Artificial Intelligence: Learning and Limitations

ALISSON PAULO DE OLIVEIRA[1] HUGO FERREIRA TADEU BRAGA[2]

Graduate Program in Metallurgical, Materials, and Mining Engineering
Universidade Federal de Minas Gerais[1]
Av. Antônio Carlos, 6627 - Campus da UFMG – Pampulha, Escola de Engenharia, BRAZIL.

Innovation Center[2]
Fundação Dom Cabral
Avenida Princesa Diana, 760, Alphaville, Lagoa dos Ingleses, BRAZIL.

Received: June 5, 2020. Revised: July 17, 2020. Accepted: July 21, 2020. Published: July 22, 2020.

Abstract: Artificial Intelligence, IA, is a new technology with enormous potential to change the world forever as we know it. It finds applications in many fields of human activity, including services, industry, education, social networks, transportation, among others. However, there is little discussion about the accuracy and reliability of such technology, which has been used in situations where human life depends on its decision-making process, which is the result of its training, one of the stages of development. It is known that the learning process of an Artificial Intelligence, which can use the Artificial Neural Networks technology, presents an error of the predicted value in relation to the real value, which can compromise its application, being more critical in situations where the user's security is a major issue. In this article, we discuss the main technologies used in AI, their development history, considerations about Artificial Neural Networks and the failures arising from the training and hardware processes used. Three types of errors are discussed: The Adversarial Examples, the Soft Errors and the Errors due the lack of Appropriate Training. A case study associated with the third type of error is discussed and actions based on Design of Experiments are proposed. The objective is to change the way the AI models are trained, to add some rare conditions, and to improve their ability to forecast with greater accuracy in any situation.

Keywords: Artificial Intelligence; Artificial Neural Networks; Deep Learning; Machine Learning; Adversarial Examples; Soft Errors.

1. Introduction

The Artificial Intelligence, AI, is a technology related to the ability of a digital computer to perform tasks commonly associated with intelligent beings. The term is often applied to systems development projects supported with human-characteristic technological processes, such as reasoning skills, meaning discovery, or learning from past experiences. Thus, such technology can replicate human behavior in matters related to the interpretation of situations (duly converted into data) and subsequent decision based on the generation of information. Moreover, this technology proves to be able to predict future situations from past experiences. However, the accuracy (or error obtained from an actual value) is far from being desired for critical applications where the user safety is a key factor [1].

Today, there is a boom in the development of a variety of artificial intelligence applications, from image recognition to even self-driving cars, to the suggestion of medical treatments. Unfortunately, however, there are cases of failure of an autonomous driving system where apparently it was not unable to identify when to stop the and this lead to an accident with loss of life [2], a case where an AI system suggested a wrong treatment for a cancer patient what could worsen his condition [3], an AI chatbot assuming unwanted "personalities" and assuming the worst tendencies from the Internet [4], among others. Some types of errors (Adversarial examples, soft errors and errors due the lack of appropriate training) are discussed. This paper focus, as major contribution, on a specific type of error originated from the lack of appropriate training. This error is discussed with a real problem faced during the development of a prediction model of steel properties. It is known [5, 6] that the error in the training of artificial intelligence algorithms is reduced with large availability of data regarding a situation where an AI with minimal error is desired. So, necessarily, regions with little existing data lead to poor AI learning with considerable response errors. This article seeks to explore the learning mechanism on which Artificial Intelligence is based and its main limitations. A relationship between the error magnitude and the amount of data available to training is also discussed and possible mitigation actions are suggested with the objective to reduce this error by designing an appropriate experiment to generate data with low frequency.

2. Overview of Artificial Intelligence Technologies

2.1. History of Artificial Intelligence

Recently, fast progress has been made in the fields related to Neuroscience and Artificial Intelligence (AI). At the beginning of the computer age, work on artificial intelligence was interrelated with neuroscience and psychology, and many of the pioneer authors traveled both fields, with collaborations between these disciplines proving to be very productive [7]. More recently, however, this interaction has become much less common, as both subjects have grown a lot and disciplinary boundaries have solidified. The essence of Artificial Intelligence techniques in engineering problem solving is learning by presenting real examples of input data (Example: Images captured on autonomous steering systems) and outputs (Example: Steering angle, braking and acceleration at autonomous driving systems) presented to them in such a way that subtle functional relationships between data are captured, even if the underlying relationships are unknown or the physical meaning is difficult to explain [8].

In a linear regression statistical model, the function f can be obtained by changing the slope of the curve, tan (Φ) and intercept β of the straight line of Figure 1 (a), such that the errors between the actual outputs and the straight line outputs are minimized.

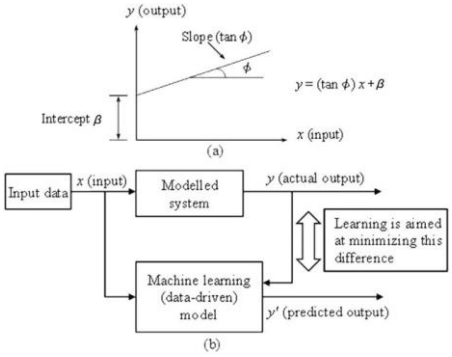

Figure 1: Linear regression versus artificial intelligence (AI) model. (a) Linear regression modeling; (b) Data-driven AI modeling [8].

The same principle is used in AI models. Artificial Intelligence can form a simple linear regression model from one input and one output, Figure 1 (b), and use available data to map between system inputs and corresponding outputs using repeated display machine learning examples of model inputs and outputs (Training) in order to find the Y = f (X) function that minimizes the error between historical (actual) outputs and AI model predicted outputs [8].

As common examples of artificial intelligence use it is possible to mention:

- Manufacturing robots;
- Autonomous driving system;
- Smart assistants;
- Social media monitoring;
- Natural Language Processing (NLP) tools;
- Proactive healthcare management;
- Conversational marketing bots;
- Prediction models;
- Image reconnaissance tools.

There are many other applications of AI technology.

2.2. Artificial Neural Networks

Artificial Neural Networks, or ANN's, are reliable tools for predicting experimental data. They could use their computational power in learning and generalization. They are widely used for solving nonlinear problems. Prediction and generalization performance depend on network training. Generalization is related to ANN performance with inputs that were not used during their training. The inspiration for artificial neural networks is the human brain, which uses parallel processing networks used to control complex relationships between input and output variables [9]. In Figure 2 a timeline of the development of deep learning and neural networks is shown in the upper panel and the lower panel numerous machine learning algorithms.

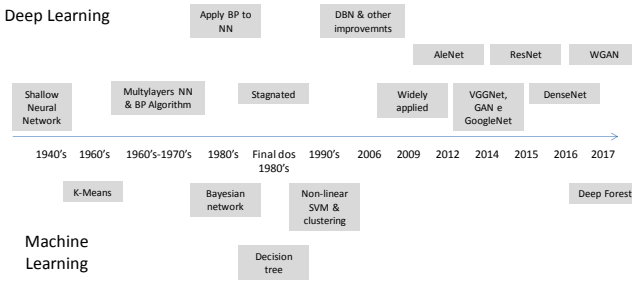

Figure 2: Deep learning development timeline and commonly used machine learning algorithms. Source: Adapted from [10].

In ANN, simple artificial nodes, known as neurons, processing elements, or units, are connected to build a network that mimics the biological neural network of humans. Back-propagation ANN is a well-known type of neural network, with multilayer perceptron architecture with error backpropagation for supervised learning and is particularly powerful for nonlinear prediction [11].

2.3. General Considerations on Artificial Neural Networks

The attraction of Artificial Neural Networks comes from the remarkable information processing characteristic of biological systems, such as nonlinearity, high parallelism, robustness, fault tolerance, learning, ability to handle

inaccurate information and their ability to generalize [12]. Its ability to learn and processing information classifies it as a form of Artificial Intelligence [13]. The most notable feature of this technology is that it can be applied to a wide variety of problems, many of which were extremely complex or lacking in more sophisticated theoretical models [14]. Figure 3 illustrates a scheme of a biological neuron with its three major functional units: dendrite, cell body, and axon. The cell body has a nucleus that contains information about hereditary characteristics and a plasma in which resides the molecular equipment used by the neuron in the process of communication with other neurons [15].

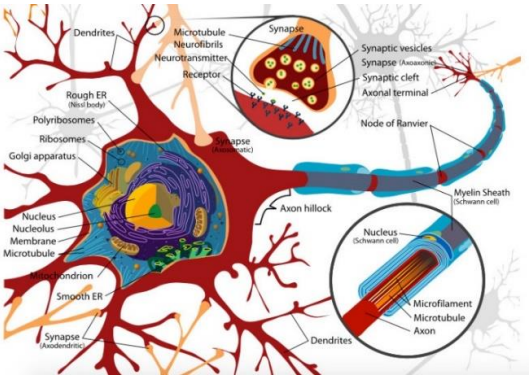

Figure 3: The biological neuron [16].

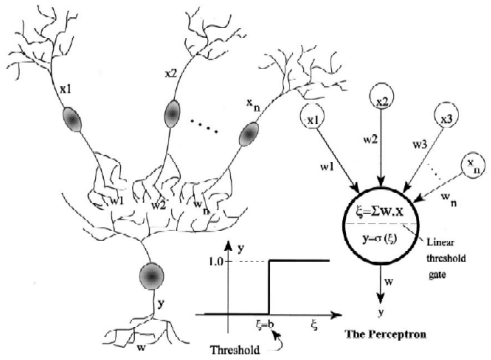

Figure 4: Signal interactions from n neurons and analogy to the sum of signals in a single-layer artificial neuron [15].

A schematic illustration of the signal transfer between two neurons across the synapse is shown in Figure 4. An impulse, in the form of an electrical signal, travels within the dendrite and through the cell body toward the presynaptic membrane of the synapse. Once the membrane is reached, a chemical neurotransmitter is released from the vesicles in amounts proportional to the strength of the newly arrived signal. The neurotransmitter diffuses into the synaptic breach toward the postsynaptic membrane and eventually into the dendrites of neighboring neurons, thereby forcing them (depending on the minimum value required for stimulating the receiving neuron) to generate a new electrical signal [15]. The system comprised of an artificial neuron and the inputs, as shown in Figure 4, is called Perceptron and is analogous to the biological neuron shown in Figure 3. It establishes a mapping between input activities (Stimulus) and the output signal. Perceptron presents n biological neurons with various signals of intensity x and synaptic strength w feeding a neuron with a minimum required stimulus value of b and the equivalent artificial neuron system. The Artificial Neural Network and the Biological Network learn by adjusting the magnitude of synapse weights or forces [15].

2.4. Deep Learning and Machine Learning

In a simple definition, machine learning or deep learning refers to the use of an artificial neural network with multiple layers of hidden nodes between output and input, as shown in Figure 5, where deep architectures are built on multiple levels on nonlinear operations [17].

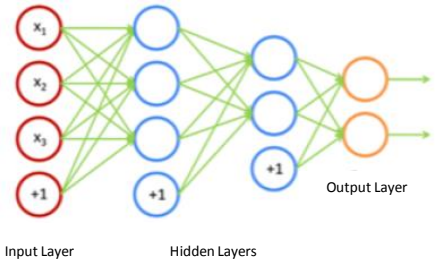

Figure 5: Artificial Neural Network with multiple layers. Source: Adapted from [17].

As shown in Figure 5, the deep network has the same architecture of a traditional neural network, but with a greater number of hidden layers. The main difference between a deep network and a traditional neural network is the development of algorithms for deep architecture training, which are faster and yield stronger results. Deep learning includes representation learning algorithms that transform raw characteristics into high-level abstractions using a deep network composed of several hidden layers. In other words, deep learning applies computational approaches, which have nonlinear multiple transformations to train data representation through various levels of abstraction [17].

3. Failures in Artificial Intelligence

3.1. Adversarial Examples

Applications in deep learning-driven systems are steadily increasing in the real world. For example, information technology and auto companies (Google, Tesla, Mercedes-Benz and Uber) are testing autonomous cars, which require a fullness of deep learning techniques such as object recognition, reinforcement learning and multimodal learning [18]. Opposing examples in conventional machine learning models have been discussed for decades. Machine learning based systems with hand-made features are the primary targets such as

spam filters, intruder detection, biometric authentication, fraud detection, and so on. The problem of Adversarial Example was formulated as a match between the adversary and the classifier, both of which were cost sensitive [18]. Attack and defense in opposing examples have become an iterative game. Given the high generalization capacity of deep neural networks, the entire research community was surprised by finding that these deep neural networks can be fooled by subtly disturbing input data [19]. This type of data is sound input data for machine learning models, but they are intentionally designed by an attacker to cause the model to fail. Figure 6 demonstrates the idea of what the opposing examples look like. Starting with the image of a panda, the attacker adds a small disturbance that has been calculated to make the image look like a gibbon. Opposing examples have the potential to be dangerous. For example, attackers may target autonomous vehicles using stickers or paints to create an opposing stop signal that the vehicle could interpret as a preference signal or other traffic signal [20].

Attackers can generate opposing commands against automatic speech recognition models in speech-controllable systems such as Apple Siri, Amazon Alexa and Microsoft Cortana [18]. The cause of these opposing examples was a mystery, and speculative explanations suggested that it was due to the extreme nonlinearity of deep neural networks, perhaps combined with the insufficient average of the model and insufficient regularization of the purely supervised learning problem.

A variety of intriguing properties of artificial neural networks and related models have been demonstrated [21]. But the consequences of errors increase dramatically when technology companies start using deep learning algorithms in applications such as two-ton machines moving on high-speed highways. A wrong decision made by autonomous artificial intelligence can lead the car to collide with the guardrail, another vehicle or run over pedestrians or cyclists [22].

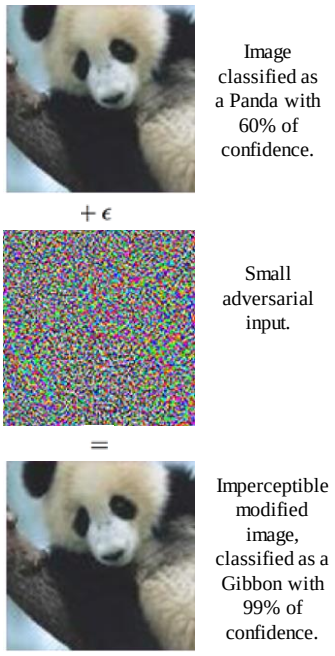

Figure 6: An opposing example constructed by modifying a panda's image in such a way that a machine learning model thinks it is a gibbon [20].

3.2. Soft Errors

The ever-increasing miniaturization of semiconductors leads to major advances in mobile, cloud and network computing. However, this has made electronic devices less reliable and microprocessors more susceptible to transient errors. These intermittent failures do not cause permanent damage but may result in program execution by changing the transfer of stored signals or values. These transient faults are also called soft errors. As technology continues to escalate, industry experts project that the problem of soft errors will become increasingly important [23]. Deep learning neural network-based applications are widely used in high performance computing systems and data centers.

While the performance of deep learning accelerators and applications has been extensively studied, the implications for reliability in their use are not well understood. One of the biggest sources of unreliability in modern systems are the soft errors. Such mild errors can cause application failures and result in violations of safety and reliability specifications. For example, in autonomous cars, a slight error can lead to misclassification of objects, resulting in a wrong action taken by the car. In fault insertion experiments, many cases were found where a truck could be misclassified under a soft error, as shown in Figure 9.

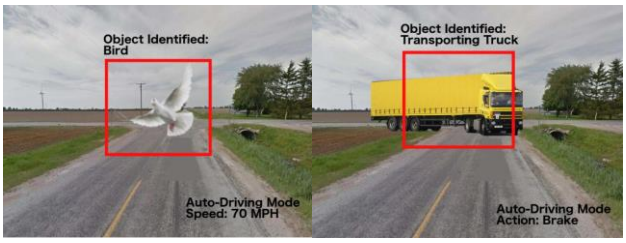

Figure 7: (Left) Fault Free Execution: A truck is identified by the deep neural network and then the brakes are applied. (Right) The truck is incorrectly identified as a bird and the brakes may not be applied [24].

The deep neural network in the car should classify the object ahead as a haul truck in a fault-free execution and then apply the brakes in time to avoid a collision. However, due to a slight error in the deep neural network, the truck is misclassified as a bird, and the braking action may not be applied in time to avoid a collision, especially when the car is operating at high speeds [24].

3.3. Errors due the lack of appropriate training

The lack of enough data for appropriate neural network training is associated with lower prediction accuracy or higher error. The section 4 will discuss a real case where the maximum and the minimum of the data, represented by the tails of a histogram, presents a bigger error if compared with the average data, represented by the central region of the histogram.

4. Development of a mathematical model based on Artificial Intelligence

In an artificial intelligence model developed for modeling the mechanical properties of hot rolled steel structural beams, a larger prediction error was found in regions with few occurrences of training data. The model was built from three artificial neural networks of the Reverse Propagation type, using the MATLAB software. The variable chosen for modeling is the Tensile Strength, TS, one of the strength measures of materials. Figure 8 shows the evolution of the Sum of Squared Errors (SSE) calculated for the Neural Network training data configured with 6 neurons to predict Tensile Strength.

Figure 8: ANN training using MATLAB for TS prediction (Oliveira, 2008).

The table 1 summarizes all the criteria used to build the Artificial Neural Networks for the prediction of the tensile Strength.

Table 1: Summary of Characteristics of Artificial Neural Networks

Characteristic	Criteria	MatLab
Partition of data set	Training set = 75%, Validation set = 25%.	RANPERM
Normalization	Formula	
Net weigh initialization	-	INITNW
Net learning ratio	-	TRAINGDX
Transfer Function	-	TANSIG
Convergence Criteria	-	
Minimum error aimed	0,001	-
Number of training cycle	700	-
Training mode	BT	-
Number of hidden layers	1	-
Size of hidden layer	6	-
Net training mode	-	TRAINBR

The figure 9 compares the actual and simulated values by the 111 samples model (25% of the original database) used in the artificial neural network validation tests. It is noticed that the neural network better simulates the values close to the TS average, not being perfectly accurate in the simulation of data that deviate much from the average (larger difference between simulated and measured points). This fact is due to the training process being effective in the regions with the highest data concentration [5].

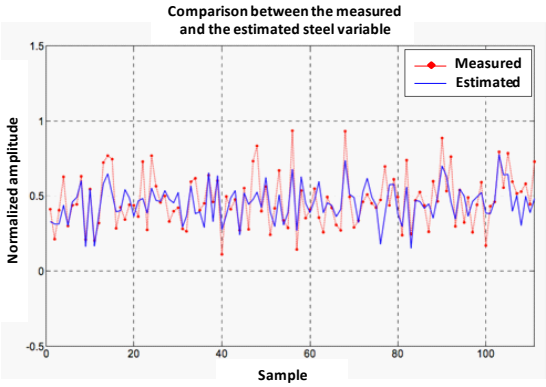

Figure 9: Result of ANN Validation for TS prediction [5].

As the histogram in Figure 10 shows, the central region concentrates most occurrences of learning data. In the tails, due to the absence of larger amount of data, as in any normal distribution, the learning does not occur as in the central region, which is evidenced by the biggest error and smaller error of the simulated data, respectively. As seen in section 3.1, where autonomous vehicle accidents were discussed under rare environmental conditions, the example shown in section 4 is similar. The tails of the

histogram correspond to the rare conditions in which there was insufficient neural network training to modelling an important feature. Due to this fact, the error found in this region is significantly higher when compared to the central region of the histogram, as can be seen in Figure 9. In this case, the estimated TS, represented by the blue curve, as a function of the sample, presents greater error in regions coincident with the highest normalized amplitudes (corresponding to the maximum and minimum) of the measured TS, represented by the red curve.

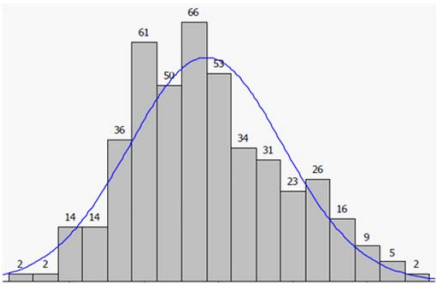

Figure 10: Histogram of the measured values of the Tensile Strength, [5].

It is proposed that, in the planning stage of the Artificial Neural Networks model for the applications discussed, either in autonomous cars or in predictive models for mechanical properties, it should be considered an experimental design phase (Design of Experiments, DOE), where the data related to the tails of the histogram are intentionally generated and therefore would provide greater learning on specific and rare situations. For example, on the case of a recent car accident the autonomous driving system was not capable to take any action to avoid the collision with an overturned truck and, also, ignored a pedestrian, at the same time [25]. Some extreme and possible situations, like an overturned truck and similar objects lying down on the road, could be on the training data to allow the car to learn that on situations like that, where the action of breaking is one of the possible measures to avoid an accident. But it appears not to be the case on this specific example. For the practical example discussed in section 4, simply design the DOE such that the maximum and minimum regions of the input variables have enough occurrences of training data. For this, the process must be modified in its origin, thus forcing the occurrence of these data [5].

5. Conclusion

This paper discussed the main technologies used in Artificial Intelligence, their development history, considerations about Artificial Neural Networks and the failures arising from the training processes, hardware and lack of enough training data. It has been seen that these models learn as they are exposed to input and output data of any phenomenon that one wishes to have predictive ability. There are numerous technologies in the area, however the Artificial Neural Networks stand out and are the basis of the so-called Deep Learning. These networks simulate the learning process of the human brain and learn through training with historical data. Technologies used in artificial intelligence tools have been found to be subject to errors, known as adversary examples and subtle errors, which are critical in certain uses where safety is a primary issue and lives may be subject to decisions made by algorithms. An example is the advent of autonomous cars, subject to both adversarial examples and subtle errors. The lack of enough data on extreme situations are also critical for training, prediction and subsequent action for uses where safety is critical. A real case of building an artificial intelligence for industrial application was discussed and the possible measures to reduce the prediction error obtained were pointed out. This article is expected to contribute to the technical and professional growth of the readers.

6. References

[1] B. J. Copeland. (2019) Artificial intelligence. https://www.britannica.com/technology/artificial-intelligence;

[2] Danny Y., & Dan T. 2016). Tesla driver dies in first fatal crash while using autopilot mode. https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk;

[3] Chen, A. (2018). IBM's Watson gave unsafe recommendations for treating cancer. https://www.theverge.com/2018/7/26/17619382/ibms-watson-cancer-ai-healthcare-science;

[4] Vincent, J. (2016). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist;

[5] Oliveira, A. P. Modelo de Previsão de propriedades mecânicas de perfis estruturais laminados a quente: uma abordagem em redes neurais artificiais. (Dissertação, Mestrado em Engenharia Metalúrgica e de Minas). Biblioteca Digital da Universidade Federal de Minas Gerais, UFMG, 2008. http://hdl.handle.net/1843/MAPO-7RLKBJ;

[6] Lohr, Steve. (2018). Vencendo os limites da pesquisa no campo da inteligência artificial. https://internacional.estadao.com.br/noticias/nytiw,vencendo-os-limites-da-pesquisa-no-campo-da-inteligencia-artificial,70002399992;

[7] Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M. Neuroscience-Inspired artificial intelligence. Neuron, 95, 2017, 245–258. http://dx.doi.org/10.1016/j.neuron.2017.06.011;

[8] Shahin, M. A. State-of-the-art review of some artificial intelligence applications in pile foundations. Geoscience Frontiers 7, 2014, 33-44. http://dx.doi.org/10.1016/j.gsf.2014.10.002;

[9] Muhammad Shahbaz, Syed A. Taqvi, Adrian Chun Minh Loy, Abrar Inayat, Fahim Uddin, Awais Bokhari, Salman Raza Naqvi. Artificial neural network approach for the steam gasification of palm oil waste using bottom ash and CaO. Renewable Energy 132, 2019, 243-254;

[10] Cao, C.; Liu, F.; Tan, H.; Song, D.; Shu, W.; Li, W.; Zhou, Y.; Bo, X.; Xie, Z. Deep Learning and Its Applications in Biomedicine. Genomics, Proteomics & Bioinformatics, 16, 2018, 17–32;

[11] Chia-Yen, L., & Tsung-Lun, T. Data science framework for variable selection, metrology prediction, and T process control in TFT-LCD manufacturing. Robotics and Computer Integrated Manufacturing, 55, 2018, 76–87. https://doi.org/10.1016/j.rcim.2018.07.013;

[12] Tu, J. V. Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes. Journal of Clinical Epidemiology, 49, 11, 1996, 1125-1231; https://doi.org/10.1016/S0895-4356(96)00002-9;

[13] Korczak, P., Dyja H., Labuda E. Using Neural Netwoks Models for Predicting Mechanical Properties after Plate Rolling Processes. Journal of Materials Processing Technology, Poland, vol.80, n.81, 1998, 481-486;

[14] Dyja H., & Korczak P. The thermal-mechanical and microstructural model for the FEM simulation of hot plate rolling. Journal of Materials Processing Technology, 92-93, 1999 463-467. https://doi.org/10.1016/S0924-0136(99)00215-0;

[15] I.A. Basheer, M. Hajmeer. Artificial Neural Networks: Fundamentals, Computing, Design and Application, Journal of Microbiological Methods, vol.43, 2000, 3-31;

[16] Castrounis. (2016). Artificial Intelligence, Deep Learning, and Neural Networks, Explained. https://www.kdnuggets.com/2016/10/artificial-intelligence-deep-learning-neural-networks-explained.html;

[17] Milad Zafar Nezhad, Najibesadat Sadati, Kai Yang, Dongxiao Zhu. A Deep Active Survival Analysis approach for precision treatment recommendations: Application of prostate cancer. Expert Systems with Applications 115, 2018, 16–26;

[18] Xiaoyong Yuan, Pan He, Qile Zhu, Rajendra Rana Bhat, Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. arXiv preprint arXiv:1712.07107, 2018;

[19] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security, October 2016. DOI: http://dx.doi.org/10.1145/2976749.2978392;

[20] Goodfellow, Ian; Papernot, Nicolas; Huang, Sandy; Duan, Yan; Abbeel, Pieter; Clark, Jack. (2017). “Attacking Machine Learning with Adversarial Examples.” OpenAI. https://blog.openai.com/adversarial-example-research/;

[21] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 2015. arXiv:1412.6572;

[22] Hsu, J. (2017). A New Way to Find Bugs in Self-Driving AI Could Save Lives. https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/better-bug-hunts-in-selfdriving-car-ai-could-save-lives;

[23] Q. Shi, H. Omar, and O. Khan, “Exploiting the tradeoff between program accuracy and soft-error resiliency overhead for machine learning workloads,” CoRR, vol. abs/1707.02589, 2017. [Online]. Available: http://arxiv.org/abs/1707.02589;

[24] Li, G., Hari, S. K. S., Sullivan, M., Tsai, T., Pattabiraman, K., Emer, J., & Keckler, S. W. (2017). Understanding error propagation in deep learning neural network (DNN) accelerators and applications. Paper presentend on: International Conference for High Performance Computing, Networking, Storage and Analysis, Denver, Colorado. https://doi.org/10.1145/3126908.3126964;

[25] Templeton B. (2020). Tesla in Taiwan Crashes Directly into Overturned Truck, Ignores Pedestrian, With Autopilot On. https://www.forbes.com/sites/bradtempleton/2020/06/02/tesla-in-taiwan-crashes-directly-into-overturned-truck-ignores-pedestrian-with-autopilot-on/#5ad97d7758e5.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)