Using Retrieval-based Voice Conversion in Educational Video Materials

Tibor Szabó[1], Rastislav Žitný[2], Ildikó Pšenáková[3], Peter Pšenák[4]

DOI: https://doi.org/10.53349/re-source.2025.is1.a1407

Abstract

The rapid and unforeseen rise of artificial intelligence (AI) raises numerous questions regarding its applicability across various fields. Consequently, it is essential to consider its impact on education. This study primarily examines the utilisation of AI within the context of the flipped classroom teaching model, with a particular focus on the audio components of educational video materials. For instance, AI can facilitate replacing a speaker's voice with that of another individual or translating it into a different language while nearly maintaining the original speaker's voice. The application of Retrieval-Based Voice Conversion (RVC) models can significantly aid this process.

Keywords: Artificial Intelligence, Retrieval-Based Voice Conversion, Educational Video Materials, Flipped Classroom

1 Introduction

Artificial Intelligence (AI), as an emerging phenomenon in modern technology, is increasingly permeating various areas of life, including almost all branches of industry and the field of education. While this development is perceived positively, there are also numerous concerns regarding the use of AI. This article contributes to clarifying the principles of AI through special algorithms in text processing, presents the RVC (Retrieval-based Voice Conversion) technology for processing natural speech and outlines the research methodology for processing teaching

[1] Constantine the Philosopher University in Nitra, Faculty of Central European Studies, Drážovská 4, 949 01 Nitra, Slovakia.
E-Mail: tszabo@ukf.sk

[2] Constantine the Philosopher University in Nitra, Faculty of Central European Studies, Drážovská 4, 949 01 Nitra, Slovakia.

[3] Trnava University, Faculty of Education, Priemyselná 4, P. O. BOX 9, 918 43 Trnava

[4] Comenius University Bratislava, Faculty of Management, Odbojárov 10, P. O. BOX 95, 820 05 Bratislava 25, Slovakia.

materials for the needs of flipped classroom educational technology. In our article, we therefore consider AI as a tool for possible improvement of voice reproduction in the preparation of educational video materials.

2 Flipped Classroom Method

We present the Flipped classroom method as an innovative method of education, although according to Lage, Platt and Treglia, it is not a new recent method, as shown by the definition from 2000: *"New learning technologies it possible for events such as lectures, which have traditionally taken place inside the classroom, to occur outside the classroom to and events which possibly occurred outside the classroom to occur inside the classroom under the guidance of the instructor."* (Lage, Platt and Treglia, 2000, p. 41).

Research showed that the use of the flipped classroom method has a positive impact on students' learning, as it increases their interest in the subject, improves their motivation and confidence, helps in developing their critical thinking, and thus ultimately leads to an overall improvement in their academic performance (Pšenáková at al., 2024).

The right use of the flipped classroom method has many advantages, the most important is individual tempo of learning, flexibility, more efficient use of class time, and development of critical thinking.

Outside the classroom, students prepare for the lesson from a variety of resources, such as instructional video materials, learning texts, have space for online discussion, etc.

We will only cover educational video materials. But the truth is, the flipped classroom method to be effective, we cannot rely on video materials alone, but on a combination of different resources. If we only provided students with videos for preparation, it might carry several disadvantages, such as: *"poor interaction, poor collaboration, low video quality, lower concentration, technical problems"* (Bui, 2021, p. 275).

3 AI in Education

Collecting large volumes of data in databases enables knowledge extraction. This knowledge can be extracted using artificial intelligence algorithms. Extracted knowledge can be used for concentrated support of education. We can analyse the knowledge stored in big data using machine learning and use artificial intelligence solutions to support activities and decision-making processes in education or to convert natural speech into desired patterns.

The teacher has limited control over how well and to what extent the students have studied the materials before the lesson. If students do not absorb information well, classroom activities can be less effective and ineffective. Therefore, when it comes to the educational process, AI can support learning outside the classroom with the help of prepared educational materials.

3.1 Incidence of Algorithms

AI represents the potential, that can search connections between existing knowledge and enables search for relative and relevant connections from available sources. From the point of view of knowledge management is important for us to provide relevant and adequate details for the needs of a given subject or field. The goal is to gain knowledge through searching from various sources based on artificial intelligence algorithms. Therefore, artificial intelligence, based on algorithms, search information based on required variables, identifies existing patterns from available sources and offers them as complete files.

Many new research articles are published every day, in which different artificial intelligence techniques (e.g., neural networks, fuzzy logic, clustering algorithms, and evolving computing) are applied to various tasks and applications related to opinion mining. (Serrano-Guerrero et al., 2021, p. 1). We present some known AI algorithms, e.g. Artificial neural networks, Support vector machine, Fuzzy logic, Genetic algorithm, Tree-based assembles, Hybrid and ensemble procedures, Deep learning, Bayesian networks, etc. (Nguyen et al., 2023, pp. 4–10).

For the needs of our study, we focus on clarifying the processes for creating natural language with algorithms supporting NLP (Natural Language Processing). Various TTS (Text to Speech) models are known. In our paper, we deal with the RVC (Retrieval-based Voice Conversion) model and try to outline the perspective of text-to-speech research for the needs of creating educational materials in the conditions of university education and supporting flipped classroom educational technology.

For this reason, we are investigating the best ways to provide students with quality delivery of the curriculum using professional voice modulation with respect to voice timbre, emphasizing important parts and essential data, facts or information. The research methodology will be based on training artificial speech, embedding and analysing scenes and testing speech samples on prepared scenes in the form of processed natural language (NLP).

The goal of the models is to perform binary classification of the speech, learning whether the audio is speech spoken naturally by a human being, or has been tampered with by retrieval-based voice conversion. The models trained are: Extreme Gradient Boosting (XGBoost), Random Forests, Quadratic and Linear Discriminant analyses, Ridge Regression (linear regression with L2 regularisation), Gaussian and Bernoulli Naive Bayes, K-Nearest Neighbours, Support Vector Machines, Stochastic Gradient Descent, and Gaussian Process. The study showed that Convolutional Neural Networks and Long-Short-Term-Memory neural networks could score around 97-99% accuracy when dealing with recognition. (Bird–Lotfi, 2023, P. 2).

3.1.1 Artificial Neural Networks

In this section, for the sake of illustration, we mention only some AI algorithms that are used in the application of speech models. A recurrent neural network (RNN) is any network whose neurons send feedback signals to each other (Grossberg, 2013, p. 1888). Recurrent neural networks have been commonly used in earlier work to model contextual information in dialogues.

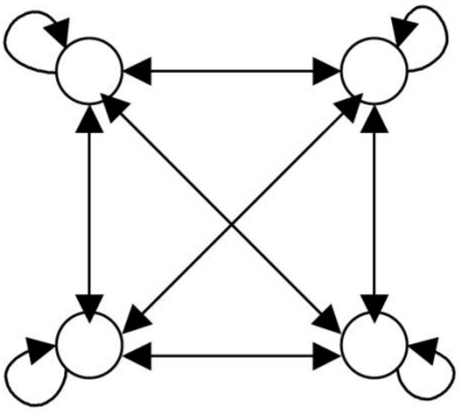


Figure 1: An example of a fully connected recurrent neural network (Medsker–Jain, 2001, P. 14).

Graphs are a kind of data structure which models a set of objects (nodes) and their relationships (edges). Graph neural networks are deep learning-based methods that operate on graph domain. (Jie et al., 2020, P. 57).

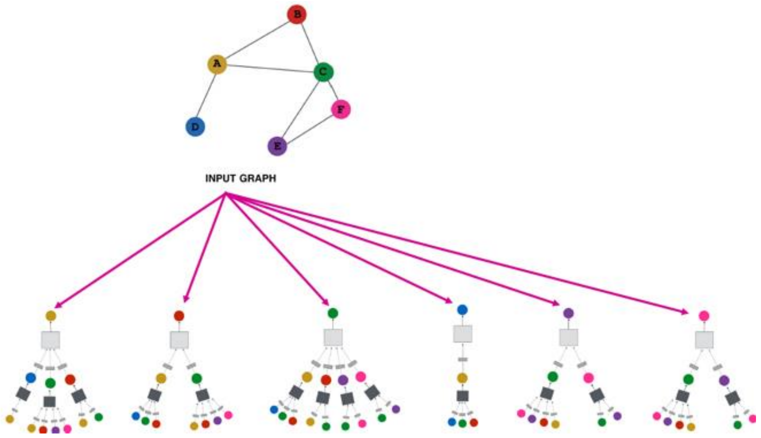


Figure 2: Each node of the Graph has its own neural network architecture (Taneja, 2024, P. 1).

Sparse volumetric data are ubiquitous in many fields including scientific computing and visualization, medical imaging, industrial design, rocket science, computer graphics, visual effects, robotics, and more recently machine learning applications. Neural networks, on the other hand, can be designed to discover such hidden features and can infer values without reconstructing the entire data set. (Kim et al., 2024, P. 2).

3.1.2 Support Vector Machine

The main motivation of Support Vector Machine is to separate several classes in the training set with a surface that maximizes the margin between them. Figure 3 shows the support vectors separate data set of two classes.

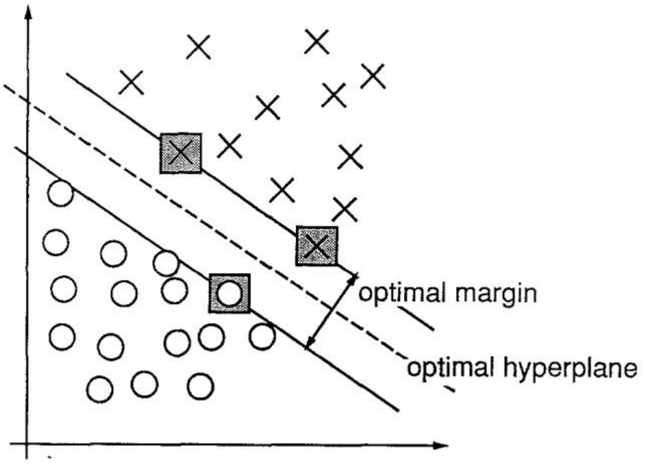
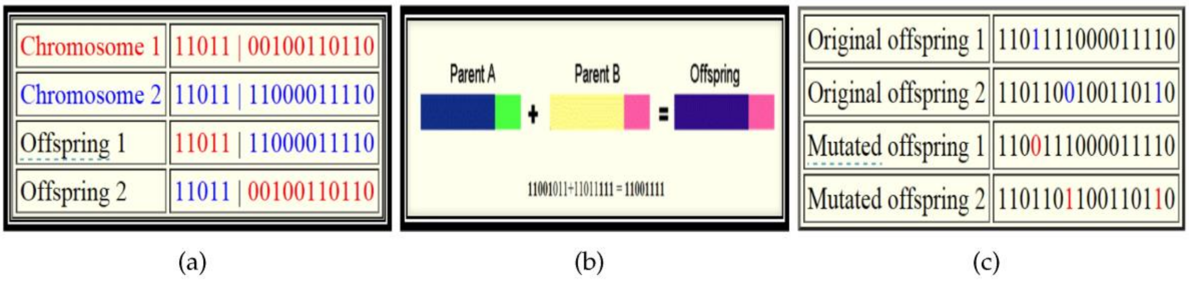


Figure 3: An example of a separable problem in a 2-dimensional space (Cortes et al., 1995, P. 5).

3.1.3 Genetic Algorithms

Genetic algorithms imitate the Darwinian theory of survival of the fittest in nature. Basic elements of genetic algorithms are chromosome representation, fitness selection and biologically inspired manipulations.

Genetic algorithm is an artificial intelligence search method that uses the process of evolution and natural selection theory and is under the umbrella of evolutionary computing algorithm. It is an efficient tool for solving optimization problems. (Hassanat, 2019, P. 1).



Chromosome 1	11011 | 00100110110
Chromosome 2	11011 | 11000011110
Offspring 1	11011 | 11000011110
Offspring 2	11011 | 00100110110

Original offspring 1	1101111000011110
Original offspring 2	1101100100110110
Mutated offspring 1	1100111000011110
Mutated offspring 2	1101101100110110

Figure 5: Examples of (a) crossover, (b) one-point crossover, (c) mutation operator in genetic algorithms (Hassanat, 2019, P. 4).

3.1.4 Tree-based Machine Learning

Tree-based machine learning methods are built by recursively partitioning a sample using different features at node from the dataset. Classification and regression trees are effectively used for predictions. According to Wei-Yin Loh classification and regression trees are machine learning methods for constructing prediction models from data. (Wei-Yin Loh, 2011, P. 1).

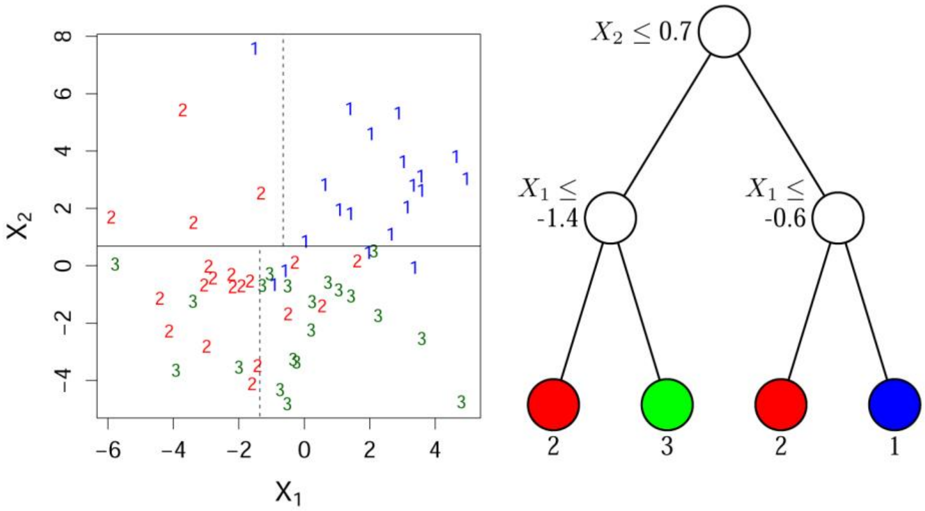


Figure 6: Decision classification and regression tree model with classes. (Wei-Yin Loh, 2011, P. 2).

3.1.5 Deep Learning

In machine learning and data science, high-dimensional data processing is a challenging task for both researchers and application developers. Thus, dimensionality reduction, which is an unsupervised learning technique, is important because it leads to better human interpretations, lower computational costs, and avoids overfitting and redundancy by simplifying models. (Sarker, 2021, P. 5).

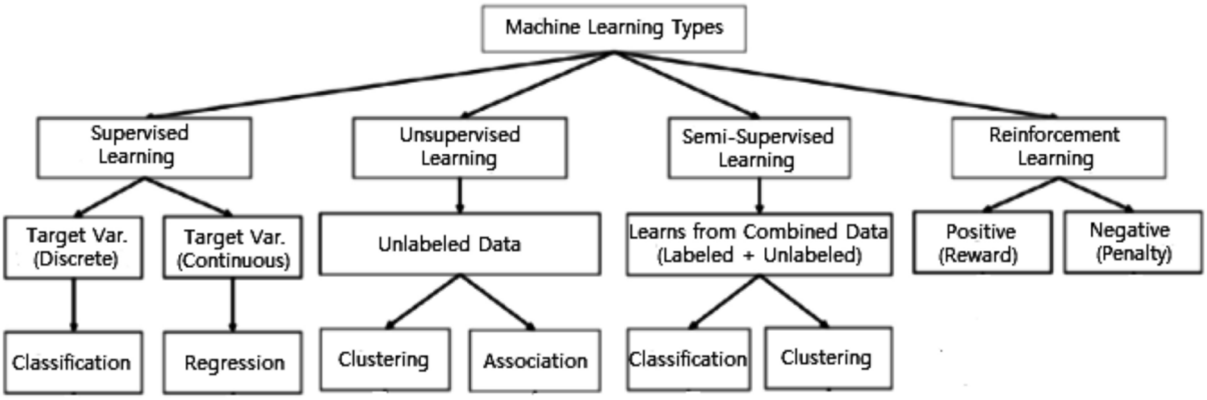


Figure 4: Various types of machine learning techniques. (Sarker, 2021, P. 4).

3.1.6 Bayesian Networks

Bayesian networks are based on the probability of dependent and independent variables. When determining the Bayesian rule, we start from the conditional probability of the phenomenon A, which is dependent on the phenomenon B, that is, the probability of the phenomenon A is equal to the joint probability of the values a, b divided by the probability of the value b.

$$P(a \mid b) = P(a,b)/P(b)$$

P is probability
a – values of A
b – values of B
From the mentioned relationship, we derive the Bayesian rule:

$$P(b \mid a) = P(a \mid b) * P(b) \ / \ P(a)$$

The Bayesian models represent a well-studied and effective way to describe and reason about problems that include uncertainty. (Kukačka, 2010, P. 30).

3.2 AI Generated Voice in Education

3.2.1 Retrieval-based Voice Conversion

RVC is characterized as an area of deep sound processing technology that offers solutions for correct prosody, pronunciation from graphemes, appropriate intonation with stress and overall control of speech style.

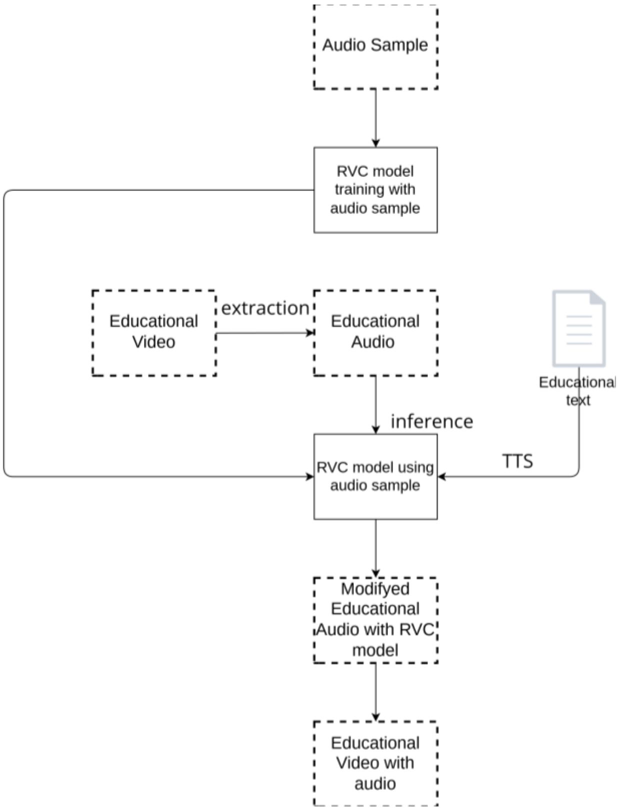

Audio Sample
RVC model training with audio sample
Educational Video
extraction
Educational Audio
Educational text
inference
TTS
RVC model using audio sample
Modifyed Educational Audio with RVC model
Educational Video with audio

Figure 5: Diagram of educational video material processing. Own elaboration.

The RVC technology is based on the process of training and creating a speech model, in which we train the system on the unique voice of an individual and thus create our own model. Next, from the process of inference, i.e. the application of the model created in this way, where we try to apply such a unified model for educational materials and create educational video materials for the method of education, in our case for the flipped classroom method.

Speech processing with RVC technology can also prepare a model for a multilingual education system. We will provide the educational materials prepared in this way to the community of students in real teaching for sound verification. Figure 5 shows a scheme for preparing video material by modifying an audio recording.

RVC has the following characteristics: even on relatively poor graphics cards can be fast training; Training with a small amount of data can also get better results (it is recommended to collect at least 10 minutes of low-noise speech data). The timbre can be changed by model fusion. RVC also cannot intelligently select the number of rounds with the highest similarity to the training sound source in the process of model training, so it needs to manually customize the number of training rounds and provide multiple model results for manual screening (Zhongxi Ren, 2024, p. 468).

3.2.2 Using of RVC in Educational Video Materials

RVC tools provide processes which we consider from point of view of preparation educational video materials for using in flipped classroom method.

Text to speech: speech is generated from the text you have written. The voice can remain original if we have an RVC model from the original speaker, or we can use a different RVC model, for example a voice that is distinctive, interesting for the listeners.

Let's imagine a situation where we teach groups in the national language, but at the same time we teach a group in English for the same subject. For both groups we want to make the video teaching material available, this feature is usable as a dubbing based on the written text. We are limited by the visual content of the video, for example the content presented in Slovak language. However, there are some video materials that are universal in terms of visual content, e.g. explanation of a mathematical example etc.

On the other hand, not every creator of learning video materials likes his own voice, he can change it to a more interesting, more distinctive, i.e. more 'ideal' one. In foreign languages, it is advisable to use the voice of a native speaker.

Speech to speech: Transforming original speech to speech with a different voice. This function can be used to replace the teacher's original voice in video teaching materials, for example, the one that we consider more interesting, and which can help attract students' attention. We can use speech to speech tool in cases like TTS. The quality of the converted voice is greatly affected by the quality of the original voice, in addition, noise should also be removed from the original.

Voice Cloning: This is the process of making a clone voice based on a sample voice of an individual person (approx. 10 minutes is enough). In this case we consider clone voice as an

RVC model. This model can then be used for both TTS and STS. The quality of the model depends on the quality of the sample.

Various RVC technologies are not perfect yet, but they are getting better. There are many websites that offer such services, free for trial. The Applio project (https://applio.org/) offers these tools for free, and we use them for our purpose. Using the Applio and installation of it requires advanced knowledge in computer science. There are various trained RVC models available on the Internet, created by the community of users. Training of RVC model, which is based on deep learning, is hardware intensive. For training, a high-quality graphics card is required to perform the computations (for testing, we used the following hardware configuration: Intel Core i5-10400, DDR4 16GB RAM, GIGABYTE GeForce RTX 3070 GAMING OC 8G rev. 2.0). Setting some parameters in bulk can affect the result when using these tools. Of course, RVC technology also carries dangers if it is used by people for illegal, criminal purposes. Public figures have various video recordings and radio recordings available to them, their voice can be cloned (training the RVC model). This model then can be used in videos, which will be a completely different speech, but identical voice - video with deepfake voice. Attackers can also use voice spoofing when making phone calls. New tools are also needed to counter these unusual methods of spoofing. For example, the possibility of protection has also been discussed by Zhao et al. (2022), Jordan and Lotfi (2023).

4 Conclusion

AI-generated voice technology has significant potential across various sectors, including healthcare, services, and education. In the context of the flipped classroom teaching model, video learning materials play a crucial role. These materials should engage the listener not only through content but also by employing an appropriate speaker's voice. The aforementioned applications of AI in voice generation can assist in enhancing educational videos. As AI technology continues to develop, it is anticipated to become more refined in the future. This study aims to investigate the effects of modifying the original voice to improve video learning materials for use in the flipped classroom method.

In this paper, we introduce the flipped classroom model as an innovative teaching method that incorporates educational video materials. We highlight that Artificial Intelligence (AI) employs deep learning algorithms to process human sensory information. Additionally, we present Retrieval-Based Voice Conversion (RVC) technology, which processes natural AI-generated speech, as a tool for preparing educational video materials within the flipped classroom framework.

Acknowledgments

This research was funded by the grant system of the Ministry of Education, Research, Development and Youth of the Slovak Republic, under the number KEGA 015TTU-4/2024, titled "The Implementation of the Flipped Classroom Method in Higher Education."

References

Bird, J. & Lotfi, A. (2023). Real-time detection of ai-generated speech for deepfake voice conversion. *arXiv preprint arXiv:2308.12734.* https://dx.doi.org/10.48550/arXiv.2308.12734.

Bui, L. (2021). Using Lecture Videos in Flipped Classroom Model. *Conference: Proceedings of the 18th International Conference of the Asia Association of Computer-Assisted Language Learning.*, pp. 269—277. https://dx.doi.org/10.2991/assehr.k.211224.026.

Cortes, C., Vapnik, V. (1995). Support-Vector Networks. Kluwer Academic Publishers, Boston. *Machine Leaming, 20,* pp. 273–297. https://ise.ncsu.edu/wp-content/uploads/sites/9/2022/08/Cortes-Vapnik1995_Article_Support-vectorNetworks.pdf. (Accessed on 15 September 2024).

Grossberg, S. (2013). Recurrent Neural Networks. *Scholarpedia, 8*(2), pp. 1888. https://doi:10.4249/scholarpedia. 1888.

Hassanat, A., Almohammadi, K., Alkafaween, E., Abunawas, E., Hammouri, A., & Prasath, V., B., S. (2019). Choosing mutation and crossover ratios for genetic algorithms—A review with a new dynamic approach. *Information 2019, 10*(12). https://doi.org/10.3390/info10120390.

Jie, Z., Ganqu, C., Shengding, H., Zhengyan, Z., Cheng, Y., Zhiyuan, L., Lifeng, W., Changcheng, L., & Maosong, S. (2020). Graph neural networks: A review of methods and applications. *AI Open, Volume 1,* 2020, Pages 57—81. https://doi.org/10.1016/j.aiopen.2021.01.001.

Katoch, S., Chauhan, S.S., & Kumar, V. A review on genetic algorithm: past, present, and future. *Multimed Tools Appl 80,* pp. 8091—8126. https://doi.org/10.1007/s11042-020-10139-6.

Kim, D., Lee, M., & Museth, K. (2024). NeuralVDB: High-resolution Sparse Volume Representation using Hierarchical Neural Networks. *ACM Transactions on Graphics, 43*(2), Article No.: 20, 2024. pp. 1—27. https://doi.org/10.1145/364181.

Kong, S., Ch. (2014). Developing information literacy and critical thinking skills through domain knowledge learning in digital classrooms: An experience of practicing flipped classroom strategy. *Computers & Education, Vol. 78*, pp. 161—173. https://dx.doi.org/10.1016/j.compedu.2014.05.009.

Kukačka, M. (2010). Bayesian Methods in Artificial Intelligence. In Šafránková, J., and Pavlů, J. *WDS'10 Proceedings of Contributed Papers, Part I, 25—30, 2010. Proceedings of the 19th Annual Conference of Doctoral Students - WDS 2010.* Prague, 1st June – 4th June 2010, pp. 25—30. https://physics.mff.cuni.cz/wds/proc/pdf10/WDS10_104_i1_Kukacka.pdf. (Accessed on 8 November 2024).

Lage, M., J., Platt, G., & Treglia, M. (2000). Inverting the Classroom: A Gateway to Creating an Inclusive Learning Environment. *The Journal of Economic Education, 31*(1), pp. 30—43. https://doi.org/10.2307/1183338.

Medsker, L., R., & Jain, S., C. (2001). Recurrent Neural Networks. Design and Applications. CRC Press LCC. Boca Raton London New York Washington, D.C.: 2001. https://www.academia.edu/download/31279335/___Recurrent_Neural_Networks_Design_And_Applicatio(BookFi.org).pdf. (Accessed on 15 September 2024).

Nguyen, T., Cherif, R., Mahieux, P., Y., Lux, J., Aït-Mokhtar, A., & Bastidas-Arteaga, E. (2023). Artificial intelligence algorithms for prediction and sensitivity analysis of mechanical properties of recycled aggregate concrete: A review. *Journal of Building Engineering, 66. Elsevier,* 1 May 2023, pp. 1—20. https://doi.org/10.1016/j.jobe.2023.105929.

Pšenáková, I., Pšenák, P.& Szőköl, I. (2024). Flipped Classroom in Pedagogical Practice. In: Auer, M., Cukierman, U., R., Vidal, E., V., Caro, E., T. *Towards a hybrid, flexible and socially engaged higher education: Proceedings of the 26th international conference on interactive collaborative learning (ICL2023).* Volume 4. Cham: Springer Nature, 2024, pp. 279—290.

Sarker, I.H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI. 2*, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x.

Serrano-Guerrero, J., Romero, F., P., & Olivas, J., A. (2021). Fuzzy logic applied to opinion mining: A review. *Knowledge-Based Systems. Volume 222,* 21 June 2021, pp. 1—20. https://doi.org/10.1016/j.knosys.2021.107018.

Taneja, A. (2024). Deep Learning with Graphs: Part 9 of my Graph Series of blogs. LinkedIn. 28 April 2024. https://www.linkedin.com/pulse/deep-learning-graphs-part-9-my-graph-series-blogs-ajay-taneja-bzgaf/. (Accessed on 8 November 2024).

Wei-Yin Loh. (2011). Classification and regression trees. *Wires Data Mining and Knowledge Discovery.* Volume1, Issue1 January/February 2011, pp. 14—23. https://doi.org/10.1002/widm.8.

Zhongxi Ren. (2024). Selection of Optimal Solution for Example and Model of Retrieval Based Voice Conversion. *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023).* Atlantis Press 1951-6851, pp. 468—475. ISBN 978-94-6463-370-2. https://doi.org/10.2991/978-94-6463-370-2_48.

Zhou, J., Hai, T., Jawawi, D., Wang D., Ibeke, E. & Biamba, D. (2022). Voice spoofing countermeasure for voice replay attacks using deep learning. *Heidelberg, 11*(1) https://doi.org/10.1186/s13677-022-00306-5.