World Journal of Innovation and Modern Technology, Vol. 7, Issue 2, (Apr) **2024**
ISSN 2682-5910
DOI: 10.53469/wjimt.2024.07(02).05

Drug Screening and Target Prediction Based on Machine Learning

Zhengrong Cui[1,*], Luqi Lin[2], Yizhi Chen[3], Sihao Wang[4], Yanqi Zong[5]

[1]Software Engineering, Northeastern University, Shanghai, China
[2]Software Engineering, SunYat-sen University, Shanghai, China
[3]Information Studies, Trine University, Allen Park, MI, USA
[4]Mathematics, Southern Methodist University, Dallas, TX
[5]Information Studies, Trine University, PhoenixAZ, USA
[*]*Corresponding author: cuizhengrong@outlook.com*

Abstract: ***With the development of artificial intelligence information technology in biomedicine, the traditional practice of drug development is gradually replaced by machine learning. In drug development, drug screening and target protein identification are the keys to modern drug development. However, the traditional comparative experiments are limited by the high throughput, low precision and high cost of biological experimental methods, and the screening and prediction of a large number of drug targets have a certain degree of blindness, which is difficult to be widely carried out in practical applications. In contrast, machine learning has powerful intelligent information processing technology capabilities such as data mining and mathematical statistics. By simulating and predicting the interaction between drugs and targets through machine learning, it can reduce research and development costs, shorten the time of new drug development, and reduce the blindness of new drug development, which is of great significance for the development of new drugs and the improvement of human medical level. In this paper, a powerful deep learning classification model (SVM) is proposed to predict drug-target interactions, achieve efficient prediction of multi-source and high-dimensional data, and make great progress in reducing research and development costs, shortening research and development cycles, and improving experimental accuracy.***

Keywords: Machine learning; Drug targeting; Predicts drug screening; Support vector machines (SVM).

1. INTRODUCTION

Drugs have always been the most important means to prevent and treat diseases, but the discovery of drug targets and the correct screening of drugs is a very difficult process. This reflects the strength of a country's medical science and technology. How to use machine learning for drug screening and targeted prediction has a positive role in promoting the country's medical status in the world, and also plays a supporting role in the steady development of social health [1-3]. At present, all kinds of epidemic viruses threaten the life and health of people around the world, and the discovery of drug targets of viruses as soon as possible is also of vital importance for the research and development of virus specific drugs.

Traditional methods have not been able to make drug screening discovery as expected, but science and technology are changing with each passing day. With the rapid development of information processing technology, intelligent computing technology has emerged [4], and neural networks, pattern recognition and other technologies have emerged. Many researchers are beginning to use machine learning and statistical methods to solve drug design problems, which is called computer-aided drug design and has played a large role in the development of new drugs. Using computer virtual technology to discover the relationship between drug screening and target prediction as the theoretical basis for the synthesis of new drugs, so as to solve the drawbacks of long time, high cost and low success rate of traditional pharmaceutical.

In the technical [5-7] background of computer-aided drug research, with the continuous accumulation of biomedical data, it is possible to apply machine learning to the prediction of drug-drug interaction. In recent years, machine learning has been successfully applied in the fields of image recognition and natural language processing, has also made good achievements in the field of bioinformatics, and has shown good application potential in drug-drug interaction prediction tasks. With the continuous accumulation of biomedical data, a large number of biomedical data sets have become increasingly rich, so that machine learning methods can rely on rich data sets to automatically learn features, transform simple features into rich and effective features, and use these effective features to more accurately predict drug-drug interactions. This avoids the cumbersome, time-consuming and costly disadvantages of clinical trials, making it possible to use previously accumulated biomedical data to predict potential drug-drug interactions quickly and efficiently.

2. RELATED WORK

In the field of bioinformatics, a drug target is a biomacromolecular structure that can interact with drug molecules and has pharmacodynamic functions. The traditional research method is to screen the molecular information of drug targets and analyze the complex process of protein information through bioinformatics and protein chip technology. With the rise of bioinformatics, the combination of traditional biological experiments and computer intelligence information processing technology [8] provides a new direction for the prediction of drug-target interaction.

Now this paper will use SVM technology to carry out experiments based on the background of machine learning.

2.1 Machine learning

Machine learning core is to make machines mimic humans and have a human way of thinking. The operation mode of the machine learning can be summarized as two steps, first of all the given training data set $\mathcal{X}train$,through learning system model $\int$(x), then use models to test data set $\mathcal{X}test$ forecast, its purpose is to find a particular pattern$\int$ hidden in the data ($\mathcal{X}$) And uses this particular model to make inferences about the new data $\int$(xnew). Machine learning can be divided into supervised learning and unsupervised learning by training whether the data set has annotation information. The representative models [9-10] of supervised learning include classification and regression tasks, and the typical models of unsupervised learning include clustering tasks. Semi-supervised learning is a learning method that uses labeled and unlabeled data to build models. Applications in this area are in great demand because in reality, a large number of unlabeled samples can often be collected and the cost of labeling these samples is huge, and semi-supervised learning can solve this problem.

Supervised learning is the most widely used and rich machine learning method at present. For example, support vector machines (SVM), logistic regression, decision trees and other common algorithms belong to supervised learning. In this paper, SVM is mainly used to develop practice. The following is the flow chart of drug screening and target prediction in machine learning.

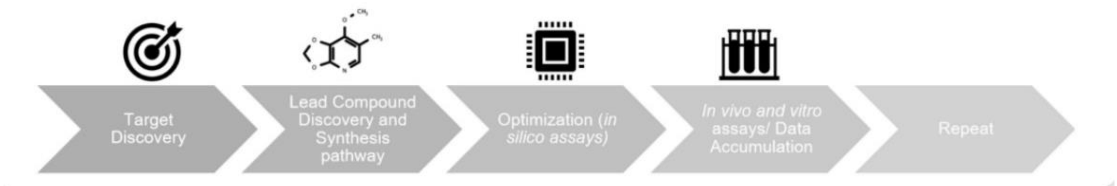

Figure 1: Machine learning flowchart in drug screening

Machine learning combines steps to collate and collect data, match suitable machine learning algorithms, build models, and test and verify the reliability of models. Currently [11], machine learning methods are mainly applied to chemical genomic databases containing standardized annotated information. Commonly used prediction methods include MT-QSAR, Nidhi's multi-class naive Bayes algorithm prediction model, Niwa's probabilistic neural network prediction model, Wale's ECFP-4 molecular fingerprint descriptor prediction method and Koutsoukas prediction model constructed by two probabilistic algorithms. An important step in the drug discovery process is determining the interaction between a drug and a target, such as a gene, which can be reliably performed through in vitro experiments.

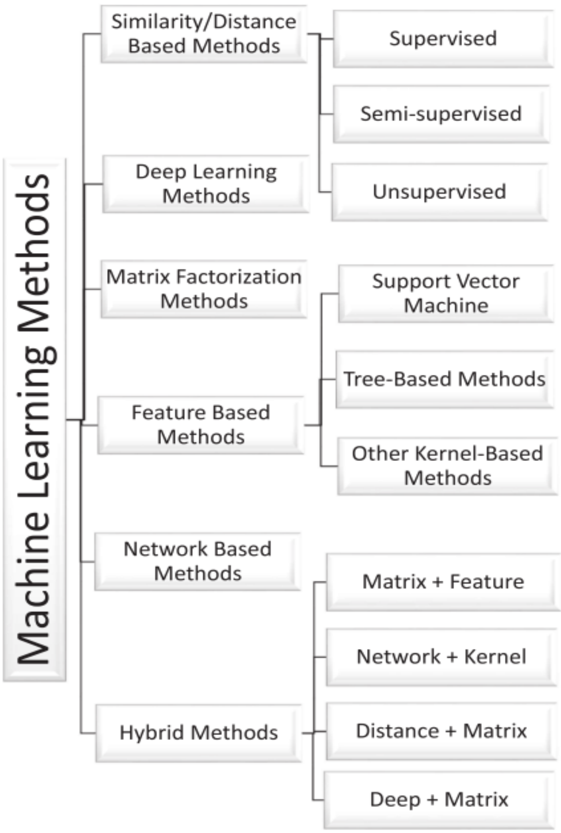

Figure 2: Machine learning main method frame diagram

Machine learning can reduce time and money costs, rather than depleting in-vitro search, virtual screening is performed first, followed by experimental validation of potentially viable options.Machine learning is good to reduce the cost of human and material resources and accelerate the speed of drug research and development, and is one of the most commonly used and important technologies in the field of drug research and development.

2.2 Comparison of traditional methods and machine learning in drug screening and targeted prediction

Traditional drug screening uses allopathic therapy, a treatment that has been highly successful in fighting the disease, but it also comes with high drug development costs that place a heavy financial burden on patients and health care systems. Because the candidates for potential drugs are so diverse, drug development requires a lot of synthesis, trial and error. Not only are costs increasing dramatically all the time, but they are also very unfriendly to the environment. Therefore, many computer-based methods [12-16] are used for lead compound discovery and optimization, including molecular docking, pharmacophore modeling, decision forest, and comparative molecular field analysis. The application of machine learning algorithms in drug discovery is not limited to specific steps, but throughout the entire process. The great potential of machine learning algorithms in drug screening and development, and current extensions of machine learning are based on the concept of multi-layer artificial neural networks that accelerate the drug discovery process. No doubt this could help chemists in their quest for new drugs. Overall, machine learning is a highly regarded approach in the field of modern drug screening and development, and as the machine learning toolbox continues to evolve, drug screening is also becoming more economical and environmentally friendly.

Targeted therapies are specific biomacromolecules within cells that can interact with drugs, and more than 98% of drug targets are proteins. However, due to the lack of understanding of the target and other reasons, although the number of drug targets in the human genome is [17-19] 6000-8000, which is difficult for traditional targeted prediction methods, machine learning is able to find information in complex data and make predictions on large-scale data. Machine learning-based approaches can infer function directly from sequence and structural information of drug molecules or target proteins. Molecules with similar functions often exhibit common properties in sequence and structure, and machine learning is very good at capturing the relationships between these features and functions. Machine learning can simulate human thinking, constantly learn the relationship between drugs and targets from existing DTI, build a model, and then use the model to make predictions about new DTI. Therefore, more and more researchers have developed a variety of DTI analysis methods based on the principle of machine learning to predict the pharmacological effects of drugs and target molecules.

2.3 Support vector machines (SVM)

SVM is a supervised learning algorithm that maps instances (or data points) to a high-dimensional space so that different classes of instances can be separated by as wide a gap as possible. SVM has been widely used in the field of gene classification, such as identifying tumor-related genes and predicting gene function [20]. By combining with DNA sequence features and protein sequence features, SVM can realize efficient gene classification. SVM can be used to predict the structure and function of proteins, such as secondary structure prediction and subcellular localization prediction. By encoding the protein sequence and combining information such as the physical and chemical properties of amino acids, SVM can accurately predict the structure and function of proteins. SVM can be used in biomedical text mining, such as gene named entity recognition, drug interaction prediction and so on. By extracting features from biomedical texts, combined with natural language processing techniques, SVM can help researchers quickly obtain relevant information.

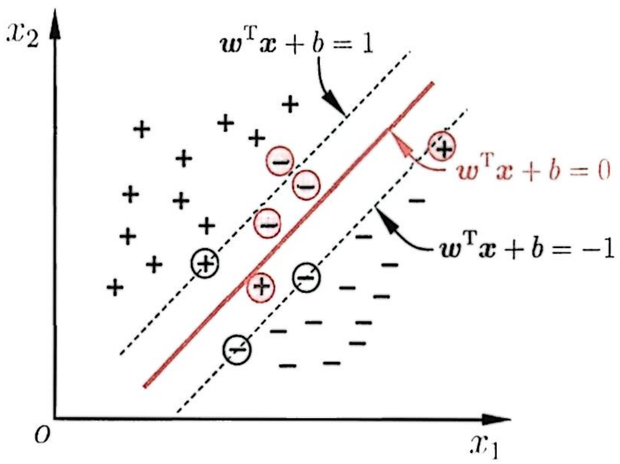

Figure 3: Support vector machine (SVM)hyperplane example diagram

3. METHODOLOGY

Topic: Based on machine learning, support vector machine model was used for drug screening and target prediction experiments.

3.1 Experimental model

Support vector Machine (SVM): In this experiment, we used a simple support vector machine model to perform protein-protein interaction experiments in drug screening and target prediction. The experiment is divided into training set and test set. The training set is composed of 70% positive and negative samples, and the test set is composed of 30% positive and negative samples [21]. In this experiment, the training set will be cross-validated 10 times, and the optimal parameters of SVM will be determined, and the key characteristic attributes of protein-protein interaction will be found through special diagnosis, which will help drug screening and target prediction. Finally, on the basis of determining the combination of parameters and optimal protein features, the experimental results were evaluated by SVM model and test set.

3.2 Experimental process

Data preparation: (from the ProtInDB database (http://protindb.cs.iastate.edu). For ligand-protein complexes, this paper integrated Binding MOAD database, PDBbind database, and 2P2I (version1). The UniProt database extracts the data and downloads the sequence information corresponding to the protein.

Data preprocessing: LPC calculation method was used to obtain the binding sites of small drug molecules and proteins. Compare the protein interaction interface with the ligand-protein interaction site. If there is overlap between the interface and the ligand protein binding site, this set of data is retained. In the end, by mapping the protein-protein complex to the protein-protein interaction, we ended up with 210 pairs of data as our positive sample containing 301 protein-protein complexes [22], 466 protein chains, and 282 small molecules. BLASTP algorithm was used to calculate the sequence similarity of protein-protein to PDB complexes in protein-protein interactions (parameter Settings: ties> 80, e-value < 1e-5). Ultimately, And 1212 pairs of protein-protein interactions were retained as negative samples for the study.

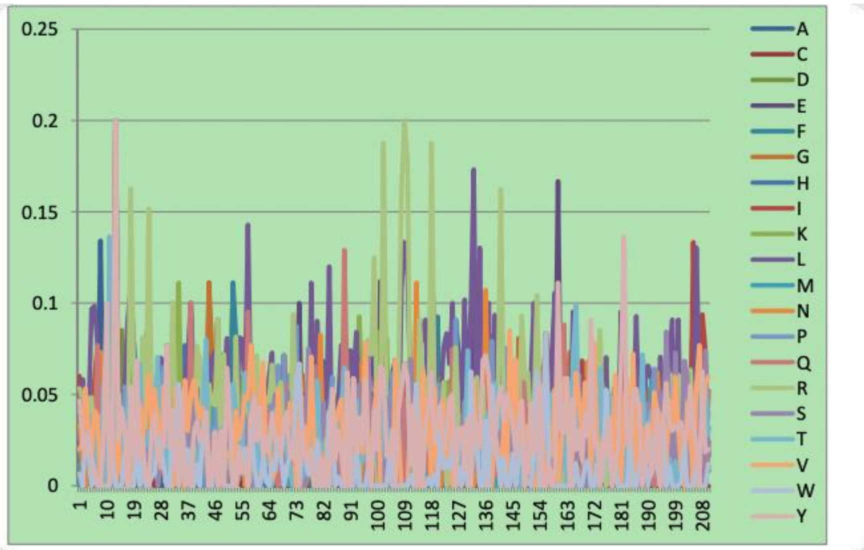

Figure 4: Frequency distribution of 20 amino acids at the protein-interaction interface in a positive sample

Build models: Build support vector machine models using Python and deep learning frameworks.

Network training: Input pre-processed positive and negative sample data into support vector machine model for training. Optimize the model parameters by minimizing the loss function. The samples were divided into training set and test set, in which the training set was composed of 70% of positive and negative samples, and the test set was composed of 30% of positive and negative samples. The training set is cross-validated by 10 times to determine the optimal parameters of support vector machine.

Model evaluation: Evaluate the trained SVM [24] model using a separate set of retained data from the test set and calculate the accuracy of the SVM model.

3.3 Experimental algorithm

We vectored the targetable protein-protein interaction $(\rho^1\rho^2)$ to f $(\rho^1\rho^2)$, where f $(\rho^1\rho^2)$ represents the characteristics of the protein-protein complex involved in the protein-protein interaction.

The characteristics we selected were: amino acid frequency characteristics; Physical and chemical characteristics; Structural characteristics; Temperature factor characteristics; Drug action protein characteristics. Jacobi similarity of the respective drug target proteins.

Formula:

$$\mathrm{ASA_{RC}} = (\mathrm{ASA_M} - \mathrm{ASA_C})/\mathrm{ASA_M} \tag{1}$$

$$\mathrm{DI_{RC}} = (\mathrm{DI_C} - \mathrm{DI_M})/\mathrm{DI_C} \tag{2}$$

$$\mathrm{PI_{RC}} = (\mathrm{PI_M} - \mathrm{PI_C})/\mathrm{PI_M} \tag{3}$$

ASA_M represents the polar surface area before protein chain binding, and ASA_{Rc} represents the relative polar surface area. DI_M represents the depth index before protein chain binding, DI_C represents the depth index after protein chain binding, DI_{Rc} represents the relative depth index. PI_M protein chain before binding the lower bulge index, PI_M protein chain after binding the lower bulge index, PI_{Rc}representing the relative bulge index.

According to the results of the prediction model, we need to establish evaluation indicators for the results. Here, we use the following indicators: accuracy, recall rate, F1 rating:

$$\text{Precison} = \frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{5}$$

$$\text{F1} = \frac{2+\text{TP}}{2*\text{TP+FP+FN}} \tag{6}$$

$$\text{Accuracy} = \frac{\text{TN+TP}}{\text{TP+FP+FN+FP}} \tag{7}$$

3.4 Practical Conclusions

In this practice, SVM and F-value feature selection methods are used to identify key feature attributes that predict the targetability of protein-protein interactions. By comparing the learning results of all features, the top 40 features of F-value ranking and the 40 feature vectors mapped by principal component analysis, it can be seen that the prediction accuracy of the top 40 features of F-value ranking is 5% lower than that of learning features, and the recall rate is 5% lower than that of learning all features.

Table 1: Sample classification results based on different characteristics

Trait	Precision	Recall	F1	Accuracy
All characteristics	0.92	0.92	0.92	0.92
Feature score (Top 40)	0.87	0.87	0.87	0.87
Feature score (Top 60)	0.90	0.90	0.90	0.90
Principal Component Analysis (Top 40)	0.87	0.88	0.87	0.88
Principal Component Analysis (top 60)	0.89	0.90	0.90	0.90

The first three F-values are drug targets acting on proteins, temperature factors of amino acids at the protein-interaction interface, hydrogen bonding at the protein-interaction interface, and hydrophobic interactions.
It can be inferred that if the protein does not have a three-dimensional structure, then based on the similar target of drug action, we can study the targeting of protein-protein interactions by analyzing the drug target of the protein and using the three-dimensional structure of the target. Greatly shorten the drug development cycle.

Compared with the traditional comparative experimental method, the SVM model using machine learning has higher accuracy, more accurate practical data and greatly shortened cycle [23-25]. Standardized high-precision algorithms and models and standardized target prediction data can explain the interactions between different drug proteins, which is more conducive to drug screening and accelerate the time of drug development. However, the research and development of traditional drugs often requires a lot of manpower and material costs and a long time of comparative practice, but the application of machine learning in medical information technology has solved such problems well.

In conclusion, machine learning research in drug targets and drug screening has significantly reduced the cost and cycle time of biological experiments. It also has the advantages of higher accuracy, efficiency, standardized reporting and personalized interpretation.

4. CONCLUSION

In conclusion, with the accumulation and availability of more and more data on the molecular structure of drugs, target activity and other properties, as well as the development of computers, powerful data mining capabilities based on machine learning are used to overcome traditional drug screening methods, which has the potential to help doctors more accurately screen drugs and target prediction, greatly improving the efficiency and progress of pharmaceutical research and development. Machine learning algorithms can learn and mine data from a large number of medical literature data, drug molecular structure, targeting activity, and protein interactions to help research and development workers better screen drugs through algorithms. Machine learning-driven drug research and development more effectively shortens the drug research and development cycle and reduces the cost of drug research and development. Improve the success rate of late drug development. Compared with traditional experimental methods used to study drug targets and their interactions, traditional research methods are both expensive and time-consuming [26-27]. Due to the multi-mapping of drug-target interactions and the confounding nature of such interactions, the research and development cost, research and development cycle and research and development accuracy are faced with great difficulties. Computational predictors at the human proteomic level through machine learning can reduce the time and cost of identifying drug targets. The SVM used in this practice is particularly useful in the identification of drug targets and the prediction of their interactions.

In the future, with the continuous progress of artificial intelligence, its application in the medical field will be more extensive and in-depth. Machine learning method is better than the method based on sequence alignment, because it takes into account the hidden similarities between features when generating prediction models, and integrates different types of features or attributes to get more accurate predictions. Developing computational methods that can accurately predict drug targets is one of the initial stages of drug development and will also help to reap benefits at a later stage. So the future with machine learning in predicting drug targets and their interactions is very bright.

ACKNOWLEDGEMENT

I would like to express my sincere thanks to Professor Zhou Yanlin for his outstanding contributions and profound insights in the fields of machine learning and bioengineering. His literature and core theories provide valuable theoretical support and inspiration for the research of this paper. Special thanks to him for his article "Utilizing AI-Enhanced Multi-Omics Integration for Predictive Modeling of Disease Susceptibility in Functioning. Phenotypes provide important reference materials for this study and key guidance for the theoretical framework and practical application of my paper. In addition, I would like to thank Professor Zhou Yanlin for his guidance and support in academic exchange and research cooperation. Thank you for your valuable advice and encouragement, which have played an immeasurable role in promoting my research work. Finally, I would like to express my heartfelt thanks and high respect to Professor Weng Huiying again, and look forward to more opportunities for academic exchanges and cooperation in the future.

In addition, the article thanks Weixiang Wan for his research contributions to machine learning and artificial intelligence related fields. During the writing process, the article consulted Professor Weixiang Wan for many times, and completed the article with his careful guidance and guidance, and Weixiang Wan also gave great help in the experiment process. Thanks to Weixiang Wan for his contributions to the academic community.

REFERENCES

[1] Wu, Yichao, et al. "LoRA-SP: Streamlined Partial Parameter Adaptation for Resource-Efficient Fine-Tuning of Large Language Models." arXiv preprint arXiv:2403.08822 (2024).

[2] Liu, Beichang, et al. "Machine Learning Model Training and Practice: A Study on Constructing a Novel Drug Detection System." International Journal of Computer Science and Information Technology 1.1 (2023): 139-146.

[3] Xu, Jingyu, et al. "Automated Scoring of Clinical Patient Notes using Advanced NLP and Pseudo Labeling." 2023 5th International Conference on Artificial Intelligence and Computer Applications (ICAICA). IEEE, 2023.

[4] Wang, Yixu, et al. "Exploring New Frontiers of Deep Learning in Legal Practice: A Case Study of Large Language Models." International Journal of Computer Science and Information Technology 1.1 (2023): 131-138.

[5] Zhou, Yanlin, et al. "Utilizing AI-Enhanced Multi-Omics Integration for Predictive Modeling of Disease Susceptibility in Functional Phenotypes." Journal of Theory and Practice of Engineering Science 4.02 (2024): 45-51.

[6] Ji, Huan, et al. "Utilizing Machine Learning for Precise Audience Targeting in Data Science and Targeted Advertising." Academic Journal of Science and Technology 9.2 (2024): 215-220.

[7] Chen, Y., Wang, S., Lin, L., Cui, Z., & Zong, Y. (2024). Computer Vision and Deep Learning Transforming Image Recognition and Beyond. International Journal of Computer Science and Information Technology, 2(1), 45-51.
[8] Qian, Wenpin, et al. "Clinical Medical Detection and Diagnosis Technology Based on the AlexNet Network Model." Academic Journal of Science and Technology 9.2 (2024): 207-211.
[9] Zeng, Q., Sun, W., Xu, J., Wan, W., & Pan, L. (2024). Machine Learning-Based Medical Imaging Detection and Diagnostic Assistance. International Journal of Computer Science and Information Technology, 2(1), 36-44.
[10] Zhang, Y., Gono, R., & Jasiński, M. (2023). An Improvement in Dynamic Behavior of Single Phase PM Brushless DC Motor Using Deep Neural Network and Mixture of Experts. IEEE Access.
[11] Wang, H., Bao, Q., Shui, Z., Li, L., & Ji, H. (2024). A Novel Approach to Credit Card Security with Generative Adversarial Networks and Security Assessment.
[12] Wu, C., Cui, J., Xu, X. et al. The influence of virtual environment on thermal perception: physical reaction and subjective thermal perception on outdoor scenarios in virtual reality. Int J Biometeorol 67, 1291–1301 (2023). https://doi.org/10.1007/s00484-023-02495-3
[13] Zhu, Mingwei, et al. "Enhancing Collaborative Machine Learning for Security and Privacy in Federated Learning." Journal of Theory and Practice of Engineering Science 4.02 (2024): 74-82.
[14] Yang, Le, et al. "Research and Application of Visual Object Recognition System Based on Deep Learning and Neural Morphological Computation." International Journal of Computer Science and Information Technology 2.1 (2024): 10-17.
[15] Qian, Jili, et al. "A Liver Cancer Question-Answering System Based on Next-Generation Intelligence and the Large Model Med-PaLM 2." International Journal of Computer Science and Information Technology 2.1 (2024): 28-35.
[16] Bao, Qiaozhi, et al. "Exploring ICU Mortality Risk Prediction and Interpretability Analysis Using Machine Learning." (2024).
[17] Cheng, Q., Gong, Y., Qin, Y., Ao, X., & Li, Z. (2024). Secure Digital Asset Transactions: Integrating Distributed Ledger Technology with Safe AI Mechanisms. Academic Journal of Science and Technology, 9(3), 156-161.
[18] Zhang, Ye, et al. "Optimizing science question ranking through model and retrieval-augmented generation." International Journal of Computer Science and Information Technology 1.1 (2023): 124-130.
[19] Xiang, Yafei, et al. "Integrating AI for Enhanced Exploration of Video Recommendation Algorithm via Improved Collaborative Filtering." Journal of Theory and Practice of Engineering Science 4.02 (2024): 83-90.
[20] Zhu, M., Zhang, Y., Gong, Y., Xing, K., Yan, X., & Song, J. (2024). Ensemble Methodology: Innovations in Credit Default Prediction Using LightGBM, XGBoost, and LocalEnsemble. arXiv preprint arXiv:2402.17979.
[21] Zhu, Mengran, et al. "Enhancing Credit Card Fraud Detection: A Neural Network and SMOTE Integrated Approach." Journal of Theory and Practice of Engineering Science 4.02 (2024): 23-30.
[22] Su, G., Wang, J., Xu, X., Wang, Y., & Wang, C. (2024). The Utilization of Homomorphic Encryption Technology Grounded on Artificial Intelligence for Privacy Preservation. International Journal of Computer Science and Information Technology, 2(1), 52-58.
[23] Akbar, A., Peoples, N., Xie, H., Sergot, P., Hussein, H., Peacock IV, W. F., & Rafique, Z. . (2022). Thrombolytic Administration for Acute Ischemic Stroke: What Processes can be Optimized?. McGill Journal of Medicine, 20(2).
[24] Zhou, Yanlin, et al. "A Protein Structure Prediction Approach Leveraging Transformer and CNN Integration." arXiv preprint arXiv:2402.19095 (2024)