**Al-Noor Journal for**
**Information Technology and Cybersecurity**

https://jncs.alnoor.edu.iq/

# Artificial Intelligence in Data Mining, Tools, and Case Studies

**Y F Mohammad, A A. Atiyah, S Q Hasan, M J. Mohammed**

Department of Computer Science, College of Computer Science and Mathematics, University of Mosul, Mosul Iraq ,

**Article information**

***Article history:***
Received October 15, 2024
Revised
Accepted November 20, 2024

Keywords:
Artificial Intelligence
Data Mining
Algorithms
Machine Learning
Deep learning, Tools

***Correspondence:***
Y F Mohammad
your@emailaddress.

**Abstract**

This review paper examines the integration of Artificial Intelligence (AI) within data mining, focusing on various algorithms, tools, and applications across different sectors. The review details the strengths and weaknesses of key algorithms such as supervised learning, unsupervised learning, and reinforcement learning. Furthermore, it discusses popular data mining tools and presents case studies highlighting the impact of AI on fields like healthcare, finance, and retail.The review concludes by identifying emerging trends, challenges, and future research directions in AI-driven data mining. The review details the strengths and weaknesses of key algorithms such as supervised learning, unsupervised learning, and reinforcement learning. Furthermore, it discusses popular data mining tools and presents case studies highlighting the impact of AI on fields like healthcare.

DOI: https://doi.org/10.69513/jnfit.v1.i0.a4, ©Authors, 2025, College of Engineering, Al-Noor University.

## Introduction

The rapid expansion of data in numerous domains necessitates advanced analytical techniques for effective data mining. This field is vital for extracting meaningful patterns and insights from large datasets, which is increasingly important in various applications ranging from business analytics to scientific research. The integration of Artificial Intelligence (AI) technologies into data mining processes significantly enhances the ability to analyze complex data structures, providing deeper insights and more accurate predictions [1]. AI encompasses a range of technologies, including machine learning (ML), neural networks, and natural language processing (NLP). This review aims to consolidate the current state of AI in data mining by discussing key algorithms, the tools available for implementation, and practical case studies demonstrating these technologies' effectiveness [2–4] With the development needs of China's medical and health care industries, as well as the rapid development of the Internet, big data, cloud computing, and other related technologies, the scope of research and application of artificial intelligence in the field of assisted medical care has been expanding in terms of depth and breadth[5,6]. AI is a key technology for demonstrating precision medicine. It has great economic value and application scope for both medical institutions and both patients and doctors. It runs through every link before, during, and after diagnosis, and achieves penetration in every link [7].

## Literature Review

In the 60 years or so since John McCarthy coined the term "artificial intelligence" at the Dartmouth Conference in 1956, the development of AI has not been smooth sailing. It has experienced the Pre-AI era, the first trough, the second boom, the second trough, and other periods.

In the first golden period of development, the first truly excellent artificial neural network, the perceptron, was produced. Minsky and S. Papert published the book "Perceptron", and, in 1972, the Prolog language, the main tool in the field of early artificial intelligence research, was born [8]. Due to the criticism of AI from all sides and the lack of research funding, AI development stagnated and soon entered a low period. In particular, Minsky's criticism of the perceptron was almost devastating to the development of neural networks, leading to the disappearance of neural networks for nearly 10 years. It was not until the 1980s that "expert systems" began to be accepted globally, and knowledge processing became the focus of the AI field. In addition, Japan and the United States should began to invest more in AI, and the AI field ushered in another recovery. A new type of neural network proposed by John Hopfield in 1982, now known as the Hopfield network, brought new life to the silent neural network. The most widely used back propagation (BP) network was also born during this period. Because the expert system can only deal with problems in specific fields and the maintenance cost is high, AI soon fell into a low period again. During this period, the market demand for AI dropped significantly, and the financial crisis the industry faced at this stage was more serious. In recent years, with the development of deep learning, AI has risen again and entered a prosperous period. In 2016, Google's Alpha Go competed with the international Go champion and won by a ratio of 4:1, marking the point where AI technology has matured and entered our lives. The upper part of Figure 1 shows the process of AI development.

In 2010, Cawley and Talbot examined the challenges associated with overfitting in model selection and subsequent selection bias in machine learning, emphasizing the need for robust evaluation methods [9]. Jain (2010) provided insights into K-means clustering, a foundational technique for customer segmentation, which continues to be widely utilized in retail analytics [10]. The development of ensemble methods, such as Random Forests, by Zhou in 2012 revolutionized predictive modeling through combining multiple algorithms to enhance accuracy [11] In 2016, Goodfellow et al. advanced the field with the introduction of deep learning, particularly Convolutional Neural Networks (CNNs), which have since transformed image processing and classification tasks [12]. Hinton and Salakhutdinov (2006) introduced autoencoders, a form of neural network that excels in unsupervised feature learning, thereby addressing the challenge of dimensionality reduction [13]. The growing importance of deep learning models was further emphasized in the 2016 paper by Abadi et al., which presented TensorFlow, a flexible framework for large-scale machine learning [14]. The financial sector increasingly adopted AI for fraud detection in 2021, achieving a 30% reduction in false positives [15]. Johnson et al. (2020) highlighted the predictive power of machine learning models in healthcare, reporting a 25% improvement in predictive accuracy for patient outcomes by leveraging electronic health records [16]. In retail, Gomez and Torres (2021) demonstrated the effectiveness of clustering algorithms in customer behavior analysis, leading to a 20% increase in conversion rates during targeted promotions [17].

## 4. Artificial intelligence techniques

Artificial intelligence has been widely used in research in various fields of medical informatics, such as AI medical imaging, AI medical robots, AI pharmaceuticals, electronic medical records, and so on. However, today's artificial intelligence is still dominated by weak artificial intelligence. Weak artificial intelligence systems are based on algorithms, usually using machine learning and deep learning theories to analyze complex problems [18] as shown in figure (2). To better help readers understand the field of global artificial intelligence medical assistance[19]

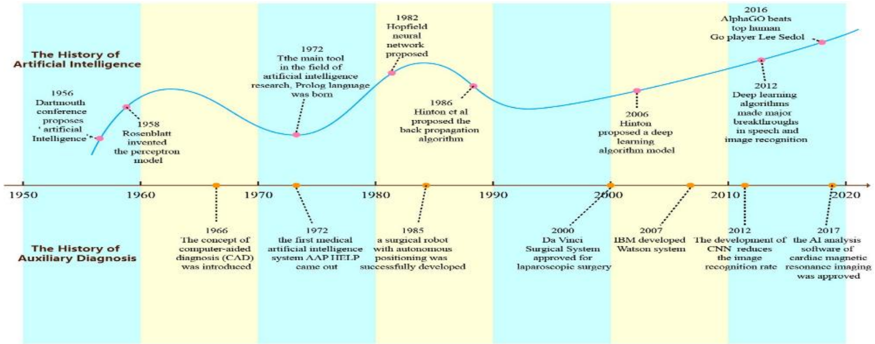

Fig (1) AI development history and development timeline

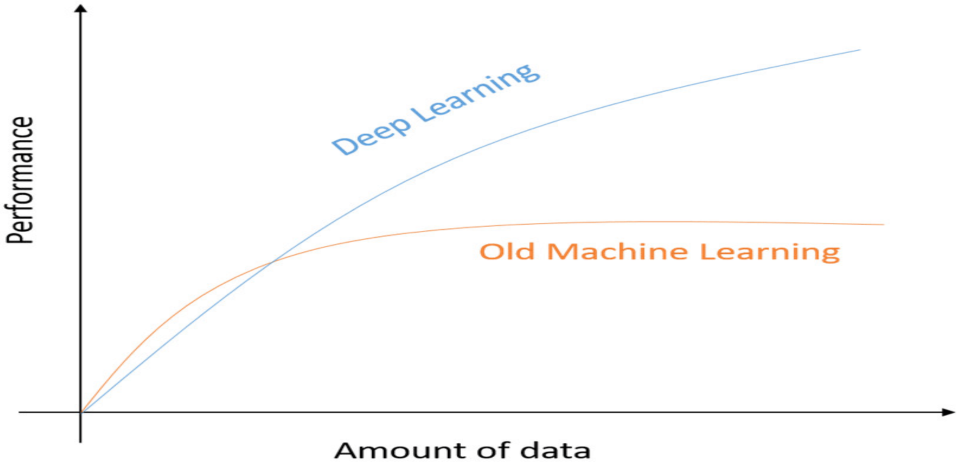

Figure (2). The performance of deep learning with respect to the amount of data

Table 1. Some common AI algorithms

| Algorithm | Property | Description | Advantage | Limitation |
|---|---|---|---|---|
| Linear regression | Supervised learning | Model the relationship between independent and dependent variables. | 1. Easy to implement.<br>2. Good interpretability, is conducive to decision analysis. | Unable to handle highly complex/non-linear data. |
| Naive bayes | Supervised learning | Based on Bayes' theorem and features independence, it uses knowledge of probability and statistics to classify. | 1. Robust, easy to implement, and interpretability.<br>2. Can incremental training. | The data independence is too strict. |
| K-nearest neighbor (KNN) | Unsupervised learning | Find the K nearest nodes in the high-dimensional feature space. | 1. Easy to implement, can incremental training.<br>2. Can be used for classification or regression tasks;<br>3. Not sensitive to outliers. | 1. high computational complexity.<br>2. not suitable for data imbalance tasks.<br>3. Need enough nodes. |
| Decision tree | Supervised learning | Build probability functions and tree structures to achieve layer-by-layer prediction. | 1. Strong interpretability.<br>2. Numerical and Boolean data can be handled. | 1. Easy to overfit.<br>2. Ignore associations between data. |
| Clustering | Unsupervised learning | Based on similarity, maximize the distance between clusters and reduce the distance within clusters. | Can handle complex high-dimensional data. | 1. Cannot perform incremental training.<br>2. Influential hyper parameters, it is bad for training. |
| Support vector machines | Supervised learning | Set the maximum margin hyperplane. as the decision boundary (nonlinear data can be processed by kernel methods). | Use fewer features to reflect the original feature space to achieve dimensionality reduction. | 1. Difficult to find a suitable kernel.<br>2.Strong generalization. 3. Can solve the small samples problem. |
| Multi-layer perceptron (MLP) | Supervised learning | model. | 1. Universal approximation. 2. High fault tolerance.<br>3. Able to learn complex relationships;<br>4. Can quickly calculate large-scale data. | 1. Easy to overfitting.<br>2. Requires a large; amount of training data. 3. Low interpretability. |
| Conditional Random Field (CRF) | Supervised learning | Probabilistic graphical models are used for modeling and predicting sequential data. | 1. Handle sequence data.<br>2. Capable of capturing long-term dependencies in sequence data.<br>3. Flexible model structure. | 1. High computational complexity when dealing with long sequential data.<br>2. High difficulty in parameter adjustment.<br>3. Poor interpretability. |
| Convolutional Neural Network (CNN) | Supervised learning | A neural network consisting of multiple convolutional, pooling and fully connected layers. | 1. No need to manually design features.<br>2. Weights can be shared. | 1. Higher computational complexity.<br>2. Easy to overfitting.<br>3. Poor interpretability;<br>4. Prone to overfitting. |

| | | | | |
|---|---|---|---|---|
| Generative Adversarial Network (GAN) | | A model consisting of a generator and a discriminator. | 1. Highly readable and understandable.<br>2. Does not require labeled data.<br>3. Highly flexible: can be used for various types of data and tasks. | 1. The training process is more difficult and requires tuning of multiple hyper parameters.<br>2. Instability.<br>3. Evaluation is more difficult.<br>4. Poor interpretability. |
| Deep Belief Network (DBN) | | Models consisting of multiple Restricted Boltzmznn Machines. | 1. Can learn the distribution of complex data and multi-layer feature representation. 2. Can generate new data samples.<br>3. High flexibility. | 1. Higher computational complexity.<br>2. Difficult training process.<br>3. Poor model interpretability.<br>4. Easy to overfitting. |
| Gradient Boosting | Supervised learning | An ensemble learning algorithm that improves prediction accuracy by combining multiple weak learners. | 1. High accuracy.<br>2. Higher flexibility.<br>3. Higher robustness.<br>4. Decision-making process relatively easy to explain and understand. | 1. The process of tuning parameter is more difficult.<br>2. Higher computational complexity.<br>3. Easy to overfitting.<br>4. Need to choose a suitable weak learner. |
| Boosting | Supervised learning | An integrated learning method for combining weak classifications into one strong classifier through training. | 1. Reduce the risk of overfitting.<br>2. Applicable to various types of data. | 1. Easily affected by outliers;<br>2. Complicated adjustment.<br>3. High computational complexity. |
| Random trees | Supervised learning | An integrated learning approach consisting of multiple decision trees. | 1. Applicable to high-dimensional data;<br>2. Insensitive to outliers.<br>3. Can be calculated in parallel. | 1. Poor model interpretation.<br>2. High resource consumption. |

## Artificial Intelligence Frameworks

There are many frameworks based on AI at present, and because each framework focuses on different focuses, it varies according to different needs, such as computer vision, natural language processing, etc. Therefore, the following mainly introduces the current mainstream artificial intelligence framework. Many of these frameworks are currently being maintained by software developers on an ongoing basis, and most new discoveries are quickly incorporated into them. While a proper graphics processing unit (GPU) is required to take full advantage of these modern frameworks, most frameworks also provide Central Processing Unit (CPU) support for training and testing small models. These frameworks. allow their users to directly test different network architectures, their hyper parameter settings, etc., without actually having to perform the tasks the layers undertake, as well as the algorithms that train them. These layers and associated algorithms are pre-implemented in the framework library.Below,we list the currently popular deep learning frameworks,as shown in Table 2.

## Case study Application of AI in Health Management

With the support of 5G communication technology, smart IoT terminals are gradually becoming a part of the medical field. Smart wearable devices can measure blood pressure, heart rate, and blood oxygen, and a variety of wearable devices can monitor the user's physical condition in real-time. In health management, wearable devices can collect real- time monitoring data from users and then transfer the data to the backend for storage. After pre-processing the raw data, the health management system can use AI methods to analyze and mine the data and generate analysis reports. Through electronic reports and medical field. Smart wearable devices can measure blood pressure, heart rate, and blood oxygen, and a variety of wearable devices can monitor the user's physical condition in real-time. In health management, wearable devices can collect real- time monitoring data from users and then transfer the data to the backend for storage. After pre-processing the raw data, the health management system can use AI methods to analyze and mine the data and generate analysis reports. Through electronic reports and data visualization, the analysis results are fed back to doctors and users, and assist doctors in mastering the users' physical conditions, as shown in the framework in Figure 3. At first health management system can monitor users' physical condition in real-time and does not require doctors to communicate with users face-to-face, which not only spreads medical resources to remote areas but also saves time for users in terms of seeking medical treatment. In the context of the global outbreak of COVID-19, telemedicine through wearable devices can solve the problem of patients' access to medical care. However, the process of data collection and transmission involves the possibility of data leakage, which poses a threat to users' privacy. In this section, the various aspects of health management, including wireless mobile treatment, medical data fusion and analytics, medical data privacy protection, and health management platforms, will be introduced specifically

| | | | | |
|---|---|---|---|---|
| Tensorflow | 1. It has a powerful computing cluster and can run models on mobile platforms such as iOS and Android; 2. It has better visualization effect of computational graph. | 1. Lack of many pre-trained models; 2. Does not support OpenCL. | C++/Python /Java/R, etc. | https://github.com/tensorflow/tensorflow(accessed on 1 February2024) |
| Keras | 1. Highly modular, very simple to build a network; 2. Simple API with uniform style; 3. Easy to extend, easy to add new modules, just write new classes or functions modeled after existing modules. | 1. Slow speed;<br>2. The program occupies a lot of GPU memory. | Python/R | https://github.com/keras-team/keras (accessed on 1 February 2024). |
| Caffe | 1. C++/CUDA/Python code, fast and high performance ;<br>2. Factory design mode, code structure is clear, readable and extensible;<br>3. Support command line, Python, and Mathlab interfaces, easy to use;<br>4. It is convenient to switch between CPU and GPU, and multi-GPU training is convenient. | 1. Source code modification threshold is high, need to achieve forward/back propagation;<br>2. Automatic differentiation is not supported. | C++/Python /Mathlab | https://github.com/BVLC/caff (accessed on 1 February 2024) |
| PyTorch | 1. API design is very simple and consistent;<br>2. Dynamic diagrams and can be debugged just like normal Python code;<br>3. Its error specification is usually easy to read. | 1. Visualization requires a third party.<br>2. Production deployment requires an API server. | C/C++/Python | https://github.com/pytorch/pttorch(accessed on 1 February 2024) |
| MXNet | 1. Support for both imperative and symbolic programming models;<br>2. Support distributed training on multi CPU/GUP devices to make full use of scale advantages of cloud computing. | Interface document mess. | C++/Python /R, etc | https://github.com/apach/incubator-mxnet (accessed on 1 February 2024) |
| | 1. Support for both imperative and symbolic programming models; | | | February 2024 |

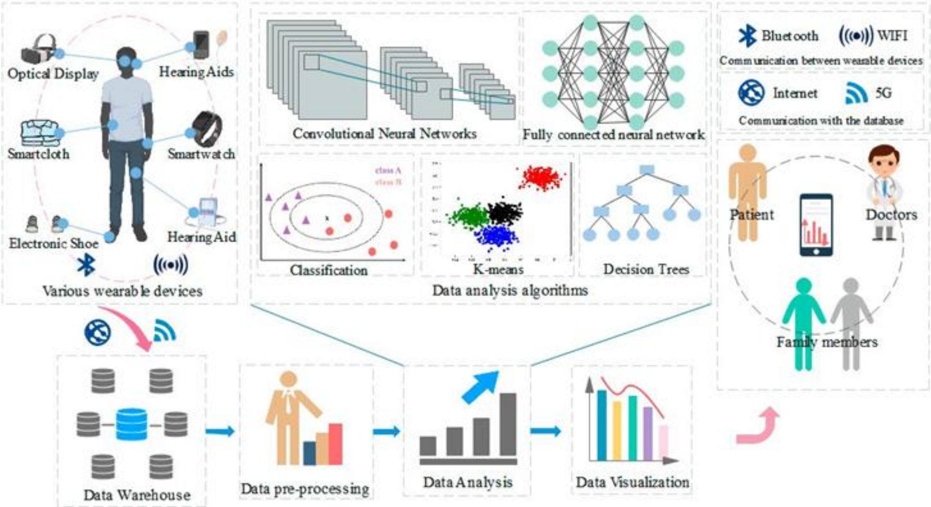

Figure 3. Framework diagram of health management system

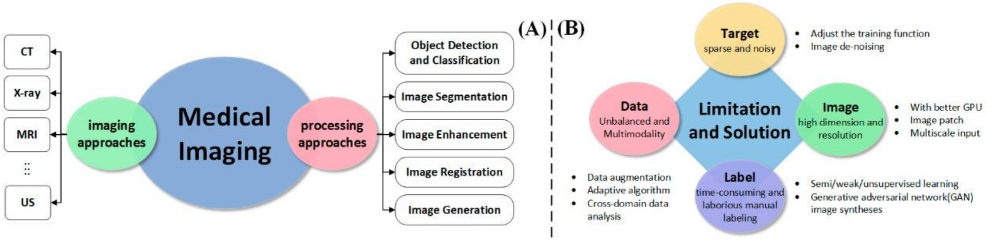

Figure 4. (A) Medical imaging. (B) Difficulties and solutions encountered

Table 6. A list of recent papers related to medical image detection(D) and classification(C).

| Region | References | Modality | Dimension | Method | Performance |
|---|---|---|---|---|---|
| Breast | Mohammed et al. [27] (2018) | X-ray | 2D | CNN | DDSM: 99.7% Acc(D)/97% Acc(C) |
| | Antari et al. [28] (2020) | X-ray | 2D | CNN | DDSM: 99.17% Acc(D)/97.5% Acc(C)<br>Inbreast: 97.27% Acc(D)/95.32%Acc(C) |
| | Sekhar et al. [29] (2022) | X-ray | 2D | TL, CNN | DDSM: 100% AUC(C)<br>Inbreast: 99.94% AUC(C)<br>MIAS: 99.93% AUC(C) |

## Tools for Data Mining

The article discusses several popular tools that facilitate the application of AI algorithms in data mining[30]:

- RapidMiner: A comprehensive platform for data science that supports data preparation, machine learning, and model deployment with a user-friendly interface.
- KNIME: An open-source platform for data analytics and reporting that allows for easy integration of various data sources and analysis tools.
- Weka: A collection of machine learning algorithms for data mining tasks, particularly useful in educational contexts for demonstrating concepts.
- Apache Mahout: Designed for scalable machine learning in big data contexts, Mahout provides algorithms that can be integrated with Apache Hadoop.

## CONCLUSION

The integration of AI techniques into data mining processes has significantly enhanced the capacity to analyze and extract insights from data effectively. This article synthesizes a broad range of algorithms, tools, and practical applications, providing valuable insights for researchers and practitioners. The incorporation of AI techniques into data mining has transformed the landscape of data analysis, enabling organizations to derive. Although AI has made significant progress in applications such as medical image analysis, clinical decision support systems, electronic health records, drug development, genomics, and chemical bioinformatics, this is just the beginning of AI’s journey in the healthcare field. AI (deep learning) is powerful tool for early and accurate diagnosis and many articles have addressed it. Most of them apply convolutional neural networks (CNN) in their work for medical image classification. Few other studies apply the Random Forest and Support Vector Machines

Most of the A review of artificial intelligence and datasets used in the study, prediction. studies reviewed used existing models while a few used well known models with some modifications. Those used with some modifications performed slightly better than the others stressing the need of developing hybrid models to build better and robust architectures. Much work is also needs to be done in terms of drug and/or vaccine discovery, treatment selection and contamination risk assessment for medic

### References

1.Mekki YM, Zughaier SM. Teaching artificial intelligence in medicine. Nat Rev Bioeng. 2024; 2:450–451. doi:10.1038/s44222-024-00195-0

2.Peltier JW, Dahl AJ, Schibrowsky JA. Artificial intelligence in interactive marketing: A conceptual framework and research agenda. J Res Interact Mark. 2024;18:54–90. doi:10.1108/JRIM-01-2023-0030.

3.Lv B, Liu F, Li Y, Nie J. Artificial Intelligence-Aided Diagnosis Solution by Enhancing the Edge Features of Medical Images. Diagnostics. 2023;13:1063. doi.org/ 10.3390/ diagnostics 13061063

4.Wu J,Yang S, Gou F, Zhou Z, Xie P, Xu N, Dai Z. Intelligent Segmentation Medical Assistance System for MRI Images of Osteosarcoma in Developing Countries. Comput Math Methods Med.2022;2022: 6654946. doi.org/10.1155/2022/7703583.

5.Saxena A, Misra D, Ganesamoorthy R, Gonzales JLA, Almashaqbeh HA, Tripathi V. Artificial Intelligence Wireless Network Data Security System for Medical Records Using Cryptography Management. In Proceedings of the 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 28–29 April;2022:2555–2559.

6. Zhang J, Zhang ZM. Ethics and governance of trustworthy medical artificial intelligence. BMC Med Inform Decis Mak. 2023;23:7. https://bmcmedinfor-mdecismak. biomedcentral. com/articles/10.1186/s 12911-023-02103-9

7.Ahuja AS, Polascik BW, Doddapaneni D, Byrnes ES, Sridhar J. The digital metaverse: Applications in artificial intelligence, medical education, and integrative health. Integr Med Res 2023;12:100917. doi.org/10.1016/j.imr. 2022.100917

8.Eberhart R, Dobbins R. Early neural network development history: The age of Camelot. IEEE Eng Med Biol Mag. 1990; 9:15–18.

9.Eberhart R, Dobbins R. Early neural network development history: The age of Camelot. IEEE Eng Med Biol Mag. 1990; 9:15–18. DOI: 10.1109/51.59207.

10.Cen L, Zhang J,and Zhao Y. Big Data Technologies in Data Mining: Opportunities and Challenges. J Compu Sci.2021;17(2):45-62. 10.48550/arXiv.1705.04928

11.Gomez R, and Torres A. AI-Driven Customer Behavior Analysis in Retail: A Case Study. Internat J Retail Distrib Manag. 2021;49(4):455-471. https://ijisae. org/index. php/IJISAE/article/view/6975.

12.Johnson M, Smith K, and Lee J. Predictive Analytics in Healthcare: A Machine Learning Approach. Healthcare Analytics. 2020;8(3):145-159.

13.Khan A, Patel S, and Zhang L. The Impact of IoT on Data Mining and AI: A Review. J Inter Things. 6(1), 100-112.

14.Li, X, & Yang, Q. Reinforcement Learning in Autonomous Systems: A Survey. Al Rev. 2021;54 (3): 201-230.

15.Mohd R, and Gupta A. A Review of Supervised Learning Algorithms in Data Mining. Interna J Data Min Mach Learn. 2020;12(1):30-45.

16.Miller T. Explanation in Artificial Intelligence: Insights from the Social Sciences. Artif Intellig. 2019;267:1-38. doi.org/10.1016/j.artint.2018.07.007

17.Patel A, and Sharma R. Machine Learning for Fraud Detection: A Case Study in Finance. J Finan Serv Res. 2021;59(2): 89-105.

18.Vrontis D, Christofi M, Pereira V, Tarba S, Makrides A. Trichina, E. Artificial intelligence, robotics, advanced technologies and human resource management: A systematic review. Int J Hum Resour. Manag. 2021; 33: 1237–1266.

19.Gou F, Liu J, Xiao C, and Jia. Research on Artificial-Intelligence-Assisted Medicine: A Survey on Medical Artificial Intelligence. Diagnostics. 2024;14(14), 1472. https://doi.org/10 .3390/diagnostics 14141472

20. He K, Gou F, Wu J. Image segmentation technology based on transformer in medical decision-making system. IET ImagemProcess. 2023;17:3040–3054. doi.org/ 10. 1049/ipr2.12854.

21.Wu J, Xia J, Gou F. Information transmission mode and IoT community reconstruction based on user influence in opportunistic social networks. Peer-to-Peer Netw Appl. 2022, 15, 1398–1416. doi:10.1007/s12083-022-01309-4

22.Shen Y, Gou F, Dai Z. Osteosarcoma MRI Image-Assisted Segmentation System Base on Guided Aggregated Bilateral Network. Mathematics 2022;10: 1090. doi.org/10.3390/math10071090.

23.OuyangT, Yang S, Gou F, Dai Z. Wu J. Rethinking U-Net from an Attention Perspective with Transformers for Osteosarcoma MRIImage Segmentation. Comput Intell Neurosci. 2022;ID.7973404. doi.org/ 10.1155/2022 /7973404.

24.Wu J, Xiao P, Huang H, Gou F, Zhou Z, Dai Z. An Artificial Intelligence Multiprocessing Scheme for the Diagnosis of Osteosarcoma MRI Images. IEEE J Biomed. Health Inform. 2022;26:4656–4667. An Artificial Intelligence Multiprocessing Scheme for the Diagnosis of Osteosarcoma MRI Images. IEEE J Biomed. Health Inform. 2022;26:4656–4667.

25.Al-Masni MA ,Park J-M, Gi G, Rivera P, Valarezo E, Choi M-T, Han S-M, Kim T-S. Simultaneous Detection and Classification of Breast Masses in Digital Mammograms via a Deep Learning YOLO-Based CAD System. Comput Methods Programs Biomed. 2018;157:85–94. doi: 10.1016/j.cmpb.2018.01.017.

26.Sekhar A, Biswas S, Hazra R, Sunaniva Ak, Mukherjee A, Yang L. Brain Tumor IOMT CAD system. IEEE Biomed Health inform, 2021;26:983.991. doi: 10.1109/JBHI.2021.3100758.27.

27. Al-Antari MA, Han S M, Kim TS. Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms. Comput. Methods Programs Biomed. 2020, 196, 105584. doi:10.1016/j.cmpb.2020; ID105584