Integrating Multimodal Inputs such as Text, Audio, and Visuals in Learning by Reading Systems to Enhance Long-Term Knowledge Retention

Ramon Gatchalian[1] and Isabella Leon[2]

[1]Pangasinan State University, College of Computing Studies, Alvear Street, Lingayen, Pangasinan, Philippines
[2]University of Southeastern Philippines, School of Computer Studies, Bo. Obrero Street, Davao City, Philippines

2023

Abstract

Multimodal learning by reading systems leverage textual, auditory, and visual information to facilitate improved comprehension and retention of complex knowledge. This approach rests upon the premise that interlinking multiple sensory channels can yield more robust cognitive representations, thereby enhancing memory consolidation processes over extended time spans. The incorporation of text, audio, and visual cues can encode concepts through complementary paths, which reduces the likelihood of memory decay and allows for more effective retrieval during high-level reasoning tasks. Recent advances in representation learning, attention mechanisms, and distributed architectures provide opportunities to unify heterogeneous data streams and automatically infer latent structures that guide learners' comprehension. Underlying these technologies are mathematical frameworks that accommodate asynchronous and synchronous modalities, enabling flexible training scenarios for various application domains such as scientific education, domain-specific skill acquisition, and interactive tutoring systems. By bridging signal processing, linguistic analysis, and psycholinguistic principles, multimodal learning by reading systems address challenges related to information overload, ambiguity, and the abstract nature of textual resources. Moreover, integrating memory-centric models with knowledge-based reasoning techniques can improve users' ability to apply acquired knowledge in real-world settings. This paper explores how the fusion of text, audio, and visuals in reading systems can strengthen long-term knowledge retention and examines potential research directions for optimizing the design of such systems.

1 Introduction

The design of learning systems that enhance long-term knowledge retention requires careful consideration of the cognitive factors influencing how learners encode, store, and retrieve information [1]. By introducing multimodal inputs such as text, audio, and visuals, researchers have sought to discover new methods of leveraging multiple channels of perception to foster deeper comprehension and more robust memory traces. These memory traces involve the formation of associations between an entity's features in semantic, episodic, and procedural contexts [2]. In formal terms, if the space of concepts can be denoted by $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$, then the encoding of these concepts into long-term memory corresponds to embedding them into a high-dimensional cognitive space $\mathcal{H}$, where multimodal cues contribute to constructing distributed representations. The fundamental cognitive models posit that connections established between different sensory representations of a concept can reinforce recall, reduce interference, and improve the learner's capacity for extrapolation.

Advanced computation techniques have amplified the possibilities of fusing textual, auditory, and visual data streams [3]. Let $\mathbf{x}_t \in \mathbb{R}^{d_t}$ represent the textual feature vector of a concept, $\mathbf{x}_a \in \mathbb{R}^{d_a}$ represent its auditory feature vector, and $\mathbf{x}_v \in \mathbb{R}^{d_v}$ represent its visual feature vector. The objective within many contemporary architectures is to learn a joint transformation $\mathbf{z} = f(\mathbf{x}_t, \mathbf{x}_a, \mathbf{x}_v)$ mapping into a latent space $\mathcal{Z} \subseteq \mathbb{R}^d$. The dimension d may be selected based on the complexity of the concept domain, balancing model capacity with the constraints of learning stability [4]. The function f may be decomposed into separate modules $f_1(\mathbf{x}_t)$, $f_2(\mathbf{x}_a)$, $f_3(\mathbf{x}_v)$, which are then integrated through summation, concatenation, attention-based fusion, or factorization. For instance, one may

define $\mathbf{z} = \sigma(W_t f_1(\mathbf{x}_t) + W_a f_2(\mathbf{x}_a) + W_v f_3(\mathbf{x}_v))$, where W_t, W_a, W_v denote weight matrices and σ is a non-linear activation. The interplay among these transformations reflects how diverse modalities contribute complementary information regarding the same concept.

In reading-based educational contexts, a key challenge arises from the necessity of effectively synchronizing textual narratives with audio commentary and related visuals [5]. For extended periods, reading systems were designed as text-centric platforms, while audio and visual elements were often supplemental or absent [6]. However, as neural networks have made it feasible to align multiple streams automatically, developers can now produce more cohesive learning experiences [7]. Observational evidence indicates that concepts communicated via a blend of textual description, spoken explanation, and illustrative visuals may have a higher likelihood of retention, especially in tasks requiring a detailed understanding of structure or functionality [8], [9]. Additionally, it is crucial to take into account the temporal characteristics of audio data and the segmentation of visual data [10]. Let there be a sequence of time steps $\{1, 2, \ldots, T\}$ over which auditory signals $\mathbf{x}_a^{(t)}$ and textual segments $\mathbf{x}_t^{(t)}$ are aligned. One can formally treat this alignment process as minimizing a distance measure $\delta(\mathbf{x}_t^{(t)}, \mathbf{x}_a^{(t)})$, ensuring that their respective content is matched. Similarly, if visual data are represented by $\mathbf{x}_v^{(t)}$ for frames or images at discrete time steps, the system must handle potential mismatches caused by asynchronous sampling and semantic discontinuities.

Parallel to these computational concerns, the cognitive basis for multimodal integration emphasizes the significance of schema theory, working memory constraints, and hierarchical storage of knowledge [11]. A thorough analysis involves the interplay between working memory buffers, which temporarily hold incoming information, and long-term memory stores, which encode permanent representations. If we let $\mathcal{W}$ represent the working memory store, then we may define a transition kernel $p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{u}_t)$ that models how an internal state $\mathbf{s}_t$ evolves over time given an external input $\mathbf{u}_t$. The vector $\mathbf{u}_t$ encompasses integrated features from text, audio, and visuals at time t. This dynamic state-space perspective helps formalize how new sensory observations update knowledge structures, thereby laying the groundwork for analyzing memory consolidation, retrieval, and the formation of robust mental models. [12]

Integrating multimodal cues into a reading-based platform also poses new technical challenges related to data representation, alignment, and the design of neural or symbolic architectures for knowledge inference [13]. Structured knowledge representation schemes, such as directed labeled graphs or grounded relational frames, can be particularly useful in capturing semantic relationships across modalities [14]. For instance, let $G = (V, E)$ be a graph whose nodes V represent concepts (or entities) and whose edges E represent relationships [15]. A node $v \in V$ might store attributes derived from $\mathbf{x}_t, \mathbf{x}_a, \mathbf{x}_v$, while edges reflect how certain concepts are connected in the domain. The synergy between graph-based symbolic representations and learned embeddings can potentially support inductive reasoning, question answering, and advanced retrieval scenarios [16], [17]. In a logic-based formalism, one might encode relational facts as $\mathrm{Rel}(c_i, c_j)$ for concepts c_i and c_j, which can then be fused with distributed encodings through specialized modules.

Moreover, the capacity for an individual to retrieve learned knowledge after extended intervals is profoundly influenced by the organization of conceptual knowledge in memory [18]. If we denote by $\mu(\cdot)$ the memory trace strength function, then the probability of successful retrieval of concept c after time interval τ may be expressed as $\Pr(\text{retrieve}(c, \tau)) = \phi(\mu(c) - \beta\tau)$, where β is a parameter capturing forgetting dynamics and ϕ is a monotonically increasing mapping. A multimodal learning by reading system can bolster the function $\mu(c)$ by strengthening associations that are less susceptible to decay. Hence, the design of the system must not only incorporate advanced fusion methods but also track how experiences (reading sessions, audio commentary, and visual demonstrations) affect memory trace strength and recall latency [19].

This paper aims to elucidate how text, audio, and visuals can be systematically combined within the context of learning by reading systems and why these combinations yield superior retention of knowledge [20]. There is a growing recognition that purely textual approaches, while effective for many tasks, can be augmented through integrative designs that expand learners' sensorimotor resonance, help them build more detailed mental images, and promote deeper understanding of cause–effect relationships [21]. The subsequent sections will examine fundamental aspects of system architectures and multimodal fusion, the cognitive foundations that motivate these designs, experimental frameworks to validate their effectiveness, and an overview of potential limitations and future research avenues. [22]

2 System Architecture and Multimodal Fusion

Designing systems capable of blending textual, auditory, and visual data necessitates specialized models and computational pipelines that facilitate feature extraction, alignment, and joint encoding [23]. The process begins with the individual feature transformations for each modality [24]. If $\mathbf{x}_t \in \mathbb{R}^{d_t}$ corresponds to embeddings derived from advanced linguistic encoders, one might generate them by a function $g_t(w_1, w_2, \ldots, w_n)$ mapping tokens $(w_1, \ldots, w_n)$ to a continuous feature space that captures syntactic and semantic properties of the text. Meanwhile,

$\mathbf{x}_a \in \mathbb{R}^{d_a}$ might correspond to features captured by a time-frequency representation of audio signals, such as mel-spectrograms or hidden states of a recurrent network processing voice data. Visual data can be extracted with a convolutional neural network or a vision transformer to produce $\mathbf{x}_v \in \mathbb{R}^{d_v}$.

The alignment challenge arises because textual segments, audio waveforms, and visual frames may not have a one-to-one correspondence in either time or semantic granularity. A typical approach is to transform each modality into a fixed-dimensional representation for each unit of meaning or event, denoted as $\mathbf{x}_t^{(i)}, \mathbf{x}_a^{(i)}, \mathbf{x}_v^{(i)}$ for $i \in \{1, \ldots, N\}$. Then, the system attempts to match these units using a cross-modal similarity objective [25]. One can define an alignment cost [26]

$$\mathcal{L}_{\text{align}} = \sum_{i=1}^{N} \sum_{j=1}^{N} \ell\big(\mathbf{x}_t^{(i)}, \mathbf{x}_a^{(j)}, \mathbf{x}_v^{(j)}\big),$$

where ℓ is a function measuring the correspondence between textual, auditory, and visual encodings [27]. Minimizing this cost encourages the model to discover consistent patterns across modalities, so that an idea expressed in the text is linked to a particular segment of speech and the relevant visual cues [28].

Once these individual features are aligned, a variety of fusion strategies can integrate the signals [29]. A frequently employed method is the concatenation approach, where fused embeddings [30]

$$\mathbf{z}^{(i)} = \Big[W_t \mathbf{x}_t^{(i)}; W_a \mathbf{x}_a^{(i)}; W_v \mathbf{x}_v^{(i)}\Big],$$

are further processed by a fully connected layer or a recurrent module. Alternatively, attention-based fusion can learn weights that signify the relative importance of each modality under specific contexts [31]. By letting $\alpha_t^{(i)}, \alpha_a^{(i)}, \alpha_v^{(i)}$ represent attention weights, one can produce a unified embedding

$$\mathbf{z}^{(i)} = \alpha_t^{(i)} \mathbf{u}_t^{(i)} + \alpha_a^{(i)} \mathbf{u}_a^{(i)} + \alpha_v^{(i)} \mathbf{u}_v^{(i)},$$

where $\mathbf{u}_t^{(i)}, \mathbf{u}_a^{(i)}, \mathbf{u}_v^{(i)}$ are intermediate modality-specific representations. This approach captures how certain cues might override others depending on the informational content and ensures that textual or visual signals can dominate if they provide a clearer basis for comprehension. [32]

In reading by learning scenarios, knowledge retention can benefit from iterative refinement and memory-based modules that preserve relevant information over extended sequences [33]. Let $\mathbf{h}_i$ denote a hidden state in a recurrent architecture, encoding the integrated representation after the i-th segment. Then the update equations might take the form [34]

$$\mathbf{h}_{i+1} = \text{GRU}(\mathbf{z}^{(i)}, \mathbf{h}_i),$$

where GRU stands for gated recurrent unit or a similar recurrent function. The final hidden state could be used to predict whether the concept is well-understood or whether additional information is necessary for reinforcement [35], [36]. By associating $\mathbf{h}_i$ with intermediate or final outputs, the system might generate clarifications, recommendations of further reading materials, or revision strategies to enhance memory consolidation.

Architectural design decisions also extend to the choice of how to incorporate domain knowledge or symbolic constraints [37]. In some systems, a knowledge graph is used to encode explicit relationships between entities, thereby guiding the model to check for consistency or coverage of relevant topics. For instance, if the reading material covers a particular scientific process, the system might ensure the correctness of references to causal relationships by verifying edges in the graph [38]. Formally, if $\text{Causal}(c_x, c_y)$ is a relational edge in a knowledge graph, the system can cross-reference textual statements $\mathbf{x}_t^{(i)}$ that suggest a causal link. If $\neg\text{Causal}(c_x, c_y)$ is encoded in the knowledge graph, then the system should lower the alignment score for any textual, audio, or visual references that contradict this established fact. Logical consistency checks can thus be integrated with the learning of representation in a holistic pipeline. [39]

Crucially, these system architectures are not static; they adapt through training and continuous usage [40]. If a learning by reading system tracks user behavior, including reading speed, replays of audio clips, or repeated viewing of certain images, it can glean feedback signals about which aspects of the material remain unclear [41]. Let $\pi(\mathbf{a}_t|\mathbf{h}_t)$ be a policy that selects actions $\mathbf{a}_t$ — for instance, presenting additional figures, restating certain text, or playing an audio summary. A reinforcement learning approach might optimize a reward function [42]

$$r(\mathbf{a}_t, \mathbf{x}_t^{(t)}, \mathbf{x}_a^{(t)}, \mathbf{x}_v^{(t)}) = \text{RetScore}(\text{post-test results}) - \gamma \cdot \text{TimeCost}(\mathbf{a}_t),$$

seeking to balance knowledge retention improvements against the temporal and cognitive overhead of presenting supplementary materials [43]. In this manner, the architecture evolves to more efficiently integrate text, audio, and visuals, customizing the learning experience for the individual to achieve higher retention levels with minimal additional cost.

In summary, system architectures for multimodal learning by reading revolve around the extraction, alignment, and fusion of features from textual, auditory, and visual channels, with subsequent modules leveraging memory-based mechanisms, symbolic constraints, and adaptive policies [44]. These structures collectively address the complexities inherent in orchestrating multiple data streams, laying the groundwork for more cognitively coherent experiences that can bolster the long-term consolidation of knowledge. [45]

3 Cognitive Foundations for Multimodal Integration

The theoretical rationale behind multimodal integration in learning by reading systems rests on well-established principles from cognitive psychology and neuroscience, particularly regarding working memory and the encoding specificity of knowledge [46]. Cognitive load theory, for instance, maintains that informational complexity can overwhelm a learner's capacity if not managed effectively, and that well-orchestrated presentation of different sensory channels can avoid overloading a single modality [47]. Formalizing the phenomenon, let $\mathcal{C}(m)$ denote the cognitive load incurred by modality m. If the total load is [48]

$$\mathcal{C}_{\text{total}} = \sum_{m \in \{,audio,visual\}} \mathcal{C}(m),$$

then one might seek to minimize $\mathcal{C}_{\text{total}}$ subject to the constraint that all essential information is conveyed. By distributing the load across modalities, learners can exploit parallel channels of processing, which fosters deeper encoding of the content.

Another cognitive principle relevant to multimodal learning is the dual-coding theory, which posits that verbal and nonverbal information is processed and stored through distinct cognitive systems [49]. Representing these systems as sets $\mathcal{V}$ (verbal) and $\mathcal{I}$ (imagistic), the theory suggests that concepts encoded in both $\mathcal{V}$ and $\mathcal{I}$ have more robust memory traces and are thus retrieved more easily. The presence of an auditory channel, often overlapping with the verbal system, further contributes to elaboration [50]. The combined encoding can be examined through a set-based approach: [51]

$$\text{Encoded}(c) = \begin{cases} 1 & \text{if } c \in \mathcal{V} \cup \mathcal{I} \\ 0 & \text{otherwise} \end{cases}$$

and if $\text{Encoded}(c) = 1$ under multiple modalities, the recall probability tends to increase. This phenomenon provides a psychological foundation for leveraging text, audio, and visuals to yield more concrete conceptual structures in memory. [52]

From a neuroscience perspective, multisensory integration has been associated with specific cortical areas that combine inputs from distinct sensory modalities [53]. Formally, one might model these areas as specialized transformations [54]

$$\mathbf{a} = A(\mathbf{x}_a), \quad \mathbf{v} = V(\mathbf{x}_v),$$

and

$$\mathbf{t} = T(\mathbf{x}_t),$$

where A, V, T stand for neural encoding functions for auditory, visual, and textual signals. The integrated representation [55]

$$\mathbf{m} = M(\mathbf{t}, \mathbf{a}, \mathbf{v})$$

corresponds to the neuronal processes found in multisensory convergence zones [56]. Learning by reading systems that incorporate these channels emulate, in part, the brain's natural capacity for combining distinct forms of sensory input [57]. While artificially constructed computational models cannot fully replicate biological processes, they can be inspired by them to structure integration so as to emulate more natural, intuitively comprehensible experiences [58], [59].

Memory theories, including the levels-of-processing framework, likewise provide insights into how different modes of information presentation may encourage deeper semantic processing [60]. When a learner reads text, listens to an explanatory clip, and observes related visuals, they may be forming multiple associative links with prior knowledge [61]. Let the set of prior knowledge elements be denoted by $\mathcal{K}$. If each new concept c is linked to elements $\{k_1, k_2, \ldots, k_r\} \subseteq \mathcal{K}$, we can define an association function

$$\text{Assoc}(c, \mathcal{K}) = \sum_{j=1}^{r} \text{str}(k_j, c),$$

where str is a function measuring the strength of the link between k_j and c. The process of forming visual, textual, and auditory associations expands $\text{Assoc}(c, \mathcal{K})$ and thereby makes c more accessible in subsequent recall tasks.

This synergy can be especially potent in domains where reading alone might be insufficient for constructing mental imagery or for illustrating dynamic processes that demand temporal and spatial cues.

Error-driven learning mechanisms and generative models of cognition further support the value of multimodal training signals [62]. In the generative view, the learner constructs hypotheses $\hat{h}_1, \hat{h}_2, \ldots$ about the material's meaning and tests them against the available sensory evidence. If the textual information suggests a particular interpretation but the audio or visual channels introduce contradictory or refined cues, the learner experiences a prediction error [63]

$$\varepsilon(\hat{h}, \mathbf{x}_a, \mathbf{x}_v),$$

which triggers hypothesis revision [64]. This iterative cycle continues until the evidence across modalities is reconciled, resulting in a more refined understanding that has better retention properties [65]. Learning by reading systems, in turn, can exploit this phenomenon by selectively presenting contradictory or complementary cues to prompt deeper processing [66].

On the practical side, educators have long noted that diverse forms of input can cater to different learning preferences [67]. However, a more rigorous computational perspective suggests that combining modalities leads to an overlap in representational encodings that insulates knowledge traces from partial forgetting. If the textual representation becomes partially inaccessible due to interference, a strongly associated audio or visual trace can aid in retrieval by reactivating the relevant connections [68]. In formal terms, let $\text{prune}(\mathbf{x}_t)$ denote a partial interference that impairs the textual representation. The presence of $\mathbf{x}_a$ and $\mathbf{x}_v$ can restore the original concept by forming a synergy

$$\mathbf{x}'_t = h(\text{prune}(\mathbf{x}_t), \mathbf{x}_a, \mathbf{x}_v).$$

One might model h as a neural network that attempts to reconstruct textual embeddings from audio and visual signals in the event of missing or degraded textual inputs [69].

These theoretical considerations inform design principles for multimodal reading-based learning environments [70]. They highlight the importance of balancing sensory channel usage, avoiding cognitive overload, and orchestrating the timing of cues to foster deep semantic encoding [71]. The synergy of text, audio, and visuals should be harnessed to enrich comprehension, rather than burdening the learner with extraneous or redundant information [72]. Thus, from a cognitive perspective, multimodal integration entails harnessing the complementary strengths of distinct sensory encodings to generate robust memory traces, while mitigating interference and overload effects [73].

4 Experimental Results and Analytical Framework

Practical demonstrations of multimodal reading systems typically involve controlled evaluations that quantify retention, comprehension, and transfer of knowledge. In a standard experiment, one might recruit participants, divide them into groups, and present reading material under different modality configurations: text-only, text-plus-audio, text-plus-visuals, or text-plus-audio-plus-visuals [74]. Subsequent recall tests are administered after varying time intervals to assess the rate of forgetting [75]. Let $\text{Score}(\tau)$ denote the average recall or comprehension score at time τ post-study. The difference [76]

$$\Delta_{\text{modality}}(\tau) = \text{Score}_{\text{multimodal}}(\tau) - \text{Score}_{\text{text-only}}(\tau)$$

provides a measure of the performance gap between multimodal and traditional reading approaches [77]. Significant positive values of $\Delta_{\text{modality}}(\tau)$ across multiple intervals underscore the efficacy of multimodal designs.

During the experiment, researchers may also capture secondary variables, such as reading time, number of replays for audio clips, or gaze-tracking data that reveal where participants focus visually [78]. Statistical models might be employed to account for individual differences and potential confounding factors [79]. For instance, a hierarchical model can be proposed as

$$\text{Score}_i(\tau) = \beta_0 + \beta_1 \cdot \text{Condition}_i + \beta_2 \cdot \text{ReadingTime}_i + \beta_3 \cdot \text{CognitiveIndex}_i + \epsilon_i,$$

where Condition_i might be an indicator variable for the modality condition of participant i, ReadingTime_i measures how long participant i spent on the material, and CognitiveIndex_i could account for prior domain knowledge or general aptitude. Such a model can reveal how much of the variance in scores is attributable to the presence of audio and visuals, after controlling for other influences. [80]

Analyses have consistently shown that combined audio-visual-textual presentations can yield higher immediate comprehension and slower decay rates for learned information [81]. Let β_{decay} represent the forgetting parameter, such that

$$\text{Score}(\tau) = S_0 \cdot \exp(-\beta_{\text{decay}}\tau).$$

Lower values of β_{decay} indicate more persistent retention. Empirical findings suggest that β_{decay} tends to be smaller for multimodal conditions than for text-only settings, consistent with the principle that multiple encodings offer redundancy and deeper semantic links. The effect sizes, however, can vary depending on factors such as the complexity of the material, the coherence between the modalities, and the learners' familiarity with the domain [82].

Additional analyses often examine transfer tasks, where participants apply the studied knowledge to new problems or scenarios [83]. If the reading system fosters a strong conceptual understanding, the improved retention should manifest as superior performance in tasks requiring inference or problem-solving, not just recall of specifics [84]. Formally, if TransferScore is a function evaluating how well participants generalize their knowledge to novel tasks, one might define

$$\text{TransferScore} = \sum_{k=1}^{M} I\big(\text{correct application of } c_k\big),$$

where I is an indicator function, and c_k are relevant concepts from the domain [85]. Multimodal conditions, by offering multiple representations of each concept, are hypothesized to yield higher TransferScore. Some studies confirm that participants who learn via combined modalities can more flexibly manipulate the learned knowledge in new contexts, an outcome that underscores the inherent synergy of multiple sensory encodings.

Beyond human-subject experiments, analytical frameworks are employed to dissect the computational underpinnings of the observed advantages [86]. In synergy with the principles of representation learning, one can define a distribution $p(\mathbf{z}|c)$ that describes how the latent embedding $\mathbf{z}$ is formed for concept c. Different modalities shape this distribution by providing complementary constraints [87]. The posterior distribution after observing textual data $\mathbf{x}_t$ might be $p(\mathbf{z}|\mathbf{x}_t)$, which can be further refined upon observing audio $\mathbf{x}_a$ to form $p(\mathbf{z}|\mathbf{x}_t, \mathbf{x}_a)$. A suitable measure of the synergy might be the mutual information [88], [89]

$$I(\mathbf{x}_t, \mathbf{x}_a, \mathbf{x}_v; \mathbf{z}),$$

which quantifies how much the modalities collectively inform the latent representation [90]. If $\mathbf{z}$ captures essential features of a concept, then higher mutual information should correlate with stronger retention. This aligns with the perspective that multiple channels of evidence reduce uncertainty in the internal model, leading to embeddings that are more stable and less susceptible to drift or forgetting. [91]

Complementary lines of research address how best to structure the presentation of multimodal content in a reading interface. In some experiments, visuals are introduced incrementally, providing partial glimpses that the learner integrates over time [92]. Textual highlights or audio prompts can direct the user's attention to critical areas in an image or a diagram, systematically reinforcing the mapping between textual assertions and their visual referents [93]. In an ideal scenario, if Focus(t) denotes the user's point of attention at time t, the system can adapt the sequence of visuals to ensure that Focus(t) aligns with the relevant part of the text or audio track, ensuring coherence. The synergy arises when the system exploits user interaction data — such as eye-tracking, mouse movements, or time spent on a particular paragraph — to guide the arrangement of visual or audio cues [94]. The result is a dynamic reading environment that orchestrates text, audio, and visuals in a manner that maximizes both immediate comprehension and long-term retention. [95]

5 Discussion of Model Limitations and Future Directions

While empirical and theoretical evidence supports the efficacy of integrating text, audio, and visuals, several limitations persist, and further research is required to refine and expand the capabilities of multimodal learning by reading systems [96]. One constraint is the inherent variability of users' cognitive styles [97]. Although averaging across a large population might reveal benefits from multimodal approaches, certain individuals may become overwhelmed or distracted by extraneous cues. Formally, we can denote Distract(m) as the distraction factor for modality m. If Distract() $<$ Distract($,audio, visual$) for a subset of learners, these learners might experience hindered comprehension in the fully multimodal condition. An adaptive mechanism that personalizes the level of multimodality based on user feedback and performance metrics is thus desirable. [98]

Another challenge arises from content alignment [99]. For highly abstract or symbolic domains, it may be difficult to produce meaningful visuals or auditory cues that truly capture the essence of the material [100]. The audio track might do little more than restate the text, and visuals might be purely ornamental if not carefully chosen to illuminate the textual discussion [101]. This alignment problem can be framed as an optimization of the form [102], [103]

$$\max_{A,V} \Big\{ \text{Relevance}(A, V, \mathbf{x}_t) - \text{Cost}(A, V) \Big\},$$

where Relevance measures how much audio (A) and visuals (V) contribute to clarifying $\mathbf{x}_t$, and Cost represents the effort needed to produce or process these modalities. Future developments might focus on automated generation

of relevant diagrams, animations, or conceptual maps from textual input, leveraging advanced image or video synthesis techniques [104]. However, these methods must ensure fidelity to the source material and avoid the risk of generating misleading or inaccurate visuals.

In terms of computational modeling, the majority of current work on multimodal fusion relies on data-driven neural architectures [105]. While highly flexible, these architectures often lack transparent interpretability [106]. A black-box model might be quite effective at predicting test scores or optimizing certain objective functions, but it can be difficult to understand precisely how the system combines text, audio, and visuals or to verify its adherence to logical constraints [107]. Symbolic methods could address this gap by enabling the system to express intermediate inference steps, but incorporating them in large-scale neural frameworks remains non-trivial [108]. One line of future research might explore hybrid systems that combine differentiable modules for high-dimensional feature processing with rule-based or logic-based modules that can verify constraints and explain decisions in a structured, interpretable format [109]. In such systems, statements might be represented as logical formulas $\Phi(c_i, c_j)$ that specify relationships among concepts, and the neural embedding modules might be expected to align with these formulas in a consistent manner. [110]

Another avenue for advancement is the integration of user interaction as a critical feedback loop. While many systems rely on one-shot presentations of text, audio, and visuals, real-world learning is iterative and adaptive [111]. If we define user interaction variables $\mathbf{u}_i$ capturing highlights, questions, or corrections, the system can refine its future presentations. For instance, if the user frequently queries a certain concept, the system can infer that the concept is insufficiently explained or remains cognitively opaque [112]. Then it might generate new visual examples, slower-paced audio elaborations, or textual expansions tailored to that concept [113]. Over time, this adaptivity could lead to a reading environment that continuously personalizes multimodal content, maximizing retention while respecting the user's bandwidth and preferences [114]. However, implementing such a system requires a more complex control policy that balances the immediate desire to clarify material with the overall progression through the content. [115]

Scalability also remains a concern, particularly for domains requiring extensive audiovisual support [116]. Generating or curating large-scale corpora of aligned, high-quality text, audio, and visual data can be expensive. Automated approaches for annotation and alignment continue to improve, but perfect alignment remains elusive, especially for open-ended reading materials that cover a broad range of concepts [117]. Additionally, verifying the correctness of automatically generated content is non-trivial [118]. For complex technical or scientific domains, an error in the visuals or a misalignment in the audio track might have severe consequences for the learner's understanding [119]. Thus, rigorous quality control, possibly via human experts or advanced validation protocols, is indispensable. [120]

Lastly, broader research needs to examine the potential for multimodal reading systems to integrate with extended educational ecosystems, including interactive simulations, collaborative problem-solving platforms, and large-scale digital libraries [121]. A comprehensive environment might combine reading-based instruction with interactive modules that allow learners to manipulate variables, run experiments, or engage in structured debates [122]. In such scenarios, text, audio, and visual inputs constitute just one facet of a broader ecosystem designed to optimize knowledge acquisition and retention. Linking a reading-based system to these external tools could further strengthen the learner's mental models, bridging the gap between theoretical knowledge and practical application. [123]

6 Conclusion

In conclusion, integrating text, audio, and visuals in learning by reading systems holds promise for enhancing long-term knowledge retention and deepening comprehension [124]. By capitalizing on the complementary strengths of multiple sensory channels, these systems can foster richer mental representations that prove more resilient to forgetting [125]. Theoretical underpinnings from cognitive psychology, such as cognitive load theory, dual-coding theory, and the levels-of-processing framework, align well with the computational approaches that unify different modalities into shared latent spaces [126]. Empirical studies further substantiate the efficacy of such systems, demonstrating improved recall rates, better transfer to novel tasks, and a reduced rate of decay in memory traces. [127]

The underlying computational techniques revolve around advanced feature extraction, alignment procedures, attention-based fusion, and memory-enriched architectures [128]. These components ensure that text, audio, and visuals can be effectively synchronized, preserving semantic coherence and allowing for dynamic adjustments based on learner interaction. Symbolic representations or logic-based constraints can be layered atop neural embeddings to ensure logical consistency, interpretability, and domain-specific correctness [129]. The resulting multimodal environment not only supplies learners with concrete, illustrative depictions of abstract concepts but also compensates for partial memory degradation by offering alternative pathways for retrieval. [130]

Nonetheless, adopting these multimodal approaches brings to the fore several open questions [131], [132]. Variations among learners may necessitate adaptive or personalized systems that calibrate the levels of multimodality to user preferences and cognitive capacities [133]. Automated generation of high-quality audio and visual content must also be approached carefully, balancing the benefits of enriched presentations against the risks of inaccuracies or information overload [134]. Interpretability and alignment with symbolic knowledge structures represent further frontiers that invite hybrid solutions merging neural architectures with explicit reasoning. By systematically addressing these concerns, future research can refine multimodal learning by reading systems into powerful platforms that efficiently consolidate information, maintain conceptual clarity across extended time spans, and ultimately transform how individuals assimilate and retain complex knowledge. [135]

References

[1] A. D. Paola, P. Ferraro, G. L. Re, M. Morana, and M. Ortolani, "A fog-based hybrid intelligent system for energy saving in smart buildings," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 7, pp. 2793–2807, Jul. 26, 2019. DOI: 10.1007/s12652-019-01375-2.

[2] O. Ozyurt and A. Ayaz, "Twenty-five years of education and information technologies: Insights from a topic modeling based bibliometric analysis.," *Education and information technologies*, vol. 27, no. 8, pp. 11 025–11 054, Apr. 28, 2022. DOI: 10.1007/s10639-022-11071-y.

[3] Z. Batmaz, A. Yürekli, A. Bilge, and C. Kaleli, "A review on deep learning for recommender systems: Challenges and remedies," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 1–37, Aug. 29, 2018. DOI: 10.1007/s10462-018-9654-y.

[4] S. Biswas, P. Riba, J. Lladós, and U. Pal, "Beyond document object detection: Instance-level segmentation of complex layouts," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 24, no. 3, pp. 269–281, Jul. 21, 2021. DOI: 10.1007/s10032-021-00380-6.

[5] D. Anderson, "Boinc: A platform for volunteer computing," *Journal of Grid Computing*, vol. 18, no. 1, pp. 99–122, Nov. 16, 2019. DOI: 10.1007/s10723-019-09497-9.

[6] L. T. D. Paolis, C. Gatto, L. Corchia, and V. D. Luca, "Usability, user experience and mental workload in a mobile augmented reality application for digital storytelling in cultural heritage," *Virtual Reality*, vol. 27, no. 2, pp. 1117–1143, Nov. 15, 2022. DOI: 10.1007/s10055-022-00712-9.

[7] S. A. Khowaja, P. Khuwaja, K. Dev, and G. D'Aniello, "Virfim: An ai and internet of medical things-driven framework for healthcare using smart sensors.," *Neural computing & applications*, vol. 35, no. 22, pp. 1–18, Sep. 2, 2021. DOI: 10.1007/s00521-021-06434-4.

[8] J. Nayak and B. Naik, "A novel honey-bees mating optimization approach with higher order neural network for classification," *Journal of Classification*, vol. 35, no. 3, pp. 511–548, Oct. 1, 2018. DOI: 10.1007/s00357-018-9270-1.

[9] M. Kansara, "Cloud migration strategies and challenges in highly regulated and data-intensive industries: A technical perspective," *International Journal of Applied Machine Learning and Computational Intelligence*, vol. 11, no. 12, pp. 78–121, 2021. DOI: 10.17613/677yg-gks37.

[10] C.-H. Lin, "Linear permanent magnet synchronous motor drive system using aaennb control system with error compensation controller and cpso," *Electrical Engineering*, vol. 102, no. 3, pp. 1311–1325, Feb. 21, 2020. DOI: 10.1007/s00202-020-00953-4.

[11] M. Carabantes, "Black-box artificial intelligence: An epistemological and critical analysis," *AI & SOCIETY*, vol. 35, no. 2, pp. 309–317, Apr. 12, 2019. DOI: 10.1007/s00146-019-00888-w.

[12] J. C. Hofmeister, J. Siegmund, and D. V. Holt, "Shorter identifier names take longer to comprehend," *Empirical Software Engineering*, vol. 24, no. 1, pp. 417–443, Apr. 26, 2018. DOI: 10.1007/s10664-018-9621-x.

[13] S. Cong and Y. Zhou, "A review of convolutional neural network architectures and their optimizations," *Artificial Intelligence Review*, vol. 56, no. 3, pp. 1905–1969, Jun. 22, 2022. DOI: 10.1007/s10462-022-10213-5.

[14] M. Khodabakhsh, M. Kahani, and E. Bagheri, "Predicting future personal life events on twitter via recurrent neural networks," *Journal of Intelligent Information Systems*, vol. 54, no. 1, pp. 101–127, Aug. 15, 2018. DOI: 10.1007/s10844-018-0519-2.

[15] B. Kasdorf, "We are at an inflection point for publishing and the web," *Publishing Research Quarterly*, vol. 36, no. 3, pp. 313–322, May 11, 2020. DOI: 10.1007/s12109-020-09727-z.

[16] X. Chen, D. Zou, and H. Xie, "A decade of learning analytics: Structural topic modeling based bibliometric analysis," *Education and Information Technologies*, vol. 27, no. 8, pp. 10 517–10 561, Apr. 18, 2022. DOI: 10.1007/s10639-022-11046-z.

[17] K. D. Forbus, K. Lockwood, A. B. Sharma, and E. Tomai, "Steps towards a second generation learning by reading system.," in *AAAI Spring Symposium: Learning by Reading and Learning to Read*, 2009, pp. 36–43.

[18] M. Repetto, D. Striccoli, G. Piro, A. Carrega, G. Boggia, and R. Bolla, "An autonomous cybersecurity framework for next-generation digital service chains," *Journal of Network and Systems Management*, vol. 29, no. 4, pp. 1–34, May 24, 2021. DOI: 10.1007/s10922-021-09607-7.

[19] A. Khanna and S. Kaur, "Internet of things (iot), applications and challenges: A comprehensive review," *Wireless Personal Communications*, vol. 114, no. 2, pp. 1687–1762, May 28, 2020. DOI: 10.1007/s11277-020-07446-4.

[20] X.-B. Zhou, Z.-Q. Wang, X.-W. Liang, M. Zhang, and G.-D. Zhou, "Neural emotion detection via personal attributes," *Journal of Computer Science and Technology*, vol. 37, no. 5, pp. 1146–1160, Sep. 30, 2022. DOI: 10.1007/s11390-021-0606-7.

[21] L. V. Astakhova, "Visualizing information resources in the conditions of digitalization of the field of knowledge: An overview," *Scientific and Technical Information Processing*, vol. 46, no. 1, pp. 20–27, May 4, 2019. DOI: 10.3103/s0147688219010052.

[22] C. P. C. Chanel, A. Albore, J. T'Hooft, C. Lesire, and F. Teichteil-Königsbuch, "Ample: An anytime planning and execution framework for dynamic and uncertain problems in robotics," *Autonomous Robots*, vol. 43, no. 1, pp. 37–62, Feb. 10, 2018. DOI: 10.1007/s10514-018-9703-z.

[23] D. Liu, Z. Ren, Z.-T. Long, G. Gao, and H. Jiang, "Mining design pattern use scenarios and related design pattern pairs: A case study on online posts," *Journal of Computer Science and Technology*, vol. 35, no. 5, pp. 963–978, Sep. 30, 2020. DOI: 10.1007/s11390-020-0407-4.

[24] J. Jiang, Y. Xiong, and X. Xia, "A manual inspection of defects4j bugs and its implications for automatic program repair," *Science China Information Sciences*, vol. 62, no. 10, pp. 200 102–, Sep. 6, 2019. DOI: 10.1007/s11432-018-1465-6.

[25] M. Y. Landolsi, L. Hlaoua, and L. B. Romdhane, "Information extraction from electronic medical documents: State of the art and future research directions.," *Knowledge and information systems*, vol. 65, no. 2, pp. 463–516, Nov. 8, 2022. DOI: 10.1007/s10115-022-01779-1.

[26] G. Garg and M. Juneja, "Particle swarm optimization based segmentation of cancer in multi-parametric prostate mri," *Multimedia Tools and Applications*, vol. 80, no. 20, pp. 30 557–30 580, Jun. 18, 2021. DOI: 10.1007/s11042-021-11133-2.

[27] A. Strebinger and H. Treiblmaier, "Profiling early adopters of blockchain-based hotel booking applications: Demographic, psychographic, and service-related factors," *Information Technology & Tourism*, vol. 24, no. 1, pp. 1–30, Jan. 3, 2022. DOI: 10.1007/s40558-021-00219-0.

[28] C. Zhou, B. Li, X. Sun, and B. Lili, "Why and what happened? aiding bug comprehension with automated category and causal link identification," *Empirical Software Engineering*, vol. 26, no. 6, pp. 1–36, Aug. 25, 2021. DOI: 10.1007/s10664-021-10010-8.

[29] M. Farina, P. Zhdanov, A. Karimov, and A. Lavazza, "Ai and society: A virtue ethics approach," *AI & SOCIETY*, vol. 39, no. 3, pp. 1127–1140, Sep. 10, 2022. DOI: 10.1007/s00146-022-01545-5.

[30] M. M. M. Mufassirin, M. A. H. Newton, and A. Sattar, "Artificial intelligence for template-free protein structure prediction: A comprehensive review," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 7665–7732, Dec. 17, 2022. DOI: 10.1007/s10462-022-10350-x.

[31] J. Uyheng, I. J. Cruickshank, and K. M. Carley, "Mapping state-sponsored information operations with multi-view modularity clustering.," *EPJ data science*, vol. 11, no. 1, pp. 25–, Apr. 18, 2022. DOI: 10.1140/epjds/s13688-022-00338-6.

[32] X. Dong, Y. Lian, Y. Chi, X. Tang, and Y. Liu, "A two-step rumor detection model based on the super-network theory about weibo," *The Journal of supercomputing*, vol. 77, no. 10, pp. 1–25, Apr. 1, 2021. DOI: 10.1007/s11227-021-03748-x.

[33] D. Fogli, A. Arenghi, and F. Gentilin, "A universal design approach to wayfinding and navigation," *Multimedia Tools and Applications*, vol. 79, no. 45, pp. 33 577–33 601, Dec. 19, 2019. DOI: 10.1007/s11042-019-08492-2.

[34] M. Shahin, M. Zahedi, M. A. Babar, and L. Zhu, "An empirical study of architecting for continuous delivery and deployment," *Empirical Software Engineering*, vol. 24, no. 3, pp. 1061–1108, Sep. 8, 2018. DOI: 10.1007/s10664-018-9651-4.

[35] D. Spadini, M. Aniche, M. Bruntink, and A. Bacchelli, "Mock objects for testing java systems why and how developers use them, and how they evolve," *Empirical Software Engineering*, vol. 24, no. 3, pp. 1461–1498, Nov. 6, 2018. DOI: 10.1007/s10664-018-9663-0.

[36] K. Forbus, K. Lockwood, A. Sharma, and E. Tomai, "Steps towards a 2nd generation learning by reading system," in *AAAI Spring Symposium on Learning by Reading, Spring*, 2009.

[37] "Cars 2021: Computer assisted radiology and surgery proceedings of the 35th international congress and exhibition munich, germany, june 21-25, 2021.," *International journal of computer assisted radiology and surgery*, vol. 16, no. Suppl 1, pp. 1–119, Jun. 4, 2021. DOI: 10.1007/s11548-021-02375-4.

[38] C. Dudschig, B. Kaup, M. Liu, and J. Schwab, "The processing of negation and polarity: An overview.," *Journal of psycholinguistic research*, vol. 50, no. 6, pp. 1–15, Nov. 17, 2021. DOI: 10.1007/s10936-021-09817-9.

[39] N. Nelson, C. Brindescu, S. McKee, A. Sarma, and D. Dig, "The life-cycle of merge conflicts: Processes, barriers, and strategies," *Empirical Software Engineering*, vol. 24, no. 5, pp. 2863–2906, Feb. 5, 2019. DOI: 10.1007/s10664-018-9674-x.

[40] N. Komwad, P. Tiwari, B. Praveen, and C. R. Chowdary, "A survey on review summarization and sentiment classification," *Knowledge and Information Systems*, vol. 64, no. 9, pp. 2289–2327, Aug. 1, 2022. DOI: 10.1007/s10115-022-01728-y.

[41] H. Zhong and H. Mei, "Learning a graph-based classifier for fault localization," *Science China Information Sciences*, vol. 63, no. 6, pp. 162 101–, May 9, 2020. DOI: 10.1007/s11432-019-2720-1.

[42] K. O. Bazilevych, D. I. Chumachenko, L. F. Hulianytskyi, I. S. Meniailov, and S. V. Yakovlev, "Intelligent decision-support system for epidemiological diagnostics. i. a concept of architecture design.," *Cybernetics and systems analysis*, vol. 58, no. 3, pp. 343–353, Sep. 1, 2022. DOI: 10.1007/s10559-022-00466-x.

[43] K. Kuru, D. Ansell, J. M. J, *et al.*, "Intelligent autonomous treatment of bedwetting using non-invasive wearable advanced mechatronics systems and mems sensors," *Medical & biological engineering & computing*, vol. 58, no. 5, pp. 943–965, Feb. 24, 2020. DOI: 10.1007/s11517-019-02091-x.

[44] A. M. Aleesa, B. B. Zaidan, A. A. Zaidan, and N. B. M. Sahar, "Review of intrusion detection systems based on deep learning techniques: Coherent taxonomy, challenges, motivations, recommendations, substantial analysis and future directions," *Neural Computing and Applications*, vol. 32, no. 14, pp. 9827–9858, Oct. 18, 2019. DOI: 10.1007/s00521-019-04557-3.

[45] A. Ed-daoudy and K. Maalmi, "A new internet of things architecture for real-time prediction of various diseases using machine learning on big data environment," *Journal of Big Data*, vol. 6, no. 1, pp. 1–25, Nov. 27, 2019. DOI: 10.1186/s40537-019-0271-7.

[46] T. K. Gunning, X. A. Conlan, P. K. Collins, *et al.*, "Who engaged in the team-based assessment? leveraging edtech for a self and intra-team peer-assessment solution to free-riding," *International Journal of Educational Technology in Higher Education*, vol. 19, no. 1, Jul. 13, 2022. DOI: 10.1186/s41239-022-00340-y.

[47] R. Mulero, V. Urosevic, A. Almeida, and C. Tatsiopoulos, "Towards ambient assisted cities using linked data and data analysis," *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 5, pp. 1573–1591, Jun. 28, 2018. DOI: 10.1007/s12652-018-0916-y.

[48] V. Piantadosi, F. Fierro, S. Scalabrino, A. Serebrenik, and R. Oliveto, "How does code readability change during software evolution," *Empirical Software Engineering*, vol. 25, no. 6, pp. 5374–5412, Sep. 24, 2020. DOI: 10.1007/s10664-020-09886-9.

[49] C. Fields and J. F. Glazebrook, "Do process-1 simulations generate the epistemic feelings that drive process-2 decision making?" *Cognitive processing*, vol. 21, no. 4, pp. 533–553, Jun. 30, 2020. DOI: 10.1007/s10339-020-00981-9.

[50] P. O'Connor, "Visible learning and whole language: Revisiting the 'garbage in, garbage out' problem," *The Australian Journal of Language and Literacy*, vol. 43, no. 2, pp. 141–151, Jun. 1, 2020. DOI: 10.1007/bf03652050.

[51] W. Said, J. Quante, and R. Koschke, "Mining understandable state machine models from embedded code," *Empirical Software Engineering*, vol. 25, no. 6, pp. 4759–4804, Sep. 8, 2020. DOI: 10.1007/s10664-020-09865-0.

[52] J. A. H. López and J. S. Cuadrado, "An efficient and scalable search engine for models," *Software and Systems Modeling*, vol. 21, no. 5, pp. 1715–1737, Dec. 27, 2021. DOI: 10.1007/s10270-021-00960-4.

[53] S. Strauss and G. B. Kim, "Teaching korean as a foreign language: Theories and practices," *The Korean Language in America*, vol. 25, no. 1, pp. 79–86, Dec. 1, 2021. DOI: 10.5325/korelangamer.25.1.0079.

[54] A. Leporati, G. Mauri, and C. Zandron, "Spiking neural p systems: Main ideas and results," *Natural Computing*, vol. 21, no. 4, pp. 629–649, Aug. 18, 2022. DOI: 10.1007/s11047-022-09917-y.

[55] X. Jian and T. Zou, "A review of locomotion, control, and implementation of robot fish," *Journal of Intelligent & Robotic Systems*, vol. 106, no. 2, Sep. 27, 2022. DOI: 10.1007/s10846-022-01726-w.

[56] A. Bernasconi, G. Guizzardi, O. Pastor, and V. C. Storey, "Semantic interoperability: Ontological unpacking of a viral conceptual model.," *BMC bioinformatics*, vol. 23, no. Suppl 11, pp. 491–, Nov. 17, 2022. DOI: 10.1186/s12859-022-05022-0.

[57] P. Fraternali, F. Milani, R. N. Torres, and N. Zangrando, "Black-box error diagnosis in deep neural networks for computer vision: A survey of tools," *Neural Computing and Applications*, vol. 35, no. 4, pp. 3041–3062, Dec. 10, 2022. DOI: 10.1007/s00521-022-08100-9.

[58] N. Rios, R. O. Spínola, M. Mendonça, and C. Seaman, "The practitioners' point of view on the concept of technical debt and its causes and consequences: A design for a global family of industrial surveys and its first results from brazil," *Empirical Software Engineering*, vol. 25, no. 5, pp. 3216–3287, Jun. 13, 2020. DOI: 10.1007/s10664-020-09832-9.

[59] A. Sharma and K. Forbus, "Automatic extraction of efficient axiom sets from large knowledge bases," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 27, 2013, pp. 1248–1254.

[60] Y. Xiang, N. Dalchau, and B. Wang, "Scaling up genetic circuit design for cellular computing: Advances and prospects.," *Natural computing*, vol. 17, no. 4, pp. 833–853, Oct. 5, 2018. DOI: 10.1007/s11047-018-9715-9.

[61] P. D. U. Coronado, J. A. A. Demeneghi, H. Ahuett-Garza, P. O. Castañon, and M. M. Martínez, "Representation of machines and mechanisms in augmented reality for educative use," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 16, no. 2, pp. 643–656, Mar. 12, 2022. DOI: 10.1007/s12008-022-00852-x.

[62] A. H. Ababneh, J. Lu, and Q. Xu, "An efficient framework of utilizing the latent semantic analysis in text extraction," *International Journal of Speech Technology*, vol. 22, no. 3, pp. 785–815, Aug. 21, 2019. DOI: 10.1007/s10772-019-09623-8.

[63] E. Oztemel and S. Gürsev, "Literature review of industry 4.0 and related technologies," *Journal of Intelligent Manufacturing*, vol. 31, no. 1, pp. 127–182, Jul. 24, 2018. DOI: 10.1007/s10845-018-1433-8.

[64] S. Zhuo, J. Zhang, and C. Zhang, "A novel data augmentation method for chinese character spatial structure recognition by normalized deformable convolutional networks," *Neural Processing Letters*, vol. 54, no. 6, pp. 5545–5563, Aug. 30, 2022. DOI: 10.1007/s11063-022-10873-y.

[65] N.-V. Nguyen, C. Rigaud, and J.-C. Burie, "Comic mtl: Optimized multi-task learning for comic book image analysis," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 22, no. 3, pp. 265–284, Jul. 17, 2019. DOI: 10.1007/s10032-019-00330-3.

[66] B. Gezici and A. K. Tarhan, "Systematic literature review on software quality for ai-based software," *Empirical Software Engineering*, vol. 27, no. 3, Mar. 17, 2022. DOI: 10.1007/s10664-021-10105-2.

[67] C. M. C. Silva, M. Galster, and F. Gilson, "Topic modeling in software engineering research," *Empirical Software Engineering*, vol. 26, no. 6, pp. 1–62, Sep. 6, 2021. DOI: 10.1007/s10664-021-10026-0.

[68] C. F. Hayes, R. Rădulescu, E. Bargiacchi, *et al.*, "A practical guide to multi-objective reinforcement learning and planning," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 1, Apr. 13, 2022. DOI: 10.1007/s10458-022-09552-y.

[69] A. Appice, G. Andresini, and D. Malerba, "Clustering-aided multi-view classification: A case study on android malware detection," *Journal of Intelligent Information Systems*, vol. 55, no. 1, pp. 1–26, May 4, 2020. DOI: 10.1007/s10844-020-00598-6.

[70] B. Zhang, J. Ye, R. Li, C. Feng, Y. Su, and C. Tang, "Pusher: An augmented fuzzer based on the connection between input and comparison operand," *Frontiers of Computer Science*, vol. 16, no. 4, Dec. 3, 2021. DOI: 10.1007/s11704-021-0075-8.

[71] F. Abuzaid, P. Kraft, S. Suri, *et al.*, "Diff: A relational interface for large-scale data explanation," *The VLDB Journal*, vol. 30, no. 1, pp. 45–70, Sep. 30, 2020. DOI: 10.1007/s00778-020-00633-6.

[72] S. Groppe, "Emergent models, frameworks, and hardware technologies for big data analytics," *The Journal of Supercomputing*, vol. 76, no. 3, pp. 1800–1827, Feb. 20, 2018. DOI: 10.1007/s11227-018-2277-x.

[73] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 161–184, Aug. 27, 2020. DOI: 10.1007/s11263-020-01369-0.

[74] H. Gaspar, L. Morgado, H. S. Mamede, T. Oliveira, B. F. Manjón, and C. Gütl, "Research priorities in immersive learning technology: The perspectives of the ilrn community," *Virtual Reality*, vol. 24, no. 2, pp. 319–341, Jul. 22, 2019. DOI: 10.1007/s10055-019-00393-x.

[75] Z. Hammami, M. Sayed-Mouchaweh, W. Mouelhi, and L. B. Said, "Neural networks for online learning of non-stationary data streams: A review and application for smart grids flexibility improvement," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 6111–6154, May 17, 2020. DOI: 10.1007/s10462-020-09844-3.

[76] C. Mallika and S. Selvamuthukumaran, "A hybrid crow search and grey wolf optimization technique for enhanced medical data classification in diabetes diagnosis system," *International Journal of Computational Intelligence Systems*, vol. 14, no. 1, pp. 1–18, Sep. 1, 2021. DOI: 10.1007/s44196-021-00013-0.

[77] M. Bodini, "Opinion mining from machine translated bangla reviews with stacked contractive auto-encoders," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 9, pp. 12 119–12 131, Feb. 25, 2022. DOI: 10.1007/s12652-022-03760-w.

[78] M. M. Mostafa, A. Feizollah, and N. B. Anuar, "Fifteen years of youtube scholarly research: Knowledge structure, collaborative networks, and trending topics.," *Multimedia tools and applications*, vol. 82, no. 8, pp. 12 423–12 443, Sep. 19, 2022. DOI: 10.1007/s11042-022-13908-7.

[79] S.-Y. Lin, L. Zhang, J. Li, L.-l. Ji, and Y. Sun, "A survey of application research based on blockchain smart contract," *Wireless Networks*, vol. 28, no. 2, pp. 635–690, Jan. 17, 2022. DOI: 10.1007/s11276-021-02874-x.

[80] G. Governatori, F. Idelberger, Z. Milosevic, R. Riveret, G. Sartor, and X. Xu, "On legal contracts, imperative and declarative smart contracts, and blockchain systems," *Artificial Intelligence and Law*, vol. 26, no. 4, pp. 377–409, Mar. 5, 2018. DOI: 10.1007/s10506-018-9223-3.

[81] U. M. R. Paturi, S. Cheruku, and N. S. Reddy, "The role of artificial neural networks in prediction of mechanical and tribological properties of composites—a comprehensive review," *Archives of Computational Methods in Engineering*, vol. 29, no. 5, pp. 3109–3149, Jan. 29, 2022. DOI: 10.1007/s11831-021-09691-7.

[82] J. Zhou, Y. Jiang, A. A. Pantelous, and W. Dai, "A systematic review of uncertainty theory with the use of scientometrical method," *Fuzzy Optimization and Decision Making*, vol. 22, no. 3, pp. 463–518, Sep. 13, 2022. DOI: 10.1007/s10700-022-09400-4.

[83] G. C. Zutin, G. F. Barbosa, P. C. de Barros, E. B. Tiburtino, F. L. F. Kawano, and S. B. Shiki, "Readiness levels of industry 4.0 technologies applied to aircraft manufacturing-a review, challenges and trends.," *The International journal, advanced manufacturing technology*, vol. 120, no. 1-2, pp. 927–943, Feb. 8, 2022. DOI: 10.1007/s00170-022-08769-1.

[84] A. Babaeinesami, H. Tohidi, P. Ghasemi, F. Goodarzian, and E. B. Tirkolaee, "A closed-loop supply chain configuration considering environmental impacts: A self-adaptive nsga-ii algorithm," *Applied Intelligence*, vol. 52, no. 12, pp. 13 478–13 496, Jan. 21, 2022. DOI: 10.1007/s10489-021-02944-9.

[85] W. P. N. dos Reis, G. E. Couto, and O. M. Junior, "Automated guided vehicles position control: A systematic literature review," *Journal of Intelligent Manufacturing*, vol. 34, no. 4, pp. 1483–1545, Jan. 5, 2022. DOI: 10.1007/s10845-021-01893-x.

[86] L. Tan, G. Wang, F. Jia, and X. Lian, "Research status of deep learning methods for rumor detection.," *Multimedia tools and applications*, vol. 82, no. 2, pp. 2941–2982, Apr. 21, 2022. DOI: 10.1007/s11042-022-12800-8.

[87] N. M. Hassan, S. Hamad, and K. Mahar, "Mammogram breast cancer cad systems for mass detection and classification: A review," *Multimedia Tools and Applications*, vol. 81, no. 14, pp. 20 043–20 075, Mar. 11, 2022. DOI: 10.1007/s11042-022-12332-1.

[88] E. Ntentos, U. Zdun, K. Plakidas, *et al.*, "Detector-based component model abstraction for microservice-based systems," *Computing*, vol. 103, no. 11, pp. 2521–2551, Aug. 28, 2021. DOI: 10.1007/s00607-021-01002-z.

[89] A. Sharma and K. Forbus, "Graph traversal methods for reasoning in large knowledge-based systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 27, 2013, pp. 1255–1261.

[90] X. Dong, T. Li, R. Song, and Z. Ding, "Profiling users via their reviews: An extended systematic mapping study," *Software and Systems Modeling*, vol. 20, no. 1, pp. 49–69, Mar. 19, 2020. DOI: 10.1007/s10270-020-00790-w.

[91] K. S. Kappel, T. M. Cabreira, J. L. Marins, L. Brisolara, and P. Ferreira, "Strategies for patrolling missions with multiple uavs," *Journal of Intelligent & Robotic Systems*, vol. 99, no. 3, pp. 499–515, Sep. 12, 2019. DOI: 10.1007/s10846-019-01090-2.

[92] M. R. da Silva, F. Marques, M. T. da Silva, and P. Flores, "A comprehensive review on biomechanical modeling applied to device-assisted locomotion," *Archives of Computational Methods in Engineering*, vol. 30, no. 3, pp. 1897–1960, Nov. 27, 2022. DOI: 10.1007/s11831-022-09856-y.

[93] N. Ali, Z. Halim, and S. F. Hussain, "An artificial intelligence-based framework for data-driven categorization of computer scientists: A case study of world's top 10 computing departments," *Scientometrics*, vol. 128, no. 3, pp. 1513–1545, Dec. 31, 2022. DOI: 10.1007/s11192-022-04627-9.

[94] A. C. Oran, G. Santos, B. Gadelha, and T. Conte, "A framework for evaluating and improving requirements specifications based on the developers and testers perspective," *Requirements Engineering*, vol. 26, no. 4, pp. 481–508, Jun. 25, 2021. DOI: 10.1007/s00766-021-00352-6.

[95] Y. Balashov, "The translator's extended mind," *Minds and Machines*, vol. 30, no. 3, pp. 349–383, Sep. 19, 2020. DOI: 10.1007/s11023-020-09536-5.

[96] J. Chambua and Z. Niu, "Review text based rating prediction approaches: Preference knowledge learning, representation and utilization," *Artificial Intelligence Review*, vol. 54, no. 2, pp. 1171–1200, Jul. 28, 2020. DOI: 10.1007/s10462-020-09873-y.

[97] A. Ghorbel, N. B. Amor, and M. Abid, "Gpgpu-based parallel computing of viola and jones eyes detection algorithm to drive an intelligent wheelchair," *Journal of Signal Processing Systems*, vol. 94, no. 12, pp. 1365–1379, Jul. 1, 2022. DOI: 10.1007/s11265-022-01783-2.

[98] O. Haddad, F. Fkih, and M. N. Omri, "Toward a prediction approach based on deep learning in big data analytics," *Neural Computing and Applications*, vol. 35, no. 8, pp. 6043–6063, Nov. 13, 2022. DOI: 10.1007/s00521-022-07986-9.

[99] T. Abbasi-khazaei and M. H. Rezvani, "Energy-aware and carbon-efficient vm placement optimization in cloud datacenters using evolutionary computing methods," *Soft Computing*, vol. 26, no. 18, pp. 9287–9322, Jun. 30, 2022. DOI: 10.1007/s00500-022-07245-y.

[100] R. Loughran and M. O'Neill, "Evolutionary music: Applying evolutionary computation to the art of creating music," *Genetic Programming and Evolvable Machines*, vol. 21, no. 1, pp. 55–85, Feb. 6, 2020. DOI: 10.1007/s10710-020-09380-7.

[101] C. Coupette, D. Hartung, J. Beckedorf, M. Böther, and D. M. Katz, "Law smells," *Artificial Intelligence and Law*, vol. 31, no. 2, pp. 335–368, Jun. 6, 2022. DOI: 10.1007/s10506-022-09315-w.

[102] L. Corti, N. D. Stefano, and M. Bertolaso, "Artificial emotions: Toward a human-centric ethics," *International Journal of Social Robotics*, vol. 15, no. 12, pp. 2039–2053, Jun. 11, 2022. DOI: 10.1007/s12369-022-00890-1.

[103] A. Sharma, M. Witbrock, and K. Goolsbey, "Controlling search in very large commonsense knowledge bases: A machine learning approach," *arXiv preprint arXiv:1603.04402*, 2016.

[104] Y. Chen, H. Wang, B. Zhang, and W. Zhang, "A method of measuring the article discriminative capacity and its distribution," *Scientometrics*, vol. 127, no. 6, pp. 3317–3341, Apr. 18, 2022. DOI: 10.1007/s11192-022-04371-0.

[105] A. G. Nath, S. S. Udmale, and S. Singh, "Role of artificial intelligence in rotor fault diagnosis: A comprehensive review," *Artificial Intelligence Review*, vol. 54, no. 4, pp. 2609–2668, Sep. 21, 2020. DOI: 10.1007/s10462-020-09910-w.

[106] T. Binalialhag, J. Hassine, and D. Amyot, "Static slicing of use case maps requirements models," *Software & Systems Modeling*, vol. 18, no. 4, pp. 2465–2505, Jun. 7, 2018. DOI: 10.1007/s10270-018-0680-7.

[107] A. C. Vidaurre, E. C. López, J. P. S. Alcocer, and A. Bergel, "Testevoviz: Visualizing genetically-based test coverage evolution," *Empirical Software Engineering*, vol. 27, no. 7, Sep. 28, 2022. DOI: 10.1007/s10664-022-10220-8.

[108] null Santhosh Kumar H S and K. Karibasappa, "An approach for brain tumour detection based on dual-tree complex gabor wavelet transform and neural network using hadoop big data analysis," *Multimedia Tools and Applications*, vol. 81, no. 27, pp. 39 251–39 274, Apr. 28, 2022. DOI: 10.1007/s11042-022-13016-6.

[109] D. Chehal, P. Gupta, and P. Gulati, "Implementation and comparison of topic modeling techniques based on user reviews in e-commerce recommendations," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 5, pp. 5055–5070, Apr. 16, 2020. DOI: 10.1007/s12652-020-01956-6.

[110] M. T. Suleman and S. M. Awan, "Optimization of url-based phishing websites detection through genetic algorithms," *Automatic Control and Computer Sciences*, vol. 53, no. 4, pp. 333–341, Sep. 16, 2019. DOI: 10.3103/s0146411619040102.

[111] M. Coppola, J. Guo, E. Gill, and G. de Croon, "The pagerank algorithm as a method to optimize swarm behavior through local analysis," *Swarm Intelligence*, vol. 13, no. 3, pp. 277–319, Aug. 23, 2019. DOI: 10.1007/s11721-019-00172-z.

[112] M. M. Nasiri, S. Salesi, A. Rahbari, N. S. Meydani, and M. Abdollai, "A data mining approach for population-based methods to solve the jssp," *Soft Computing*, vol. 23, no. 21, pp. 11 107–11 122, Nov. 28, 2018. DOI: 10.1007/s00500-018-3663-2.

[113] F. Piccialli, Y. Yoshimura, P. Benedusi, C. Ratti, and S. Cuomo, "Lessons learned from longitudinal modeling of mobile-equipped visitors in a complex museum," *Neural Computing and Applications*, vol. 32, no. 12, pp. 7785–7801, Feb. 27, 2019. DOI: 10.1007/s00521-019-04099-8.

[114] S. Martínez-Municio, L. Rodriguez-Benitez, E. Castillo-Herrera, J. Giralt-Muiña, and L. Jimenez-Linares, "Linguistic modeling and synthesis of heterogeneous energy consumption time series sets," *International Journal of Computational Intelligence Systems*, vol. 12, no. 1, pp. 259–272, 2018. DOI: 10.2991/ijcis.2018.125905639.

[115] Y. Francillette, E. Boucher, N. Bier, *et al.*, "Modeling the behavior of persons with mild cognitive impairment or alzheimer's for intelligent environment simulation," *User Modeling and User-Adapted Interaction*, vol. 30, no. 5, pp. 895–947, Jun. 20, 2020. DOI: 10.1007/s11257-020-09266-4.

[116] F. J. Gallego-Durán, R. Satorre-Cuerda, P. Compañ-Rosique, C. Villagrá-Arnedo, R. Molina-Carmona, and F. Llorens-Largo, "A low-level approach to improve programming learning," *Universal Access in the Information Society*, vol. 20, no. 3, pp. 479–493, Nov. 15, 2020. DOI: 10.1007/s10209-020-00775-y.

[117] M. Miquel-Ribé and D. Laniado, "The wikipedia diversity observatory: Helping communities to bridge content gaps through interactive interfaces," *Journal of Internet Services and Applications*, vol. 12, no. 1, pp. 10–, Nov. 1, 2021. DOI: 10.1186/s13174-021-00141-y.

[118] E. Sulis, I. A. Amantea, M. Aldinucci, *et al.*, "An ambient assisted living architecture for hospital at home coupled with a process-oriented perspective.," *Journal of ambient intelligence and humanized computing*, vol. 15, no. 5, pp. 1–2755, Sep. 21, 2022. DOI: 10.1007/s12652-022-04388-6.

[119] X. Nie, X. Liu, H. Yang, *et al.*, "Fully automatic identification of post-treatment infarct lesions after endovascular therapy based on non-contrast computed tomography," *Neural Computing and Applications*, vol. 35, no. 30, pp. 22 101–22 114, Dec. 12, 2022. DOI: 10.1007/s00521-022-08094-4.

[120] X. Gao and Y. Wang, "Optimized integration of traditional folk culture based on dsom-fcm," *Personal and Ubiquitous Computing*, vol. 24, no. 2, pp. 273–286, Nov. 27, 2019. DOI: 10.1007/s00779-019-01336-8.

[121] L. Rajaonarivo, E. Maisel, and P. D. Loor, "An evolving museum metaphor applied to cultural heritage for personalized content delivery," *User Modeling and User-Adapted Interaction*, vol. 29, no. 1, pp. 161–200, Feb. 25, 2019. DOI: 10.1007/s11257-019-09222-x.

[122] I. Lytra, C. Carrillo, R. Capilla, and U. Zdun, "Quality attributes use in architecture design decision methods: Research and practice," *Computing*, vol. 102, no. 2, pp. 551–572, Oct. 1, 2019. DOI: 10.1007/s00607-019-00758-9.

[123] N. O. Ezeamuzie, "Project-first approach to programming in k–12: Tracking the development of novice programmers in technology-deprived environments," *Education and Information Technologies*, vol. 28, no. 1, pp. 407–437, Jun. 30, 2022. DOI: 10.1007/s10639-022-11180-8.

[124] H. Ahuett-Garza, P. D. U. Coronado, J. N. Velasco, E. D. de León López, B. Markert, and T. R. Kurfess, "Train the trainers in industry 4.0: A model for the development of competencies in non-synchronous environments," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 16, no. 2, pp. 775–789, May 12, 2022. DOI: 10.1007/s12008-022-00901-5.

[125] F. A. Somogyi and M. Asztalos, "Systematic review of matching techniques used in model-driven methodologies," *Software and Systems Modeling*, vol. 19, no. 3, pp. 693–720, Nov. 1, 2019. DOI: 10.1007/s10270-019-00760-x.

[126] B. Xue, H. Xu, X. Huang, K. Zhu, Z. Xu, and H. Pei, "Similarity-based prediction method for machinery remaining useful life: A review," *The International Journal of Advanced Manufacturing Technology*, vol. 121, no. 3-4, pp. 1501–1531, Jun. 13, 2022. DOI: 10.1007/s00170-022-09280-3.

[127] S. Mujahid, G. Sierra, R. Abdalkareem, E. Shihab, and W. Shang, "An empirical study of android wear user complaints," *Empirical Software Engineering*, vol. 23, no. 6, pp. 3476–3502, Mar. 23, 2018. DOI: 10.1007/s10664-018-9615-8.

[128] L. Belcastro, R. Cantini, F. Marozzo, A. Orsino, D. Talia, and P. Trunfio, "Programming big data analysis: Principles and solutions," *Journal of Big Data*, vol. 9, no. 1, Jan. 6, 2022. DOI: 10.1186/s40537-021-00555-2.

[129] M. Khan, J. Iqbal, M. Talha, M. I. Arshad, M. Diyan, and K. Han, "Big data processing using internet of software defined things in smart cities," *International Journal of Parallel Programming*, vol. 48, no. 2, pp. 178–191, Apr. 25, 2018. DOI: 10.1007/s10766-018-0573-y.

[130] Y. Sheng, Z. Xu, Y. Wang, and G. de Melo, "Multi-document semantic relation extraction for news analytics," *World Wide Web*, vol. 23, no. 3, pp. 2043–2077, May 18, 2020. DOI: 10.1007/s11280-020-00790-2.

[131] S. P. Perumal, G. Sannasi, and K. Arputharaj, "An intelligent fuzzy rule-based e-learning recommendation system for dynamic user interests," *The Journal of Supercomputing*, vol. 75, no. 8, pp. 5145–5160, Feb. 23, 2019. DOI: 10.1007/s11227-019-02791-z.

[132] A. Sharma and K. M. Goolsbey, "Learning search policies in large commonsense knowledge bases by randomized exploration," 2018.

[133] M. A. Teruel, A. Maté, E. Navarro, P. González, and J. Trujillo, "The new era of business intelligence applications: Building from a collaborative point of view," *Business & Information Systems Engineering*, vol. 61, no. 5, pp. 615–634, Feb. 14, 2019. DOI: 10.1007/s12599-019-00578-3.

[134] Y. Sun and Y. Zhai, "Mapping the knowledge domain and the theme evolution of appropriability research between 1986 and 2016: A scientometric review," *Scientometrics*, vol. 116, no. 1, pp. 203–230, Apr. 19, 2018. DOI: 10.1007/s11192-018-2748-0.

[135] L. David, A. Thakkar, R. Mercado, and O. Engkvist, "Molecular representations in ai-driven drug discovery: A review and practical guide," *Journal of cheminformatics*, vol. 12, no. 1, pp. 56–56, Sep. 17, 2020. DOI: 10.1186/s13321-020-00460-5.