Postprint of: Kostek, B. (2023). Data, Information, Knowledge, Wisdom Pyramid Concept Revisited in the Context of Deep Learning. In: Czarnowski, I., Howlett, R., Jain, L.C. (eds) Intelligent Decision Technologies. KESIDT 2023. Smart Innovation, Systems and Technologies, vol 352. Springer, Singapore. DOI: 10.1007/978-981-99-2969-6_1

Data, Information, Knowledge, Wisdom Pyramid concept revisited in the context of deep learning

Bożena Kostek[1][0000-0001-6288-2908]

Gdańsk University of Technology, ETI Faculty, Audio Acoustics Laboratory
Narutowicza 11/12, 80-232 Gdańsk, Poland
bokostek@audioakustyka.org

Abstract. In this paper, the data, information, knowledge, and wisdom (DIKW) pyramid is revisited in the context of deep learning applied to machine learning-based audio signal processing. A discussion on the DIKW schema is carried out, resulting in a proposal that may supplement the original concept. Parallels between DIWK pertaining to audio processing are presented based on examples of the case studies performed by the author and her collaborators. The studies shown refer to the challenge concerning the notion that classification performed by machine learning (ML) is/or should be better than human-based expertise. Conclusions are also delivered.

Keywords: Data, Information, Knowledge, Wisdom (DIKW) Pyramid, audio signals, machine learning (ML)

1 Introduction

Data, information, knowledge, and wisdom (DIKW) are typically presented in the form of a pyramid [1-7]. When looking at this schema, several thoughts may occur. The first refers to what is not included in it, i.e., intuition and perception. These two notions are certainly needed in acquiring knowledge and wisdom. Moreover, the Cambridge dictionary definition of wisdom shows: "the ability to use your knowledge and experience to make good and judgments." So, other missing factors are experience and judgment. Then, what about courage, pushing the boundaries of believing in what one is doing, perseverance, and expertise, which is not an exact synonym for knowledge or wisdom but competence, proficiency, or aptitude. Then, one may look for humility in wisdom to know one's limits. What about understanding and intelligence? Where do they fit in this schema?

Furthermore, we cannot treat the pyramid as an equation in which a sum of data, information, and knowledge equals wisdom. So, maybe this concept should not be presented as a pyramid but rather as a chaotic mixture of all those factors mentioned blending into each other (see Fig. 1). Figure 1 illustrates some of the notions recalled above. Of course, there is sufficient ocean-like space to contain more ideas and beliefs to supplement such a visualization concept.

On the contrary, the hierarchical presentation has advantages, as it shows the direction from raw data through information and knowledge toward wisdom [6,7], measured as a level of insight. Even though there is an observation that the levels of DIKW hierarchy wisdom are fuzzy, moreover, the path from data to wisdom is not direct. This is further expanded in a strong thread on DIKW within the rough set community [8-9]. A wistech (wisdom technology) was introduced, defined as a salient computing and reasoning paradigm for intelligent systems [8].

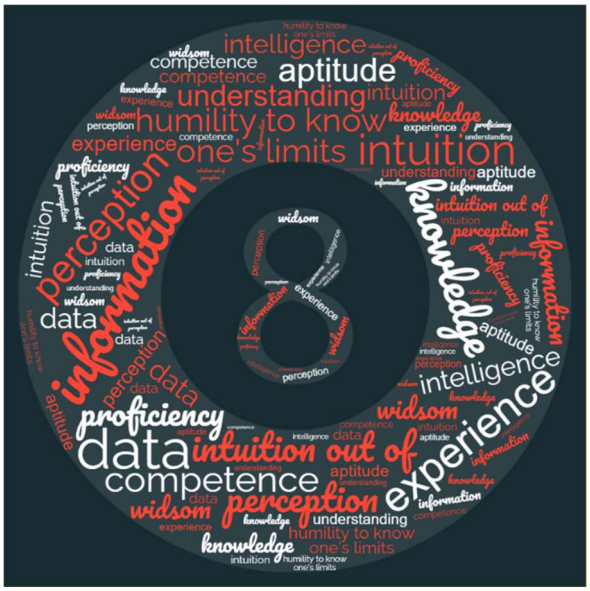


Fig. 1. Word cloud translating data, information, knowledge, and wisdom (DIKW) pyramid to the chaotic concept of acquiring knowledge and wisdom (created with the use of https://www.wordclouds.com/).

In that respect, all these questions pertain to machine learning (ML) and artificial intelligence (AI). Nowadays, the first attempts that employ learning algorithms are called conventional or baseline. Still, when an artificial neuron was modeled by McCulloch and Pitts [10], and Rosenblatt later proposed a perceptron to classify pictures, there were already ambitious plans for what could be done with such an approach. A well-known Rosenblatt's statement envisioned that perceptron would be able to "recognize people and call out their name," "instantly translate speech in one language to speech or writing in another language," "be fired to the planets as mechanical space explorers," but also "reproduce itself" and be self-conscious in the future [10,11]. However, before this belief came true, several decades passed, and technology had to change to employ graphical units instead of CPU (central processing unit), resulting in data-hungry deep learning [12]. A very apt statement of Vandeput on the last decade's machine learning evolvement refers to the "deep learning tsunami" [10]. However, already in the '50s of the previous century, a need for data was recognized [10]. Even though one may discern a difference between machine learning and deep learning, the first regarded as

prediction, classification, etc., and the second considered as "algorithms inspired by the structure and function of the brain called artificial neural networks" [11]; they both need data.

In most cases, data should be structured and annotated by experts. The latter notion, however, may no longer apply as synthesized data may substitute carefully crafted datasets [13]. Moreover, a knowledge-based approach to machine learning may not be necessary as relevant features are extracted automatically in some deep model structures [13,14]. Contrary, the notion of imperfect data in the sense of incomplete, too small a size, unbalanced, biased, or unrepresentative [15] is still valid.

In this paper, examples of work performed by the author and her collaborators are shown further on. This short review of study encompasses intelligent audio processing.

2 Intelligent music and speech processing

Even though music information retrieval (MIR) is a well-established area encompassing musical sound, melody and music genre recognition, music separation and transcription, music annotating, automatic music mixing, music composing, etc. [16-18], there is a void between human- and machine-based processing, which is sometimes referred to as a semantic gap or bridging a semantic gap [19], i.e., finding interconnections between human knowledge, content collections, and low-level signal features. There are two layers to music services, i.e., a general map of the relation between songs − interconnections between the users and songs, and a personalization layer − information from the above analysis is confronted with the user's music preferences, mood, emotions, and not only what the particular user listens to but what songs they like to combine. In the dictionary of terms related to MIR, one should include music representation, which may be obtained by automatic tagging using metadata (ID3v2), included in, e.g., Gracenote or FreeDB databases; manual tagging by experts or social tagging; content-based; low-level description of music (feature vectors based on MPEG-7 standard, Mel-Frequency Cepstral Coefficients (MFCC), and dedicated descriptors), or 2D maps as features (e.g., spectrograms, mel-spectrograms, cepstrograms, chromagrams, MFCCgrams, designated for deep learning [20]. One should not forget collaborative filtering in Music Recommendation Systems (MRS) that creates maps based on neighbors or taste compatibility.

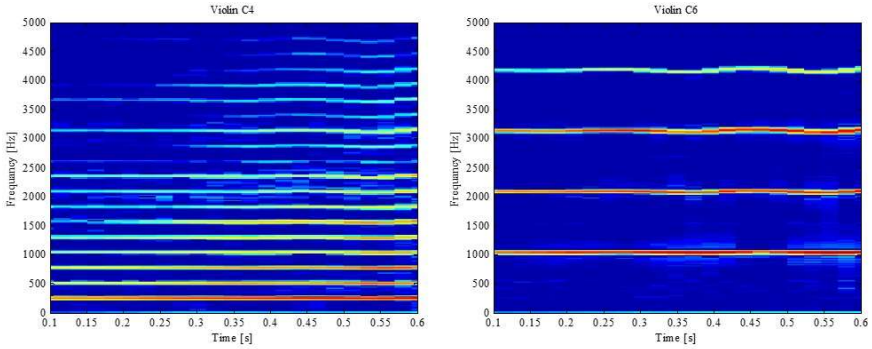
None of the mentioned representations is devoid of problems. For example, if there are millions of songs in a music service, then even very active users cannot listen to 1% of the music sources; thus, this may result in an unreliable recommendation if the co-occurrence-based method is considered. Contrary to the above consideration, low-level descriptors seem a straightforward representation. However, when comparing time- or time-frequency representation of music/speech signals, one may notice that sounds of the same instrument differ regardless of their representation (see Fig. 2). Obviously, male and female voices uttering the same sentence also differ (see Fig. 2). This may cause identification problems.

From the derivation of signal representation (as shown in Fig. 2) to much more sophisticated tasks is not so far. This may be illustrated based on the identification process

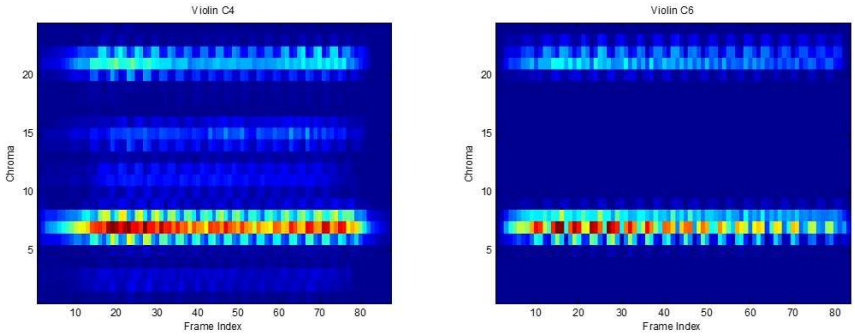
of mixed and often overlapped instruments in a music piece to decide which classes are contained in the audio signal [21]. This task was performed on Slakh dataset [22], designated for audio sources separation and multi-track automatic transcription, consisting of 2,100 songs. In most cases, there are four instruments in a song in Slakh. This concerns piano, bass, guitar, and drums.

In Fig. 3, a block diagram of the deep model designated for musical instrument identification is shown [21].

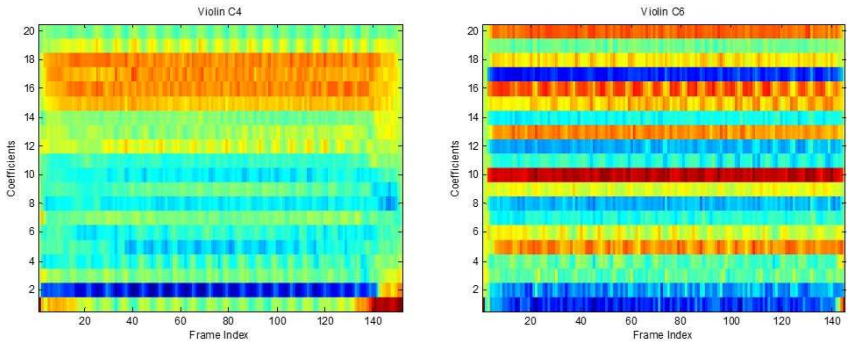
a.



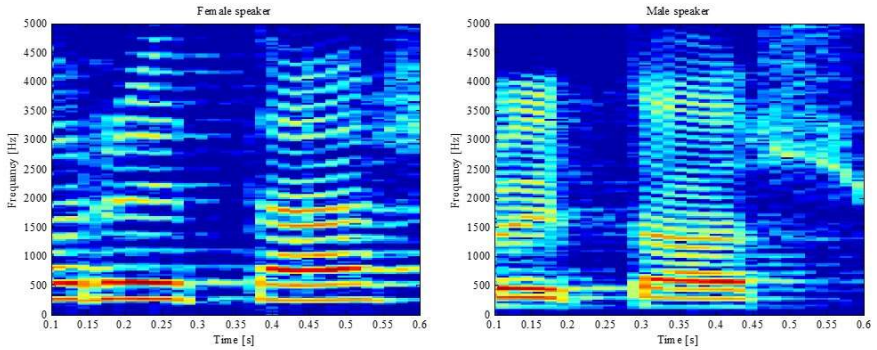
b.



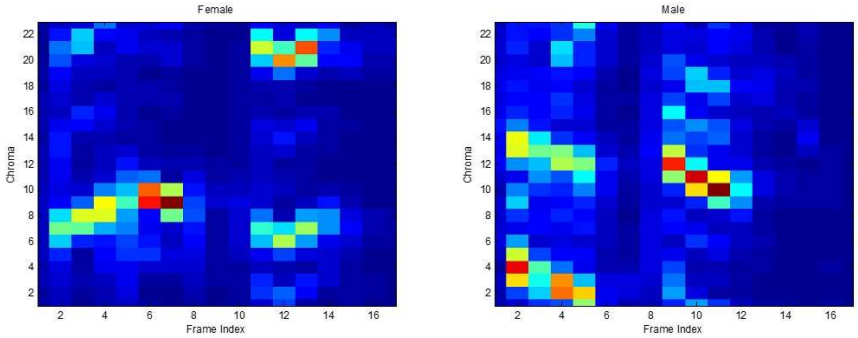
c.



d.



e.



f.

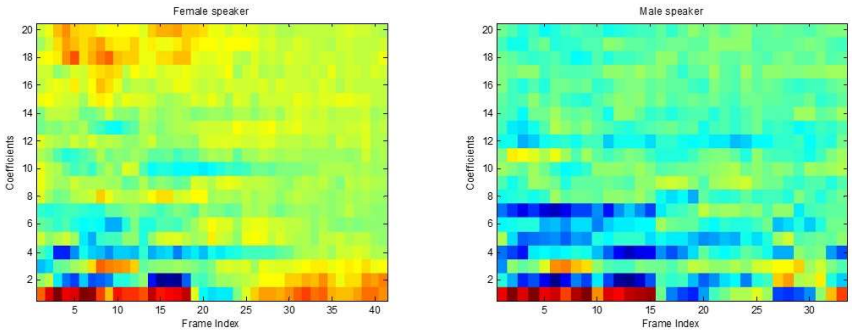


Fig. 2. Violin C4 and C6: (a) spectrograms, (b) chromagrams, (c) MFCCgrams; male/female utterance: (d) spectrograms, (e) chromagrams, (f) MFCCgrams.

The dataset for music identification was divided into three parts: training set – 116,369 examples; validation set – 6,533; evaluation set – 6,464. Figure 4 refers to the metrics obtained during the training and validation processes [21]. In Fig. 5, a histogram of instruments contained in a music piece, identified by the deep model, is presented.

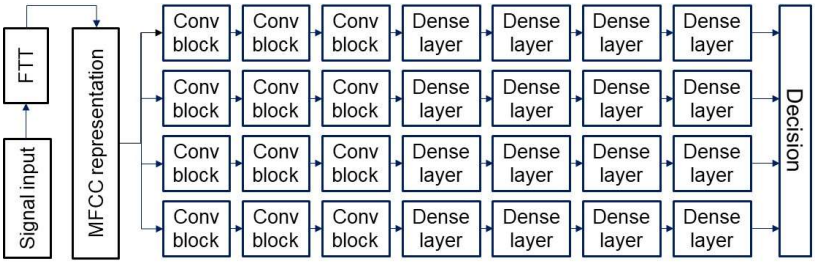


Fig. 3. A block diagram of the deep model designated for musical instrument identification based on Mel Frequency Cepstral Coefficients (MFCCs).

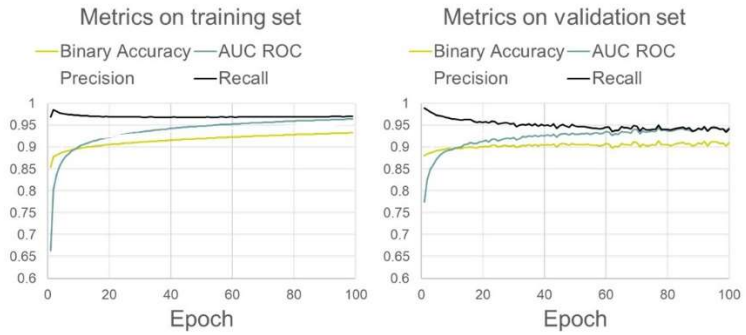


Fig. 4. Metrics on validation and training sets [21].

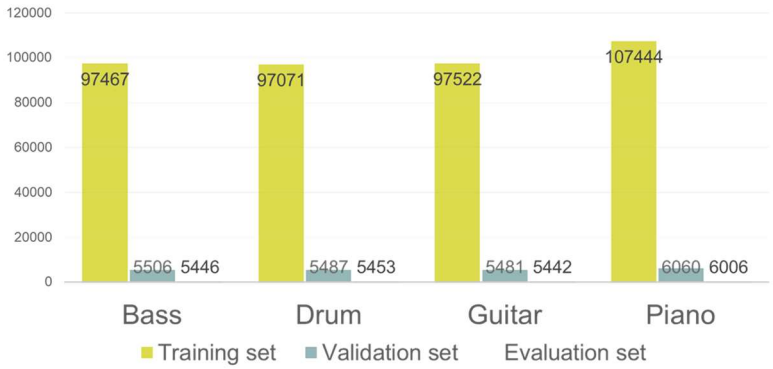


Fig. 5. Histogram of instruments contained in a music piece [21].

Overall results were as follows: precision – 0.95; recall – 0.94; AUC ROC (area under the ROC curve) – 0.94; true positive – 21,064; true negative – 2,470; false positive – 1,283; false negative – 1,039.

Another example concerns the autonomous audio mixing using the wave U-Net deep model [23]. The signal waveform and music genre label are provided at the net input.

Individual models are mixed to a stem (stem-mixing is a method of mixing audio material based on creating groups of audio tracks and processing them separately prior to combining them into a final master mix). Then, stems are mixed within the given genre to the entire mix.

In Fig. 6, spectrograms resulting from a mix prepared by a professional mixer and that of the deep U-Net model are shown. One can see that these signal representations are visually indistinguishable, which was further confirmed by the outcome of listening tests (see Fig. 7).

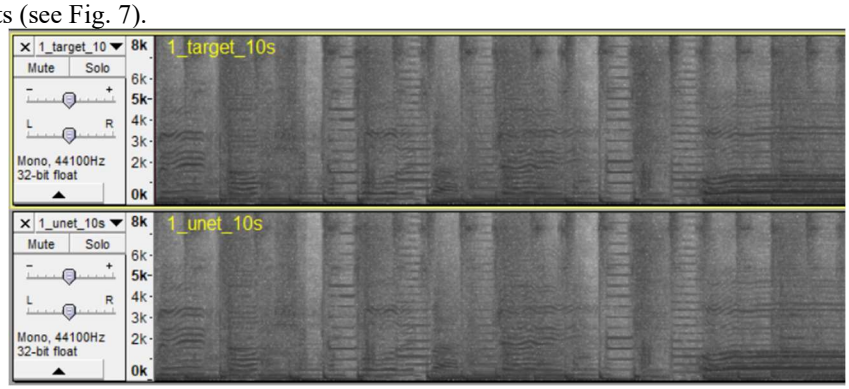


Fig. 6. Spectrograms resulted from autonomous audio mixing using mixes prepared by a professional mixer and the U-Net deep model.

In Fig. 7, the results of the subjective tests checking the quality of mixes prepared by autonomous deep model, technology-based (Izotope), anchor (filtered, low-quality sound), and reference mixes, the last one referring to professionally created mix [23], are shown. Listeners correctly identified both the reference and anchor signals. The U-Net model, in the listeners' opinion, is almost as good as the reference signal and is much better than state-of-the-art-based technology [23].

In music processing – information is often provided by tagging music and its user's behavior and actions (i.e., creating an ecosystem); it is contained in music services (MRS) within the frameworks of music ecosystem (music+users of music services); music content is analyzed at the low-level features, or there is a mixture of approaches.

In speech processing – datasets are collected by, e.g., automatically extracting speech and conversations from TV, radio, Facebook, YouTube, and other resources, as well as listened to and recorded by Alexa, Siri, Google, WhatsApp, etc. Indeed, there exist (and are still created) resources prepared manually dedicated to a particular problem [20]; however, as already mentioned, synthesized data may fill in these needs [13].

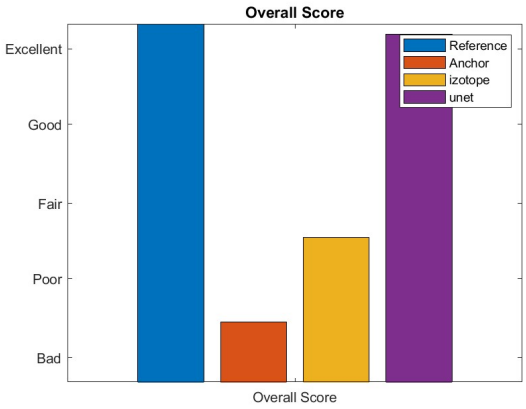


Fig. 7. Outcome of the subjective tests checking the quality of mixes prepared by autonomous deep mode, technology-based (Izotope), anchor (low-quality filtered signal), and the reference mixes, the last one referring to professionally created mix [23].

In the speech area, several applications may be discerned, e.g., speech recognition enabling communication, healthcare assistance, etc.; voice recognition/authentication systems; emotion recognition in speech and singing; voice cloning (testing vulnerability to attack speaker verification system); automatic aging of biometric voice; pronunciation learning by 2L (second language) speakers; automatic diagnosis or computer-aided diagnosis based on speech characteristics retrieved from the patient's voice (voice, speech and articulation disorders, Parkinson disease, dementia, dysarthria, etc.).

Voice authentication (VA), i.e., testing vulnerability to attack speaker verification system, based on DeepSpeaker-inspired architecture models using various parametrization approaches, brought high values of accuracy and a low level of equal error rate. The outcome of such a study for voice authentication based on the DeepSpeaker-inspired model, along with various representations, such as VC (vocoder), MFCC (Mel Frequency Cepstral Coefficients); GFCC (Gammatone Frequency Cepstral Coefficients), and LPC (Linear Predictive Coding Coefficients) is shown in Table 1. Depending on what criterion is more important, i.e., equal error rate (EER) or the number of epochs (each epoch took between 7 and 30 minutes), one may optimize the approach to VA.

Table 1. Outcomes of a DeepSpeaker-inspired model, along with various representations, such as VC (vocoder), MFCC (Mel Frequency Cepstral Coefficients); GFCC (Gammatone Frequency Cepstral Coefficients), and LPC (Linear Predictive Coding Coefficients).

Representation/model	F-score	Accuracy	EER	Epochs
VC model (MFCC)	0.875	0.997	0.0208	895
MFCC	0.784	0.8641	0.0829	400
GFCC	0.732	0.8132	0.1378	400
LPC	0.741	0.8216	0.0936	400

3 Conclusions

Challenges that could be identified within audio technology are related to the role of human factors such as, for example, the user's personality and experience, emotions in the user's models, and personalized services. Emotions are one of the most important aspects of interpersonal communication, as spoken words often − in addition to their content − contain additional, more deeply hidden meanings. Recognizing emotions, therefore, plays a crucial role in accurately understanding the interlocutor's intentions in all human-computer (and vice versa) technology. When searching for the keyword "emotion recognition in speech" on Google in December, the number shown was 17,500,000 results; today, as of January 8th, the value increased by almost 2 million. This shows the extremely high and growing importance of this issue, which can also be observed within the scientific community [24,25]

Moreover, speech signal contains phonemic variation, temporal structure, prosody, timbre, and voice quality. It also includes various aspects of the speaker's profile. State-of-the-art methods employ deep learning to recognize all these components in audio signals. One may say that what is easily discerned and analyzed by a human may no longer escape an ML-based approach, as this is already happening.

Finally, the author hopes that this paper is another voice in the discussion regarding whether this is already the stage when algorithms gain wisdom on their own.

References

1. Liew, A.: DIKIW: Data, Information, Knowledge, Intelligence, Wisdom and their Interrelationships. Business Management Dynamics (2), 49-62 (2013).
2. Rowley, J.: The wisdom hierarchy: Representations of the DIKW hierarchy. Journal of Information Science 33(2), 163–180 (2007). https://doi.org/10.1177/0165551506070706
3. Tuomi, I.: Data is more than knowledge: Implications of the reversed knowledge hierarchy for knowledge management and organizational memory. Journal of Management Information Systems 16(3), 103–117 (1999).
4. Wood, A. M.: The wisdom hierarchy: From signals to artificial intelligence and beyond. A framework for moving from data to wisdom; https://www.oreilly.com/content/the-wisdom-hierarchy-from-signals-to-artificial-intelligence-and-beyond/ [accessed Dec. 29, 2022].
5. Barlow, M.: Learning to Love Data Science, 2015, O'Reilly Media, Inc., ISBN: 9781491936580 [accessed Dec. 29, 2022].
6. Mahmood, I., Abdullah, H.: WisdomModel: convert data into wisdom. Applied Computing and Informatics, (2021). https://doi.org/10.1108/ACI-06-2021-0155 https://www.emerald.com/insight/content/doi/10.1108/ACI-06-2021-0155/full/html
7. Van Meter H. J.: Revising the DIKW pyramid and the Real Relationship Between Data, Information, Knowledge and Wisdom, Law, Technol Humans. (2), 69-80 (2020) doi: 10.5204/lthj.1470.
8. Jankowski, A., Skowron, A., Swiniarski, R.: Interactive Rough-Granular Computing in Wisdom Technology, In: 2013 INTERNATIONAL CONF. ON ACTIVE MEDIA TECHNOLOGY (AMT 2013), October 29-31, 2013, Maebashi, Gunma, Japan, Springer, Heidelberg, 1–13, (2013).

9. Skowron, A., Jankowski, A.: Toward W2T Foundations: Interactive Granular Computing and Adaptive Judgement In: Wisdom Web of Things (W2T), Springer International Publishing Switzerland, Cham, Switzerland (3), 47–71 (2016).
10. Vandeput N.: A Brief History Of Neural Networks from Data Science for Supply Chain Forecasting, https://medium.com/analytics-vidhya/a-brief-history-of-neural-networks-c234639a43f1 [accessed Dec. 29, 2022].
11. New Navy Device Learns By Doing; Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser. https://www.nytimes.com/1958/07/08/archives/new-navy-device-learns-by-doing-psychologist-shows-embryo-of.html [accessed Dec. 29, 2022].
12. Leung, K.: How to Easily Draw Neural Network Architecture Diagrams, https://towardsdatascience.com/how-to-easily-draw-neural-network-architecture-diagrams-a6b6138ed875 [accessed Dec. 29, 2022].
13. Korzekwa, D., Lorenzo-Trueba, J., Drugman, T., Calamaro, S., Kostek, B.: Weakly-Supervised Word-Level Pronunciation Error Detection in Non-Native English Speech. In: INTERSPEECH (2021). https://doi.org/10.21437/interspeech.2021-38.
14. W.-K. Leung, X. Liu, and H. Meng, "Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis. In: ICASSP (2019) 8132–8136.
15. Unquestioned assumptions to imperfect data [https://heyday.xyz/blog/research-project-challenges /] [accessed Dec. 29, 2022].
16. Moffat, D., Sandler, M. B.: Approaches in intelligent music production. Arts (8), 5, 14, September (2019).
17. De Man, B., Reiss, J.D.: A knowledge-engineered autonomous mixing system. Audio Engineering Society Convention 135, October (2013).
18. Martinez-Ramírez, M. A., Benetos, E., Reiss, J.D.: Automatic music mixing with deep learning and out-of-domain data. In: 23rd INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL CONF. (ISMIR), December (2022), 10.3390/app10020638.
19. Celma, O., Herrera, P., Serra, X.: Bridging the Music Semantic Gap, In: Bouquet P, Brunelli R, Chanod JP, Niederée C, Stoermer H, editors. In: WORKSHOP ON MASTERING THE GAP, FROM INFORMATION EXTRACTION TO SEMANTIC REPRESENTATION, with the EUROPEAN SEMANTIC WEB CONF.; Budva, Montenegro (2006) Jun 11-14.
20. Kostek, B.: Towards searching the Holy Grail in automatic music and speech processing - examples of the correlation between human expertise and automated classification. In: SIGNAL PROCESSING: ALGORITHMS, ARCHITECTURES, ARRANGEMENTS, AND APPLICATIONS (SPA) 16, (2022). doi: 10.23919/SPA53010.2022.9927877.
21. Slakh | Demo site for the Synthesized Lakh Dataset (Slakh). Available online: http://www.slakh.com/. [accessed: Dec. 29, 2022].
22. Blaszke, M., Kostek, B.: Musical Instrument Identification Using Deep Learning Approach. Sensors 22(8), 3033 (2022). https://doi.org/10.3390/s22083033.
23. Koszewski, D., Görne, T., Korvel, G., Kostek B.: Automatic music signal mixing system based on one-dimensional Wave-U-Net autoencoders. EURASIP, 1 (2023). https://doi.org/10.1186/s13636-022-00266-3.
24. Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar M. H., Alhussain, T.: Speech Emotion Recognition Using Deep Learning Techniques: A Review. IEEE Access (7), 117327-117345 (2019). doi: 10.1109/ACCESS.2019.2936124.
25. Konangi, U.M. Y., Katreddy, V. R., Rasula, S. K., Marisa, G., Thakur, T.: Emotion Recognition through Speech: A Review. In: 2022 INTERNATIONAL CONF. ON APPLIED ARTIFICIAL INTELLIGENCE AND COMPUTING (ICAAIC), 1150-1153 (2022). doi: 10.1109/ICAAIC53929.2022.9792710.