SDIRSACR

Official Journal of the Serbian Association for Cancer Research *001.89:004.8*

Radulovic M. Oncol Insights 2024 No. 2
Manuscript received on: Sep 09, 2024.
Manuscript revised on: Nov 15, 2024.
Manuscript accepted on: Nov 19, 2024.

Oncology Insights

SDIR

PERSONAL POINT OF VIEW ***Open access***

The Point of View: Will Artificial Intelligence Soon Become the Main Driver of Scientific Innovation?

Marko Radulović[1]

[1]*Department of Experimental Oncology, Institute of Oncology and Radiology of Serbia, Belgrade, Serbia*
**Correspondence: marko@radulovic.net; phone: +381 65 460 1082*

Abstract

In the pre-AI world, scientific progress was incremental, with gradual discoveries leaving many questions unanswered. Is rapidly advancing AI positioned to solve all the remaining questions within the next several decades? Only two years ago, writing an original essay seemed uniquely human. Then ChatGPT emerged, demonstrating that AI can produce human-like essays. Now, we wonder: How far will this go? What will AI be able to accomplish in the future?

This perspective article discusses a fundamental and imminent shift in the way science is conducted, as research is set to become the first intellectual activity largely transitioned from humans to machines. This is because all intellectual activities require reasoning, creativity and imagination, while science uniquely demands a comprehensive knowledge of the vast and ever-growing body of scientific data as a foundation for effective experiment planning. Humans, however, are inherently limited in their capacity to acquire and retain such extensive knowledge. This is precisely where AI is expected to gain a decisive advantage over humans, as it can rapidly assimilate information.

However, full advantage of the future AI ability to brainstorm through the entire scientific knowledge can be reached only when its reasoning abilities at least reach human intellect. Such convergence of comprehensive scientific knowledge and more advanced reasoning is projected to create a superior AI scientist, likely by the end of this decade, when AI is expected to match human-level reasoning. Moreover, AI is even expected to surpass human reasoning ability in the following decade, further emphasizing its superiority in scientific discovery and innovation.

As AI is poised to rapidly permeate scientific research, this article not only presents a point-of-view but also introduces the fundamental concepts of artificial intelligence that all scientists, regardless of their field, urgently need to master.

What is Artificial Intelligence?

In the pre-AI world, scientific progress was incremental, with gradual discoveries leaving many questions unanswered. Is rapidly advancing AI positioned to solve all the remaining questions within the next several decades? Only two years ago, writing an original essay seemed uniquely human. Then ChatGPT emerged, demonstrating that AI can produce human-like essays. Now, we wonder: How far will this go? What will AI be able to accomplish in the future?

Understanding the current stage of AI development is crucial for get a clear grasp of its present capabilities and future potential. AI evolution is usually conceptualized into three major stages [1]:

1) Weak AI: Performs only some tasks better and faster than humans. This type of AI is already available (Gemini, ChatGPT, Copilot, xAI, Grok).
2) Strong AI: Capable of performing intellectual tasks as well as a human, including reasoning, critical thinking, creativity, adaptability and understanding complex concepts. This stage has not yet been achieved but is expected by the end of this decade.

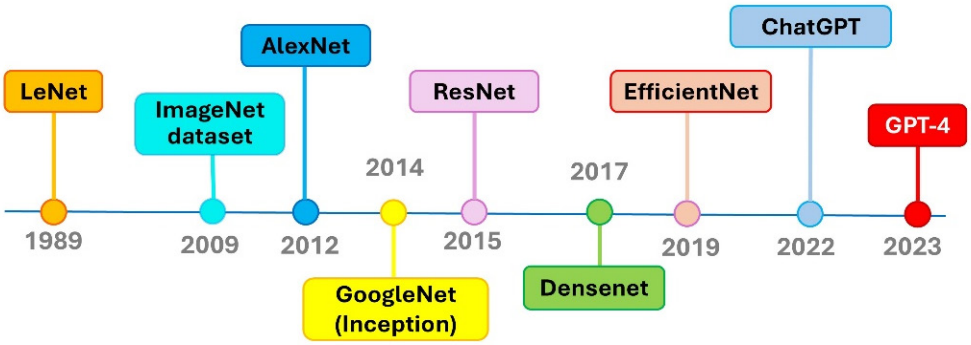

Figure 1. A brief history of ANNs.

3) Super AI: Surpasses human intelligence in all areas and capable of experiencing emotions and desires. This is a more futuristic concept projected to emerge within next decade or beyond.

Currently, narrow or weak AI is the only available form of artificial intelligence. Despite being in its early stages, AI is becoming a clinically valuable tool [2, 3] with the U.S. Food and Drug Administration (FDA) having already authorized 950 AI/ML-enabled medical devices, mainly in radiology (https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices).

What are Artificial Neural Networks and what is behind their rapid advancement?

Artificial Neural Networks (ANNs) are computational models inspired by the neural circuitry of the brain. Although the renewed interest in ANNs is very recent, this concept has been in development since Frank Rosenblatt's work in 1958 [5]. In 1989, Yann LeCun et al. already introduced the first commercially applied convolutional neural network (CNN) named LeNet, for use in recognizing handwritten postal codes on envelopes [6].

The release of ImageNet dataset containing 15 million high-resolution images for CNN training, in 2009 further fuelled the advancement of CNNs [7]. AlexNet, which won the 2012 ImageNet competition, demonstrated the power of deep CNNs and GPU acceleration, sparking widespread interest in deep learning[8]. Despite consisting of only eight layers, five convolutional and three fully connected, it greatly outperformed other image recognition models (Figure 1).GoogleNet (Inception), introduced in 2014, featured novel Inception modules that optimized the computational efficiency of deep networks.In 2015, ResNet introduced residual learning, which allowed for the training of extremely deep networks by addressing the vanishing gradient problem (Figure 1). DenseNet (2017) further improved this architecture by connecting each layer to every other layer, enhancing feature reuse and reducing the number of parameters. EfficientNet (2019) utilized a compound scaling method to balance network depth, width and resolution, achieving state-of-the-art performance with less resources (Figure 1). ChatGPT (2022) and GPT-4 (2023), based on the novel transformer architecture,revolutionized natural language processing (NLP), enabling more sophisticated conversational AI and pushing the boundaries of what AI can achieve in understanding and generating human language (Figure 1).

The rapid development of ANNs was primarily driven by advances in graphical processing unit (GPU) hardware, which was originally designed for efficient processing of computer graphics and images. GPUs conveniently provided the massive computing power required for ANN training due to their architecture, which is optimized for parallel processing. Unlike CPUs, which have only several large cores and are thus better suited for sequential processing, GPUs feature thousands of smaller cores that are capable of handling multiple tasks simultaneously. This parallel processing capability makes GPUs ideal for large-scale matrix operations and tensor calculations in deep learning. Additionally, GPUs have higher memory bandwidth, enabling them to process vast amounts of data more effectively. Before the use of GPUs, training on large image datasets was prohibitively time-consuming and limited to low-resolution images, and large language models (LLMs), such as ChatGPT, were not feasible [9]. AI has primarily focused on text and image inputs, though applications in sound processing are also emerging. LLMs, like ChatGPT, are trained on billions of words from sources such as articles, books, and internet-based content.

What are the main ANN architectures?

Artificial Neural Networks (ANNs) consist of multiple layers, each containing numerous nodes that mimic biological neurons. These nodes are connected to input features and other nodes through weights mimicking neuron's dendrites

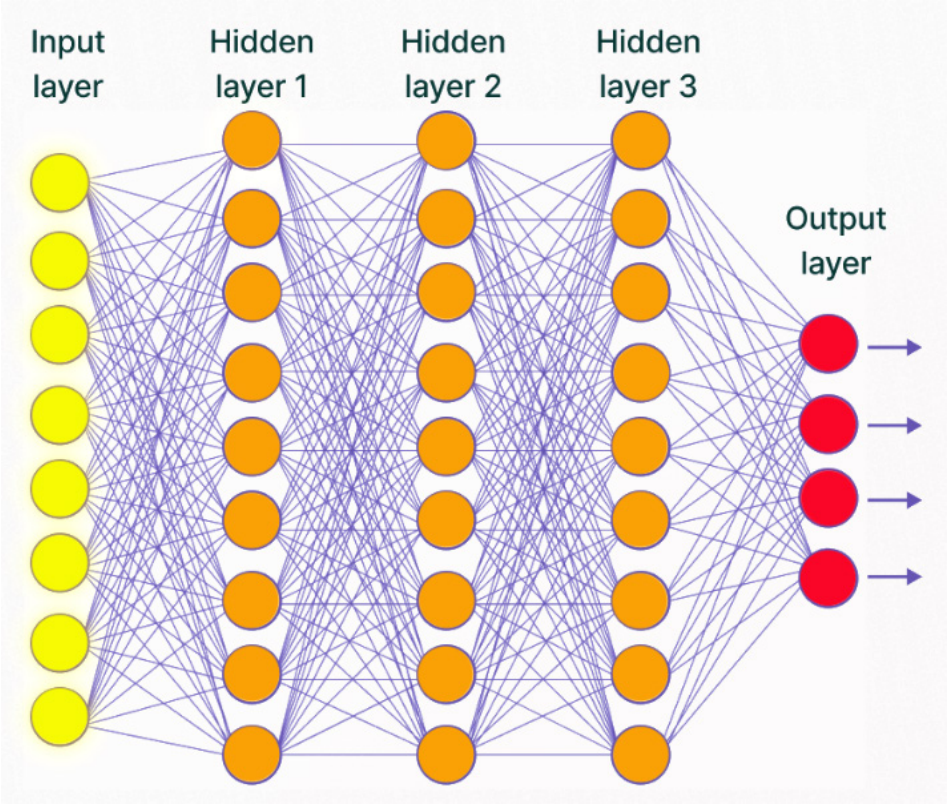

Figure 2. General ANN architecture

and synapses. ANNs typically include an input layer, one or more hidden layers, and an output layer (Figure 2).
Input features are passed from the input layer to the hidden layers (Figure 2), where they are modified by weights, biases, and activation functions (Figure 3).
Each node functions as a simple processing unit, receiving inputs, modifying them and producing an output. The number of hidden layers and neurons can vary, resulting in different architectures, ranging from shallow networks with a single hidden layer to deep learning models with many hidden layers [10].
The dominant contemporary ANN architectures are:

- **Transformer**: used by LLMs, such as ChatGPT
- **Convolutional**: used for image analysis and classification
- **Diffusion**:Used for image generation

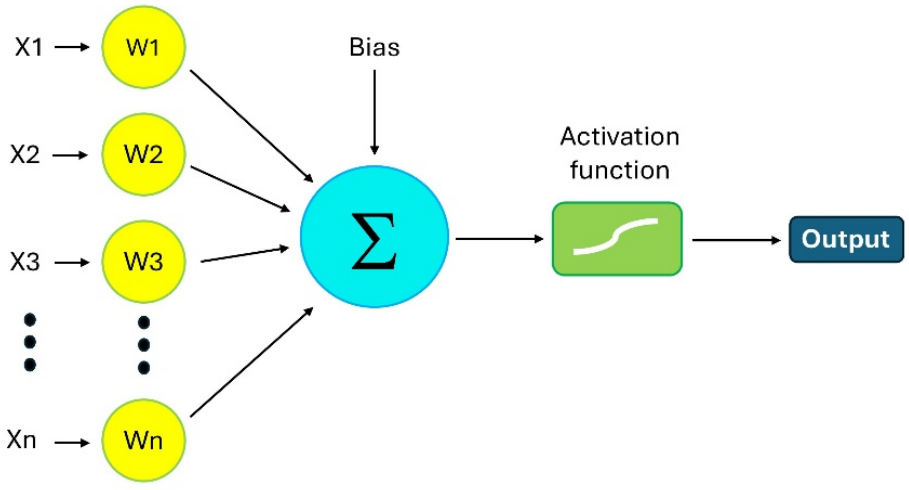

Figure 3. Schematic representation of a Neuron in a Neural Network: This diagram illustrates the functioning of a neuron, depicting inputs x, weights w, bias b, and activation function. The inputs xare multiplied by corresponding weights w and summed together along with the bias b. The resultant weighted sum is then passed through the activation function introduce non-linearity and produce the neuron's output. This process enables the neuron to learn and model complex patterns from the data.

Transformers are designed for sequential data processing, particularly forNLP which is a subfield of artificial intelligence that enables computers to understand and communicate in human language. Sequential data processing is crucial for NLP because natural language itself is inherently sequential. Another essential component of transformer-based models is the self-attention mechanism which determines the relative importance of different words in a sentence. The transformer architecture consists of multiple layers of encoders and decoders. GPT-4, for instance, has a total of 120 layers. First introduced in 2017, transformers gained prominence with the release of ChatGPT by OpenAI in November 2022[11].

Convolutional Neural Networks are primarily applied to image-related tasks such as object detection, image classification, segmentation and generation. They have a hierarchical architecture composed of three types of layers, starting with convolutional layers, which extract and learn local features from the input image; pooling layers, which reduce dimensionality; and fully connected layers, which integrate the features to produce the final output [12]. This architecture enables CNNs to learn increasingly abstract features at each layer, from simple edges in the initial layers to more complex patterns in the deeper layers.

Diffusion models gained prominence in image generation with models like DALL-E in 2021. They typically use U-Net architectures with encoder-decoder structures for the denoising processgenerating high-quality images from random noise. Their clinical application includes tumor segmentation in medical images[13].

Vision-Language Models (VLM) allow users to upload an image and engage in textual conversations with the model,giving instructions or asking questions based on the image. VLMs typically combine CNN architectures for processing images and transformer architectures for text processing[14].

How does ANN learn?

Learning in neural networks is based on weights and biases assigned to each node, following this formula[15]:

$$z= x_1w_1+ x_2w_2+ \ldots x_nw_n+b \quad (1)$$

z = weighted sum
x = input values
w = weights or “parameters”

The weighted sum (z) depends on input values (x1 and x2), which are either features from the dataset or outputs from neurons in the previous layer. The weights, w1 and w2, are parameters that the model learns during training, while the bias, b, is added to the weighted sum of inputs for each node. This bias allows the node to activate even when the weighted sum of the inputs is zero. Further downstream, the activation function introduces non-linearity to the output signal, enabling the network to learn complex patterns. This non-linearity is essential because linear models can only capture linear relationships, which are insufficient for solving real-world problems that predominantly involve non-linear patterns.

Learning in ANNs is achieved through the training process, which begins with small random values for weights and biases— the key trainable parameters. These values are gradually adjusted to minimize the error between the predicted and desired outputs, using optimization algorithms that involve backpropagation and gradient descent[16].

What can ANNslearn?

ANNs can be trained with text, images or sound[16].

1) Text: remarkably, “labelled” text, structured as pairs of questions and answers allows for most efficient training, while human-like learning of unlabelled text is less efficient for ANNs. This is because the question-answer format allows the network to compare its predicted answer with the correct answer, calculate its error and adjust its parameters during several epochs to minimize that error. Once trained, the network should generalize to new, unseen data, enabling it to respond accurately to differently formulated questions.
2) Images: Training with imagesalso involves using pairs of images and their descriptions, allowing the network to learn to recognize and interpret image content.

The initial training of a large ANN requires immense hardware resources. For example, GPT-4 was trained on 25,000 Nvidia A100 GPUs simultaneously for approximately 100 days. Each A100 GPU has 80 GB of VRAM and costs $15,000. GPUs were used instead of CPUs because the A100 GPU has 6,912 cores, whereas the most powerful CPU in 2024 (Xeon)

has only 288 cores. Although GPU cores are smaller, their large quantity results in significantly higher performance for AI training and inference compared to CPUs, especially when the trained model is used to make predictions based on new data. Inference refers to the process of using a trained machine learning model to make predictions or decisions based on unseen data.

ANNs in Science: Transfer learning using Pretrained networks

Because ANNs need thousands or millions of examples to efficiently learn,they are mainly applied in science through "transfer learning"[17].This technique uses the knowledge a model has learned from one task as a starting point for a second, related task. Transfer learning provides a significant head start, enabling the application of ANNs to experiments with small sample sizes by transferring pre-acquired knowledge to new tasks. For example, a CNN pre-trained on the ImageNet dataset, which contains millions of general images, can be further trained with only 100+ specific medical images. Transfer learning is crucial in fields like cancer research, where assembling patient cohorts with thousands of cases is difficult.

Expansion of a networkknowledgecan be achieved by learning through:

1) **Fine-tuning**
2) **Embeddings**

Pre-trained networks are typically trained on broad, general knowledge, while researchers often need detailed, specialized knowledge for specific fields, such as oncology. The ability to expand a network's knowledge is of critical practical importance because it allows scientists to take a pre-trained language model (LLM) like GPT-4, or freely available models like LLAMA, MISTRAL, FALCON, or BERT, and feed it with expertise in their specific area of interest. Such acquisition of the large amount of specialised knowledge by LLMs has a potential to finally integrate the scientific knowledge, which isof crucial importance because the scientific knowledge is highly fragmented due to the fact thatthe rate of scientific publication is expanding at a high speed while the human ability of information intake rate cannot be improved. The expertise of a pre-trained network can be expanded by inputting PDFs of research articles, either through fine-tuning the model on this data or by using embeddings:

Fine-tuningis a technique within transfer learning which takes a pre-trained model and continues its training process on a new dataset. Rather than training the entire model from scratch, it starts with the pre-trained weights thatalready incorporate general reasoning, logic, knowledge, language and conversational skills,and adjusts them to fit the new data.This process typically involves unfreezing some layers of the pre-trained model and training them alongside any newly added layers.Fine-tuning is computationally very intensive, requiring large hardware resources, especially for deep learning models with many layers and parameters, as it involves updating the neural network's weights[18].

Embeddings provide a simple, fast method to expand a neural network's pre-trained knowledge without extensive re-training and massive hardware resources[19].More than a basic database, embeddings capture semantic relationships, allowing for easy retrieval or clustering of similar data based on vector representations. They transform high-dimensional data, such as words, into lower-dimensional vectors that still retain semantic meaning and relationshipsbetween words. For example, words are considered high-dimensional because if a vocabulary contains 10,000 words, each word would be represented as a 10,000-dimensionalvector, with only one value being 1 and the rest 0. Embeddings reduce this to 50-300 dimensions, making text analysis more efficient for machine learning models. Consequently, embeddings serve as an efficient contextual database for machine learning applications and can also represent images as continuous numerical vectors that capture key image features.

Continuous Learning

Contrary to popular belief, even the latest models like GPT-4 do not have the ability for continuous learning. Once trained and deployed, these models do not update their knowledge or learn from new inputs in real-time. Any updates or improvements require a separate fine-tuning process on new data in an additional training phase.

Applications of ANNs in Oncology Research

The most common use of AI in science over the past decade has been in supervised learning, where a model is "trained" on retrospective data that has been annotated with the correct answers and then expected to generalize by providing correct predictions on new, unseen data. The main goal is to use available data to capture prognostic clues that can reliably

predict future events. Specifically, in cancer research, AI is primarily applied to tumour detection and the prognosis of chemotherapy response and disease outcomes [20]. These tasks are typically achieved through computational AI analysis of medical images, including tumour histopathology, MRI, CT scans, X-rays, and PET scans[17, 21].Improving disease outcome prognosis and therapy response prediction is expected to enhance patient survival by personalizing treatment plans. Additionally, AI may assist surgeons in planning complex oncological surgeries by providing detailed 3D models of tumours. AI's role could also extend to drug discovery and development through "virtual compound screening," which predicts the biological activity of chemical compounds against cancer targets [22].Another application is molecular docking, where AI models drug-target interactions to optimize drug design[23].

Monitoring data from wearable devices using historical and real-time activity/vitalmetrics to detect early signs of cancer progression is an obvious application of AI[24]. Wearables have been successfully used for early detection of conditions such as COVID-19 [25], heart failure [26] and seizures in epilepsy patients [27]. Common physiological metrics measured include blood pressure, oxygen saturation, body temperature, heart rate, and respiratory rate. Additionally, devices can monitor galvanic skin response, blood glucose levels, electromyography, electroencephalography and electrocardiography. Physical activity and sleep patterns are often tracked using accelerometer-derived data, with several studies indicating that low physical activity correlates with worsening cancer symptoms [28]. When combined with gyroscopes and magnetometers, accelerometers form an inertial motion unit (IMU), which can detect changes in walking style and posture, both valuable for identifying cancer-related deterioration. IMUs also assist LED optical photoplethysmography sensors in heart rate monitoring, which detect blood volume changes in the microvascular bed of tissues. AI could thus support decision-making by flagging potential early signs of cancer during routine check-ups.

The Point of View - Will AI Soon Become the Main Driver of Scientific Innovation?

The AI applications in clinical practice and medical research mentioned above are anticipated, as they merely enhance existing computational methods that did not use AI. However, a truly surprising and groundbreaking application would be AI's independent role in the most creative aspects of research—planning experiments and formulating novel approaches that could enable and accelerate scientific breakthroughs.

Effective and innovative scientific experiment planning depends on several key factors: robust reasoning and analytical skills, combined with a comprehensive understanding of the scientific literature. Although the common sense and logical reasoning abilities of even the most advanced AI models, like GPT-4, currently fall short of those of an average human, this gap is expected to disappear in the coming years. With each new generation, AI models are trained on exponentially larger datasets and demonstrate progressive improvements in reasoning capabilities[29].

Even more staggering is the fact thatAI will surpass humans by billions-foldin the above crucial aspect:reading and memorizing of scientific articles. AI can already learn from scientific text millions of times faster than humans, the only problem is that scientific articles are not available in the form of clean text but PDF files which contain scientific text mixed with information about authors,publisher,legal formulations, page numbers, acknowledgements, footnotes, tables and figures. Extracting clean scientific text from this mixed format remains surprisingly difficult, even for powerful AI models.

Even more astonishing is that AI will surpass humans by billions-fold in a critical area: reading and memorizing scientific articles. AI can already learn scientific texts millions of times faster than humans. However, a significant challenge remains: scientific articles are typically available in PDF format, which combines scientific content with extraneous information—such as author details, publisher information, legal text, page numbers, acknowledgments, footnotes, tables and figures. Extracting clean scientific text from this mixed format is surprisingly difficult, even for advanced AI models.

This challenge in comprehending scientific articles is further compounded by AI's limited ability to interpret data in tables and even more so in figures. Additionally, ANNsstruggle with unsupervised learning directly from raw text; instead, text often needs to be converted into a "question-answer" format for supervised learning[30]. Although AI can perform this transformation quickly, there is a risk that the original meaning may be altered in the process, requiring human verification to ensure accuracy.

In unsupervised learning, the input data does not include the desired output, requiring the network to independently identify meaningful patterns[31].Advances in self-supervised learnin, where models learn from raw data, will greatly improve AI's ability to learn scientific information efficiently [32].Once AI becomes proficient in unsupervised learning, including extracting scientific content from research paper PDFs and interpreting tables and figures, it will inevitably surpass humans in the amount of scientific knowledge it can acquire by several additional orders of magnitude, ultimately exceeding human capacity for creative scientific thinking [33].

This shift of science away from humans may come as a surprise to most scientists who were not aware of the impact of comprehensive literature knowledge as a basis for scientific innovation.

Conclusion

AI is positioned to enhance both scientific research and the medical treatment of cancer patients. The main current and potential applications of AI in cancer research include:

- Improving early tumour detection and diagnosis through the analysis of tumour imaging, genomic data, real-time body activity monitoring, and comprehensive integration of medical data.
- Providing more accurate cancer prognosis and predicting chemotherapy outcomes, enabling personalized treatment plans and thus improving patient care and survival.

- Advancing scientific understanding far beyond human capabilities, based on AI's ability to quickly and efficiently learn from the vast and growing body of published scientific literature. This is crucial as the rate of scientific publication continues to exponentially accelerate, while human information processing capacity remains limited.

- Acceleration of scientific breakthroughs by planning innovative experimental strategies based on AI's broad and detailed knowledge of scientific literature, combined with its analytical and reasoning skills.

References:

1. Holl, C., The content intelligence: an argument against the lethality of artificial intelligence. Discov. Artif. Intell., 2024. 4(1).
2. Yang, D., et al., Advances in artificial intelligence applications in the field of lung cancer. Front Oncol, 2024. 14: p. 1449068.
3. Jiang, B., et al., Deep learning applications in breast cancer histopathological imaging: diagnosis, treatment, and prognosis. Breast Cancer Res, 2024. 26(1): p. 137.
4. Kolla, L. and R.B. Parikh, Uses and limitations of artificial intelligence for oncology. Cancer, 2024. 130(12): p. 2101-2107.
5. Rosenblatt, F., The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev, 1958. 65(6): p. 386-408.
6. LeCun, Y., et al., Backpropagation applied to handwritten zip code recognition. Neural Comput., 1989. 1(4): p. 541-551.
7. Deng, J., et al., ImageNet: A large-scale hierarchical image database, in 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009, IEEE.
8. Krizhevsky, A., I. Sutskever, and G.E. Hinton, ImageNet classification with deep convolutional neural networks. Commun. ACM, 2017. 60(6): p. 84-90.
9. Nickolls, J. and W.J. Dally, The GPU Computing Era. IEEE Micro, 2010. 30(2): p. 56-69.
10. Guo, Y., et al., Deep learning for visual understanding: A review. Neurocomputing, 2016. 187: p. 27-48.
11. Sanderson, K., GPT-4 is here: what scientists think. Nature, 2023. 615(7954): p. 773.
12. Ibrahim, R. and M.O. Shafiq, Explainable Convolutional Neural Networks: A taxonomy, review, and future directions. ACM Comput. Surv., 2022.
13. Dong, Y. and K. Gong, Head and neck tumor segmentation from [(18)F]F-FDG PET/CT images based on 3D diffusion model. Phys Med Biol, 2024. 69(15).
14. Zhang, J., et al., Vision-Language Models for vision tasks: A survey. IEEE Trans. Pattern Anal. Mach. Intell., 2024.

46(8): p. 5625-5644.

15. Han, S.H., et al., Artificial Neural Network: Understanding the Basic Concepts without Mathematics. Dement Neurocogn Disord, 2018. 17(3): p. 83-89.

16. Alzubaidi, L., et al., Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data, 2021. 8(1): p. 53.

17. Jiang, X., et al., Deep Learning for Medical Image-Based Cancer Diagnosis. Cancers (Basel), 2023. 15(14).

18. Thirunavukarasu, A.J., et al., Large language models in medicine. Nat Med, 2023. 29(8): p. 1930-1940.

19. Dai, B., X. Shen, and J. Wang, Embedding Learning. J Am Stat Assoc, 2022. 117(537): p. 307-319.

20. Maier, H.R., et al., Exploding the myths: An introduction to artificial neural networks for prediction and forecasting. Environmental Modelling & Software, 2023. 167: p. 105776.

21. Avanzo, M., et al., The Evolution of Artificial Intelligence in Medical Imaging: From Computer Science to Machine and Deep Learning. Cancers (Basel), 2024. 16(21).

22. Lv, Q., et al., Artificial intelligence in small molecule drug discovery from 2018 to 2023: Does it really work? Bioorg Chem, 2023. 141: p. 106894.

23. Bande, A.Y. and S. Baday, Accelerating Molecular Docking using Machine Learning Methods. Mol Inform, 2024. 43(6): p. e202300167.

24. Birla, M., et al., Integrating AI-driven Wearable Technology in Oncology Decision Making: A Narrative Review. Oncology, 2024.

25. Cheong, S.H.R., et al., Wearable technology for early detection of COVID-19: A systematic scoping review. PrMed, 2022. 162: p. 107170.

26. Singhal, A. and M.R. Cowie, The Role of Wearables in Heart Failure. Curr Heart Fail Rep, 2020. 17(4): p. 125-132.

27. Verdru, J. and W. Van Paesschen, Wearable seizure detection devices in refractory epilepsy. Acta Neurol Belg, 2020. 120(6): p. 1271-1281.

28. Ortiz, B.L., Data Preprocessing Techniques for Artificial Learning (AI)/Machine Learning (ML)-Readiness: Systematic Review of Wearable Sensor Data in Cancer Care. JMIR Mhealth Uhealth, 2024.

29. Hao, S., et al., Reasoning with language model is planning with world model, in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023, Association for Computational Linguistics.

30. Orescanin, M., et al., Editorial: Deep learning with limited labeled data for vision, audio, and text. Front Artif Intell, 2023. 6: p. 1213419.

31. Zhao, T., et al., Deep Bayesian Unsupervised Lifelong Learning. Neural Netw, 2022. 149: p. 95-106.

32. Nadif, M. and F. Role, Unsupervised and self-supervised deep learning approaches for biomedical text mining. Brief Bioinform, 2021. 22(2): p. 1592-1603.

33. Kariri, E., et al., Exploring the advancements and future research directions of artificial Neural Networks: A text mining approach. Appl. Sci. (Basel), 2023. 13(5): p. 3186.