ISSN: 2959-6386 (Online), Volume 3, Issue 3, September 2024

Journal of Knowledge Learning and Science Technology

Journal homepage: https://jklst.org/index.php/home

ENHANCING FINANCIAL SERVICES THROUGH BIG DATA AND AI-DRIVEN CUSTOMER INSIGHTS AND RISK ANALYSIS

Tianyi Yang[1*], Qi Xin[2], Xiaoan Zhan[3], Shikai Zhuang[4], Huixiang Li[5]

*[*1] Financial Risk Management, University of Connecticut, Stamford CT, USA*
[2]Management Information Systems, University of Pittsburgh, Pittsburgh, PA, USA
[3] Electrical Engineering, New York University, NY, USA
[4] Electrical Engineering,University of Washington, Seattle, WA, USA
[5] Information Studies, Trine University, AZ, USA

Abstract

The article discusses the integration of big data and artificial intelligence (AI) technologies in the financial sector, focusing on supervised learning for pricing models to enhance customer identification and targeting. It details the construction of customer feature systems, including attributes like debit and credit card transactions, loan applications, and online behavior. By leveraging AI, financial institutions aim to accurately profile customers, boost consumption, and improve price management, ultimately aiding risk management and loan approval decisions. The article also covers related work in financial risk monitoring and machine learning in credit risk modeling, highlighting advancements and challenges in these areas.

Keywords: Big Data; Artificial Intelligence; Financial Risk Monitoring; Machine Learning

Article Information:
Received: 22-April-24 *Accepted: 20*-May-24 *Online:* 11-June -24 *Published:* 25-Sep-24
DOI: https://doi.org/10.60087/jklst.vol3.n3.p53-62

[i] **Correspondence author:** Tianyi Yang **Email:** donaldyan3@gmail.com

1. INTRODUCTION

With the wide application of big data and artificial intelligence technology in all walks of life, we have witnessed the explosive growth of data and the flood of information. In the financial field, customers' various transaction behaviors, product ownership information and marketing activity participation have built a huge database of information, laying the foundation for digital applications. [1]Compared with traditional measurement methods and statistical models, machine learning has significant advantages in processing massive data and discovering potential laws. Therefore, we choose supervised learning in machine learning as the basis of artificial intelligence pricing model, aiming to realize the combination of man-machine decision-making, so as to achieve the purpose of accurately identifying customers and effectively reaching the target.

The foundation of machine learning is built on structured databases. To this end, we first dig into customer information, and combined with business experience to build six characteristic system. These characteristics mainly include customer attributes, debit card transactions, credit card installments, loan applications, trend characteristics, and visit behavior to product pages. [2]Among these characteristics, debit card and credit card transaction characteristics mainly focus on the transaction frequency, consumption amount and credit

overdraft of customers in different time Windows; The trend features focus on the consumption characteristics of individuals in the past month and the borrowing ratio. [3]Through the above data processing and feature construction, our goal is to use artificial intelligence to achieve an accurate portrait of customers, so as to stimulate the consumption potential of customers and enhance the leverage of price management. This will provide financial institutions with more effective risk management and loan approval decision support, and promote the digital transformation and innovative development of the financial services industry.

2. RELATED WORK

2.1 Financial Risk Monitoring Technology

Financial risk, the risk generated in financial activities, refers to any potential threat that may lead to financial losses for businesses or institutions. [4-6]Financial risk can be categorized into several aspects, including market risk, credit risk, liquidity risk, operational risk, and legal risk. Market risk involves fluctuations in financial market prices, which may arise from changes in interest rates, stock prices, and other factors. Credit risk refers to potential losses due to borrowers or counterparties failing to fulfill their contractual obligations. Operational risk involves errors and losses caused by internal processes, systems, or human factors. Liquidity risk refers to losses incurred when assets cannot be effectively bought or sold within a specific period. Legal risk involves potential legal disputes and regulatory changes in financial activities. [7]These risk components intertwine to form the complex and diverse risk landscape in the financial system, requiring effective risk management measures from financial institutions and regulatory authorities to mitigate potential adverse impacts.

Financial risk monitoring plays a crucial role in today's economic system. It is not only a key to maintaining financial system stability but also essential for protecting investor interests, ensuring market confidence, enhancing market efficiency, and improving financial regulatory frameworks. Effective risk monitoring techniques can alert potential financial risks, prevent major economic impacts, optimize resource allocation for financial institutions, and enhance overall market operational efficiency [8][9]. However, financial risk monitoring faces many challenges, including handling large and complex financial data, adapting to rapidly changing market environments, technological innovations, preventing financial fraud risks and security threats, and addressing challenges posed by globalization and diversified risks.

Traditional financial risk monitoring techniques mainly include credit scoring, market risk analysis, liquidity risk management, operational risk assessment, and compliance monitoring. These techniques rely heavily on expert experience and traditional statistical models, but they may perform poorly in fast-changing market environments and may have limitations in identifying complex business scenarios and new types of financial risks.

In the context of digital transformation, the accelerated pace of financial technology innovation and the widespread application of new technologies in the financial field have introduced new dimensions of risk, leading to new characteristics of financial risks. [10-11]Therefore, financial regulation and risk prevention face greater challenges, requiring urgent research on the application of new technologies such as artificial intelligence and big data in financial risk monitoring and making reasonable predictions and arrangements for future applications.

This paper suggests starting from the perspective of patent analysis. Patent layout plays a critical role in modern business and technological development. It not only protects innovative achievements and maintains market competition but also generates additional income for companies through authorization and transfer, promoting product research and technological progress. Additionally, patent layout helps companies effectively manage risks, avoid infringement, protect technological advantages when expanding international markets, enhance corporate image, increase brand value, and meet specific industry policy drivers and compliance requirements. A strong patent portfolio is an indispensable part of corporate strategic planning.

2.2 Machine Learning and Bank Credit

Over the past two decades, automated underwriting and account management systems have become commonplace in the retail credit industry. In traditional interview-based underwriting systems, loan officers use subjective criteria to measure a customer's creditworthiness. Instead, banks automate this process with evaluation models that use data extracted from external credit department data and internal account management data to predict the probability of a "bad" customer – see Thomas (2000)[12-13]. These models typically score customers to quantify the probability that a customer is a "bad customer," which is defined by some metric related to delinquency or default.

While the customer's level of risk is an important factor for institutions to consider, lenders are primarily interested in maximizing profits. As we all know, risk and profit are not necessarily monotonously correlated. For example, a credit card revolving customer may have a small possibility of default, but need to pay a lot of financial fees; If they pay off the balance each cycle, no fees accumulate. The difference between risk models and the profit motive of lenders has been noted in papers such as Finlay (2008), but models predicting account level profits are still relatively rare in the industry due to data limitations and the potential complexity of the models. However, with the advent of machine learning methods in credit risk modeling [14][15], financial institutions have come to rely on models to estimate increasingly complex relationships, leading to an increase in account-level profit models for a variety of use cases.

A range of literature in decision science and operations research studies such models. Finlay (2008) and Finlay (2010) compare linear and machine learning-style approaches in neural network algorithms, demonstrating that a continuous financial behavior model is superior to a binary model of customer default. Fitzpatrick and Mues (2021) extend a set of algorithms for predicting profitability in the context of P2P lending [16]. Verbraken et al. (2014) point out that profit-based modeling allows the calculation of an optimal profit-based threshold that would otherwise not fully approximate the average profit using default scores and segment levels.

2.3 Risk control solution based on big data and machine learning technology

2.3.1. Face recognition technology

At present, face recognition technology has been applied in all aspects of life, such as facial payment, high-speed rail station and even the public security Skynet system[17]; In the financial business, credit card processing, online loan applications are also widely used face

recognition: compare the applicant's photo with the ID card in the public security system to judge the similarity of the two photos, which can effectively avoid the problem of non-application.

Similar to human face recognition, when we humans are familiar with a person, we can easily judge whether the person is this person based on his facial features and the features stored in our mind in the past (although there will be misjudgments), similarly, behind the face recognition technology is also a set of deep learning algorithms. The process of our thinking into a model algorithm, there are a number of commercial application companies that provide face recognition services on the market, but the essence of the algorithm is basically the same, the core idea of face recognition is that different faces are composed of different features.

Where do the vast array of features used to represent faces come from? This is where deep learning (deep neural networks) comes into play[18]. After learning and training on tens of millions or even billions of level face databases, it will automatically summarize the face features that are most suitable for computer understanding and differentiation.

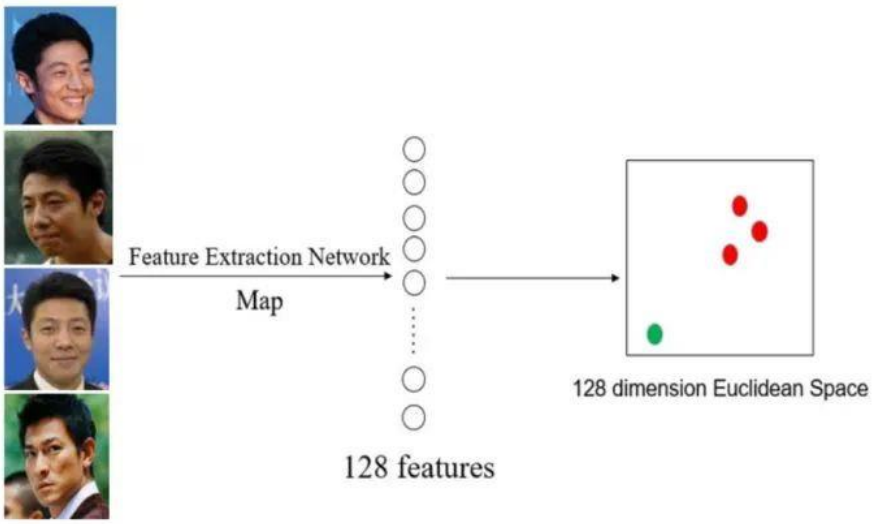

Figure 1. Principle of machine learning face recognition technology

After elucidating that different faces are composed of different features, we have enough knowledge to analyze faces, and algorithm engineers usually need certain visualization methods to know what features the machine has learned to distinguish between different people [19]: the features extracted from different photos of the same person are very close in the feature space, and different people are far apart in the feature space.

2.4 Identification of data forgery

Checking user information can be used to determine whether the borrower may be at risk of fraud, using the relationship graph to do cross-checking, although not guaranteed 100% accuracy, but it is a powerful reference in the manual audit. The personal information that users fill out is usually false. For example, the borrower Zhang SAN and the borrower [20]Li Si fill in the same company phone number, but the company Zhang SAN fills in is completely different from the company Li Si fills in, which becomes a risk point. By visualizing the

relationship graph data, we can intuitively find the contradiction between the two, and we can determine that at least one of them is at risk of fraud.

2.5 Analysis of gang fraud

In the credit scenario, the losses caused by gang fraud are more severe, but it is also difficult to detect gangs in the complex data. Based on the knowledge graph, we usually intuitively analyze data at multiple levels, including first degree association, second degree association, third degree association, and even more dimensional association. Although the gangs use false information for credit and use, they usually have common information, such as the same WIFI and the same area. LOUVAIN, LPA, SLPA [21] and other community discovery algorithms and label propagation algorithms can effectively and quickly find gangs.

3. CUSTOMER CHURN PREDICTION MODEL IN TELECOM INDUSTRY

This paper draws a sample of bank accounts from multiple datasets and analyses 150,000 loans originated in 2012, tracking them until December 2015. In order to preserve bank-level anonymity while still capturing meaningful product- and strategy-level heterogeneity, this paper divides accounts into "integrated loan portfolios". For this purpose, the paper classifies banks as a) "high liquidity" banks (Spender, Spender), b) "high finance charge" banks (Revolver, Revolver), or c) neither (Middle, Middle) [22]. The paper then runs a multi-class classification problem to predict whether an account belongs to a), b) or c). The critical value of the score determines the structure of the overall credit portfolio. Figure 1 shows a comparison of loan portfolios by typical credit sector attributes, and shows that spending banks tend to consist of wealthier, higher credit quality, and more creditworthy borrowers than weekly transition banks.

3.1. Experimental Data

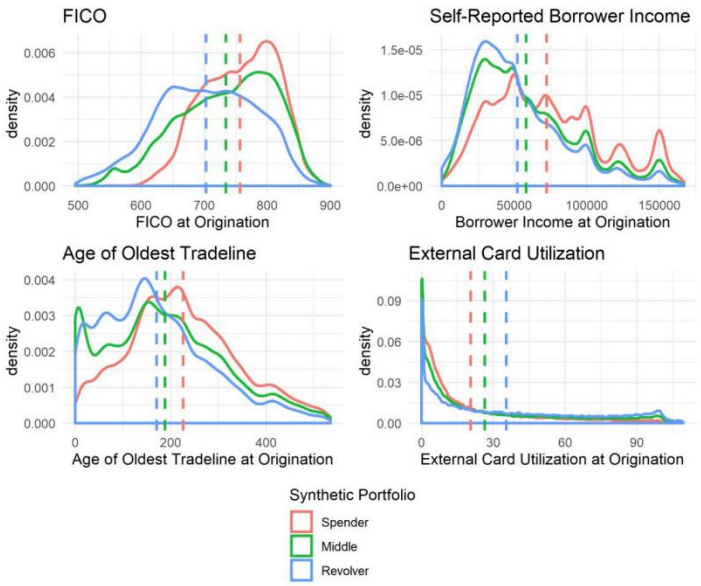

Figure 2. Experimental data

All the observed components of income and expenditure are added together to produce a profit graph. Since net credit losses increase monotonically with risk, subtracting net credit losses from income will result in a significant reduction in the risk "sweet spot" in terms of credit risk, as shown in the "actual value" in Figure 2. We can see that the hump shape of the curve is most pronounced in the turnover portion, where the credit portfolios of spenders and intermediaries are more closely related to risk.

This suggests that it is more important for institutions with such credit portfolios to understand risk-independent profits; At the same time, doing so is more challenging because of its non-monotonic relationship with risk. Relatingly, across all portfolios, the profit quartile spread across risk ranges increases with the level of risk, indicating the potential challenge of capturing all changes in profits in high-risk areas.

3.2 Risk Model

After studying the relationship between risk and experiential profit, we can further study the relationship between risk and predicted profit. This requires a shift from empirical profit to predictive profit, which in turn increases noise proportional to the predictability of the information set. Therefore, this paper establishes the profit model of related credit portfolio.

When modeling the behavioral subcomponents of profit, there are trade-offs in detail. One can take a highly refined approach and build account-level models for a customer's monthly balance, monthly repayment percentage, monthly tendency to pay late, and monthly tendency to write off, and then aggregate these models across time at the account level. [23]Large quantum models will increase the degree of freedom of aggregation models, enabling them to better simulate relationships in NPV that are non-linear or have higher-order interactions. This may also increase transparency, as one can find out which sub-models contribute the most to customer profit calculations. However, a large number of submodels can also increase the potential sources of model risk and multiply errors. In addition, if developers use more complex modeling methods, such as machine learning methods, they will have the ability to fit highly nonlinear relationships and higher-order interactions that may occur in more aggregated modeling structures.

The actual model algorithm used for the estimation is[24] "Extreme Gradient Boosting", also known as XgBoost. In the initial model exploration phase, this paper tested the driver model structure at different levels of granularity and determined that the approach with the best model performance was at a relatively aggregated level. Optimal model performance is achieved by modeling revenue, default probabilities, and default loss rates at the account level. However, this method is relatively difficult to explain because it does not distinguish between different sources of income. Due to this difficulty in interpretation, this paper chooses to model the revenue drivers separately (but still at the account level), and the decline in out-of-sample model performance is almost insignificant.

Submodel	η	Tree depth	Num trees
Finance Charge	0.009	6	2135
Interchange Income	0.001	2	6793
Late Fees	0.005	8	2056
Other Fees	0.014	7	928
PD	0.005	9	455
LGD	0.004	5	2696

Note: Calculated on 20% holdout set.

After the structure selection is complete, the next step is the complete model estimation process. For each submodel, the paper uses a range of hyperparameter search and variable selection techniques to improve performance and interpretability, and reduce potential overfitting. Firstly, grid search is carried out to adjust the tree depth and learning rate hyperparameters. Control the number of trees with 5x cross validation and early stop functions for AUC. Because of the high risk of overfitting of this sample, the learning rate was limited to a relatively small range (between 0.001 and 0.01); The number of trees is also high, reaching into the thousands. This paper selects the best hyperparameters according to the out-of-sample mean square error of the regression submodel and the out-of-sample classification error of the default probability model, and then finds the first 30 variables of each submodel according to XGBoost. After that, we re-estimate all submodels on these constrained feature Spaces through another complete hyperparameter search. The results of these searches are shown in Table 1.

3.3 discussion

As for model performance, it can be seen from the previous article that the forecasting performance of revenue model and total profit model is poor, but the ranking performance is better. Without richer account performance data, it is difficult to accurately estimate the net present value of customers to issuers. For example, the model has difficulty capturing changes in experiential profits within the risk range and cannot capture increases in experiential profit changes within the higher risk range. [25-27]This may have implications for the reliability of profit forecasts in the higher risk range, where profit-based modeling provides more information than risk-based modeling.

However, the model still does a good job of separating more profitable customers from less profitable customers, and the hump-shaped relationship between the profit component and risk is retained between the predicted curve and the actual curve. This shows that the average profitability of the portfolio obtained by ranking by this score is significantly higher than the average profitability when the risk score is used alone.

4. CONCLUSION

This paper describes the application of big data and artificial intelligence technology in the financial field to achieve accurate customer identification and target marketing through supervised learning models. Six feature systems, including customer attributes, debit and credit card transactions, loan applications, trend characteristics and product page visit behavior, are constructed, and artificial intelligence technology is used for data processing and feature construction, and finally to achieve an accurate portrait of customers[28-31]. This will not only

stimulate the consumption potential of customers, but also increase the leverage of price management, provide financial institutions with more effective risk management and loan approval decision support, and promote the digital transformation and innovation development of the financial services industry.

In the future, with the continuous advancement of technology and rapid changes in the financial environment, financial risk monitoring and customer behavior prediction will face more challenges and opportunities. The further application of artificial intelligence and big data technologies is expected to improve the accuracy and efficiency of risk forecasting, especially in dealing with complex financial risks and changing market environments. In addition, by improving model algorithms and data processing methods, financial institutions can better maximize profits while reducing risks, further promoting the innovation and development of fintech, and bringing new growth points to the industry.

Acknowledgement

We would like to extend our sincere thanks to Yadong Shi, Jiaqiang Yuan, Peiyuan Yang, Yufu Wang and Zhou Chen for their article [1]《Implementing Intelligent Predictive Models for China Patient Disease Risk in Cloud Data Warehousing.》 Excellent work. We would like to thank their research for providing important inspiration and help to our article. Yadong Shi provided key guidance on the overall research direction and methodology; Jiaqiang Yuan played a central role in data analysis and model building; Peiyuan Yang to ensure the accuracy and completeness of data; Yufu Wang contributed valuable technical support in algorithm optimization and performance improvement; Zhou Chen provided critical review and feedback at all stages of the study. Thanks to their hard work and outstanding contributions, our research can proceed smoothly and make important progress.

We sincerely thank Huixiang Li, Ang Li, Yuning Liu, Yiyu Lin and Yadong Shi for their article [2]"AI Face Recognition and Processing Technology Based on AI. Outstanding work done in GPU Computing. We would like to thank their research for providing important inspiration and help to our article. The expertise and innovative contributions in information research by Huixiang Li, business analytics by Ang Li and Yuning Liu, computer science and engineering by Yiyu Lin, and facial recognition and processing technology by Yadong Shi have provided us with valuable guidance and support. Thanks to their hard work and outstanding contributions, our research can proceed smoothly and make important progress.

REFERENCES

[1] Shi, Y., Yuan, J., Yang, P., Wang, Y., & Chen, Z. *Implementing Intelligent Predictive Models for Patient Disease Risk in Cloud Data Warehousing*.

[2] Li, Huixiang, et al. "AI Face Recognition and Processing Technology Based on GPU Computing." *Journal of Theory and Practice of Engineering Science* 4.05 (2024): 9-16.

[3] Zhan, Xiaoan, Chenxi Shi, Kangming Xu, Lianwei Li, and Haotian Zheng. "Aspect category sentiment analysis based on multiple attention mechanisms and pre-trained models." *Applied and Computational Engineering* 71 (2024): 21-26.

[4] Xu, J., Wu, B., Huang, J., Gong, Y., Zhang, Y., & Liu, B. (2024). *Practical Applications of Advanced Cloud Services and Generative AI Systems in Medical Image Analysis*. arXiv preprint arXiv:2403.17549.

[5] Huang, J., Zhang, Y., Xu, J., Wu, B., Liu, B., & Gong, Y. *Implementation of Seamless Assistance with Google Assistant Leveraging Cloud Computing*.

[6] Liang, Penghao, et al. "Automating the Training and Deployment of Models in MLOps by Integrating Systems with Machine Learning." arXiv preprint arXiv:2405.09819 (2024).

[7] Zhan, Tong, et al. "Optimization Techniques for Sentiment Analysis Based on LLM (GPT-3)." arXiv preprint arXiv:2405.09770 (2024).

[8] Zhang, Y., Liu, B., Gong, Y., Huang, J., Xu, J., & Wan, W. (2024). *Application of Machine Learning Optimization in Cloud Computing Resource Scheduling and Management*. arXiv preprint arXiv:2402.17216.

[9] Gong, Y., Huang, J., Liu, B., Xu, J., Wu, B., & Zhang, Y. (2024). *Dynamic Resource Allocation for Virtual Machine Migration Optimization using Machine Learning*. arXiv preprint arXiv:2403.13619.

[10] Chen, B., Zhu, Y., Ye, S., & Zhang, R. (2018). *Structure of the DNA-binding domain of human myelin-gene regulatory factor reveals its potential protein-DNA recognition mode*. *Journal of Structural Biology*, 203(2), 170-178.

[11] Shi, Y., Yuan, J., Yang, P., Wang, Y., & Chen, Z. *Implementing Intelligent Predictive Models for Patient Disease Risk in Cloud Data Warehousing*.

[12] Huang, J., Zhang, Y., Xu, J., Wu, B., Liu, B., & Gong, Y. *Implementation of Seamless Assistance with Google Assistant Leveraging Cloud Computing*.

[13] Lei, Han, et al. "Automated Lane Change Behavior Prediction and Environmental Perception Based on SLAM Technology." arXiv preprint arXiv:2404.04492 (2024).

[14] Fan, C., Li, Z., Ding, W., Zhou, H., & Qian, K. *Integrating Artificial Intelligence with SLAM Technology for Robotic Navigation and Localization in Unknown Environments*.

[15] Ding, W., Tan, H., Zhou, H., Li, Z., & Fan, C. *Immediate Traffic Flow Monitoring and Management Based on Multimodal Data in Cloud Computing*.

[16] Wang, Y., Zhu, M., Yuan, J., Wang, G., & Zhou, H. (2024). *The Intelligent Prediction and Assessment of Financial Information Risk in the Cloud Computing Model*. arXiv preprint arXiv:2404.09322.

[17] Zhou, Y., Zhan, T., Wu, Y., Song, B., & Shi, C. (2024). *RNA Secondary Structure Prediction Using Transformer-Based Deep Learning Models*. arXiv preprint arXiv:2405.06655.

[18] Liu, Beichang, et al. "Precise Positioning and Prediction System for Autonomous Driving Based on Generative Artificial Intelligence."

[19] Wang, B., He, Y., Shui, Z., Xin, Q., & Lei, H. *Predictive Optimization of DDoS Attack Mitigation in Distributed Systems using Machine Learning*.

[20] Cui, Z., Lin, L., Zong, Y., Chen, Y., & Wang, S. *Precision Gene Editing Using Deep Learning: A Case Study of the CRISPR-Cas9 Editor*.

[21] Wang, Y., Zhu, M., Yuan, J., Wang, G., & Zhou, H. (2024). *The Intelligent Prediction and Assessment of Financial Information Risk in the Cloud Computing Model*. arXiv preprint arXiv:2404.09322.

[22] Li, Hanzhe, et al. "Driving Intelligent IoT Monitoring and Control through Cloud Computing and Machine Learning." arXiv preprint arXiv:2403.18100 (2024).

[23] Tian, J., Li, H., Qi, Y., Wang, X., & Feng, Y. *Intelligent Medical Detection and Diagnosis Assisted by Deep Learning*.

[24] Qi, Y., Wang, X., Li, H., & Tian, J. (2024). *Leveraging Federated Learning and Edge Computing for Recommendation Systems within Cloud Computing Networks*. arXiv preprint arXiv:2403.03165.

[25] He, Z., Shen, X., Zhou, Y., & Wang, Y. *Application of K-means Clustering Based on Artificial Intelligence in Gene Statistics of Biological Information Engineering*.

[26] Zhou, Y., Osman, A., Willms, M., Kunz, A., Philipp, S., Blatt, J., & Eul, S. (2023). *Semantic Wireframe Detection*.

[27] Qi, Yaqian, Yuan Feng, Jingxiao Tian, Xiangxiang Wang, and Hanzhe Li. "Application of AI-based Data Analysis and Processing Technology in Process Industry." *Journal of Computer Technology and Applied Mathematics* 1, no. 1 (2024): 54-62.

[28] Tian, J., Qi, Y., Li, H., Feng, Y., & Wang, X. (2024). "Deep Learning Algorithms Based on Computer Vision Technology and Large-Scale Image Data." *Journal of Computer Technology and Applied Mathematics*, 1(1), 109-115.

[29] Wang, X., Tian, J., Qi, Y., Li, H., & Feng, Y. (2024). "Short-Term Passenger Flow Prediction for Urban Rail Transit Based on Machine Learning." *Journal of Computer Technology and Applied Mathematics*, 1(1), 63-69.

[30] Feng, Y., Li, H., Wang, X., Tian, J., & Qi, Y. (2024). *Application of Machine Learning Decision Tree Algorithm Based on Big Data in Intelligent Procurement*.

[31] Tian, Jingxiao, Hanzhe Li, Yaqian Qi, Xiangxiang Wang, and Yuan Feng. "Intelligent Medical Detection and Diagnosis Assisted by Deep Learning."