AI-Powered Knowledge and Expertise Mining in Healthcare from a Field Experiment

Julia Kauper[1](✉), Susanne Franke[2], Felix Franke[2] and Steven Grieshammer[3]

[1] Staatliche Studienakademie Plauen, Schloßberg 1, 08523 Plauen, Germany
julia.kauper@ba-sachsen.de

[2] d-opt GmbH, Oberneumarker Str. 59, 08496 Neumark, Germany
s.franke@d-opt.de

[3] Heinrich-Braun-Klinikum gGmbH, Karl-Keil-Str. 35, 08060 Zwickau, Germany

Abstract. With the increasing prevalence of mobile applications across various domains, there is a growing demand for individualised and self-adaptive learning pathways. This is particularly important in the mobile health sector, where there is a critical need to investigate how expert and experiential knowledge can be acquired, digitalised and formalised into data which is subsequently processed and further used. To address this demand, our research explores how Artificial Intelligence (AI) can power this process. We developed a prototype mobile application with a standardised learning pathway that features speech-language therapy exercises of varying levels of difficulty. In a 12-week field experiment involving 21 individuals with aphasia, we analysed the results using supervised and unsupervised algorithms. Our findings suggest that AI has the potential to generate new knowledge, such as identifying features that can determine which learning words are perceived as easier or more difficult on an inter-individual basis. This knowledge enables algorithmisation and the design of standardised (database-supported) artefacts, which in turn can be used to formulate self-adaptive and individualised learning pathways. This significantly enhances the development of effective mobile applications to assist speech-language therapy.

Keywords: Artificial Intelligence · Expertise Mining · Mobile Health · Speech-Language Therapy

1 Introduction

As the world becomes increasingly digitalised, there is a growing emphasis on the human element within technological environments. Technology is being used to adapt processes in e.g., production, healthcare, and education to better serve human needs [1]. With the shift from Industry 4.0 to Industry 5.0, the focus is set on promoting human well-being and developing skills and abilities [2]. Artificial Intelligence (AI) is an important and rapidly evolving technology that can support this process. Qualification processes for humans are of particular interest and relevance, for example, in training new employees and reskilling workers in the context of work, as well as in re-

C. Zinke-Wehlmann und J. Friedrich (Hrsg.): *First Working Conference on Artificial Intelligence Development for a Resilient and Sustainable Tomorrow*, AIDRST 2023, Informatik aktuell, S. 37–49, 2024. https://doi.org/10.1007/978-3-658-43705-3_4

covery from serious illnesses or medical procedures. In all these cases, there is a need for individualised learning processes that are tailored to each individual's skill level and any relevant restrictions (e.g., health or knowledge). Our goal is to formulate individualised learning content and exercises to optimise the learning progress for each participant. In this paper, we investigate how AI can support this process by providing deeper insights into the available data base. The paper is structured as follows: Sect. 2 provides background information, Sects. 3 presents the proposed methods, Sect. 4 and 5 describe the experimental results, and Sect. 6 provides a conclusion and outlook for future work.

2 Theoretical Background and Current State of the Art

AI as a branch of computer science deals with the development of algorithms to perform tasks traditionally associated with human intelligence, like the ability to learn and solve problems. A thorough statistical basis and an introduction into Machine Learning approaches is given in [3]. In [4], the focus is set on the algorithmic viewpoint and e.g. complexity analysis, whereas [5] provides an introduction into Deep Learning. AI applications are already prevalent in diagnostics, especially in the field of computer vision for tumour identification, see e.g., the study [6]. Recent developments were achieved in disease classification based on segmentation [7] and in natural language processing [8].

Especially in the context of Industry 5.0, the focus is set on the "human in-the-loop" (HITL), whereas AI is seen as supporting technology: The human provides an extensive data base of high quality (which is required to not be biased), detects poor training results, interprets them and extracts knowledge for further use. Explainable AI for methods which do not provide such an explanation of the solution-finding process (as e.g. the decision tree algorithm does) is an avid field of research. For an overview, see [9]. However, the results are mostly theoretical and not yet in application. The choice of an algorithm highly depends on the planned usage of the results. From a practical point of view, users prefer to apply algorithms that provide a high level of (inherent) explainability. Measuring learning progress and learning analytics is an active field of research, as evidenced by numerous publications. A reference model is given by [10], whereas [11] provides a framework of quality indicators to standardise the evaluation of learning analytics tools. Various empirical studies and implementations, such as [12], also exist.

HITL and AI-enabled adaptive learning systems have several capabilities such as generating individualised learning pathways adapted to individual progress, preferences, and pace [13]. A user-centred approach enables system feedback and correction, adapting to users' learning styles and knowledge, providing individualised learning pathways aligned with specific needs and individual goals [14], ensuring higher accuracy and effectiveness by adapting to changing conditions and data [13]. Adaptive learning systems have broad applications across domains, including language education using frameworks such as the Common European Framework of Reference for Languages (CEFR) [15]. Speech-language therapy lacks such frameworks and relies on the expert and experiential knowledge of speech-language therapists (SLTs) using different treatment guidelines, techniques, and technologies based on the principles of evidence-based practice (EBP), comprehensive diagnosis, and individual therapy goals [16].

Digital applications used in speech-language therapy include high-tech augmentative and alternative communication (AAC) tools such as speech-generating devices, apps on tablets [17] and interactive tables with touchscreens [18]. There is also an increasing usage of mobile apps [19]. Machine learning can enable SLTs to automatically adjust training and rehabilitation devices remotely based on a tailored treatment plan [20], thereby providing a personalised and more effective therapy experience. Hence, there is a growing need to explore expert and experiential knowledge in data to design self-adaptive and individualised learning pathways.

3 Methodology

To create self-adaptive and individualised learning pathways for mobile applications in speech-language therapy, expert and experiential knowledge needs to be translated into data and implemented as artefacts in an app for testing and evaluation. Following an iterative approach, we combine two different methods to achieve this goal: 1) the enhanced Action Design Research (ADR) process for the design and development of the mobile app, and 2) the Cross-Industry Standard Process for Data Mining (CRISP-DM) as part of the evaluation to gain domain knowledge and thus improve the artefacts.

ADR aims to tackle a real-world problem within a specific organisational context while simultaneously designing and evaluating an IT artefact that deals with a set of issues inherent to that situation [21]. By incorporating a field experiment into the ADR process for iterative refinement and evaluation of the mobile app, we combined design science and field experiments as described in [22].

CRISP-DM is an open, industry-, software- and application-independent standard process model and provides an approach for data mining processes. The model consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Assessment and Deployment. [23]

The following sections describe the application and results of these methods in more detail.

4 Enhanced ADR and Field Experiment

The app development and the associated transformation of expert and experiential knowledge into data were guided by the enhanced ADR with the following stages:

Stage 1) Problem formulation: To understand how a mobile app can support speech-language therapy, patients' specific needs and objectives were identified following EBP as recommended by the professional guidelines for speech therapy of the German Federal Association for Speech Therapy (dbl) [24], taking into account factors such as patient expectations, therapist's clinical experience, and available scientific evidence [25]. Data was collected through scientific research and interviews with SLTs to investigate feasibility, including information on patient diagnoses and individual therapy goals. Therapy goals take into account factors like speech impairment type and severity, individual needs and preferences, and participation limitations according to the International Classification of Functioning, Disability and Health ICF [26]. Despite challenges in achieving standardisation, considerations were made towards standardisation and algorithmisation in designing a standardised database-supported portfolio that considers individual needs, the severity and manifestation of speech disorders, and different therapy goals.

Stage 2) Building, intervention, and evaluation: SLTs designed a uniform core learning pathway for all participants as a foundation for creating individualised learning pathways based on therapy goals, diagnoses, patient-specific needs and abilities, and app usage data. This core learning pathway consisted of a predefined sequence of three evidence-based speech-language exercises per day on at least five days per week such as picture-word association, spelling, and selection of the correct sentence with increasing difficulty. Semantic fields (categories) with medium to high intersubjective relevance to daily life of the participants, including food, clothing, housing, body, everyday objects, garden, illness, and personal care, taking into account the gender and age distribution of the patients, were identified. Subsequently, we selected related target words, defined data requirements for the prototypical mobile app for speech-language therapy, and evaluated user acceptance, design, as well as usability through field trials.

Stage 3) Field experiment: A 12-week field experiment was conducted from January to April 2022 with 21 post-stroke aphasia patients in Germany with varying levels of severity to assess the efficacy and usability of the app under real-world conditions. Post-stroke aphasia is a language disorder, affecting any aspect including speaking, writing, reading, and comprehension, that arises following a stroke [27] and is often accompanied by comorbidities such as motor impairments like hemiparesis and cognitive deficits like memory loss [28]. Eligibility criteria included a diagnosis of aphasia, premorbid competence in German, and sufficient auditory/visual acuity to interact with a tablet. Some eligible participants were excluded due to resource constraints, i.e., the lack of access to a device or internet. Participants' mean age was 63.6 years (range 38–81), and 67% were male. Personal information was provided through an anamnesis questionnaire, including visual and hearing impairment, handedness, and stroke history. Data on the specific type of aphasia, severity, type of stroke and co-morbidities were not collected as they were not considered relevant to the inclusion or exclusion criteria. This is because the design of the speech-language exercises aimed to take into account the heterogeneity of clinical disorders observed in people with aphasia. User feedback was collected through two questionnaires: a usability test at first use and a final questionnaire, which included self-assessment of the difficulty of the exercises. Progress was monitored, analysed, and evaluated at both individual and inter-individual level.

Stage 4) Reflection and learning: Through the evaluation of the field experiment's results, we engaged in deliberate and conscious reflection on the problem formulation, our working hypotheses, and the app artefacts that were developed. The next section presents and discusses the outcomes of the first five phases of the CRISP-DM deployment process. These results will be applied in *Stage 5* of the enhanced ADR, *Formalisation of learning*, to optimise and formalise our models and assumptions.

5 Exploration of Expert Knowledge Using CRISP-DM

To optimise the learning curve for each individual patient, the exercises have to contain words of a suitable level of difficulty. In the following, data-based evidence is extracted to evaluate the chosen level of difficulty for each word and to provide relevant features, extract potentially irrelevant features and suggest additional features to be acquired. Therefore, we applied CRISP-DM to the example of the picture-word association exercise and its difficulty factors to answer the following two questions: (1) How do the assumptions on difficulty-determining features affect the frequency of errors per word across users with heterogeneous clinical presentations in terms of "easy", "medium" or "hard"? (2) Which features lead to the classification of a target word as "easy", "medium" or "hard"?

5.1 Understanding of the Use Case

The picture-word association exercise in the app prototype presents a picture of the target word on the right side of the screen and four written words (the target word and three distractors) on the left and centre. Participants must correctly identify and match the word with the corresponding picture. The exercise's difficulty increased weekly based on factors such as the type of exercise, prototypicality and unambiguous representation of the target word in the picture, semantic field (category), and, more detailed below, target word/item and distractors.

Several factors determine the difficulty of the target word. These include individual frequency of use [29], degree of concreteness according to the Dual Coding Theory (DCT) [30], prototypicality (e.g., "Möhre" (carrot Eng.) is more prototypical than "Schwarzwurzel" (salsify Eng.) for the category *vegetables*), length and structure of the word (shorter words with simpler syllable structures (e.g., CV) are easier to learn than longer words with more syllables and complex syllable structures [31]), level of specification within a semantic hierarchy (e.g., "Hund" (dog Eng.) is easier to learn than "Dackel" (dachshund Eng.)), and properties such as dialectal peculiarities/synonyms, imageability, number of phonemes, compound words, and foreign words.

The properties of the three distractors (semantic, phonological, and non-relational) in relation to the target item also influence the difficulty level of this exercise. For *semantic distractors*, the level of difficulty increases with the closeness to the target item (equivalence in terms of prototypicality/type frequency, imageability/visual redundancy or distinctiveness, semantic/relational proximity). For example, "Keks" (cookie Eng.) is an easier semantic distractor for the target word "Möhre" in the semantic field food than "Rettich" (radish Eng.). For *phonological distractors*, it holds that the more similar the phonological-lexical properties (compared to the target word), the more difficult this distractor is [32]. For example, "Matratze" (mattress

Eng.) has the same initial letter as “Möhre”, but a different number of syllables as well as a different number of graphemes and is therefore an easy phonological distractor. “Möwe” (seagull Eng.) is a more difficult distractor, as it has the same initial letter, same number of syllables, similar syllable structure, same vowel length, and it also rhymes with “Möhre”. *Non-relational distractors* must not be in a semantic or phonological-lexical relation to the target word or to any of the other distractors, like “Goldkette” (gold chain Eng.) for “Möhre” (carrot Eng.), as frequent errors with this distractor indicate guessing.

Data collection took place during and after the 12-week test phase and included questionnaires, recorded app usage data, a vocabulary database with items (e.g., target words, and distractors) created by the SLTs, and developed attributes. The criteria that determine the level of difficulty of an item were elaborated and described in terms of their characteristics.

5.2 Data Understanding

This phase involved a thorough investigation of the provided database. The data on targets included around 25 features, with 11 having been identified as relevant for this investigation by the SLTs and data analysts. Hereby, certain features were neglected for this investigation (this mainly depended on the availability and validity of the data). Since the test phase was already completed, it wasn’t possible to acquire additional data from the test phase itself as well as patient-specific characteristics (e.g., displayed distractors which weren’t clicked; the patients’ familiarity with items from certain semantic fields).

The selected features included categorical values like difficulty and concreteness/abstractness, integer values like number of graphemes, Boolean values like compound words, and strings like structure syllables. The same attributes were provided for the distractors, with the additional feature “type” with its feature space {semantic, phonological, non-relational}. Additionally, data from the test phase was provided, including e.g., exercise ID, patient ID, time stamps of each click within the exercise-solving process by the patient, correct target value, falsely chosen distractors, used support (audio, joker). For a better understanding of the data, uni- and multivariate data analysis was performed: Histograms and boxplots allowed a deeper insight on distributions and possible outliers (which occurred, for example, regarding the duration of an exercise or the total number of errors per item). Additionally, the calculation of correlation coefficients and category-wise variance analysis provided a deeper understanding of the features’ connections (e.g., there is an obvious correlation between the length of a word and the number of its syllables, but the word type frequency is not correlated with the other item’s features and the type-specific number of errors). Appropriate visualisations, both individual and inter-individual, included e.g., the number of trainings per day and exercise (for each week) as stacked bar charts, the distribution of the types of errors (semantic, phonological, and non-relational) as pie charts as well as box plots, and the temporal distribution of the trainings as scatter plots.

5.3 Data Preparation

In the context of univariate data analysis, it became apparent that the duration of some picture-word association exercises lasted up to 23 times as long as the average time and/or were not finished. Unfinished exercises were neglected; regarding the duration, a time limit was set in coordination with the SLTs. Furthermore, plausibility checks performed in Python 3.9 showed that for exercises exceeding one day, incorrect data was stored in the database regarding the type of exercise. These mistakes were corrected. There were no missing values.

Further, it was required to integrate the data from the different sources. The characteristics of the targets and distractors were provided in various Microsoft Excel files, as they were assembled manually by the SLTs. Data from the test phase were acquired automatically and made available in a MySQL database. The relevant data was extracted and structured with regard to the analysis to be performed in a new MySQL database. This also included the calculation of new features like the type-specific (semantic, phonological, and non-relational) number of errors per exercise.

After data cleaning and integration, the data was then prepared in accordance with the AI algorithms to be applied. For the decision tree algorithm, both numerical and categorical data are allowed, and the data do not need to be normalised. Regarding clustering methods, the preprocessing process included distribution and outlier handling, transformation and normalisation of numerical data, and removing highly correlated features.

5.4 Modelling

Two different AI approaches were implemented. First, the decision tree as supervised method was applied to identify the features and their values leading to the difficulty of each target (with difficulty space {easy, medium, hard}). The main advantage of this approach were the explainable results, which means that the features and their respective values leading to the categorisation could be provided and discussed with the SLTs. As hyperparameters, the gini criterion was chosen, with a maximum depth of 10 being sufficient to provide unique classification results.

Second, clustering algorithms were applied as unsupervised methods to check which grouping of items was suggested. For centroid-based methods, the elbow method to identify the optimal number of clusters led to three (which matched the number of categories the SLTs chose). For comparison, two and four clusters were also set as hyperparameters. Additionally, connectivity-based methods were applied.

Besides the items' characteristics (which formed the basis for the SLTs to derive the difficulty of the items), the results from the test phase could also be included, i.e., the number of mistakes for each of the three difficulty types. Hereby, the types semantic, phonological, and non-relational distractors were distinguished. Both approaches (leaving out and including data from the test phase) were applied and evaluated individually.

Additionally, statistical methods for multivariate analysis were applied to compare the feature values of targets with same difficulty. This allowed the identification of conspicuous targets or peculiarities with respect to the semantic fields. For example,

Fig. 1 in Sect. 5.5 shows the comparison of selected feature values of target words with the same difficulty for the semantic fields clothing and food.

5.5 Evaluation

The results were discussed with the SLTs. Regarding the decision tree, an iterative investigation of the leaf nodes revealed items which weren't categorised easily and required deep branching. Targets which were difficult to categorise had an 85% intersection for both approaches (leaving out and including test phase data), indicating a high consistency of the SLTs' and patients' perception. If considered, the relative number of errors heavily influenced the decision nodes, but not the results as much – especially if a sufficiently high depth was chosen as hyperparameter. Tab. 1 shows the impact of features selected for the algorithm with maximum depth value of 5. Here, the relative number of misclassified targets (also specifying the respective falsely stated difficulty) is depicted for three cases of selected features. It becomes obvious that including the data acquired during the test phase with the patients significantly improves the results. Simultaneously, the typical problem of words misclassified as "medium" is still apparent.

Tab. 1. Number of misclassified targets under different features selected for depth 5

Decision tree of depth 5	Easy	Medium	Hard
Features excluding errors			
Misclassified as "Easy"	–	10%	17%
Misclassified as "Medium"	17%	–	17%
Misclassified as "Hard"	3%	8%	–
Features including errors in total			
Misclassified as "Easy"	–	12%	3%
Misclassified as "Medium"	3%	–	17%
Misclassified as "Hard"	3%	2%	–
Features including all types of errors			
Misclassified as "Easy"	–	2%	3%
Misclassified as "Medium"	0%	–	17%
Misclassified as "Hard"	0%	0%	–

Five features were identified to have major impact on the categorisation. Certain semantic fields could be classified easily as suggested by the SLTs, whereas targets with "medium" difficulty were the most problematic ones. These items were further investigated using clustering results: Changing the number of clusters (and, hence, the number of difficulty categories) to two provided a deeper insight into the connection between the item characteristics and the results from the test phase. Most of the target values with difficulty "easy" and "hard" were assigned to the two different clusters, respectively, indicating that the SLTs' classification of "easy" and "hard" was mostly in accordance with patients' perception. Target values with "medium" difficulty were reassorted. It became apparent that for each semantic field with this difficulty, either a reassignment to "easy" or "hard" heavily predominated. The results brought deeper insight into which features might be relevant for the assignment to a cluster. Creating four clusters is however challenging from the application point of view: A further division of the difficulty "medium" would be necessary. As for three clusters, about half of the targets matched the SLTs' classification, where the targets identified as critical with the decision tree method were assigned to different clusters with this method as well. These results in connection with the target analysis on targets with same difficulty provided the basis for the SLTs to find causes for targets with a high deviation from the original difficulty. Such reasons were, e.g., an unsuitable or ambiguous picture provided in the test phase.

Several working hypotheses were confirmed by cross-patient valid attribute rules. For example, the difficulty of exercises was found to depend on the type of exercise, and the percentage of phonological and semantic errors in the picture-word association exercise increased with higher difficulty, indicating appropriate distractors were developed for these levels of difficulty. Additionally, for the category food at an easy level, there was a positive correlation between word type frequency and error rate. This confirmed the hypothesis that words that occur more frequently in the German corpus (e.g., "Marmelade", jam Eng.) would elicit fewer errors than words that occur less frequently (e.g., "Käse", cheese Eng., or "Tee", tea Eng.). However, word type frequency alone was not sufficient to explain the error patterns. Type frequency from the lexical database dlexDB [33] was used, which is based on the reference corpus of the German language of the 20th century from the Digital Dictionary of the German Language (DWDS) and ranges from 1900 to 1999. Therefore, the type frequency of certain words does not allow any conclusion on the current usage frequency of these words. For instance, we found more errors for words with shorter word length than for words with longer word length within the same category and level of difficulty. This contradicted the assumption regarding word type frequency and number of syllables. Possible factors that could account for this discrepancy are syllable structure, picture quality, distractor choice, or everyday relevance (depending on each patient's perception). We also performed pair- and groupwise comparisons. For the two categories food and clothing at an easy level of difficulty, the similarity between features like word type frequency across both categories is depicted in Fig. 1. However, there were fewer phonological and semantic errors for food compared to clothing. A possible conclusion is that the visual representation of food items may show a greater degree of prototypicality and a closer alignment with everyday visual perception as compared to items of clothing. Furthermore, food might activate other sensory modalities such as taste.

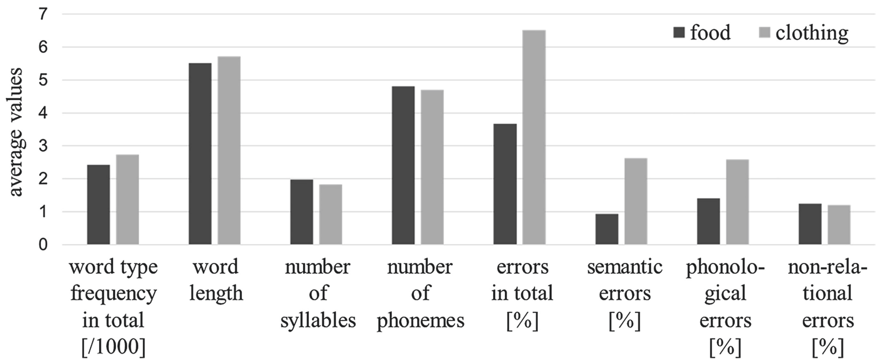

Fig. 1. Comparison of the selected feature values of target words with the same level of difficulty for the categories clothing and food

6 Conclusion and Outlook

Research on the transformation and exploration of expert and experiential knowledge into data using AI is crucial for the development of self-adaptive and individualised learning pathways in mobile applications to support patient-centred speech therapy treatments. In this study, we developed a prototype app, tested it with aphasic patients in a field experiment, and subsequently analysed and evaluated the results using different AI approaches. The study confirmed some of the assumptions and criteria made by SLTs for developing a standardised core learning pathway, while also uncovering patterns and weaknesses that can enhance the revision and improvement of the prototype.

Further development of the prototype could enable patients with speech disorders to train continuously with increasing demands under professional supervision. The individual therapy goal setting, taking into account participation goals (ICF) and inter-individual comparisons using newly developed algorithms for comparability and pattern recognition, is a pioneering basis. The significance of the current results is mainly restricted by the small number of patients and the brevity of the test phase. While the application of the decision tree algorithm allows a high transparency for the SLTs, it is highly susceptible to overfitting. Systematically selecting different features and applying suitable maximum depths helped to avoid those downfalls; with a bigger data set, random forests as an ensemble method could also be applied. In the future, we plan to test Deep Learning approaches as well to get new insights. The inherent problem of decision trees to be biased towards the dominant class in imbalanced data sets must be considered as well. Clustering methods, on the other hand, as unsupervised methods, have the disadvantage of not providing explainable results. Additionally, the quality of the results highly depends on the data points' location in the feature space, such that suitable algorithms must be chosen.

To overcome the described inaccuracies, we plan to increase the sample size for future studies, adjust the app and vocabulary databases, and revise the criteria for determining the difficulty of a target word and exercise. Additionally, a more precise description of individual and inter-individual evaluations would enable a more accurate definition of the data collection requirements, including additional patient-specific data. We also recommend developing new item-specific attributes that can adapt the difficulty of an exercise to individual learning progress.

Acknowledgments. This work was partly funded by the EFRE/ESF SMS funding program "Förderung innovativer Ansätze im Bereich der Gesundheits- und Pflegewirtschaft", project IMPULS-SACHSEN 2.0, and co-financed with tax funds on the basis of the budget approved by the Saxon State Parliament, project Sprach-Impulse. We would like to express our gratitude to all those who have contributed to this paper in various ways. We would like to thank the study participants for their time and cooperation, and all project partners for sharing ideas and data.

References

1. European Commission, Directorate-General for Research and Innovation, Breque, M., De Nul, L., Petridis, A.: Industry 5.0: towards a sustainable, human-centric and resilient European industry, Publications Office (2021). https://data.europa.eu/doi/10.2777/308407.
2. Rannertshauser, P., Kessler, M., & Arlinghaus, J. C. (2022). Human-centricity in the design of production planning and control systems: A first approach towards Industry 5.0, *IFAC-PapersOnLine 55*(10), 2641–2646. https://doi.org/10.1016/j.ifacol.2022.10.108.
3. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning* (1st edn., pp. 67–227). New York: Springer.
4. Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning. From Theory to Algorithms* (1st edn., pp. 115–284). Cambridge University Press.
5. Goodfellow, 1., Bengio, Y., & Courville, A. (2016). *Deep Learning* (pp. 162–481). The MIT Press.
6. Dembrower, K., Wåhlin, E., Liu, Y., Salim, M., Smith, K., Lindholm, P., Eklund, M., & Strand, F. (2020). Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. *Lancet Digit Health, 2*(9), 468–474. https://doi.org/10.1016/S2589-7500(20)30185-0.
7. Kok, E.Y., Pszczolkowski, S., Law, Z. E., Ali, A., Krishnan, K., Bath, P. M., ... & French, A. P. (2022). Semantic Segmentation of Spontaneous Intracerebral Hemorrhage, Intraventricular Hemorrhage, and Associated Edema on CT Images Using Deep Learning. *Radiology: Artificial Intelligence, 4*(6),3–4. https://doi.org/10.1148/ryai.220096.
8. Fink, M. A., Kades, K., Bischoff, A., Moll, M., Schnell, M., Küchler, M., ... & Kleesiek, J. (2022). Deep Learning-based Assessment of Oncologic Outcomes from Natural Language Processing of Structured Radiology Reports Radiology. *Artificial Intelligence, 4*(5),5–7. https://doi.org/10.1148/ryai.220055.
9. Gohel, P., Singh, P., & Mohanty, M. (2021). Explainable AI: Current status and future directions. *arXiv preprint* arXiv:2107.07045. https://doi.org/10.48550/arXiv.2107.07045
10. Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2013). A reference model for learning analytics. *International Journal of Technology Enhanced Learning, 4*(5–6), 318–331.
11. Scheffel, M., Drachsler, H., Stoyanov, S., & Specht, M. (2014). Quality indicators for learning analytics. *Educational Technology & Society, 17*(4), 117–132.

12. Ifenthaler, D., & Yau, J. Y. K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Education Technology Research and Development 68,* 1961–1990. https://doi.org/10.1007/s11423-020-09788-z.
13. Kabudi, T., Pappas, I., Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers in Education: Artificial Intelligence, 2,* 100017. https://doi.org/10.1016/j.caeai.2021.100017.
14. Vesin, B., Mangaroska, K., & Giannakos, M. (2018). Learning in smart environments: user-centered design and analytics of an adaptive learning system. *Smart Learning Environments, 5*(1), 24. https://doi.org/10.1186/s40561-018-0071-0
15. Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume, Council of Europe Publishing, Strasbourg, available at www.coe.int/lang-cefr, Accessed: 26. March 2023.
16. Grötzbach, H. (2018). Evidenzbasierte Praxis als Grundlage sprachtherapeutischer Entscheidungen. *Sprache-Stimme-Gehör, 42*(03), 133–139.
17. American Speech-Language-Hearing Association (ASHA): Augmentative and Alternative Communication (AAC) Homepage, http://www.asha.org/public/speech/disorders/AAC/, Accessed: 15. March 2023.
18. Plößel, L., Heide, J., & Netzebandt, J. (2021). Einsatzmöglichkeiten des ActivTable in der kognitiv orientierten Sprachtherapie: eine Exploration. In: Fritzsche, T., Breitenstein, S., Wunderlich, H., Ferchland, L. (eds.): *Spektrum Patholinguistik* (Band 14., pp. 83–92). Universitätsverlag Potsdam, Potsdam. https://doi.org/10.25932/publishup-51034.
19. Meier, B., & Leinweber, J. (2023). Ein Update zu digitalen Interventionen in der Logopädie/Sprachtherapie. *Neurologie & Rehabilitation, 29*(1), 27–33.
20. Mulfari, D., La Placa, D., Rovito, C., Celesti, A., & Villari, M. (2022). Deep learning applications in telerehabilitation speech therapy scenarios. *Computers in Biology and Medicine, 148*, 105864. https://doi.org/10.1016/j.compbiomed.2022.105864.
21. Sein, M. K., Henfridsson, O., Purao, S., Rossi, M., & Lindgren, R. (2011). Action Design Research. *MIS Quarterly, 35*, 37–56. https://doi.org/10.2307/23043488.
22. Sun, R. T., Han, W., Chang, H. L., & Shaw, M. J. (2021). Motivating adherence to exercise plans through a personalized mobile health app: enhanced action design research approach. *JMIR mHealth and uHealth, 9*(6), e19941. https://doi.org/10.2196/19941.
23. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide, Article 13.
24. Deutscher Bundesverband für Logopädie e.V. (dbl) (2012). Berufsleitlinien Logopädie. Fassung des Beschlusses der Mitgliederversammlung vom 04.06.2010. Frechen.
25. Aichert, I., & Staiger, A. (2015). Therapie der Sprechapraxie. *Forum Logopadie., 29*(5), 30–37. https://doi.org/10.1055/b-0037-144279.
26. World Health Organization (WHO): International Classification of Functioning, Disability and Health (ICF). https://www.who.int/classifications/international-classification-of-functioning-disability-and-health, Accessed: 3. March 2023.
27. Berthier, M. L., & Pulvermüller, F. (2011). Neuroscience insights improve neurorehabilitation of poststroke aphasia. *Nature Reviews Neurology, 7*(2), 86–97.
28. Ogar, J., Slama, H., Dronkers, N., Amici, S., & Gorno-Tempini, M. L. (2005). Apraxia of speech: An overview. *Neurocase, 11*(6), 427–432.
29. Cherney, L. R., Patterson, J. P., Raymer, A., Frymark, T., & Schooling, T. (2008). Evidence-based systematic review: Effects of intensity of treatment and constraint-induced language therapy for individuals with stroke-induced aphasia. *Journal of speech, language, and hearing research 2008 Oct;51*(5), 1282–99. https://doi.org/10.1044/1092-4388(2008/07-0206).
30. Paivio, A. (2014). Intelligence, dual coding theory, and the brain. *Intelligence, 47*, 141–158. https://doi.org/10.1016/j.intell.2014.09.002.

31. Nickels, L., & Howard, D. (2004). Dissociating effects of number of phonemes, number of syllables, and syllabic complexity on word production in aphasia: It's the number of phonemes that counts. *Cognitive Neuropsychology, 21*, 57–78.
32. Bi, Y., Xu, Y., & Caramazza, A. (2009). Orthographic and phonological effects in the picture–word interference paradigm: Evidence from a logographic language. *Applied Psycholinguistics, 30*(4), 637–658. https://doi.org/10.1017/S0142716409990051.
33. dlexDB Dokumentation. http://www.dlexdb.de/documentation/overview/, Accessed: 21. March 2023.