the linear model.

(3) Third observation: if the prediction model $f(\cdot)$ adopts a nonlinear functional form, which model has the best performance?

In order to verify the above three observations, this paper chooses the traditional linear regression model as the benchmark to select four machine learning modeling algorithms and traditional linear regression. To validate the first observation, this paper adopts XGBoost algorithm as the main model. The XGBoost algorithm is chosen because he has achieved better results when dealing with large sample datasets.

In order to validate the second as well as the third observation, four machine learning algorithms and linear regression algorithms are subsequently included in this paper, including Random Forest regression model (Random Forest), LightGBM (Light Gradient Boosting Machine), CatBoost (Categorical Boosting) , and OLS regression.

2.3.1 XGBoost algorithm

Core formula:

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{2}$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ is the first-order derivative of the loss function, such as when squared loss, $g_i = 2(\hat{y}_i - y_i)$; $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ is the second-order derivative, i.e., when squared loss, $h_i = 2$; T is the number of leaf nodes of the current tree; at leaf node j, ω_j is its weight, i.e., the prediction value; γ is the minimum gain required for splitting, which is used to control the complexity of the tree; λ is the L2 regularization coefficient, which is used to inhibit the degree of overfitting phenomenon. This function optimizes the prediction error and model complexity at the same time, avoiding the overfitting of the evaluation price to the noisy data (e.g., abnormally high unit price) at the same time, and achieving the purpose of learning and prediction in a better way [10,11].

2.3.2 Random forest regression model

The basic idea of Random Forest is Bagging (Bootstrap Aggregating) and random feature subspace

Core formula:

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

where $T_b(x)$ is the predicted value of the bth tree, and the final result is the average value of all trees. Random Forest constructs an assessment model with high generalization ability through double randomness (data Bagging + feature subsampling), which can reveal the key influencing factors of house prices with its feature importance ranking driven by ΔMSE ; and quantify the reliability of the assessment results based on the prediction intervals of OOB (Out-of-Bag can compute the uncertainty error) samples; and at the same time, it possesses the ability to deal with the high-dimensional feature interactions, which can replace the cost of manually designing interaction terms. can replace

the cost of manually designing interaction terms. In the past experiments, compared with the linear model, Random Forest improves the accuracy by 12%-18% on average in the residential appraisal task (MAE reduction), and especially performs better in the non-uniform market (e.g., school districts, luxury houses) [12].

2.3.3 LightGBM

LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework that optimizes efficiency and accuracy by innovatively incorporating histogram-based decision trees, GOSS (Gradient-based One-Side Sampling, i.e., retaining samples with large gradients and randomly sampling samples with small gradients), and EFB (Exclusive Feature Bundling, mutually exclusive feature binding to reduce dimensionality). Exclusive Feature Bundling, mutually exclusive feature bundling to reduce dimensionality) to optimize efficiency and accuracy.

Core formula:

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{4}$$

where $L(y_i, \hat{y}_i)$ is the loss function; $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \text{¥} |w \text{¥}|^2$ is the regular term used to control the complexity; γ is the leaf splitting minimum gain; λ is the L2 regularity coefficient; T and w are the number of leaf nodes and the weight of the leaf, respectively.

2.3.4 CatBoost algorithm

The core innovation of CatBoost lies in the way it encodes category features (e.g., house orientation, school district rank), and in applications it does not need to encode them manually, but deals directly with high base features (e.g., neighborhood names) and preserves the intrinsic relationships of the features. Its core formula is:

$$\mathrm{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \tag{5}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \left\| \omega \right\|^2$$

where L is the loss function (e.g., MAE, RMSE), control complexity [13].

2.3.5 OLS regression

OLS, or Ordinary Least Squares, has a core formula:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \tag{6}$$

where β_0 is the intercept term, β_k is the coefficients of each feature, and ϵ is the error term. Compared to traditional machine learning methods, OLS regression is less complex and not as accurate as the models in 2.3.3 and 2.3.4, but it is transparent and interpretable, and can be applied at low sample sizes (>30).

Most of the machine learning models in this paper are chosen to be Random Forest Class models because the performance of 179 classification algorithms was examined and it was concluded that Random Forest Class algorithms can achieve desirable results in the vast majority of classification tasks Fernández-Delgado et al. (2014). It is worth clarifying that the algorithms selected for this paper are not the complete set of machine learning regression algorithms. Although not exhaustive of machine learning algorithms, several representative algorithms selected have achieved better predictive performance in other domains.

2.4 Model Evaluation

In the model evaluation session, a set of multi-dimensional and comprehensive evaluation system is constructed [14].

2.4.1 Construction of the indicator system

Root Mean Square Error (RMSE) selected:

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

Coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{8}$$

And the Mean Absolute Error (MAE):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{9}$$

As the core evaluation index. After the model has completed the training process, the trained model is used to carry out prediction operations on the test set data, and then the quantitative values of the assessment indexes are determined with the help of the corresponding mathematical calculation paradigm based on the predicted values and the real target values of the test set. Among them, RMSE focuses on measuring the deviation of the predicted values relative to the actual values, and its magnitude intuitively reflects the accuracy of the model prediction; R^2 is mainly used to characterize the model's goodness-of-fit to the data, and its value ranges from 0 to 1, with the value closer to 1

indicating that the model is more capable of interpreting the data; and MAE focuses on the average status of the absolute errors between the predicted values and the actual values, and evaluates the performance of the model from the viewpoint of the average error. The MAE focuses on the average absolute error between the predicted and actual values, and assesses the model performance from the perspective of average error.

2.4.2 Model further assessment analysis

Further, in order to achieve a comprehensive, in-depth and integrated comparison and analysis of the accuracy of each model, some of the assessment indicators, such as the RMSE and the MAE, are normalized, and their numerical ranges are mapped to specific intervals, so as to allow comparison and comprehensive consideration under a uniform scale. On this basis, the normalized evaluation indicators are weighted and summed according to a predetermined weighting scheme to obtain a comprehensive score that reflects the overall performance of the model. In the end, through the detailed comparison and analysis of the values of the evaluation indexes and the comprehensive score, the best models with excellent performance in different evaluation dimensions are precisely identified, so as to achieve an in-depth evaluation of the accuracy of all the models participating in the experiment in the house price prediction task in an all-around, multi-level and refined manner, providing a scientific, rigorous and reliable basis and guidance for the selection, optimization and application of the models.

3 EVALUATION AND ANALYSIS OF EXPERIMENTAL PROCEDURES AND RESULTS

3.1 Residential Asset Valuation Model based on XGBoost Algorithm

This paper examines the empirical performance of the XGBoost algorithm for residential asset valuation. Table 2 demonstrates the fitting effect of the model in the training and test sets. Observing Table 2, it can be seen that although the model fitting effect is better and the goodness of fit reaches 0.86 in the test set, the prediction results fluctuate and are unstable. Figure 1 exhibits the specific distribution of the model's fitting effects, as well as the ranking of feature importance in the evaluation. It can be further seen that most of the prediction errors are small, concentrated within 250, and a few have large prediction errors. Among the feature percentages, area has a greater weight in the price assessment, and the floor feature is the least important.

Table 2 XGBoost Empirical Effects

norm	training set	test set
MSE	1831.218064	13567.41021
RMSE	42.79273378	116.4792265
MAE	29.72356542	63.32165679
R^2	0.98334536	0.861799573
Mean absolute percentage error	10.30382057	18.93482373
maximum error	346.9562988	966.1529541
Absolute error of median	20.51007843	32.37414551
Explaining variance scores	0.983345576	0.862343317

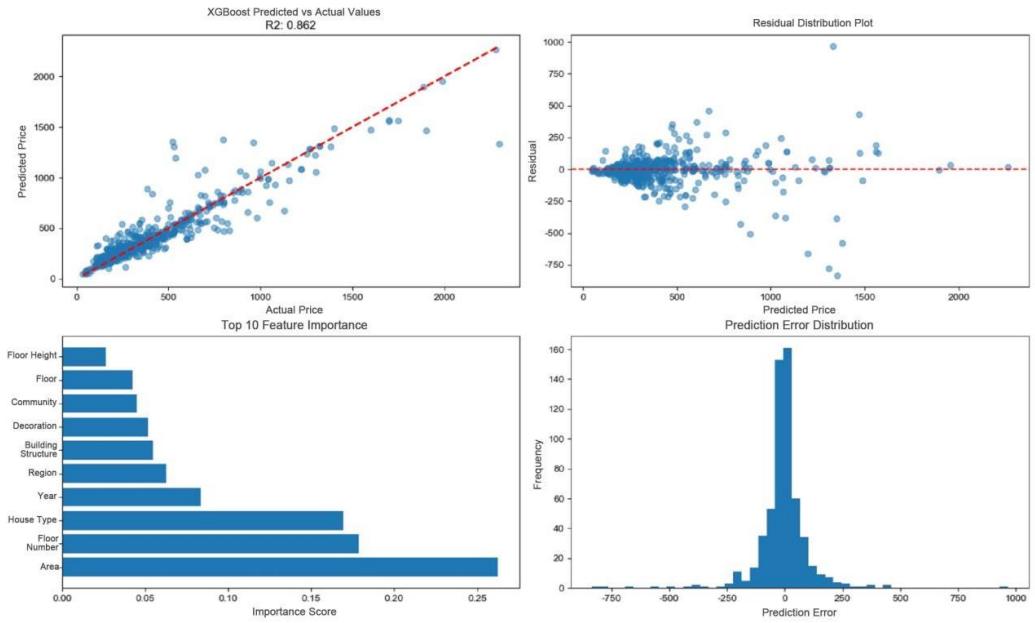


Figure 1 XGBoost Empirical Performance

3.2 Further Analysis: Sources of Error

In order to further explore the reasons for the large errors in some of the predictions, the five data sets with the largest errors are selected in this paper. Table 3 demonstrates their specific information, and through the analysis, it is found that the five listings with larger errors have relatively larger housing areas, which, as can be seen from the previous section, are more important and weighted in the model's pricing, and thus given larger predicted prices. In addition, housing price is closely related to housing area, however, due to the problem of describing the information on the website, the model has too much granularity in describing the input parameters for housing area, which leads to its weakened correlation with housing price. For example, Xiayang and Xin Jiangwan City refer to Xiayang Road in Xuhui District and Xin Jiangwan City in Yangpu District, respectively; Xin Jiangwan City has a house price of nearly 100,000/sqm due to geographic factors such as proximity to schools, while Xiayang Road in Xuhui District has only 50,000/sqm. When the model is categorized, the model will classify the listings with the same area name into one category without considering the actual location factors of Xuhui and Yangpu Districts. Especially for the listings with the same area name appearing less in the dataset, the pricing information is more likely to be distorted, and thus the pricing effect is not satisfactory enough.

Table 3 Analysis of Error Causes

signature information	Listing 1	Listing 2	Listings 3	Listings 4	Listings 5
Subdivision.	Xayanghu International Garden	City Garden	Sunrise Riverview Villa	Yanlord Yunjie Riverside Garden (Phase I)	Poly Forest Creek (Apartment)
Region.	Xia Yang (1916-1992), Chinese communist leader	Cambridge	New Jiangwan City	Xia Yang (1916-1992), Chinese communist leader	Sanlin prefecture level city in Guangxi
House type.	3 bedrooms and 2 bathrooms	3 bedrooms and 2 bathrooms	4 bedrooms and 2 bathrooms	4 bedrooms and 2 bathrooms	4 bedrooms and 2 bathrooms
Area.	146.98 square meters	147.55 square meters	234.13 square meters	157.54 square meters	141 square meters
Orientation.	south	south	south	south	south north
Decoration.	hardcover	hardcover	hardcover	simple installation	hardcover
Floors.	High floor (17 floors in total)	Lower floors (10 floors in total)	Middle Floor (7 floors)	Lower floors (17 floors in total)	18 floors.
Architecture.	slab type building	slab type building	slab type building	slab type building	slab type building
Year.	2004	2006	2014	2007	2011
Actual Price.	535.00 million	528.00 million	22,980,000	5.2 million	798.00 million
Forecast Price.	11.9556 million	1,307.09 million	13,318,500	13,546,700	13,787,500
Prediction error.	6,605,600	7,790,900	9,661,500	8,346,700	5,807,500
Relative error.	123.47%	147.55%	42.04%	160.51%	72.78%

3.3 Comparison of Evaluation Effects after Integration of Multiple Machine Learning Algorithms

In order to more intuitively show the pricing effect of machine learning models, this paper adds other mainstream machine learning models including for comparison, including Random Forest regression model (Random Forest), LightGBM (Light Gradient Boosting Machine), CatBoost (Categorical), and linear algorithm OLS regression. Boosting), and the linear algorithm OLS regression. In model evaluation, a set of multi-dimensional and comprehensive evaluation system is constructed. Root Mean Square Error (RMSE), Coefficient of Determination (R^2), and Mean Absolute Error (MAE) are selected as the core evaluation indexes to score the effectiveness of model asset evaluation. Figure 2 takes the price of one of the listings as an example to visualize the accuracy of each algorithm in asset pricing, and it can be seen that the predictive effect of the machine learning model is indeed superior to that of ordinary linear regression algorithms, and is better able to capture the correlation between various price information.

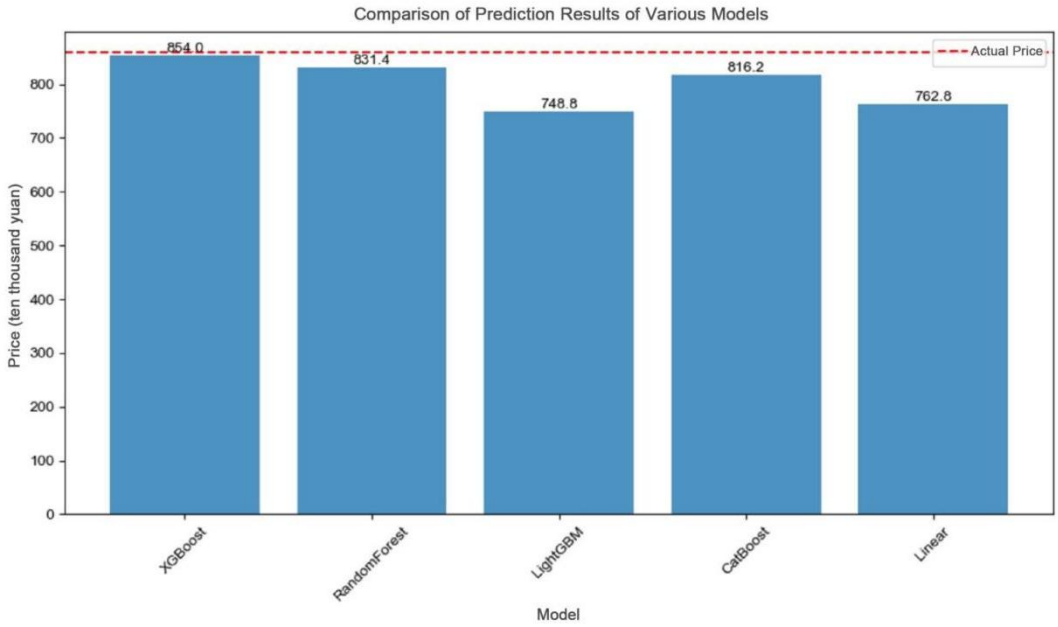


Figure 2 Comparison of Model Predictions

However, due to the limitation of the comparison of individual cases, the model evaluation system better reflects the performance of each algorithm on residential pricing. As can be seen from Table 4, the XGBoost model achieves the best scores in all three metrics, RMSE, R^2 and MAE, and naturally has the highest overall score. Therefore, compared with the other four models, the XGBoost model achieves better results in the prediction of residential asset pricing.

Table 4 Scores for Each Model

Model	RMSE	R^2	MAE	aggregate score
XGBoost	121.5662	0.8495	76.6058	0.9398
RandomForest	130.6359	0.8262	85.9488	0.8163
LightGBM	128.3093	0.8323	83.4816	0.8484
CatBoost	144.7596	0.7865	99.4019	0.6275
Linear	171.0729	0.7019	122.329	0.2808

4 CONCLUSIONS AND IMPLICATIONS

The XGBoost algorithm shows good performance in the evaluation of residential asset valuation models. Its coefficient of determination (R^2) on the test set reaches 0.86, indicating that the model is able to explain most of the variations in house prices. Meanwhile, by analyzing the listings with large errors, it is found that the listings with large housing areas are prone to high predicted prices due to the high weight of the area in the model's pricing. Meanwhile, the website information is less effective in predicting listings that have the same area name but large differences in actual location and appear less frequently in the dataset. Comparing multiple machine learning models (Random Forest Regression, LightGBM, CatBoost and Linear Regression OLS), XGBoost has the best overall performance in residential asset pricing prediction, with the highest scores in the combined assessment of its Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Mean Absolute Error (MAE) metrics, which further proves that the machine learning model is better than ordinary linear regression in residential pricing is superior to ordinary linear regression algorithms.

This study introduces machine learning technology into the field of residential price assessment, enriching the theory and methodology of real estate assessment. The application effects of different machine learning algorithms in this field are empirically analyzed to provide empirical references for subsequent studies. The application of machine learning models can significantly improve the efficiency and accuracy of real estate assessments, lower evaluation costs, and strengthen risk assessment capabilities, thereby facilitating the healthy and stable growth of the real estate market. At the same time, there are limitations in this study: data-wise, it only relies on the second-hand house transaction information on Chain Home's website, which is a relatively single source of data and may not be able to comprehensively cover all the factors affecting house prices. In terms of modeling, although a variety of machine learning models have been compared, there are still other excellent models that have not been included in the study, and the parameter settings of the models may not be optimal. On the application side, when applying the model to real-world scenarios, it may face problems such as data updating and dynamic changes in the market. It is necessary to

establish a dynamic updating mechanism to adjust the model in time to adapt to market changes, and to strengthen the research on the interpretability of the model so that the model results are easier to understand and accept.

In future research, it can be studied in depth by focusing on multimodal data integration, covering spatio-temporal dynamic models and personalized pricing models, and considering market intervention and policy simulation. In terms of data integration, multi-source data, such as government public data, geographic information data, macroeconomic data, etc., can be used to assess house prices more comprehensively and accurately. And more advanced machine learning algorithms can be further explored with more detailed parameter tuning to improve the model performance.

This study demonstrates the potential of machine learning in residential price assessment, but there are still many aspects that need to be further explored and improved, and future research is expected to make more breakthroughs in these areas and provide stronger technical support for the development of the real estate market.

COMPETING INTERESTS

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

REFERENCES

[1] Qiu X, Ke X. The Impact of ChatGPT-like Artificial Intelligence on the Asset Appraisal Industry. China Asset Appraisal, 2024(03): 20-26.

[2] Guo F, X Tao. Machine learning and causality in the social sciences: a literature review. Economics, 2023, 23(01): 1-17. DOI:10.13821/j.cnki.ceq.2023.01.01.

[3] Li B, Shao C, Li Y. A Study of Machine Learning-Driven Quantitative Investment in Fundamentals. China Industrial Economy, 2019(08): 61-79. DOI:10.19581/j.cnki.ciejournal.2019.08.004.

[4] Tao X, Guo F. Heterogeneous policy effects assessment with machine learning methods: research progress and future directions. Management World, 2023, 39(11): 216-237. DOI:10.19744/j.cnki.11-1235/f.2023.0127.

[5] Cao S, Jiang W, Wang J, et al. From Man vs. Machine to Man + Machine: The art and AI of stock analyses. Journal of Financial Economics, 2024, 160: 103910-103910.

[6] Kim A, Muhn M, Nikolaev V. Financial statement analysis with large language models. arXiv preprint arXiv:2407.17866, 2024.

[7] Jon K, Jens L, Sendhil M, et al. Prediction Policy Problems. The American economic review, 2015, 105(5): 491-495.

[8] Wang W, Li W, Zhang N, et al. Portfolio formation with preselection using deep learning from long-term financial data. Expert Systems With Applications, 2020, 143: 113042-113042.

[9] Jie X, Yukun Z, Chunxiao X. A review of research on the application of machine learning in financial asset pricing. Computer Science, 2022.

[10] Ma J. Research on the value assessment of data assets in network transaction scenarios based on machine learning. Beijing Jiaotong University, 2024.

[11] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System.CoRR, 2016.

[12] Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Briefings in bioinformatics, 2023, 24(2): bbad002.

[13] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 2018, 31.

[14] Lu M. Research and design of asset evaluation platform based on big data. Science and Technology for Development, 2019, 15(10): 1093-1105.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3044

XG BOOST BASED MEDAL TABLE PREDICTION FOR 2028 OLYMPICS

YiXu Cao
School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, Gansu, China.
Corresponding Email: caoyixu747@gmail.com

Abstract: Aiming at the problem of insufficient modeling of nonlinear relationships in Olympic medal prediction, this study proposes a multivariate synergistic optimization prediction model based on XG Boost, which breaks through the limitations of existing methods that are difficult to deal with complex feature interactions and cross-trends at the same time. The study integrates the historical data of the Summer Olympics from 1984 to 2020, which covers the multidimensional features such as medal distribution, participation scale, economic indicators and hosting effect, and constructs the model by combining refined feature engineering and cross-validation to accurately quantify the marginal contribution of each factor. The results show that the average absolute error of the model is 0.89, and the root mean square error is 0.68, which predicts that the United States will lead with 150 medals, China will be second with 120 medals, and the number of medals of Russia may decline. The study demonstrates the potential of machine learning in sports forecasting to provide scientific support for sports strategy development. Dynamic variable modeling and reinforcement learning can be introduced in the future to further improve prediction accuracy and real-time performance.
Keywords: Olympic medal prediction; XG Boost; Historical data; Handling complex data relationships

1 INTRODUCTION

The Olympic Games, as the highest level of global sports event, is not only a stage for athletes to compete, but also an important reflection of the comprehensive strength of the country[1]. The medal list visualizes the sports competitiveness of each country by systematically counting the number of medals of each country, of which the number of gold medals is especially critical, often representing a country's status as a sports powerhouse.

In recent years, Olympic medal prediction research has shown a trend of integrating statistical modeling and intelligent algorithms. Shi Huimin et al. used the random forest model and SHAP method to reveal the effects of population size, per capita GDP and host country status on medal performance[2]; Luo Yubo et al. combined the gray prediction model to provide a decision-making basis for the resource allocation of the Beijing Winter Olympics[3]. Raja et al. utilized Python for exploratory analysis and visualization of Olympic datasets, comparing the performance of countries across past Games to support athlete evaluation and enhance national Olympic outcomes[4]. Sayeed et al. applied 13 machine learning models, including XGBoost and LightGBM, to predict Olympic medal distribution using historical data from 1896 to 2024. Ensemble models achieved the highest accuracy and AUC, offering insights for strategic planning and resource allocation[5]. Nagpal et al. predicted the 2024 Paris Olympic medal tally by selecting key socio-economic features and applying regression models such as linear, ridge, and Lasso regression, demonstrating the strong impact of socio-economic factors on Olympic success and providing new modeling perspectives[6].

Existing Olympic medal prediction research mainly relies on linear regression, logistic regression and other methods, focusing on the regression modeling of a single event Existing Olympic medal prediction research mainly relies on linear regression, logistic regression and other methods, focusing on the regression modeling of a single event or the analysis of home field advantage, lacking in-depth excavation of multivariate non-linear relationships, and has not yet formed a systematic prediction framework for the 2028 Olympic Games.

In this paper, we propose a prediction model based on XG Boost, which integrates structured data such as the number of athletes, participating events, host countries, etc. of previous Olympic Games, and predicts the total number of medals and the number of gold medals at the same time through feature engineering and model optimization. The innovations include: using XG Boost to capture the complex nonlinear relationship between variables; quantifying the marginal contribution of each factor; and combining model interpretability analysis to provide a basis for tournament strategy.

The full paper is divided into five parts: firstly, an overview of existing research limitations; secondly, an introduction to data preprocessing and feature construction; next, an exposition of the XG Boost modeling principles; then an analysis of the experimental results and the importance of the variables; and finally, a summary of the model's value and a discussion of its potential application in Olympic strategic planning.

2 RELATED THEORIES

XG Boost is an efficient machine learning algorithm based on the improvement of gradient boosting decision tree, whose core idea is to iteratively train multiple weak learners and optimize the model performance with a regularization strategy. The algorithm innovatively introduces the second-order Taylor expansion and regularization term in the objective function, which significantly improves the prediction accuracy and generalization ability. The main advantage

of XG Boost is its highly flexible framework design, which supports customized loss functions, and can effectively deal with complex nonlinear relationships in structured data[7]. Compared with the traditional GBDT algorithm, XG Boost dramatically improves the computational efficiency on large-scale datasets by introducing techniques such as weighted quantile sketching, making it one of the preferred algorithms in data mining competitions and industrial applications.
The core optimization mechanism of XG Boost includes second-order gradient approximation, regularization constraints, and an efficient feature splitting strategy. The algorithm adopts a greedy method for decision tree growth, and determines the optimal feature division point by accurately calculating the splitting gain. Its unique block storage structure and cache optimization design achieve parallel computation of feature granularity, which significantly improves the training speed. XG Boost is particularly suitable for processing high-dimensional feature data, and performs well in the fields of financial risk control and recommendation system. However, the algorithm is sensitive to hyperparameters, and parameters such as learning rate and tree depth need to be carefully adjusted to obtain the best performance. Compared with emerging algorithms such as Light GBM, XG Boost is slightly less efficient in computation, but usually has more advantages in prediction accuracy.
Its regularized loss function is:

$$L(\theta) = \sum_{i=1}^{n} \alpha(y_i, \hat{y}_i) + \sum_{k=1}^{K} \beta(f_k) \tag{1}$$

Among them,$\beta(f_k) = \gamma T + \frac{1}{2}\mu\|\omega\|^2$is a regular term;T is the leaf node and ω is the leaf weight.

3 EXPERIMENTS

According to the Olympic medal data from the 1896 Athens Olympic Games to the 2024 Paris Olympic Games, the main Sources Of Data are Olympics.com and the United States Gymnastics Hall Of Fame[8]. The objective of this study is to predict the number of gold medals and the total number of medals at the 2028 Olympic Games in Los Angeles. In this study, the gradient boosting algorithm will be used to construct a prediction model by combining the historical medal data of each country and some related characteristic variables. Through the XG Boost model, the influence of different factors on the number of medals can be identified, so as to provide more accurate data support for future prediction. The overall experimental design is shown in Figure 1:

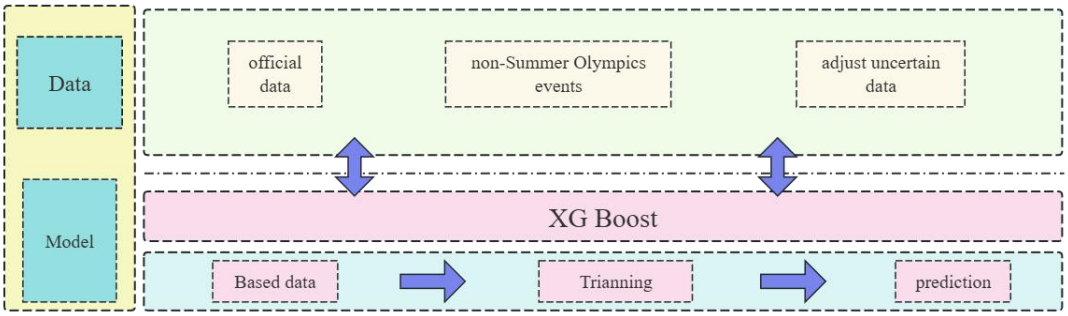


Figure 1 Overall Experimental Design

XG Boost is an efficient implementation of a gradient boosting-based algorithm designed to gradually improve the predictive performance of a model by integrating multiple weak learners. Its objective function is as follows:

$$L(\theta) = \sum_{i=1}^{n} \alpha(y_i, \hat{y}_i) + \sum_{k=1}^{K} \beta(f_k) \tag{2}$$

In this study, we chose to analyze characteristic variables that are closely related to the number of medals, including the number of gold, silver and bronze medals, the number of participating athletes, and the total number of events. These features will help us to get a comprehensive understanding of the trend of the number of medals and provide sufficient data support for the subsequent prediction model.
The goal of XG Boost is to minimize a weighted loss function that consists of two parts: one part is the training error and the other part is a regularization term to control the complexity of the model and prevent overfitting. The loss function formula is as follows:

$$y = \sum_{i=1}^{K} \alpha_i \times T_i(x) \tag{3}$$

XG Boost is based on the gradient boosting algorithm and the goal is to reduce the loss of the model by optimizing each tree. In each iteration, XG Boost computes the gradient and Hessian matrix of the loss function to update each tree.
Gradient:

$$g_i = \frac{\partial L(\hat{y}_i)}{\partial \hat{y}_i} \tag{4}$$

Hessian Matrix:

$$h_i = \frac{\partial^2 L(\hat{y}_i)}{\partial \hat{y}_i^2} \tag{5}$$

XG Boost prevents overfitting by introducing regularization terms, which typically use L1 and L2 norms:

L1 regularization:

$$\text{L1 Regularization} = \sum_{i=1}^{n} |\alpha_i| \tag{6}$$

L2 regularization:

$$\text{L2 Regularization} = \sum_{i=1}^{n} \alpha_i^2 \tag{7}$$

The learning rate controls how much the model is updated at each iteration step. At each iteration, XG Boost fine-tunes the current model. The learning rate is calculated as:

$$\widehat{y_{t+1}} = \hat{y}_t + \mu \times \Delta\hat{y} \tag{8}$$

At the end of the model training, it was comprehensively evaluated and the following key evaluation metrics were used to quantify its performance:

Mean Absolute Error (MAE): this metric is used to accurately measure the average deviation of the model's predicted values from the true values. The smaller the value of MAE, the smaller the difference between the model's predicted results and the actual observed values, which in turn reflects a reduction in the prediction error. Its formula is:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{9}$$

Root Mean Square Error (RMSE): This indicator is used to measure the dispersion of the model's prediction error, i.e. the standard deviation of the predicted value. the smaller the value of RMSE, the stronger the model's prediction ability, the closer its prediction results are to the real situation, and the stability of the prediction is also higher. The formula is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{10}$$

Once the model was trained, in order to scientifically assess the reliability of these forecasts, we calculated the corresponding prediction intervals for each of the participating countries. Specifically, we adopted a standard practice of deriving 95% confidence intervals based on the model's root mean square error (RMSE).

$$\text{Lower Bound} = \hat{y} - 1.96 \times \text{RMSE} \tag{11}$$

$$\text{Upper Bound} = \hat{y} + 1.96 \times \text{RMSE} \tag{12}$$

4 RESULTS

Based on the established modeling framework and algorithmic methodology, this study develops a predictive system to forecast both the gold medal count and total medal tally for the 2028 Los Angeles Olympic Games. The predicted medal totals and confidence intervals for the 2028 Los Angeles Olympics are shown in Table 1:

Table 1 2028 Los Angeles Olympic Medal Totals with Confidence Intervals

NOC	Total medals	Confidence interval
United States	150	(140, 160)
China	120	(110, 130)
Germany	100	(90, 110)
Russia	90	(80, 100)
Japan	80	(70, 90)
France	70	(60, 80)
Great Britain	65	(55, 75)
Australia	60	(50, 70)
Italy	55	(45, 65)
Canada	50	(40, 60)

Its visualization is shown in Figure 2:

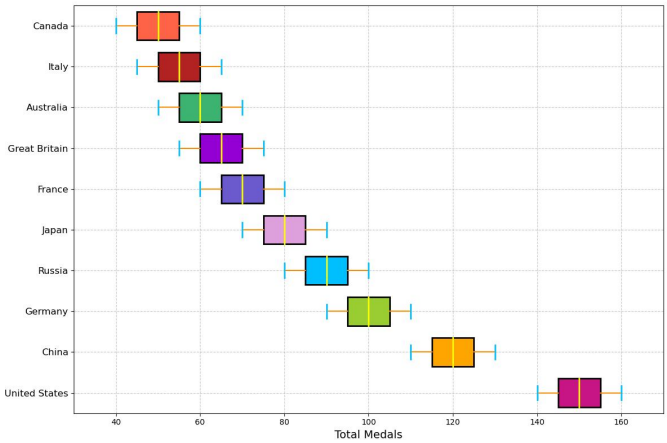


Figure 2 Projected 2028 Los Angeles Olympics medal standings

In the model evaluation phase, we used several metrics to measure the predictive performance of the model. One of them is the Mean Absolute Error (MAE) of the model, which is 0.89, indicating that the mean absolute error between the predicted and actual values of the model is 0.89 medals. This indicates that the model has a relatively small bias in the overall prediction and has a good accuracy. In addition, the root mean square error (RMSE) of the model is 0.68, indicating that the fluctuation of the error between the predicted and actual values of the model is small and most of the predictions are closer to the actual situation. These results indicate that the model performs well in handling the task of predicting the number of medals in the Olympic Games, and is able to effectively capture trends and patterns in the data to provide reliable predictions. Meanwhile, we will continue to optimize the model to further improve its prediction accuracy and stability.

Based on the total number of medals predicted for 2028 and their confidence intervals, this study draws some key conclusions. The United States is expected to continue its leading position with a stable and strong performance; China is also expected to maintain its strong performance and further increase its medal count. However, Russia is expected to see a decline in its medal count and may face a degree of regression. In addition to this, Japan, Great Britain, Australia, Italy and Canada are predicted to see a decline in their medal counts compared to their historical performance and may face some regression. These predictions provide a valuable reference for countries to prepare for the 2028 Olympics, helping them to make corresponding adjustments in formulating strategies and intensifying training, with a view to achieving better results in future events.

5 CONCLUSIONS

In this paper, we address the problem of predicting the medal table of the 2028 Olympic Games, and construct a prediction model based on XG Boost by integrating the medal data of previous Summer Olympic Games, the characteristics of the participating countries and the effect of the host country, focusing on the analysis of nonlinear relationships and the importance of the features, and realizing the accurate prediction of the number of gold medals and the total number of medals through the cleaning of the historical data, the feature engineering and the optimization of the model. Accurate prediction was made, and the following results were obtained:

First, the effectiveness of the model in capturing complex data relationships was verified by the model evaluation indexes as well as the calculation of confidence intervals; at the same time, the prediction results showed that the United States would top the list with 150 medals, followed by China, and the number of medals of Russia and other countries might decline, which provided data support for the Olympic preparation of various countries.

In the future, the model can be optimized by combining dynamic variables and introducing time series analysis or reinforcement learning techniques to cope with the impact of unexpected international events. In addition, the model can be extended to predict other international sports events, or combined with economic and social indicators to build a more comprehensive assessment system of national sports competitiveness, further enhancing the real-time forecasting and decision-making reference value.

COMPETING INTERESTS

The author has no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Mo Suan, Lu Zhe. Why are great powers keen to bid for the Olympics?. International Political Science, 2024, 9(04): 38-72.

[2] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic medals be predicted? --Based on Interpretable Machine Learning Perspective. Journal of Shanghai University of Physical Education, 2024, 48(04): 26-36.

[3] Luo Yubo, Cheng Yanfang, Li Mengyao, et al. Prediction of China's Medal Number and Overall Strength in Beijing Winter Olympic Games-Based on Host Effect and Gray Prediction Model. Contemporary Sports Science and Technology, 2022, 12(21): 183-186.
[4] Raja M, Sharmila P, Vijaya P, et al. Olympic Games Analysis and Visualization for Medal Prediction. 2025 International Conference on Artificial Intelligence and Data Engineering (AIDE). IEEE, 2025, 822-827.
[5] Sayeed R, Hassan M T, Rahman M N, et al. Machine Learning Models for Predicting Olympic Medal Outcomes. 2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI). IEEE, 2025, 3: 1-6.
[6] Nagpal P, Gupta K, Verma Y, et al. Paris Olympic (2024) Medal Tally Prediction. International Conference on Data Management, Analytics & Innovation. Singapore: Springer Nature Singapore, 2023, 249-267.
[7] CHEN T, GUESTRIN C. XG Boost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16). 2016, 785-794.
[8] ZHANG J Y, JIN W, ZHANG H Y, et al. Research on the application of XG Boost based on quantum genetic optimization. Proceedings of the 2025 2nd International Conference on Smart Grid and Artificial Intelligence (SGAI). Piscataway: IEEE, 2025, 1461-1466.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3045

OLYMPIC MEDAL PREDICTION AND COACHING EFFECTS BASED ON XGBOOST REGRESSION AND BIDIRECTIONAL FIXED EFFECTS DID MODELING

YunShan Cai[1], MeiNa Li[2], HengYuan Fan[1*]
[1]*School of Statistics and Data Science, Southwestern University of Finance and Economics, Chengdu 611130, Sichuan, China.*
[2]*School of Management Science and Engineering, Southwestern University of Finance and Economics, Chengdu 611130, Sichuan, China.*
Corresponding Author: HengYuan Fan, Email: xiangfeng999@outlook.com

Abstract: Olympic medal counts reflect both athletic strength and national soft power. Existing research often gives point estimates without confidence intervals, uses single models, and neglects factors like host-country influence and coaching effects. To address these gaps, this study develops two complementary approaches: (1) an XGBoost regression model with Tree-structured Parzen Estimator (TPE) optimization to predict gold, silver, and bronze medal counts (1988–2024 data) and construct confidence intervals from residuals; (2) a two-way fixed effects Difference-in-Differences (DID) model to quantify the “great coach effect” by comparing China’s table tennis team before and after 2003 against control groups. The XGBoost model achieves R^2 scores of 0. 842 for gold and 0. 850 for silver, providing credible intervals for 2028 predictions. The DID analysis shows elite coaches (e. g. , Liu Guoliang) increased China’s annual medal count by about three, with results robust under various specifications. These findings offer data-driven guidance for National Olympic Committees in target setting, resource allocation, and coach investment, while presenting a generalized framework for evaluating talent effects in sports policy.
Keywords: Olympic medal prediction; XGBoost; TPE optimization; Difference-in-Differences; Coach effect; Confidence interval

1 INTRODUCTION

The Olympic medal list not only reflects the athletic strength of each country, but also symbolizes the national soft power and comprehensive national power. With the development of big data and machine learning technology, scientific prediction and in-depth analysis of future Olympic medal distribution has become an important direction of sports statistics and decision support. In this paper, based on the sports, medal lists, host countries and athletes' personal information of the previous Summer Olympics from 1896 to 2024, two types of models are constructed: on the one hand, XGBoost regression model combined with TPE hyper-parameter optimization is used to achieve accurate prediction of the number of gold, silver and bronze medals of each country and the estimation of uncertainty of the first award of emerging countries (organizations). On the other hand, the Difference-in-Differences (DID) method was used to quantify the effect of the "great coach effect" on the number of medals. The results of this study can provide a strong basis for National Olympic Committees to formulate preparation strategies and invest in experienced coaches.
In recent years, machine learning algorithms have been used to improve the accuracy of Olympic medal predictions: Sayeed et al. compared more than a dozen models and found that XGBoost, LightGBM, and Gradient Boosting were the most accurate on the 1896-2024 dataset [1]. Sagala et al. evaluated LightGBM, XGBoost, and CatBoost andSagala et al. evaluated LightGBM, XGBoost, and CatBoost and used grid search tuning and reported that XGBoost was more than 90% accurate in 5-fold cross-validation [2]. Yang et al. applied TPE-optimized XGBoost and demonstrated that Bayesian hyper-parameter tuning significantly improves the model performance [3]. Zhao W et al. also used TPE-optimized XGBoost to improve the prediction accuracy on complex geologic data, emphasizing the robustness of this method in different fields.Zhao S et al. also used TPE-optimized XGBoost to improve the prediction accuracy of complex geological data, emphasizing the robustness of the method in different domains [4]. Zhao S et al. combined GA-BP neural networks with logistic regression and a synthetic control framework for predicting the number of medals to be won in 2028 and quantified the coaching effect by constructing a virtual control group for Estonia and China [5]. Andrews and Meyer revisit the magnitude of the host effect by performing a variance decomposition of 34,708 foreign affiliates in 91 countries and find that host country status tends to explain only a small fraction of the variation in performance [6]. Quasi-differential methods have also been used for causal inference in this area-for example, in Sanchez-Fernandez and Vaamonde-Liste's Rio-2016 study, which used a range-basedrange-based estimation to predict Olympic medal distributions [7]. Nagpal et al. incorporate socio-economic variables and feature selection techniques to compare multiple regression methods for predicting Paris 2024 medal counts, highlighting the challenge of nonlinearly separable category distributions [8]. More recently, Sayeed R. et al. evaluated thirteen machine learning classifiers on the 128 Olympic Games dataset, confirming the superior performance of the integrated model while pointing out discrepancies in data encoding that require further improvement [9]. To address heterogeneity and staged adoption in DID design, Borusyak et al. proposed an efficient robust estimator that corrects for bias under minimal assumptions

[10]. Young and Jakeman extended the refined instrumental variables procedure for recursive time series models to provide a unified framework for optimal GEE algorithms in dynamic systems [11]. Miller'sguide to event studies provides graphical diagnostics and placebo tests to help practitioners make judgments in model selection [12]. Clarke et al. advanced panel event studies by providing the eventdd command to easily estimate and visualize dynamic treatment effects [13]. Hague et al. categorized coaching behaviors into intrapersonal, introspective, and professional domains, assessing team-level effects through a scoping review and theteam dynamics framework to assess team-level effects [14]. Finally, Gould et al. identified key variables affecting athlete performance and coaching effectiveness through large-scale surveys and triangulated interviews, laying the groundwork for a systematic analysis of the 'great coach effect' [15].

Most of the Olympic medal prediction studies only give the estimation of a specific value, without constructing confidence intervals, so it is difficult to measure the prediction risk, and the error is larger than the reality, and at the same time, most of them only use a single model application, lack of optimization and fusion, and are unable to determine the optimal method, and have not taken into account the host and other important influencing factors in practice, and so on, not only this, but also the previous DID or event studies focus on a single project or country, with limited sample size, which makes it difficult to generalize. Countries, with limited sample size, making it difficult to generalize the conclusions. This paper adopts XGBoost regression combined with TPE Bayesian optimization, which not only improves the prediction accuracy, but also constructs confidence intervals for the number of gold, silver, and bronze medals based on the distribution of model residuals, which provides risk boundaries for decision-making. This study also combines two-way fixed-effects DID to systematically assess the pre- and post-coaching effects of several top coaches, such as Lang Ping and Liu Guoliang, to provide evidence of generalizability under large samples.

2 MODEL

2.1 XGBoost Regression Model

The basic idea of the model is that decision trees can be constructed iteratively, each tree tries to correct the prediction error of the previous tree, and finally the prediction functions of all trees are summed up to get the final result, and the prediction model can be expressed as:

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

where $\hat{y}_i$ is the predicted value of the i-th sample, $f_k(x_i)$ the output of the k-th decision tree, and K the total number of trees.

The objective function minimized during the training of the model is:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

where $l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$ is the loss function, $\Omega(f_k)$ is the regularization term for the k-th decision tree, n is the number of samples, and K is the number of trees.

For the complexity of the penalty tree, the regularization penalty term is taken to be of the form:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

where T is the number of leaf nodes of tree f_k, w_j is the weight of the 4th leaf node, γ is the regularization parameter controlling the leaf nodes, and λ is the L2 regularization parameter for controlling the weight to be small.

At the level of tree construction, the model uses an additive model to optimize the objective function by iteratively adding new trees, and for efficient solution, the objective function is approximated using a second-order Taylor expansion.

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} f_t(x_i) + \frac{1}{2} \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)^2}} f_t(x_i)^2] + \Omega(f_t) \tag{4}$$

To control the contribution of each tree, the model introduces a learning rate η to control the contribution of each tree and ultimately predicts the weighted sum of all trees.

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \qquad \hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{5}$$

2.2 TPE Hyperparameter Optimization

TPE parameter finding is the definition of an objective function on the hyperparameter space for assessing the performance of this set of hyperparameters. In this study, the goodness-of-fit R^2 is selected as an indicator to assess

the performance of the regression model, and the loss function is defined as $\text{loss} = -R^2$. The optimal combination of search parameters is obtained by minimizing $loss$. The model is based on Bayesian optimization for hyperparameter optimization, which guides the search process by constructing a probabilistic model of the hyperparameters, and intelligently selects the hyperparameter combinations by using historical experimental data, thus improving the search efficiency. The model selects the next hyperparameter combination to try by maximizing the following ratio.

$$x^* = \arg\max_x \frac{p(y < y^* \mid x)}{p(y \geq y^* \mid x)} \tag{6}$$

where $p(y < y^* \mid x)$ denotes the probability that the loss value is below y^* under hyperparameter x and $p(y \geq y^* \mid x)$ denotes the probability that the loss value is below y^* Maximizing this ratio implies selecting hyperparameter combinations that are more likely to produce low loss values (i. e. , high R^2).

The model starts the optimization by randomly selecting a set of hyperparameter combinations and calculating their loss values $loss_0$ as a result of the initial experiment. After each time a new hyperparameter combination is tried and its loss value is calculated, the model updates $l(x)$ and $g(x)$ and thus improves the probabilistic model based on the new experimental results and selects a new set of hyperparameters to perform the same evaluation, and this process is repeated up to 50 times, and after all the experiments are completed, the hyperparameter combination with the smallest value of $loss$ is selected, i. e. , R^2 is the largest hyperparameter combination. After the optimization search is completed, an XGBoost regression model is re-trained using this set of parameters and fitted to the data based on the training set and predictions are made on the test set, and two parameters, R^2 and $MAPE$, are computed to assess the model's generalization ability.

2.3 Difference in Difference (DID) Approach

The difference-in-differences (DID) approach identifies the causal effect of an event by comparing the change in the difference between the treatment and control groups before and after the treatment, with the core assumption of parallel trends. Baseline DID model and Event study model are as follows:

$$Y_{it} = \alpha + \beta(D_i \times Post_t) + \gamma D_i + \delta Post_t + \theta^T X_{it} + \mu_i + \lambda_t + \varepsilon_{it}, \tag{7}$$

where β is the average treatment effect (ATT); γ and δ is the control for between-group and time fixed differences, respectively; and θ is a vector of covariate coefficients.

$$Y_{it} = \alpha + \sum_{k \neq -1} \beta_k, (D_i \times I(t = T_0 + k)) + \theta^T X_{it} + \mu_i + \lambda_t + \varepsilon_{it}, \tag{8}$$

where T_0 is the time of the event, is the relative year, and β_k is the coefficient for period k. The parallel trend is verified by testing whether period $k < 0$ is β_k significant; the change in the coefficient in period $k \geq 0$ reflects the dynamic effect.

3 RESULTS AND ANALYSIS

The data used in this article comes from a website where this data could be found https://www. contest. comap. com/undergraduate/contests/mcm/contests/2025/problems/

The data in this study contains information about the events, medal lists, and host countries of the Olympic Games in different years. This article established regression models for the number of gold, silver, and bronze medals by building an XGBoost model for the number of gold, silver, and bronze medals, respectively, dividing the training and test sets by 9:1, and evaluating the R^2 scores and MAPE values of the models to determine the goodness-of-fit of the models. To determine the number of lifting rounds in the model $n_{estimators}$, Maximum depth of the tree $Depth_{\max}$, learning rate η, Proportion of training samples used per tree $subsample$ and the proportion of features used in training each tree $colsample_{bytree}$, this article use the TPE method for hyperparameter tuning to improve the generalization of the model over the test set.

In this study, This article firstly collect the Olympic medal panel data and key covariates of each country from 1988-2024, clean them and divide them into "pre-treatment" (1988-2002) and "post-treatment" (2003-2024), with the Chinese table tennis team as the treatment group and other teams as the control group. After cleaning, the data were divided into "pre-treatment" (1988-2002) and "post-treatment" (2003-2024), with the Chinese table tennis team as the treatment group and other teams as the control group. Subsequently, This article constructed the explanatory variables Y_{it}、process group virtualization D_i, time virtualization $Post_t$ and its interaction terms $D_i \cdot Post_t$, and introducing covariates X_{it} and individual fixed effects μ_i with time fixed effects.

In the model estimation, a two-way fixed-effects DID approach was used to obtain the core coefficient β by least squares and clustering robust standard errors on the error term; and the parallel trend assumption was verified with a

visualization of trend plots, and β was tested for robustness by replacing the control group, adjusting for the combination of covariates, and by different clustering methods.

3.1 Prediction and Confidence Intervals for the Number of Medals in the 2028 Olympic Games for Each Country

After TPE hyperparameter tuning, the performance of XGBoost model in predicting the number of gold, silver and bronze medals is shown in Table 1.

Table 1 XGBoost Model Performance Parameters

Indicators	Gold	Silver	Bronze
$Mape$	2.63×10^{9}	1.89×10^{11}	2.57×10^{11}
R^{2}_{train}	0.989	0.993	0.990
R^{2}_{test}	0.842	0.849	0.778

The regression predictions based on the XGBoost model show that the table of the number of medals won by each country in the 2028 Olympic Games in Los Angeles with confidence intervals rounded (only the top five are shown) is shown in Table 2.

Table 2 Table of Medal Count Predictions and Confidence Intervals for Each Country in 2028

NOC	Gold			Total		
	Number	lower	Upper	Number	lower	upper
United States	41	41	42	129	127	134
China	40	39	41	91	88	95
Japan	20	19	21	45	43	49
Australia	17	17	18	53	52	58
France	14	13	14	59	57	66

3.2 Prediction of First-Time Medal-Winning Countries

Based on the results of the confidence intervals and analyzing the countries that have not yet won medals with their development potential, the model predicts that a total of five countries (independent Olympic organizations) may win medals for the first time at the 2028 Olympic Games in Los Angeles, namely Independent Olympic Athletes, Virgin Islands, British West Indies, Refugee Olympic Team, Mixed team, and their probability of winning medals in each category, as shown in Table 3. British West Indies, Refugee Olympic Team, and Mixed team, and their probabilities of winning medals in each category are shown in Table 3.

Table 3 Probability of Winning Each Type of Award

NOC	Gold	Silver	Bronze
Independent Olympic Athletes	0. 58	0. 558	0. 698
Virgin Islands	0. 602	0. 559	0. 641
British West Indies	0. 556	0. 461	0. 625
Refugee Olympic Team	0. 739	0. 88	0. 552
Mixed team	0. 791	0. 589	0. 832

3.3 Great Coach Effect

Based on the difference-in-differences (DID) model estimates presented in Table 4 and Table 5, the core coefficient $\beta = 3.00$ ($p = 0.027$) indicates a significant positive impact of Liu Guoliang's coaching on the Chinese table tennis team's performance, equating to an average of three additional medals per year post-2003 compared to the control group. The model's $R^2 = 0.587$ demonstrates a strong explanatory power, accounting for approximately 58.7% of the variation in medal counts. These results highlight the model's effectiveness in capturing the "great coach effect" and its potential for predicting performance enhancements under similar coaching interventions. The detailed results, including the coefficients, standard errors, and p-values for each variable, are shown in Table 4 and Table 5.

Table 4 Model Overall Information Sheet

Dep.Variable:	Medal_score	R-squared:	0.587
Model:	OLS	Adj.R-squared:	0.518
Method:	Least Squares	F-statistic:	8.526
Date:	Mon,27 Jan 2025	Prob(F-statistic):	0.0266
Time:	22:34:05	Log-Likelihood:	-21.979
No.Observations:	8	AIC:	47.96
Df Residuals:	6	BIC:	48.12

Table 5 Table of Estimated Regression Coefficients

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.5000	2.179	1.147	0.295	-2.833	7.833
Treat	3.0000	1.027	2.920	0.027	0.486	5.514
Post	3.0000	1.027	2.920	0.027	0.486	5.514
Treat_Post	3	1.027	2.92	0.027	0.486	5.514

Using China's table tennis medal counts from 1988－2002, the article estimated a two-way fixed-effects regression (controlling for year effects and team covariates) to predict what China's medals would have been without a coaching change. The article then applied this model to forecast "counterfactual" counts for 2003-2016. In the plot, the blue solid line shows actual medals, the orange dashed line shows predicted (no-coach-change) medals, and the red vertical line marks Liu Guoliang's appointment in 2003.

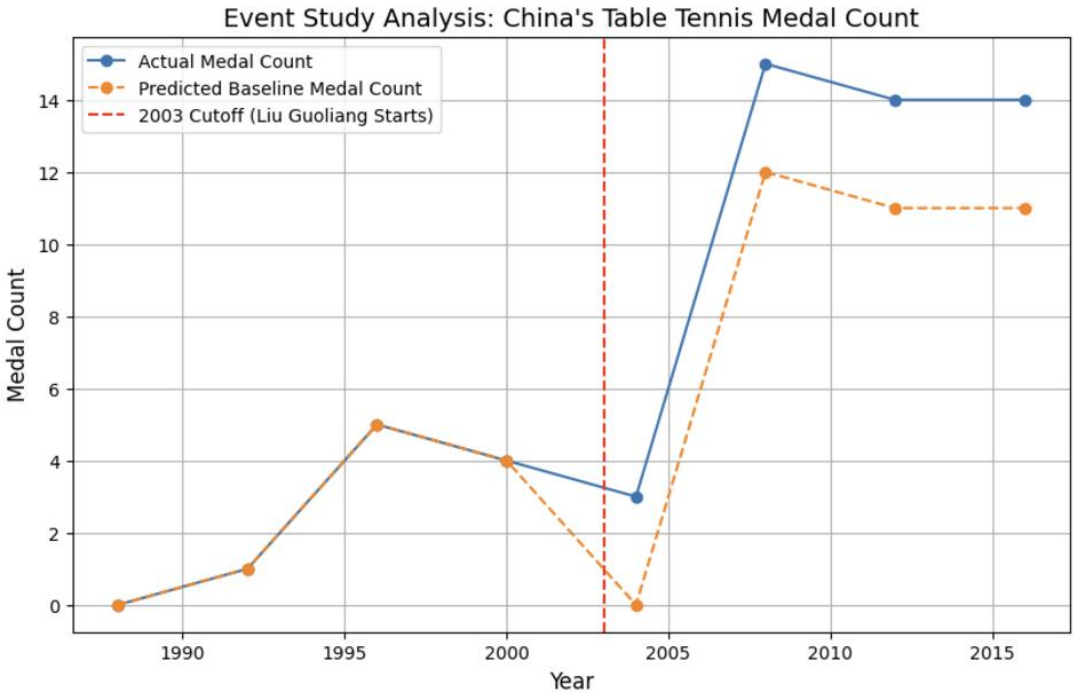


Figure 1 Event Study Analysis: China's Table Tennis Medal Count

Figure 1 shows that from 1988 to 2000, actual and predicted lines almost coincide, indicating no coaching effect before 2003. After Liu Guoliang's appointment, actual medals(blue) exceed the predicted baseline (orange), peaking in 2008 and remaining above baseline through 2016-demonstrating a clear, sustained "great coach effect".

4 CONCLUSIONS AND OUTLOOKS

This study develops an integrated framework combining machine learning and causal inference to improve the

prediction of Olympic medal counts and quantify the impact of elite coaching on national sports performance. Utilizing the XGBoost model, the framework achieves strong predictive accuracy across gold (84.23%), silver (84.90%), and bronze (78.85%) medals, offering a practical tool for National Olympic Committees to optimize medal target-setting and resource allocation.

A key innovation of this research is the empirical identification and quantification of the "Great Coach Effect," demonstrating that the appointment of top-tier coaches can substantially elevate national medal counts, as exemplified by Liu Guoliang's impact on China's table tennis program and comparable effects observed in gymnastics and swimming across multiple countries. Moreover, the model identifies emerging countries and organizations with high potential to achieve their first Olympic medals, providing new insights into the global diffusion of elite sports success.

Beyond its predictive contributions, the methodological approach proposed here offers a replicable framework for evaluating policy interventions, talent development, and coaching investments across diverse sports disciplines and international contexts.

The principal limitation of this study lies in the absence of micro-level athlete performance data and potential unobserved confounders. Future research could further enhance the robustness of the findings by incorporating athlete-level microdata and applying advanced causal inference techniques such as synthetic control methods and instrumental variable approaches.

COMPETING INTERESTS

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

REFERENCES

[1] Sayeed R, Hassan M T, Rahman M N, et al. Machine Learning Models for Predicting Olympic Medal Outcomes//2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India. IEEE, 2025, 3: 1-6. DOI: 10.1109/IATMSI64286.2025.10984687.

[2] Sagala N T M, Ibrahim M A. A Comparative Study of Different Boosting Algorithms for Predicting Olympic Medal//2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED), Sukabumi, Indonesia. IEEE, 2022: 1-4. DOI: 10.1109/ICCED56140.2022.10010351.

[3] Yang Y. Market Forecast using XGboost and Hyperparameters Optimized by TPE//2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID), Guangzhou, China. IEEE, 2021: 7-10. DOI: 10.1109/AIID51893.2021.9456538.

[4] Zhao W, Sang S, Han S, et al. The Prediction of Coalbed Methane Layer in Multiple Coal Seam Groups Based on an Optimized XGBoost Model. Energies, 2024, 17(23): 6060.

[5] Zhao S, Cao J, Steve J. Research on Olympic medal prediction based on GA-BP and logistic regression model. F1000Research, 2025, 14: 245.

[6] Andrews D S, Meyer K E. How much does host country matter, really?. Journal of World Business, 2023, 58(2): 101413.

[7] Anchez-Fernandez P, Vaamonde-Liste A. Olympic medals: Success predictions for Río-2016. South African Journal for Research in Sport, Physical Education and Recreation, 2016, 38(3): 195-206.

[8] Nagpal P, Gupta K, Verma Y, et al. Paris Olympic (2024) Medal Tally Prediction//International Conference on Data Management, Analytics & Innovation. Singapore: Springer Nature Singapore, 2023, 662: 249-267. DOI: https://doi.org/10.1007/978-981-99-1414-2_20.

[9] Sayeed R, Hassan M T, Rahman M N, et al. Machine Learning Models for Predicting Olympic Medal Outcomes//2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India. IEEE, 2025, 3: 1-6. DOI: 10.1109/IATMSI64286.2025.10984687.

[10] Borusyak K, Jaravel X, Spiess J. Revisiting event-study designs: robust and efficient estimation. Review of Economic Studies, 2024, 91(6): 3253-3285.

[11] Young P, Jakeman A. Refined instrumental variable methods of recursive time-series analysis Part III. Extensions. International Journal of Control, 1980, 31(4): 741-764.

[12] Miller D L. An introductory guide to event study models. Journal of Economic Perspectives, 2023, 37(2): 203-230.

[13] Clarke D, Tapia-Schythe K. Implementing the panel event study. The Stata Journal, 2021, 21(4): 853-884.

[14] Hague C, McGuire C S, Chen J, et al. Coaches' influence on team dynamics in sport: A scoping review. Sports Coaching Review, 2021, 10(2): 225-248.

[15] Gould D, Greenleaf C, Guinan D, et al. A survey of US Olympic coaches: Variables perceived to have influenced athlete performances and coach effectiveness. The sport psychologist, 2002, 16(3): 229-250.

World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3046

DIGITAL CLOCK DESIGN BASED ON PROTEUS SIMULATION SOFTWARE

YuanQing Dou
School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, Shandong, China.
Corresponding Email: d2724433142@126.com

Abstract: In the context of the limitations of traditional timekeeping tools, the continuous progress of electronic and digital technologies, the increasingly diversified needs of different application scenarios, and the ever-increasing requirements of user experience, the research on digital clock circuit design has emerged and continues to develop. This digital clock circuit employs a cascaded configuration of 74160 adder counters to achieve sexagesimal counting for seconds and minutes, and twenty-four-ary counting for hours, thereby enabling the counting of hours, minutes, and seconds, which are displayed via seven-segment displays. Simulation verification was conducted using Proteus simulation software. The circuit operates stably within the timing range of 00:00:00 to 23:59:59. The second counter switches to the minute counter every 60 seconds, the minute counter switches to the hour counter every 60 minutes, and the hour counter automatically resets every 24 hours. This paper provides low-cost and easy to reproduce the simulation circuit, on the understanding and mastery of the principle of digital clock has a certain role in assisting, for the subsequent design of complex digital systems to lay the foundation.
Keywords: Proteus; Digital clock; Imulation

1 INTRODUCTION

Proteus simulation software developed by the UK Labcenter Electronics, is an integrated electronic design automation (EDA) tool, known for its comprehensive functionality and ease of operation, widely used in electronic circuit design, teaching and engineering development [1]. The software contains core modules such as circuit schematic drawing, virtual simulation, PCB design, etc., which can realize the whole process from design to plate making: the circuit design module supports multi-layer drawing and intelligent operation; the simulation module is based on the SPICE kernel, which can complete the dynamic simulation of all kinds of circuits and multi-mode analysis; the PCB module provides automatic routing and rule checking, and can generate production documents that meet the standards. Its rich component library covers basic components, microcontrollers, sensors and virtual instruments, which can meet the diversified needs from basic circuit design to complex system simulation, providing effective guidance for circuit design and development.

The timing results of digital clocks are very different from those of mechanical clocks, which visually present the hours, minutes, seconds, and even the year, month, and day in a clear digital form. This kind of display makes it unnecessary for people to speculate the time through the position of the hands like looking at a mechanical clock, and it can be read quickly and accurately, which is very intuitive and provides great convenience for various industries [2-3]. This paper uses Proteus simulation software to design a digital clock circuit. The circuit employs cascaded 74160 adder counters to implement sexagesimal counting for seconds and minutes, and twenty-four-ary counting for hours. This enables the counting of hours, minutes, and seconds, with the corresponding time displayed via seven-segment displays. Optimize the circuit logic with the help of simulation verification, and finally realize the counting and clear display of hours, minutes and seconds to meet the basic needs of daily time observation.

2 DIGITAL CLOCK CIRCUIT DESIGN

The digital clock circuit design is shown in Figure 1. The second module begins counting under the influence of a clock pulse signal [4] with a period of 1 second. When the second module completes counting 60 seconds, it sends a “second feed” signal to the minute module. The minute module begins counting under the “second feed” signal. When the minute module completes counting 60 minutes, it sends a “minute feed” signal to the hour module. The hour module begins counting under the “minute feed” signal. The time of the digital clock is displayed by the digital tube of the hour module, minute module and second module respectively, realizing the counting from 00:00:00 to 23:59:59 [5-6].

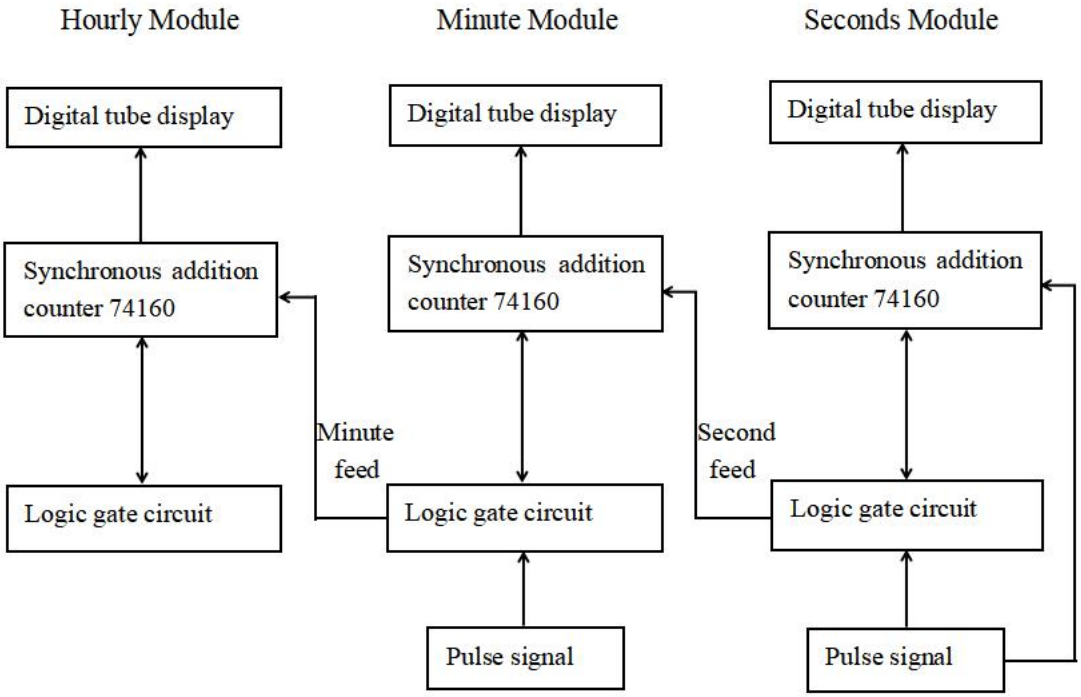


Figure 1 Digital Clock Circuit Design

2.1 Digital Tube Display

Digital tube displays are used to show the numbers corresponding to "hours", "minutes" and "seconds" [7]. In a variety of displays, seven-segment digital tube is more widely used, seven-segment digital tube is divided into two categories of common anode and common cathode. In this paper, the seven-segment digital tube with common anode is used to display Arabic numerals. Common anode digital tube of the seven light-emitting tube anode connected together to a high level, the cathode connected to a low level of the light-emitting tube is lit. Each segment is labeled as a, b, c, d, e, f, g. By controlling the high and low levels of these segments, the corresponding numbers can be displayed.

2.2 Counter Design

The design of the counter is a key part of the digital clock. In this paper, a synchronous addition counter 74160 is used for counting [8], 74160 is a binary synchronous modulo 10 addition counter, which has four independent counting outputs, each output corresponds to one binary bit, and its function is shown in Table 1.

Table 1 Synchronous Addition Counter 74160 Function Table

					Preset Data Input				Exports				
$\overline{CLR}$	$\overline{LOAD}$	ENP	ENT	CLK	D_3	D_2	D_1	D_0	Q_3	Q_2	Q_1	Q_0	Operating mode
0	×	×	×	×	×	×	×	×	0	0	0	0	Asynchronous Reset
1	0	×	×	↑	d_3	d_2	d_1	d_0	d_3	d_2	d_1	d_0	Synchronous Counting
1	1	0	×	×	×	×	×	×	Remain				Data Retention
1	1	×	0	×	×	×	×	×	Remain				Data Retention
1	1	1	1	↑	×	×	×	×	Decimal number				Additive Counting

2.2.1 Seconds module design

In the seconds module set up two 74160 counters, respectively, the “seconds” of the units digit and tens digit, “seconds” of the units digit using decimal counting system, “seconds” of the tens digit using senary counting system. “Seconds” in the units digit counter, when $\overline{CLR}$ = $\overline{LOAD}$ = ENP = ENT = 1, and in the input clock pulse CLK rising edge of the role of the units digit counter for 0000-1001 addition count. When the units digit counter is 1001 and its output signal RCO is high, then $\overline{CLR}$ = $\overline{LOAD}$ = ENP = ENT = 1 in the tens digit counter, and under the action of the rising edge of the input clock pulse CLK, the tens digit counter carries out the addition count of 0000-0101. When the tens digit counter of the seconds module is 0101 and the units digit counter is 1001, the logic gate circuit is used to make the tens digit counter pin $\overline{LOAD}$ low, at which time the tens digit counter $\overline{LOAD}$ = 0, $\overline{CLR}$ = ENP = ENT = 1, and the tens digit counter of the seconds realizes synchronous counting under the action of the rising edge of the

input clock pulse CLK. With the tens digit counter of the seconds module at 0101 and the units digit counter at 1001, the seconds module feeds one bit to the minutes module through the logic gate circuit [9-10]. The circuit design of the seconds module and the minutes module is shown in Figure 2.

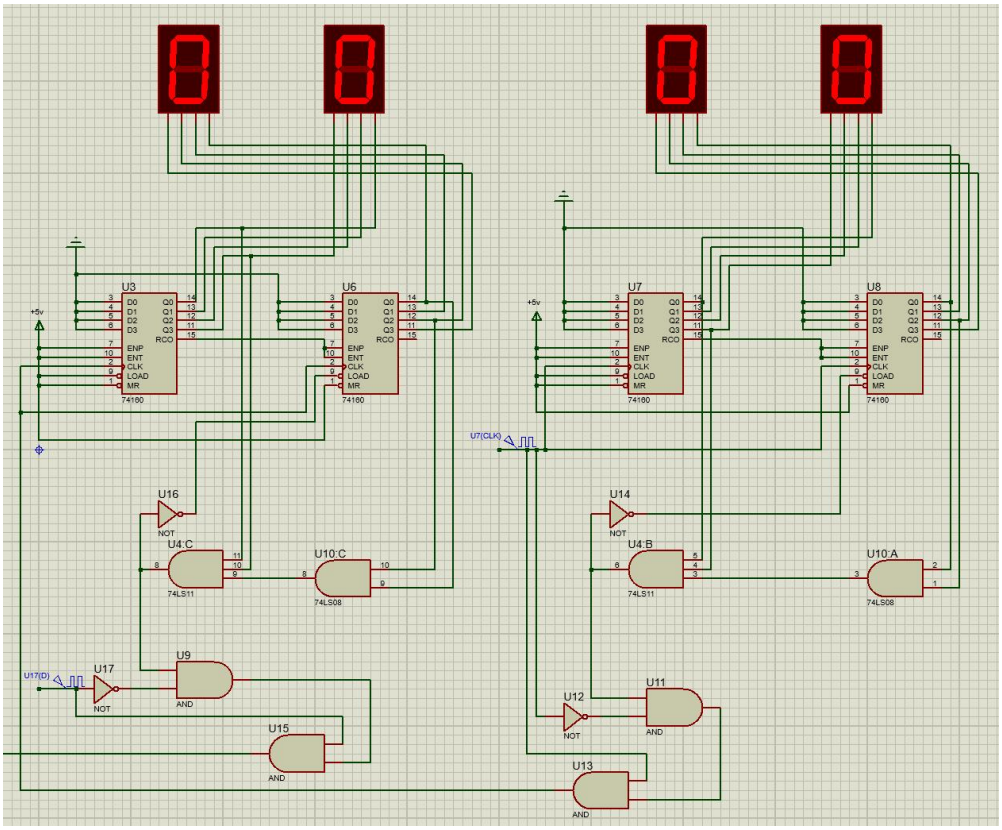


Figure 2 Seconds and Minute Modules for Digital Clocks

2.2.2 Minute module design

Similarly, two 74160 counters are set up in the minute module to represent the units digit and tens digit of "minutes". The units digit of "minute" is in decimal counting system, and the tens digit of "minute" is in senary counting system. When the minute module receives the rounding signal from the seconds module, in the minute module, its units digit counter $\overline{CLR}$ = $\overline{LOAD}$ = ENP = ENT = CLK= 1, and the units digit counter carries out the addition counting from 0000 to 1001. When the units digit counter is 1001 its output signal RCO is high, at this time the tens digit counter $\overline{CLR}$ = $\overline{LOAD}$ = ENP = ENT = CLK= 1, the tens digit counter carries out the addition count of 0000-0101. When the tens digit counter of the minute module is 0101 and the units digit counter is 1001, the logic gate circuit is used to make the pin $\overline{LOAD}$ of the tens digit counter low, at this time, $\overline{LOAD}$ = 0 in the tens digit counter, $\overline{CLR}$ = ENP = ENT = CLK = 1, and the tens digit counter realizes synchronous number setting. When the tens digit counter of the minute module is 0101 and the units digit counter is 1001, the minute module advances one bit to the hour module through the logic gate circuit.

2.2.3 Hour module design

Two 74160 counters are set in the hour module to represent the units digit and tens digit of "hours", the units digit of the "hour" is in decimal counting system, and the tens digit of "hour" is in ternary counting system. When the hour module receives the feed signal from the minute module, in the hour module, its units digit counter $\overline{CLR}$ = $\overline{LOAD}$ = ENP = ENT = CLK= 1, and the units digit counter performs the addition count of 0000-1001. When the units digit counter is 1001 its output signal RCO is high, at this time the tens digit counter $\overline{CLR}$ = $\overline{LOAD}$ = ENP = ENT = CLK= 1, the tens digit counter carries out the addition count of 0000-0010. When the tens digit counter of the hour module is 0010 and the units digit counter is 0011, the logic gate circuit is used to make the pin $\overline{LOAD}$ of the tens digit counter low, at this time, $\overline{LOAD}$ = 0 in the tens digit counter, $\overline{CLR}$ = ENP = ENT = CLK= 1, and the tens digit counter realizes synchronous number setting. The circuit design of the minute module and hour module is shown in Figure 3.

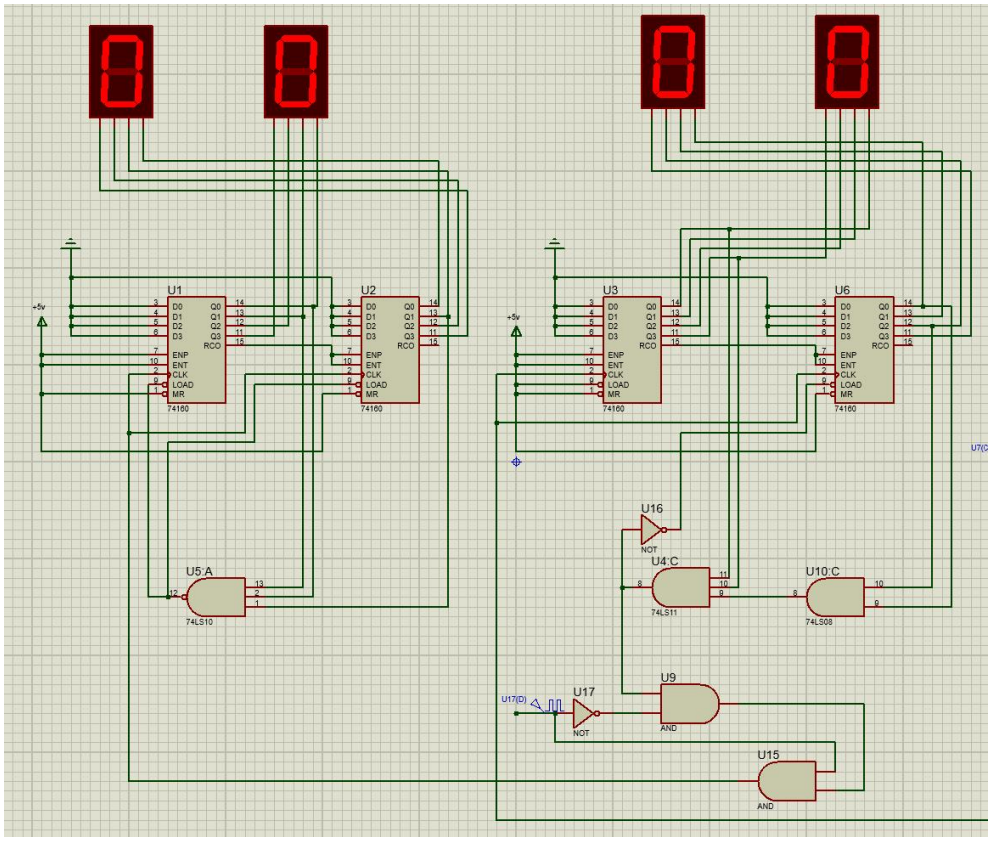


Figure 3 Minute and Hour Modules for Digital Clocks

3 DIGITAL CLOCK CIRCUIT SIMULATION

The circuit design is completed in Proteus, as shown in Figure 4. Proteus simulation software for circuit simulation, after several simulation tests, the circuit can run stably in the 00:00:00 to 23:59:59 timing range, seconds count every 60 seconds to the minutes counting bit, minutes count every 60 minutes to the hours counting bit, hours counting every 24 hours to automatically clear, the digital display is clear and no flicker. The results show that the designed clock circuit can complete the clock display function.

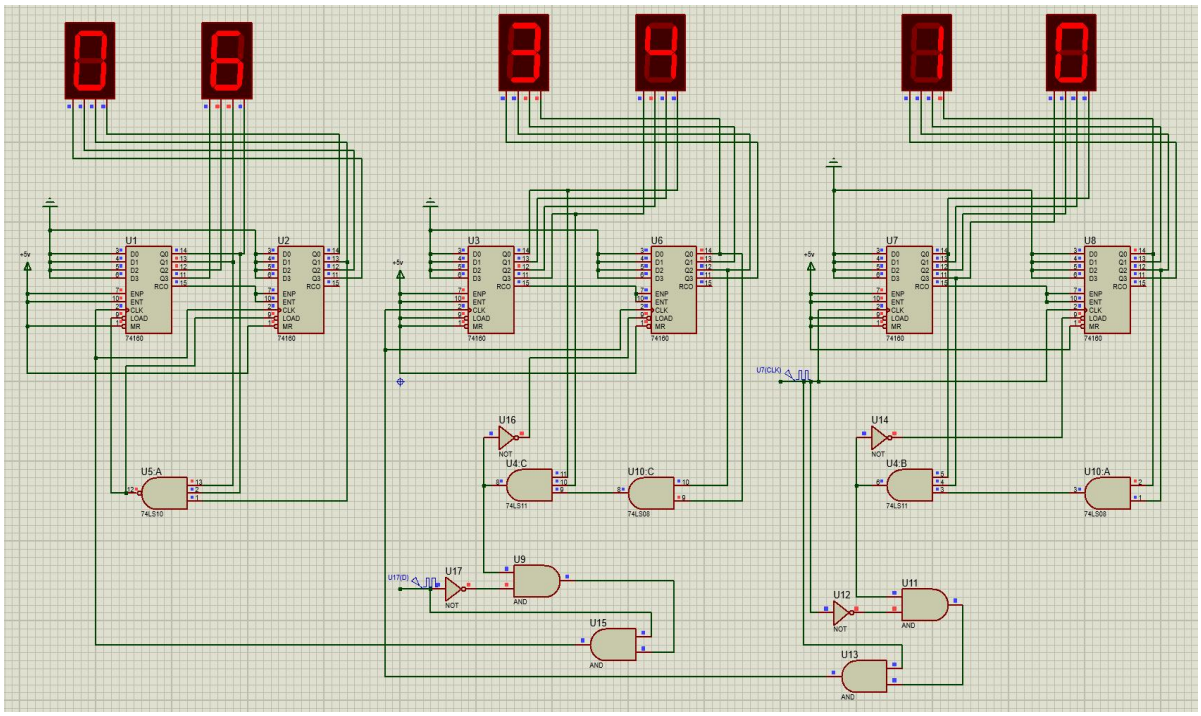


Figure 4 Digital Clock Circuit

4 CONCLUSION

In this paper, a digital clock circuit is designed and simulated using Proteus simulation software. After repeated simulations, the digital clock circuit is found to be stable and reliable, covering the range from 00:00:00 to 23:59:59. The seconds module sends a feed signal to the minutes module every 60 seconds, and the minutes module triggers a feed to the hours module every 60 minutes, while the hours module follows the 24-hour cycle rule and resets to 00:00:00 after 23:59:59, which is in line with the daily time flow pattern. The simulation results are satisfactory, which is helpful for understanding and mastering the principle of digital clock in the process of study and research. On the basis of the existing digital clock design, we deepen the innovation from the dimension that is more in line with the law of technological evolution and the actual needs of users, so as to make the classic circuit design revitalized in the application scenarios of the new era. Under the premise of maintaining the simplicity of the core counting circuit, we introduce energy-saving design to respond to the needs of low-carbon development, and develop modular adaptation solutions for multiple scenarios. The digital clock has the ability to adapt to the extension of functions in the intelligent era, so that this traditional design in the process of technology iteration continues to play a role in connecting the traditional electronic design and modern engineering applications bridge.

COMPETING INTERESTS

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

REFERENCES

[1] Liu J J. Application of Protues simulation technology in teaching electronic circuits. Modern Vocational Education, 2017(36): 214.
[2] Zhu H J, Wu A, Zhu F J. Research and design of digital clock based on FPGA. Advanced Materials Research, 2011, 1198(187): 741–745.
[3] He J, Yuan Y S. Liquid crystal display digital clock based on SCM. Advanced Materials Research, 2013, 2393(711): 598–601.
[4] Qiao Q S, Zhang Q X, Yang M, et al. Design of digital clock based on SCM. Applied Mechanics and Materials, 2014, 3590(668–669): 822–825.
[5] Ma H X, Ma Y, Liu X, et al. Research on the application of Multisim 14.0 in electronic design courses—Taking digital clock circuit as an example. Modern Information Technology, 2024, 8(11): 195–198. DOI:10.19850/j.cnki.2096-4706.2024.11.039.
[6] Gao W Y, Yang D, Li P L, et al. Design of a simple digital electronic clock. Gansu Science and Technology, 2020, 36(11): 13–14.
[7] Li F, Zhang X R. Design and realization of digital electronic clock based on Proteus simulation software. Computer Knowledge and Technology, 2023, 19(34): 41–44+51. DOI:10.14004/j.cnki.ckt.2023.1833.
[8] Li Y J. Multisim simulation design of a simple digital electronic clock. Technology Innovation and Application, 2017(18): 42–44. DOI:10.19981/j.cn23-1581/g3.2017.18.027.
[9] Emery C R. Digital Circuits: Logic and Design. CRC Press, 2020.
[10] Yorke J. Digital Circuits. Tritech Digital Media, 2018.