A Survey on Ambient Intelligence Contexts: A Context-Aware Taxonomy based on Deep Learning and Internet of Things Synergy

Parsa Pure Hamedany*
Tehran, Iran
ParsaPureHamedany@gmail.com

Abstract

Deep Learning (DL) and the Internet of Things (IoT) are critical components of modern Ambient Intelligence (AmI), which integrate state-of-the-art Artificial Intelligence (AI) and Information Communication Technologies. The recent synchronization of IoT-enabled big data generation (5G) and advanced reasoning of DL models, such as Multimodal Large Language Models, has created new opportunities and challenges for modern AmI. However, comprehensively analyzing and understanding the broader implications of AmI in the big picture is difficult because of its interdisciplinary nature. To address this intricacy, by adopting the Context-Aware Computing perspective, a systematic arrangement of AmI contexts is proposed. This survey develops a taxonomy of AmI contexts based on four key dimensions: **Human, System, Space**, and **Time**. Each of these dimensions is further split into sub-context categories. By organizing DL and IoT applications within this taxonomy, the study offers a systematic framework to understand and customize AmI systems based on requirements. The resulting context portfolios serve as a flexible conceptual-functional toolkit for researchers and practitioners, aiding them in selecting and adapting contexts for specific applications. This taxonomy aims to clarify the complex landscape of AmI and provide a foundation for future innovations in the field.

Key words: Ambient Intelligence, Deep Learning, Intelligent Environments, Context-Aware Computing, Internet of Things, Cyber-Physical Systems, Smart Environments

1 INTRODUCTION

We are living in the era of ubiquitous mobile interactive large Artificial Intelligence (AI) models in Internet of Things (IoT), learning high-level complex contexts through Context-Aware Computing as Large Language Models (LLM) in the cloud connected to mobile phones via 5G. The trend of computing advancements has been accelerated, as Information Communication Technologies (ICT) panoramic expansion by Moore's law [87]. The plethora of Big Data generation from World Wide Web (cyber) and pervasive IoTs (cyber-physical) in streams, simultaneous with huge progressions in processing capabilities in both software (e.g., algorithmic architectures) and hardware (e.g., microchip processors), made a big technological paradigm shift [133, 106]. In another view, advancements and cheapening of hardware processing capabilities have paved the infrastructural way for more affordability of heavy AI models, trained on large amounts of data like Deep Neural Networks (DNN) [306, 242]. Computer Vision (CV) and Natural Language Processing (NLP) tasks are being achieved by continuously evolving Deep Learning (DL) architectures uniquely. Whether generating realistic images, programming websites, solving mathematical problems, or telling the funny elements of memes, Multimodal Large Language Models (MLLM) handle them with only a few prompts (e.g., words or images) [301, 283, 175].

Ambient Intelligence (AmI) as a subset of AI is impacted by the paradigm shift by DL and LLMs. Prior to State-Of-The-Art (SOTA) DL, such as MLLMs, if watching AI along with IoT, the modern facets of AmI would be revealed [70]. That is observable in various applications, such as smart vehicles, smart homes, smart healthcare systems, or monitoring and controlling trajectories of billions of facilities in Industry 4.0. The term Ambient Intelligence, coined by Zelkha [303], has been defined in various ways, which in the most recent one, Dunne et al. [70] states:

> *"AmI is the combining of AI and IoT with the ubiquity of mobile devices."*

AmI can also be defined based on its features of intelligence, sensitivity, ubiquity, transparency, adaptivity, and responsibility [60]. In parallel, *Intelligent Environments* (IE) are defined by Augusto et al. [21] as, *"An Intelligent Environment is one in which the actions of numerous networked controllers (controlling different aspects of an environment) is orchestrated by self-programming pre-emptive processes (e.g., intelligent software agents) in such a way as to create an interactive holistic functionality that enhances occupants experiences."*. Whether AmI is more than programmed autonomy in an environment, it requires tasks to be performed with a degree of intelligence (e.g., reasoning) like assistance without micro-management [87, 70, 21, 60].

Apparently, despite the rapid changes in AI and DL in recent years, there is no survey covering the SOTA AmI; while LLMs have been transforming DL into not only general-purpose reasoners but also emergently zero-shot learners without requiring model parameters changing (pretraining or finetuning) by prompting [287, 286]. Despite the recent technological transformations related to AmI in the last decade, they have not yet been analyzed and interpreted contextually in the *big picture* as is depicted in **Figure 1**. This lack of research is addressed here through a comprehensive survey of SOTA trends organized by AmI contexts. Since insufficient works are titled as AmI or IE with SOTA AI inclusiveness, it is aimed to survey them by their *contexts focusing* on DL. The context philosophy, as applied in *Context-Aware Computing* [2, 65, 67, 214], is adopted here to survey AmI because direct data on this topic is lacking.

HUMAN
AmI
SYSTEM
SPACE
TIME
Types
Location
Physics
Cycle
Flow
Scope
Cyber Physicality
Ambit
Outdoor
Indoor
Monitoring properties
Perception of priors
Appearance
Sentiment
Behavior
IoT
Internet of Things
Data
collection
DL
Deep Learning
Things, Vehicle
Store, Home
Office,Gym
Mall, Hospital
Hotel, Building
Museum, School
University, Street
Port, Airport
Industries, City
Physical reality
VR, AR, MR
Environment
Ecosystem
GPS, Camera
BLE, RFID, Wi-Fi, CCTV
Observable, Latent
Modeling
data-driven
model-based
hybrid
Reasoning
Attendance
Age
Gender
Physique
Fashion
Body language
Verbal language
Physiogical states
Personality
Activity
Detection
Recognition
Tracking
Authentication
Chronological
Biological
Female, Male, Others
BMI, Pose, Shape
Style learning
Fashion synthesis
Facial expression
Hand gesture
Body posture/gesture
Eyes gaze, pupil, blink
Speech, Opinion
Brain EEG, CT, f-MRI
Circulatory
ECG, PPG, PW, BCG, BVP
Skin SKT, GSR, EDA
Eye EOG
Respiration RSP
Muscles EMG
Cues, Traits
16PF, PEN, Big-Five ,MBTI
Motion, Action, Interaction
Individual, Group, Crowd
Deterministic
Stochastic
Timescales
microsecond
years
Periodicity
Sequences
Change points
Eevents
Memory decay
dependencies
gradient vanishing
Historical profiles
Memory
Data
Spatiotemporal
Stream
Time-Series
Analysis
Retrospective
Prospective
Real-Time
Latency
Subjective
present
Perception
Network
Processing
Application
Business
Acquisition
Pre-processing
Multimodality
Learning types
Architectures
Sensors
Actuators
Devices
802.11 ,NFC, RFID, BLE
5G, LTE ,4G ,ZigBee
Wireless Sensor Networks
Communication strategies
Computing processors
CPU, GPU,TPU
Middleware
device, fog, cloud, hybrid
Healthcare, Smart Home, AAL
Smart City, Industry, Commercial
Business model, Management, Monitoring
Discovery, Generation, Augmentation
Labeling, Re-labeling, Compression
Validation, Cleaning, Integration
Modality types, Multimodal ML
Supervised, Unsupervised
Reinforcement, Semi-Supervised
Representation, Transfer, Multi-Task
Self-Supervised, Federated, Meta
Multimodal, F-ew-Shot, Zero-Shot,
In-Context, Reasoning, Planning
CNN, RNN, LSTM, GAN, DRL
Transformers, PTMs, Foundation Models
LLMs, GPT, ChatGPT, MLLMs, LVMs

Figure 1: Ambient Intelligence (AmI) contexts' taxonomy. The taxonomy of AmI contexts is extended in four directions. Each direction represents a context and its sub-contexts at three more in-depth levels. By deepening into further levels of the taxonomy, their background color gets brighter to white. The first level of the taxonomy with four contexts of Human, System, Space, and Time are the most abstract contexts. The second level is about the major categories of each context. The paper organizationally follows the taxonomy structure, whether some subordinates (e.g., the lowest level) are not titled. The taxonomy aims to illustrate a big picture of AmI contexts as its framework, to view its variety more simplified.

Various definitions, categorizations, and applications have been submitted in context and context-awareness after the initial use of the *"context-aware"* term by Schilit et al., which eventually was for an environment [244]. This paper follows Dey et al.'s [65] definition of context by its greater generality and updates over the most official one [2], which is:

> *"Any information that can be used to characterize the situation of entities (i.e., whether a person, place, or object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves. Context is typically the location, identity, and state of people, groups, and computational and physical objects."*

Context-Awareness has plenty of ways to be accomplished, as DL and IoT are its two enabling technologies [12, 214, 236, 196] contributing to awareness throughout any phase of the context lifecycle (acquisition, modeling, reasoning, dissemination) [214]. As *categorizing the contexts* of an entity clarifies the most effective parts for designers, interpreting it as a contextual taxonomy of that entity can organize high-level conceptual and practical comprehensions [2]. The first level of AmI context taxonomy, as its structure, is inspired by previous classifications, which are reviewed in the next section. AmI, an interdisciplinary subject covering heterogeneous applications, will evolve into a *transdisciplinary* subject with immense intricacies if it becomes context-aware. IE subsystems are complex entities themselves as if they are potentially independent systems, and when consolidated, the new system becomes a complex *Systems of System* (SoS) [126].
In order to overcome these challenges and overview inclusively, the *taxonomy* of that entity (AmI) by categorizing its portions, will serve as a functional method to simplify complications. The methodological idea as the framework of this survey is adopted from the Context-Aware Computing paradigm [214, 2, 67, 65] in structured systems thinking while categorizing its contexts, to observe SOTA in AmI and IEs in abstraction. The AmI context taxonomy enables a systematic and less biased survey that identifies implications of future challenges and opportunities. Moreover, reviewing survey papers that *deliberately classify* up-to-date information about their addressed topics (contexts) serves as a reliable guideline for understanding the current state of the subject. Hence, the main resources of this study are recent survey review research.
This *study aims* to propose a modern comprehensive *Ambient Intelligence context taxonomy*, as is represented in **Figure 1**, to survey recent trends and provide a toolkit for both researchers and practitioners. The approached notion here is formed through the consolidation of related cutting-edge technologies and the philosophy of Context-Aware Computing with systems thinking as its inspiration. The first contribution is to comprehend the current status per context, followed by systematically overviewing each concept and its interconnections. Then, address contexts and subsume each's incorporations. Finally, recognizing salient challenges and assessing precedents' hindsight to project future foresight extrapolation through discussion is represented. The SOTA of AI and ICT are considered respectively DL and IoT, which are mutually concatenated to Big Data.
By reviewing context types found in previous works, the nominated one is derived as *Human*, *System*, *Space*, and *Time*, which are briefly introduced here:

- **Human**: how people can be recognized (who) and predicted contextually, whether in the fashion they *appear*, in the way they *feel in* a deeper context, or in the manner they *behave*. The manifestation of their appearance, expressions, and actions are cues (why) to bring metadata about their demography, emotions, attitudes, health, habits, and characters towards interacting.

- **System**: which assumed SOTA subsystems of IE systems are characterized by data, IoT, and DL. It incorporates the attributes of *data collection*, *IoT layers*, and some impactful *DL algorithms and learning methods*.

- **Space**: what types of IEs are and how space can be perceived by AmI using positional features (where) or physical knowledge (how). Physical information from sensing to perception further enhances *spatial awareness and physically aware AI*.

- **Time**: when time-based contexts of AmI can significantly affect IEs. By the *one-way flow* of time, as in series sequences or *recurring temporal patterns*, to notice phenomena and memorize them as for reasoning, time has much information to represent.

1.1 Related works

In this section, similar survey papers will be reviewed and then the related taxonomies in context-awareness will be analyzed. The related surveys are filtered from keywords of "*Ambient Intelligence*", "*Intelligent Environments*", "*context-awareness*", and "*Smart Environment*" with more attention to their recency. While a pleasant result was not obtained, research continued in surveys composed of AI and IoT related terms. The closest survey to the mentioned Key Performance Indicators was a *CSUR* publication, Dunne et al. [70], which has considered SOTA DL and IoT into AmI. However, almost no attention was paid to context-awareness, and no categorization was presented for IEs except listing related features, as the survey was done approximately *four years ago* (before the LLM paradigm). The other related paper is Gams et al. [87] published *five years ago*, which makes it not covering SOTA. Instead of surveying AmI, it tended to define several applications in conceptual terms with an eye on the future, whether no referral to IE there was. Cai and Yang et al. [41] focused on IoT's role in AmI overviewed architectures in sensing and local processing in multiple applications, however, like the priors, context-aware computing was not investigated. An older survey (2013) while more inclusive, Perera et al. [214], was IoT-based and highly heeded context-aware computing, caused by publish time, even if ahead of its time mentioned DL, could not support SOTA. Other related works are analyzed in **Table 1** by *DL*, *IoT*, *AmI*, *IE*, *Context Awareness*, *Context Taxonomy*, and *Published Year* parameters, whether covered or not.

Table 1: Related works by covering factors which CA is context-awareness, CT is context taxonomy, and PY is published year. Indicators show 'ü' as covered, 'Ø' as not covered, and '!' as vaguely noted.

	[264]	[57]	[13]	[183]	[40]	[70]	[261]	[143]	[239]	[306]	[269]	[196]	[87]	[41]	[182]	[214]	[21]	[236]	[60]
DL	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	ø	ø	ø
IoT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	ø	ø	ø
CA	ø	ø	✓	ø	!	!	ø	ø	ø	✓	ø	!	!	!	ø	✓	✓	!	✓
AmI	!	ø	✓	ø	ø	✓	ø	ø	!	✓	ø	ø	✓	✓	ø	✓	✓	✓	✓
IE	ø	ø	ø	ø	ø	✓	ø	ø	ø	ø	ø	ø	ø	✓	ø	ø	✓	ø	!
CT	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	✓	✓	ø	✓
PY	2023	2023	2022	2022	2022	2021	2021	2021	2021	2020	2020	2020	2019	2019	2018	2013	2013	2011	2009

Selecting context categorization can prepare an important connection between reasoning and each specific application [20]. Whether the system must bring intelligence into an environment to interact with people, categorization has to be specialized for that field. In that setting, Feng et al. [82] presented context categorization for AmI by bifurcating it into two major categories of *user-centric* and *environmental* context. The user-centric context contains the users' background, dynamic behavior, physiological state, and emotional state; besides, the environmental context comprises the physical environment, social environment, and computational environment. Another related context categorization by Augusto et al. [21] in IEs divided context into *user*, *environment*, and *system*. Another influential categorization by its vision similarity is by Chen et al. [50], which categorized contexts into *user* context, *physical* context, *computing* context, and *time* context, rooted in Schilit et al. [244].

Other context categories in Context-Aware computing bring different notions of context philosophy and how its concepts can be discerned, which for more comparisons visiting Perera et al. [214] is recommended. Abowd et al. [2] classified contexts into characteristics of the entity (5Ws) that are *identity* (Who), *location* (Where), *status* (What), and *time* (When); in addition, those contexts would be used to figure out the *reason* (Why) a user took that specific action. Dourish et al. [67] observed two views of context by noting the activity-oriented aspect, *representational* and *interactional*. The representational aspect of context is about the information that is stable and separable from activities in the environment. Although content and context are two segregated entities, the interactional form "instead argues that context arises from the activity" [67]. Consequently, all activities and interactions of each user in the environment would be deemed as context. *Operational* and *conceptual* views of context are bifurcated by Van Bunningen et al. [270], which operational is about the procedure where context information is acquired, modeled, and treated. Albeit the conceptual category states the meaning and contextual interrelations by identifying user-centric contexts from ambient [82]. The operational view was based on another categorization made by Henricksen et al. [110] in four types of sensed, static, profiled, and derived. Van Bunningen et al. [270] called derived context as *high-level* context and the past three types as *low-level* context. Calling a context as a high-level can although be related to context lifecycle in an IoT-based vision as Perera et al. [214] brought up in four cyclic steps of *acquisition*, *modeling*, *reasoning*, and *dissemination*. Whether the raw data is received from sensors (context acquisition), then modeled, and after it is processed (context reasoning), can be called high-level context instead of low-level.

The exchange of views continues regarding a comprehensive context categorization while it is aimed to get ideas from the previous works to the concluded position in this paper, which is: *Human*, *System*, *Space*, and *Time*. That resembles Augusto et al. [21] if Time context is added, correspondence with Chen et al. [50] by transforming computing context to System, and more unabridged categories or less generality comparing the user-centric and environmental context types of Feng et al. [82].

In the following sections of this paper: Section (2) *Human context* in three subsections of *appearance*, *sentiment*, and *behavior* is viewed, Section (3) addresses *System context* in three subsections of *data*, *IoT*, and *DL*, Section (4) observes *Space context* in *scope*, *location*, and *physics* subsections, Section (5) *Time context* of AmI is mentioned in two *flow* (one-way) and *cycle* (two-way) forms of time, Section (6) discussion about the study's configuration is presented, highlights the *open issues* and *challenges* of this study and AmI in general, points several possible *future directions* of this context, and ultimately brings the *conclusion*.

2 HUMAN CONTEXT

Even if there is no background information for newcomers, their extractable information in the environment is the Human context's concentration. The way humans' status looks and how conjectures about their characteristics can be made by AI is discussed in the human context as is exemplified in **Figure 2**. That is perceivable by different levels of human analysis from outside to inside and deepening into

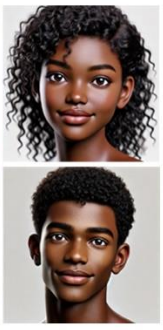
■ Appearance

■ Authentication

■ Age

■ Physique

■ Gender



■ Fashion

● Behavior

● Action

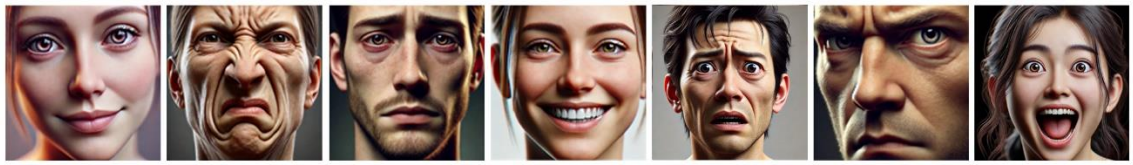
● Activity

▲ Sentiment

▲ Speech

▲ Hand Gesture

▲ Facial Expression



▲ Physiological State

▲ Body Posture

▲ Opinion

Figure 2: The Human contexts in AmI. These contexts can be diversely comprehensive as a brief presence in the AmI might bring various metadata of the Human with the available technologies such as AI and IoT. The goal of this Figure is to visualize some Human contexts as examples. All images used in this image is generated by an AI foundation model (ChatGPT 4o).

cognizance of users via SOTA technologies analyzing how people *look*, *feel*, and *act*. Here the Human contexts are classified into three aspects (1) *Appearance* (2) *Sentiment* (3) *Behavior*.

2.1 Appearance

The appearance section is how people would seem outwardly, whereby they are occurred by their semblance and appearance characteristics estimation from the first interaction to infer. This section outlines appearance analysis in five parts: (1) *attendance* (2) *age* (3) *gender* (4) *physique* (5) *fashion*.

2.1.1 Attendance

Attendance is when a person enters the environment, *recognizes* it as with object detection, counts her/him with other present people in the area, and detects motions [316]. It also may be tracking the guy by any *footprint*, as by (multi) object tracking, until exiting [178]. Each human would be detected by face (facial geometry, texture, skin color), motion (tracking), and body features [15] for identification purposes. Detection can also be accompanied by *tracking* during the presence of people to record the whole trajectories of each person spatiotemporally. A way would be *Indoor Positioning Systems* (IPS) to attain people's attendance by tracking location via smart devices owned by them, which is not considered here but in the Space context. On a larger scale, *crowd* attendance by using counting and density estimation (crowd statistics) is possible [31, 79, 252]. The detection can be done simultaneously with *authentication* in which biometric methods seem more relevant to be done via face recognition, iris recognition, fingerprint, palm-print recognition, voice recognition, keystroke touch dynamics, and physiological signals [233].

2.1.2 Age

Age group estimation approximates human age by *biometric* features [238] as the human body has two age forms, *chronological* and *biological*, to distinguish. Chronological view is how old a person is by birthdate and biological is "based on the biological quality and functioning of tissues, apparatus and organs of an individual" [216]. Because of the complexities of detecting biological age, mainly the chronological aspect is meant here. People in peer groups have more *common* than dissimilar groups, causing their needs to be akin to each other. That can be done by both image processing (facial, body) and sensors which the more attention is on facial analysis, but sensor-based approaches are feasible as gait analysis in recent years [7].

2.1.3 Gender

The conjecture and recognition of male, female, or other gender types based on *physical-biological* appearance is the point of this part. People of different genders have diverse societal roles, living experiences, and needs [268]. Gender detection can be operated with vision-based and sensor-based approaches while gender detection via image is more common and accurate. By CV, gender can be classified by face [243], body skeleton [27], and gait [5, 243]. It is also seen that tasks like this are achieved easily by DL with higher accuracy than humans [282].

2.1.4 Physique

Physical and structural features of the *body* such as *height* (stature), *weight*, *shape*, and *Body Mass Index* (BMI) estimation is the intention of the physique part. By knowing the position of the camera (height and angle) height of objects and humans' stature are measurable [198, 156]. Human body weight estimation with anthropometric features can prepare further analysis like BMI approximation and help individuals control and use it through health improvements and even clinical uses [129]. 3D models of the body have many advantages and can be used in all sections of the human context, which is possible even with a single 2D image (depth sensor data can be used too) [212] to do the human mesh recovery (3D body pose and shape estimation) [267].

2.1.5 Fashion

Detecting and classifying each individual's *clothing*, *accessories*, *hair*, and *makeup* styles is about this part. Fashion recognition can provide information about humans' lifestyles as how luxurious they prefer, which style is their type, or any sort of classification to find their needs caused by communication form of fashion [26]. In order to reach the fashion context by multimedia, after landmark detection and pose estimation in the pixel computation level, understanding and assessing the stylistic features of a person's fashion is required for attaining fashion analysis at an ultimate level [254]. In a more detailed view, Cheng et al. [54] considered four aspects for enabling intelligent fashion through fashion detection [32] at first, fashion analysis with *attribute recognition* [111] and *style learning* [141] secondly, then fashion *synthesis* with *transferring style* [122] and *pose transformation* [179, 23], finally fashion recommendation for compatibility [98], matching [255], and suggestions [48]. In almost all intelligent fashion-related works, CV is used.

2.2 Sentiment

Indeed, AI neither feels nor understands emotions like a human, nor comprehends complicated human interrelation circumstances, nor understands how the grief of losing a loved one is. However, it can bring real-time decision-making about individuals' emotional state with context-aware multimodal emotion recognition systems in *affective computing* [8, 296, 259, 281, 220], and in some ways, it can afford higher agility and ubiquity than humans in emotion recognition. According to Webster's New World College Dictionary, 4th Edition *sentiment* is:

"*a* complex combination *of feelings and opinions as a* basis *for or* judgment*;* general *emotionalized attitude".*

Emotion has been modeled in three *discrete* (categorical), *dimensional*, and *componential* manners [147]. Discrete models consider emotions as separate basic emotional state categories (e.g., Ekman's six basic emotions) [77]. Dimensional models consider emotions as cluster points in a multidimensional graph made of two dimensions (e.g., Russell's circumplex model of affect) [234] or three dimensions (e.g., Mehrabian's pleasure-arousal-dominance) [192]. Componential models are the combination of various factors that impact emotions (e.g., Plutchik's psychoevolutionary theory [217], Ortony-Clore-Collins (OCC) model of emotion [207]). All in all, each research in this field picks one of these models as a structure to classify the output of emotion recognizer. The sentiment section consists of *emotion* (feelings) and *notion* (opinions) through affective computing that relates psychology and social sciences to cognitive and computer sciences by containing both emotion recognition and opinion mining [281]. Emotion recognition methods detect how a person's emotional state is from physical and physiological contexts, however, opinion mining methods want to understand what people's thoughts and sentiments (e.g., negative, neutral, positive) are mainly from their linguistic signals [220].

2.2.1 Body Language

Humans' *postures* and *gestures* can communicate nonverbally, revealing signals to other people. It also exposes a person's emotional condition and each body movement can be interpretable [213]. Body language can communicate with diverse body parts, including *facial expressions*, *hand gestures*, *body gestures*, *postures*, and *eye movements and* any of these parts can be emotionally interpretable [204]. Although body language is vigorously culture-dependent [138] even in facial expressions [123], whether in hand gestures contradictions might be the opposite or offensive [213].

2.2.1.1 Face

Humans express their feelings and senses of their thoughts with their *facial components' configuration* and effect for reacting to stimuli, unconsciously or in a controlled manner. The face also is not only a multi-signal system but also a multi-message system, in which extracting messages brings a reliable non-verbal communication bridge [75]. The face manifests three types of signals [75]: *static* (e.g., skin pigmentation, face shape), *slow* changes by aging, and *rapid* with any movement of facial muscles (e.g., raising the eyebrows). The position of *rapid signal changes* provides *Facial Expression Recognition* input to sentimental analysis. One of the key factors in Facial Expression Recognition is the performance in a *wild* situation, which improving it needs an appropriate database [161]. That means *micro-expressions* can facilitate emotion and affect recognition [195] by quick (less than 0.5 s) facial emotional response [100] caused by time-based attributes of emotional responses [76]. All in all, detecting facial components' movements by not heeding contexts would not be enough, even if it is learned by any sophisticated computing algorithm [28]. While vision-based works are prevalent, using other sensors (depth, EEG, infrared thermal, audio) caused by problems such as illumination variation and head pose in multimodal sensor data in facial expression recognition may impact performance [240, 42].

2.2.1.2 Body

Posture is the unconscious body positioning with the *skeletal muscles' contraction* in the space [45]. It contains the configuration of the *head*, *body*, *arms*, and *legs* with each other. While the head represents emotions naturally, the body demonstrates the intensity of emotions [74]. The human body can be modeled in *part-based* or *kinematic* models by assuming the body as a configuration of separate parts or interconnected joints [204]. In each procedure for recognizing emotion from body gestures, after detecting the human body, body pose detection and tracking has to be made for representation learning and emotion recognition [204]. The *walking style* (*gait*) can even present particular gait kinematic patterns in separate emotional states and effects [295, 258]. Posture prediction can be made by anticipating the possible posture from a temporally incomplete time series to localize *joints' positions* [180]. Human body posture and gesture recognition can be done by sensor-based (e.g., depth, accelerometer) and Radio Frequency-based (e.g., Wi-Fi, RFID), besides vision-based techniques [131]; while radar-based gesture recognition is another new way for this task [250].

2.2.1.3 Hand

The way hands *move* and the *shape* of hands represent specific meanings that can express people's feelings and send expressive messages about their thoughts [138]. Hand gestures are a manual communication channel by sign language as a natural language recognizable via machine [227, 103]. Computers can recognize hand gestures and their meanings in various ways, but wearable glove-based, camera vision-based, and surface electromyography (EMG) are the mainly used approaches [208, 299]. Besides, other sensing technologies are available such as ultrasound and pressure sensors [103], while soft systems like bioelectronics, e-skin, e-tattoo, and soft circuits are playing the future role [130].

2.2.1.4 Eye

Eye tracking is the procedure of *gaze analysis* spatiotemporally by measuring where and how long the eyes are stared at and the pupil size [167, 145]. By analyzing *eye movement* and *pupil behavior's* related metrics, visual attention, emotional arousal (esp. stress), and cognitive workload can be depicted [253]. Those metrics are *fixations*, *eye movements*, *smooth pursuit*, *pupil size*, and *blinks* [253]. It can also be used to analyze mood from the reactions of individuals to each object [88]. Eye-trackers have various types that are eye-attached, optical, and electric potential measurement [167], or from another point of view, they are head-mounted or remote [55]. The main eye tracking techniques are classified as scleral search coil, infrared oculography, electrooculography (EOG), and video oculography [145].

2.2.2 Verbal Language

Language as the most effective tool of humanity for communication and cognition plays a vital role for us and what we make. The NLP incorporates both Natural Language Understanding (NLU) and Natural Language Generating (NLG). NLG is about creating meaningful statements and NLU studies sound of words (phonology), structure of words (morphology), words' arrangement (syntax), meanings referring (semantics), and contextual inferring (pragmatics) [140]. From another point of view, language analysis can be classified as discourse analysis and sentence analysis (syntax and semantics) while the theoretical aspect of language (e.g., grammar) is in a segregated category from computation linguistics [56]. NLP has various tasks, but some more advanced ones acquire more attention by their diverse usages, like Sentiment Analysis (SA) and automatic speech recognition, which are brought up here. LLMs as SOTA in NLP are task-agnostic general purpose task solvers, supporting many tasks whether by zero-shot learning or finetuning downstream NLP tasks [312].

2.2.2.1 Speech

Speech is one of the most convenient and efficient communication ways that several sorts of information can be extracted from it nowadays such as the speaker's *age*, *gender*, *language*, *accent*, *identity*, *health condition*, *emotional status*, and *subject of content* [200]. Text and speech processing except for their different input formats are distinct from paralinguistic speech variations (e.g., voice pitch, accent, speaking style, speed) [186]; while language models are almost used in both of them, speech grammar rules are not used all the time [186]. Paralanguage and language constituting speech and speech emotion recognition will be formed by considering its main features such as prosody, spectral, voice quality, and Teager-Kaiser Energy Operator features [284]. *Prosody* is the study of syllables, intonation, stress, and rhythm; *spectral* features shape of the vocal tract by its frequency; *voice quality* is how vocal waveform is; *Teager-Kaiser energy operation* measures energy and frequency of voice signal amplitude.

2.2.2.2 Opinion

To understand the meaning of a linguistic modality, many sides of NLP take part in its downstream tasks, whether in summarization, translation, conversational interactions (question-answering), or Sentiment Analysis (SA). LLMs support myriad NLU tasks such as SA and reasoning (e.g., inference, mathematics, comprehension) within conversation structure [201]. They are also capable of doing SA even if the model cannot afford its standards. With LLM's ICL the SA accuracy is raised by demonstrating a few related examples [66]. Among NLP tasks, SA is focused on comprehending the notion of a piece of information separated from reasoning in a query of facts. The SA or opinion mining field is an affective computing part which is one of the major NLP tasks. It is mostly performed on text data to understand what the author's thoughts and emotions are about by considering the *polarity* and *aspects* of text entities [49, 174].

SA incorporates different levels in three main *document*, *sentence*, and *aspect* levels [166, 35]. The most uncomplicated level is the document level, which considers documents as one entity and tries to classify the document's discourse polarity (positive, negative, neutral) or subjectivity detection. Subjectivity detection filters information related to facts to determine the opinionated information [173] for feeding polarity classification, whether syntactically or semantically [49]. At the sentence level, the document text parses into sentences and each sentence would be semantically and syntactically analyzed as a distinct entity [49]. More complicated and fine-grained aspect level analysis wants to find out what aspects are covered in the text content to recognize the opinion holder's main intention, further to polarity and subjectivity detection [245]. Apart from subjectivity detection and aspect extraction, other remarkable SA tasks include opinion spam detection, implicit language detection (sarcasm), and cross-domain and cross-language sentiment classification [285].

2.2.3 Physiological state

Human physiological processes contain plenty of reliable and unconscious information to analyze without the bias of pretense [142, 8]. Physiological information can be used for individuals' feelings recognition and monitoring from the *brain*, *heart*, *blood circulatory*, *skin*, *muscle*, *breath*, *sweating*, *temperature*, and *eyes* [71]. The brain's electrical activities and its changes are chiefly recorded and measured by Electroencephalography (EEG) signals as other methods are used in medical applications as Magnetic Resonance Imaging (MRI) Computed Tomography (CT) techniques (e.g., functional-MRI, Diffusion Tensor Imaging, Single-Photon Emission Computed Tomography, Positron Emission Tomography). The circulatory system, especially heart rate and blood state, is analyzed in various ways with electrocardiography (ECG), photoplethysmography (PPG), Pulse Wave (PW), ballistocardiography (BCG), Blood Volume Pulse (BVP), blood oxygen saturation, and glucose. Skin related signals include Skin Temperature (SKT) and Galvanic Skin Response (GSR) or electrodermal activity (EDA) which is the electrical potential changes through sweating moisture. Other techniques to mention are electromyography (EMG) for muscles and nerves' electrical activity measurements, respiration (RSP) for analyzing breathing, and electrooculography (EOG) for the eye's retinal electrical analysis.

To do emotion recognition via physiological signals, typically after *stimulation* indicated to a person (e.g., music, image, video), signals' states will get measured for classification based on the emotion model [281]. The most used signals in emotion recognition are EEG, ECG, EDA/ GSR, RSP, and EMG [153, 249, 72, 172] whereas others like PW and SKT were used less than the priors. Stress and mood recognition in everyday experience has been done from physiological signals [153, 237]. Training DL models on EEG, ECG, and EMG signals, multiple analysis tasks are applicable such as brain and heart functionality at *diseases*, *sleep stage*, *age*, *gender*, *motion*, and *emotion* [105]. Each physiological signal has its proper sensor, but their acquisition tools for emotion recognition are classified as smart wearable, mounted, and external [8]. Although physiological states might be used as a modality of multimodal affective computing, which beyond multi-physical modalities, get fused either as homogeneous multi-physiological or heterogeneous physical-physiological multimodalities [281].

2.3 Behavior

Human behavior is *taking any action* and *conducting manner* of oneself or with others in interaction with the environment (objects, system) or people [83]. In this section, the external observable and measurable aspect of behavior is considered as in the *behavior analysis* [22]. Three aspects construct this section: (1) *Human Activity Recognition* (HAR) techniques to detect and analyze human behavior from sensory data [6], (2) *personality computing* which explains behavior psychologically by computing on personality trait theories [276], and (3) *interactions* as feedback and instructions. Acts of individuals represent their behave whether the behavior is a proper tool to measure personality, although personality traits are also related to sentiments as feelings and thoughts impact behavior [293]. Relation and casual loops exist between the three of them, whereas personality computing tasks are related to how data and computation algorithms, but not the description and explanation still [215]. By recognizing activities and analyzing them in their frequency quantity and contextual parameters, behavior patterns and habits will be discovered [151]. That can result in obtaining behavior change [85] and make anomaly detection tasks more enhanced [158] by considering the routines and contexts to interpret behavior as normal or abnormal, as in Ambient Assisted Living (AAL) [58]. Since *Kurt Lewin* stated that behavior is a function of the *person* (history, personality, motivation) and *environment* (physical and social surroundings) (B = f (P, E)) [159], contextual parameters had been considered effectual on behavior.

2.3.1 Activity

Human behavior analysis tasks can be classified based on the degrees of semantics and time frame into *motion*, *action*, *activity*, and *behavior* [47]. Action and activity terms are almost interchangeable in HAR, which action is often simpler and has less duration compared to activity, whether an activity is mainly composed of actions [53]. For instance, grabbing a glass is an action but making a coffee is an activity created by a sequence of actions. In HAR when the action is done decision is made, but in many cases, the decision has to be taken before occurring the event caused by urgency. For instance, the fall prediction of the elderly or the autonomous vehicles that have to decide swiftly what action to take prior to the accident happening. The solution to that problem is action or activity prediction which is about the future state by learning from unfinished cut from the end data, to estimate what the action before it is completely performed will be [148]. This task can be achieved by human *pose estimation* if the objective is just predicting what the posture might be by analyzing the trajectory sequence of body joints' localization before the judgment execution [180]. HAR is classified into *motion-based* (e.g., people counting, motion detection, tracking), *action-based* (e.g., gesture recognition, posture recognition, behavior recognition, fall detection, activities of daily living, AAL), and *interaction-based* (e.g., human, object) categories [118]. This standpoint of HAR is also interpretable in an atomic to complicated hierarchical leveling manner [6]; following motions and gestures, action recognition takes place, then interaction detection, and finally group activities with the highest level of complexity. Interaction is a reciprocal relation between humans and objects in three forms *human-human* (e.g., kissing) [235], *human-object* (e.g., wearing glasses) [92], and *object-object* (e.g., pot on the stove) [95].

Along with human-human, human-object, and human-scene interactions [118], any action in the environment can be considered an *interaction*. Interacting with the system can be further than settings changing in a smart mobile application and getting pervasive into the environment's embedded sensors and actuators using SOTA IoT. LLM agents understand and reason our instructions, they furthermore can adjust their reactions after our behavior as feedback towards planning future actions [280]. Therefore, if they get finetuned on other behavioral attributes of users as their activities and personalities, it would bring a customized interactive framework. Whether by getting multimodal as MLLMs, many planning tasks can be achieved by interactions such as multimodal conversation (e.g., visual QA), saying story beneath images, and finding funny reasons of memes [301, 283]. So, models capable of solving high-level mathematical questions from images even without optical character recognition, might soon be able to infer human behavior into a cohesive conversation.

HAR also can be regarded with numerical scaling as *individual*, *group*, and *crowd* [294]; in which groups consist of fewer individuals and have more intercorrelations than crowd [241]. *Group activity* is a type of interaction among a group's members [294] as standing in a queue, walking together, or other social activities [73]. *Crowd behavior* has been studied in two views, microscopic (bottom-up) and macroscopic (top-down), by considering the crowd as multiple separate individuals or a monolith entity [266]. Besides how to see a crowd, its analysis would be done by crowd scene analysis and crowd statistics [31, 79, 252]. In which crowd scene analysis is integrated by subtopics of crowd behavior recognition, motion tracking analysis and prediction, and group behavior analysis [31].

The most common categorization made in the HAR field is bifurcation according to its input data format to bring sensor-based and vision-based [62] HAR, which has been stated as the main question of many works as sensor-based [52, 279, 60, 51] and vision-based [125, 309, 305, 228, 128] human action/activity recognition. Vision-based supports only RGB (Red Green Blue) and/or depth sensors (RGBD) while sensor-based contains all types of sensors that are capable of getting used in HAR whether the sensor is wearable, attached to any object, or device-free (environmental) [118].

2.3.2 Personality

Personality is the inner characteristics in behavioral patterns that are shaped (e.g., habits) and shown by people in each particular situation. Personality computing is automatic personality judgment from the main observable stable factors of persons' behaviors about how the inner personality might be, based on models of *personality trait theories* [276]. Whether personality computing as HAR does both the perception for predicting personality traits and the recognition automatically, it also includes automatic personality synthesis as a major problem [276].

There are different personality trait theories as *16PF* (Sixteen Personality Factor) [46], *MBTI* (Myers-Briggs Type Indicator) [199], and *PEN* (Psychoticism, Extraversion, and Neuroticism) [78]; but the *five-factor* model (Big Five) [189] is the most used in personality trait recognition [313]. The Big Five model contains the Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism factors and their opposites. Human personality is recognized from *behavioral cues* containing almost all the previous human contexts (e.g., face, body, speech, fashion) [193, 43]. Considering the point of view of this article's human context, personality detection has been done with all the

sentiment section's parts and fashion cues. Recognizing personalities has applied based on varied data types (text, audio, static image, video, and physiologic) [132, 185, 260, 44] while SOTA works similarly to affective computing using a compound of this information in a multimodal fashion [193, 43].

3 SYSTEM CONTEXT

A system is an organized entity composed of multiple interactive elements to achieve common goal(s) and if both the elements and their interactions have sophistication, it is a *complex system* [277]. Elements of the system might be systems themselves as subsystems of the system, and if the subsystems can be independently operable, it is a system itself and the system is a *System of Systems* (SoS) [126]. The IEs are a type of Embedded System and CPS which can be potentially a complex system, SoS, and a large-scale system, so it seems worth analyzing it as a System. The system will be built based on stakeholders' expectations and needs towards its *requirements* to be architected and designed. The system's design differs since it evolves from requirements, missions, purposes, applications, and forms which accordingly, functionalities get specified. *Functional* (logical) and *physical* requirements would lead to concept selection toward system's architecture through the nominated concepts. It is aimed by this survey that context portfolio (taxonomy) lets designers and architectures choose concepts, to provide solution-neutral space, at both abstract and technical levels, resulting in an appropriate allocation.

A traditional way for systems analysis is to understand the system's building blocks in functional and componential hierarchies, to perceive the utilities of each part in an organized arrangement. In the componential hierarchical decomposition of the system, the first element might be other systems that are independently operable each on their own. However, they are consolidated with integrated interconnections in SoS conforming to the same objective, coordinated and coherently [126]. If the system is not a SoS, is composed of subsystems that are not fully independent runnable systems; whether it would be a complex system as human body parts like the brain, which is not independently operable but complex [149]. The subsystem's subordinate, components (modules), are the constituent items of subsystem configuration, whether are made of subcomponents. Finally, in the most granularity of hierarchy, there are parts, which their functionality is only confined to be merged with others [149]. For instance, a group of people is an SoS, a human is a system, the brain is a subsystem, the cerebellum is a component, the temporal lobe is a subcomponent, and granule cells are parts.

As mentioned above, any system based on its requirements shall have a particular concept selection for the architecture and design, which among alternatives in each level of hierarchy, there are trade-offs (e.g., cost, performance, risk, robustness, schedules). As a result, even for similar applications, stakeholders' expectations, and missions, it is highly challenging (illogical) to have one specific systematic framework; although, requirements in IEs are not analogous. Albeit, as the attention of this survey is SOTA in AmI, the roles of some systems are bolded more than others. Consequently, the concentration of this context is focused on specific DL and IoT systems since they have pivotal effectivity in SOTA AI and ICT, and modern AmI [70].

A system should be organized and developed on its logical and functional requirements in its life cycle. Besides, some recent technological opportunities such as real-time wireless systems in IoT (e.g., 5G) and Pre-Trained Models (PTM) in DL can be candidates in any AmI system, because of their generalization. Thus, the candidate concepts in this context are *IoT* and *DL* as systems of the IE systems, since they are building blocks of a modern AmI system [70] as is illustrated in **Figure 3**. Further to that, by the diversity of inputs (contexts) and the importance of proper *data* in any context, informational aspects are also analyzed in the System context.

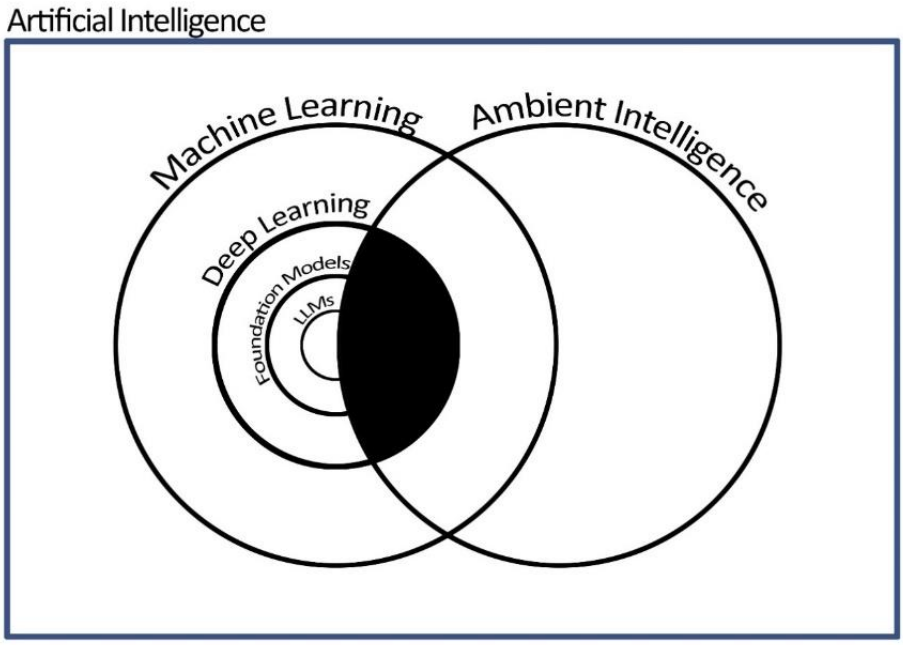


Figure 3: The covering domain of Ambient Intelligence discussed in this paper. Any part of Deep Learning as foundation models can be Ambient Intelligence enabler.

3.1 Data

Both IoT and DL are dependent on data directly and indirectly; IoT is a modern powerful tool to capture *Big Data* in high *volume*, *variety* (heterogeneity), and *velocity* (real-time) particularly where it should be utilized for obtaining contextual data [89]. Although IoT is an adjusted infrastructure for big data acquisition, noisy raw sensory data is not prepared enough to be efficient on DL models. On the other side, better performance of DL is straightly affiliated to training data's quality and quantity, therefore, developing an apt dataset is crucial; as the data-centric AI and *software 2.0* are getting heard from the AI community, a paradigm is going to shift towards there [304, 290]. This means, that if we have equal models with the same hyperparameters but enhanced data, performance rises while with the existing democratized transformable models, thinking about it would be worth indeed [210]. Albeit, without considering this scenario or model-centric AI and physics-based simulations, high-quality data is decisively impactful for AI even in the big data era [210]. Moreover, selecting data for each application differs from another to profile [280] while some processes are within the same pipelines.

To reach acceptable data, after data acquisition (if needed), data should be preprocessed and improved [290, 232]. In a general systematic view of the data development lifecycle, requirements should be analyzed as the data must work for the goal it is following, afterward there would be design, implementation, testing, and maintenance [119]. The sides of information that are discerned remarkable from the mentioned phenomena in this context, which regards data collection for DL's avail, are data *acquisition*, *pre-processing*, and *multimodality*. Without an existing proper dataset, required data should be collected. The first step in data collection is data acquisition, which is composed of discovery, augmentation, and generation of data; afterward, if necessary, data labeling and refinement must be done [232]. Multimodality brings varied types of data into an integrated system to get the use out of assorted resources of information.

3.1.1 Acquisition

When the prepared pertinent dataset is not discovered yet due to unavailability or expensiveness, the data must be generated depending on the requirements and design specifications, or be augmented which is called data acquisition [119, 232]. In *data discovery*, the goal is finding or sharing datasets from outsources such as collaborative sharing, web, and data lakes [232]. Whether the data development is task exclusive, regular ways of data generation are *experiments*, *crowdsourcing*, and *synthetic* data generation, besides *big data generators* like IoT sensors [210, 232, 89]. *IoT data* usually from various sensor nodes or in networks would generate at massive volume and velocity as stream data. *Synthetic data* can be generated via generative AI models as GANs or simulations' output data, besides data augmentation techniques [210]. *Data augmentation* is reproducing new data from the existing data by making changes to samples, or generate established on them. In NLP, data augmentation methods can be viewed in three categories, paraphrasing, noising, and sampling [160]. On the other hand, data augmentation in CV especially with image data is significant in semantic segmentation, image classification, and object detection [297]. Augmentation methods in CV would be classified into manipulation, erasing, mixing, auto augments, feature augmentations, and deep generative models [297].

3.1.2 Pre-processing

If the data is gathered, it should be preprocessed to get enhanced for higher performance of DL models [232]. Preprocessing improvements can be done by *labeling*, *validation*, *cleaning*, *sanitization*, and *integration*. As the performance of DL models is better with labeled data, *data labeling* and *annotation* is a hot topic. But doing it manually even by crowdsourcing is expensively time-consuming, so, there are several automatic procedures. To mention some bolded ML-based solutions for lack of labeled data, there would be Semi-Supervised Learning, Active Learning, Weak Supervised Learning, and Self-Supervised Learning (SSL) [304, 290]. If the quality of labels is not in the standard domain because of their noise and incorrect labels, *re-labeling* is suggested [290, 232]. Another modification is data *compression* to reduce data size to utilize storage and other processing resources more efficiently [127]. After collecting and shaping data, except labeling, other steps should take place such as *validation* (e.g., visualization), *cleaning*, *sanitization*, and *integration* [290].

3.1.3 Multimodality

Modalities are just alike human perception mechanisms rooted in the *five senses* (observing, hearing, touching, smelling, and tasting), from *raw* to *abstract* forms of the same signal content [165]. Raw modalities are aligned with raw sensory data without abstraction. For example, besides learning RGBD frames of videos that have audio in raw form, the data might have abstract modalities within it such as multi object recognition and emotion recognition (e.g., facial expression, body gesture, and speech). As a result, *each context can be a modality* and capable of getting used in a multimodal learning fashion whether in raw or abstract forms, if the abstract ones are homogeneous [24]. Data integration by fusion of multiple datasets or modalities in a multimodal fashion like we humans, can make DL (*Artificial* Neural Networks) perceive more comprehensively [24]. That similar diversity in modalities would lead to more natural interactions of machines with humans in HCI while traditional fusion techniques do not have much in common with SOTA, there are many fusion strategies. Whether in CV, a fusion of modalities leads to a higher comprehension of the model [134].

Multimodality, as mentioned in the Human *sentiment* context, is practical in myriad applications of ML such as for affective computing and smart healthcare to use multimodalities to be more inclusive [8, 296, 281, 220]. A categorization of modalities is by their *dimensionality*, as in 1D signals and sequences, 2D images and audio, or 3D videos be considered [155]. Although, the dimensionality of heterogeneous Big Data by fusion of multiple modalities can be beyond that, as an RGBD video with its audio and caption containing semantically high-level contexts like detected objects, all labeled. A multimodal ML model must handle the heterogeneity, interactions, and connections of modalities in different types of challenges as representation, alignment, generation, reasoning, transference, and fusion [165, 24]. *Representation* is

features of each modality by their variety and interconnections to be learned, while *alignment* is detecting the relations between modalities. *Generation* challenge covers the compression of multimodal data, translating by mapping each modality to others, and generating other modalities out of the existing ones. *Reasoning* with this context is earning knowledge to infer and deduce via multimodalities. The *transference* challenge is about how to transfer the knowledge of each modality toward another's exploitation through representations and models.

3.2 Internet of Things

IoT is the point of view to have objects (*things*) interconnected to a global network infrastructure (*internet*) dynamically integrated or fully embedded as one entity while ubiquitous [137]. IoT can participate in data collection real-time such as data acquisition, almost all preprocessing steps, and providing abstract multimodality inside its architecture. The objective of this section is to get familiar with the constituents of IoT and how IoT-based system components conform and consist. To look at IoT from a finer perspective systematically, an *IoT architecture's structure* can be the way to analyze the elements' arrangements of the IoT-based system. One known architecture is the basic *five-layered* made of *perception* (physical), *network* (transportation), *processing* (middleware), *application*, and *business* layers [139, 247]. Advocating an architectural style is not the subject, but to indicate different aspects of IoT systematically to comprehend each layer's functions [10]. Although *IoT is an SoS* and its architectures diverge by applications and business orientations, the components are removable from the system [86, 105].

Input of the system, *perceived* sensory data get transferred by *networks* to be gathered for *process* into information or more analysis will be utilized according to the *application* of the whole IoT system. While, that should be along with market demanding needs and *business* strategies. The processed contexts might return as adjusting feedback or operating commands to actuators, ultimately. The closest layer to the users is the perception's physical objects incorporating sensors, actuators, and devices (things). Network transports are almost through wireless internet-connected networks; the pivotal concept of Wireless Sensor Networks (WSN) is as intercommunicated sensor nodes [9]. The sensor node in WSN is a device composed of processing and transportation layers (processor, storage, and transceiver modules) beside of perception layer (sensors) to convert analog sensory data to digital and send it wirelessly [136]. The WSNs or just the perception layer after connecting to the internet will bring up the data to the gateways and middleware [105]. Middleware is the mediator software between physical objects and applications, while the processing layer's other component is computer hardware [202]. The processing would differ based on how far the computation site is; above the device, at a closer distance would be fog (edge) computing with more resource limits and more real-time, or cloud computing deeper and extended [170]. Finally, after processing if it was considered, the made decision changes will recur via actuators in the environment (application interface).

3.2.1 Perception

The perception layer is made of sensors, actuators, and devices physically in the environment for *identification*, *gathering contexts*, and *physical operations* via sensors and actuators (devices) [105]. *Sensors* measure physical parameters' qualities of the environment quantitatively by responding to changes in physical stimuli. They cover a wide range of sensing contexts' changes and reactions. The sensor is categorized into two major types if they are *invasive* or *non-invasive* sensors from an AI perspective. The invasive sensors should be wearable or attached to the human body, while non-invasive ones not [261]. Non-invasive sensors can be installed in the environment or be used in the objects, whether they are vision-based or not [51]. Vision-based approaches receive more attention in practice, caused by the heterogeneity of sensors that sophisticates the calculations. However, with RGB or RGB-D (red, green, blue, depth) data, CV offers various options; moreover, many tasks are unfeasible with vision-based approaches (e.g., EEG for brain electrical activities) and in some cases using other methods is less costly. As we are living in the Big Data era, proper data is a necessity for the system, and by providing it, objectives seem much more reachable while accomplishing each ML task is highly dependent on the data as in ImageNet [64]. In addition, with the view of WSNs, sensors can be classified into *Micro-Electro-Mechanical Systems* (e.g., gyroscopes, accelerometers, magnetometers), *Complementary Metal-Oxide-Semiconductors* (e.g., humidity sensors, temperature sensors), and *Light-Emitting Diodes* (e.g., ambient light sensors, proximity sensors) [146].

An *actuator* is any object that can convert energy's form to produce a change for an operation, such as *LEDs*, *speakers*, *displays*, *motors*, *thermostats*, and *soft actuators* [247, 162]. The actuators like sensors have plenty of usage to perform any physical action in the environment whether it is a servo motor, a digital signage, or a shapeless arm in soft robotics. A *device* is hardware that drives software while has access or is integrated with sensors and/or actuators [105]. Devices also potentially have other capabilities such as computing processors [222] but each would be considered in the related layer of this architecture. A device might cover all five layers in itself as smartphones, but in the perception layer the data-perceiving physical object as a device seems to be the subject.

3.2.2 Network layer

The network, transport, or transmission layer connects parts and provides communication for them. It transmits sensory signals to the processing layer through network channels. Alongside the IoT paradigm, network protocols to link data, shifted toward *wireless connectivity* technologies. One view to analyze wireless communication methods is by their *range* [137]. The short-range wireless methods are *Radio Frequency Identification* (RFID) and *Near-Field Communication* (NFC) which after them *Bluetooth Low Energy* (BLE) would take place. *Wireless Local Area Networks* (WLAN) are medium-range based such as 802.11 IEEE *Wi-Fi* protocols, whether *ZigBee* covers short to medium ranges. Lastly, the widest wireless range belongs to *cellular* technologies like *Global System for Mobile communications* (GSM) supporting 3G, LTE, 4G, and 5G protocols. Though, in network tradeoffs, other evaluation factors are significant as speed (latency, jitter), bandwidth, and energy consumption.

Communication types in any IoT device have four main strategies: *Device to Device* (D2D), *Device to Gateway* (D2G), *Device to Cloud* (D2C), and *Device to Application* (D2A) [257]. That classification is separated from the technological infrastructure but represents the ways a smart IoT device can interact and exchange data. After all, if the device is not able to communicate independently, a *gateway* is the destination of its data to get transferred. The gateway translates and relays data for transmission to a data center and processing layer [247]. It can be more than an intermediary proxy role and automatically does *data filtering*, *preprocessing*, *processing*, *routing*, and *management* depending on its resources [34]. With this respect, WSN is the sensor nodes' configuration with routing nodes in the environment to collect sensory inputs and transfer them to gateways [184]. WSN is not a subset of the network layer but includes both perception and network layers capable of supporting the processing layer. A WSN can use any wireless communication method in two major phases, within itself sensor node's transceiver and outside connection via gateway [146].

3.2.3 Processing layer

The processing layer contains middleware and computing processors where aptly, further to *data accumulation* in storage, *preprocessing* to prepare information, *computation*, and *analysis*, the *decision* would be made there. After the sensed data is transferred via a device or gateway to the data center, a middleware would be to manage (e.g., abstraction) and join it to the application layer [247]. A *middleware* bridges devices and software interfaces by purveying *Application Programming Interfaces* (API) and handling heterogeneities such as data types and device protocols by integration [10, 81]. The middleware should not only provide network and syntactic interoperability but also semantically. Using semantic technologies in middleware (e.g., XML, OWL) allows data to be represented and related in a way the machine understands [307, 230].

In recent years, computation can be done in multilevel layers such as *device*, *edge*, *fog*, *cloud*, or a hybrid combination [302]. The IoT devices are the nearest and the most decentralized level for storage and processing while having the most constraints in resources. Whether a device like a smart gadget is limited in memory capacities and computing processors can do simpler tasks by itself privately. But, one conventional internet-based solution for limited capabilities is *cloud computing* infrastructure by serving massive ubiquitous computation power and storage remotely connected to the internet network. That made expensive powerful hardware (e.g., server farms, quantum computers) affordable by virtualization and sharing resources to run gigantic models and processing big data just by *pay-per-use* [194]. Setting infrastructural resources in a deep centralized position makes the responding time-consuming based on the network characteristics; however, situating cloud computing resources closer to the end devices, handles that issue [302]. *Fog computing* or *edge computing* does that by putting resources horizontally with smaller scale accessible for scenarios where IoT devices cannot operate and cloud computing has higher latency than the speed standard. Where fog or edge computing are in between, edge computing seems more edge at networks layer. All in all, it looks like a funnel where cloud computing has the most processing power and is far away from users despite the IoT devices [170].

3.2.4 Application layer

The application layer is where the services are to provide the interface for users containing modules to *monitor* and *control* [229]. As IoT can act as an internet infrastructure in CPSs, its applications cover a wide range of domains related to the service types and qualities they provide [18]. IoT integration with pervasive computing, as the Internet of Everything (IoE), would be the enabling technology of various applications. One of the most important applications of IoT and AmI is *healthcare*, as patients' physiological and medical status real-time gets monitored remotely via sensor networks connected to the internet [3]. *AAL* also can deliver healthcare services besides being used in smart homes, which is another application of IoT [58]. While autonomously, it lets the elderly live independently longer and people with physical or mental impairment be safer and more empowered. Smart cities are to construct all fundamental components of an urban area more efficiently using IoT as a basis, which includes almost all other applications of IoT in itself [30].

Smart Homes using IoT further to being assistive (e.g., smart kitchen) and present living enhancements (e.g., caring), would be effective in power and energy efficiency management, even for a whole building [246]. *Smart Buildings* as an infrastructural component of a smart city, vary in applications based on their uses as a smart office for administrative, a smart home for general, or a smart shopping mall with commercial retailing uses. Two mobility IoT tasks are smart transportation (e.g., traffic monitoring, smart vehicles, smart airport, smart port) and smart logistics [223]; however, that (Mobile IoT) can be used in different modes of ground transportation, maritime (tankers), or in aviation industries. Another impressive related aspect is the *Industrial Internet of Things* (IIoT) which is applied in almost every aspect of *Industry 4.0* in products' manufacturing to distribution processes [3, 91, 40, 184]. Some important industries with bolded IIoT role in them are *energy*, *grids*, *manufacturing*, *agriculture*, and *aquaculture*. Environmental monitoring by *remote sensing* is also a major application of IoT, whether in the environment as surveillance, or ecosystem monitoring for atmospheric and biological purposes [269].

3.2.5 Business layer

The business layer takes the managerial part to handle the system on its *business model* and *market orientations*, using monitoring analytics of the application layer's data to adopt strategies [139, 30]. The business layer on the top of all layers acts as an administrator to judge and control the system's performance through indicators to improve Business Intelligence or the business model [10, 184]. As the engineering is a tradeoff between functional and financial criteria; which impacts technological tools' selection based on their benefits and disadvantages and funded capital to decide the most profitable application's services option. This aspect is not in the scope of this survey but is much important.

3.3 Deep Learning

Whether IoT plays the enabling infrastructure role for AmI, AI pushes autonomy to intelligence, while DL methods do the recognizing and understanding of highly complicated patterns apparently better than any rival. *CV* and *NLP* are two majorly important applications of ML that have various uses in modern AmI, which are essential in some cases. AI can be used in all IoT architecture layers for optimization and/or automation purposes further to privacy, security, and resource allocation uses as its effectiveness on Quality of Services (QoS) [143]. Computing frameworks would use different ML methods by their characteristics, as edge computing is more limited than cloud computing. While, DL has numerous beneficiaries for *IoT data* like time-series analysis, noise toleration, real-time streaming data analysis, and handling heterogeneity in large-scale data [239, 182]. On the other hand, as DL algorithms' performance improves with training on larger data, IoT is a Big Data provider from multiple sensor networks and pervasive devices producing nonstop real-time data streams [40].

Computing systems need embedded robust software to program, but for processing and decision making at complex problems like chess game or multi object recognition, traditional functional programming alone does not work. Although ML was a huge step forward to do those tasks by learning the data algorithmically, with no programming on issues that cannot be solved even via massive coding. ML is the main subset of AI to learn the patterns and correlations in data using mathematics and statistics for its optimization algorithms. *Classic ML algorithms* like Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree, Random Forrest, Logistic Regression, and K-means somehow were competing with each other in SOTA ML until *deep* Artificial Neural Networks (ANN) outperformed prior ML methods on large datasets and proper computing power (e.g., GPU). However, some ML methods are getting used yet as SVM for its performance, Decision Tree by its interpretability and traceability, and they require less consumption in energy, amount of data, and processors than DL methods. Multi-layered deep NNs (DL) using optimization functions (e.g., Gradient Descent, Adam, Genetic Algorithm) demonstrated that can learn complex patterns and unlike classic ML are needless to hand-crafted feature extraction by doing it automatically. In addition, DL has more *robustness*, *generalization*, *interoperability*, and *scalability*, which regarding these reasons, the focus in AI is attracted on DL hitherto [155, 12].

During writing this paper (2024) AI in practice is *Artificial Narrow Intelligence* (ANI) and there is no *Artificial General Intelligence* (AGI). That means all today's ML-based systems and models are capable of being specific on their objective function in limited subjects. They cannot learn every knowledge in a general way like an omniscient, even using new DL paradigms like In-Context Learning (ICL). Whether the ANI we have now (e.g., MLLMs) is likely the closest one to a future AGI, still there is not a clear description of AGI. But, even with today's AI at learning patterns on huge amounts of data and giant models, which can achieve multi tasks autonomously, omnipresent in real-time, scalable for new downstream tasks, with higher accuracy and precision than a human in some simpler tasks, and processing millions of rows of data in a fraction of time, are impossible for any person to do. If look at the ML just as a context reasoning method, there can be others as probabilistic logic-based, case-based, rule-based, and ontology-based [307], but are not in the scope of this study. It should be noted that there is absolutely no free lunch in AI to have a master key, and each technique must be analyzed in its context. In this section, DL, to be systematically analyzed as a conceptual and technical brief overview, is viewed by its two major aspects of (1) *learning methods* and (2) *model architectures*.

3.3.1 Learning methods

Any learning type has mechanisms to confront issues from their requirements as many learning methods try to solve data-related problems quantitatively (e.g., samples) and qualitatively (e.g., unlabeled data) to be more efficient. Some learning procedures fit better in specified circumstances as Reinforcement Learning in interactive environments or controlling multi agents autonomously, concerning overfitting. As a primary goal of most learning methods is to modify pattern recognition and feature extraction automatically, some also follow more generalization and model's knowledge comprehension in more varied areas (e.g., Transfer Learning). To do so, in this section, there are impactful *conceptual learning mechanisms* that formed SOTA AI to glimpse. Any of these kinds of learning would be engineered to do a task, whether it is detection, classification, prediction, generation, or others which might be a mix of tasks. Although a DL model can include many learning methods and paradigms integrated into one architecture, while they are not completely separated from each other.

ML mainly had been differentiated from the *labeled* or *unlabeled* training data, which are the key factors to divide between the ML algorithms. Learning on labeled data is known as *supervised learning*, which learns mainly based on classification or regression techniques, while *unsupervised learning* uses unlabeled data as clustering, generative modeling, dimensionality reduction, and anomaly detection [36]. Models' performance using labeled data usually is expected to be better than unlabeled, but as annotation and labeling costs are high, clean useful labeled data's accessibility is highly limited. Although, building models on unlabeled data is a hot topic and there are methods with sufficient efficiency as Self-Supervised Learning.

In classic ML data representations or features within the data, as a core element in ML, would be selected aided by domain expert's knowledge or other feature engineering methods manually; but **representation learning** in DL, with much less affiliation to feature engineering of labor, learns the underlying features autonomously [32]. Representation learning methods try to extract characteristics of data by finding patterns, as in manifolds or coherences, through the NNs weights computation. **Semi-supervised learning** uses *both* labeled and unlabeled data to build a more efficient learning framework relying on unlabeled data, where labeled data is rare [271]. For instance, wrapper methods are popular in semi-supervised learning, which is a supervised learning model on labeled data, predicting the labels of unlabeled data, within the *pseudo-labeling* process for further uses [271].

Reinforcement Learning (RL) framework uses one or more *agents to explore* the environment and take actions sequentially (exploration, exploitation) made on *Markov Decision Process* (MDP) optimization to learn each action state's *reward* function and maximize total reward value. The RL agent pursues a specific target to attain by exploiting the most rewarding situations according to its interactive intuitions based on its *policy*'s setting [262]. Actions' rewards depend on the dynamic contextual states related to environment and agent position; that can

be optimized at prior feedback on posterior decisions even in a delayed rewarding for long term at value function, besides immediate actions' rewards [262]. That causes the agent to choose actions based on value computations but not only the rewards as the policy determines, whether if the RL system is *model-based* to discover the environment's behavior, future prediction would be added. Models imitate the environment for planning the agent's activities, searching for the best policy via simulation, despite *model-free* approaches which learn by trial-and-error. That makes RL agents compelled to fail to learn (exploration-exploitation dilemma), which leads to RL's data inefficiency.

Deep Reinforcement Learning (DRL) is RL combined with DL methods, to benefit traditional RL in dynamic interactive sequences of decisions, and learning complex representations and optimization of DL in higher dimensional data of complex environments and complex agents' control [16, 248]. DL in DRL can act as a function approximator for any RL element, whether is a model-based or model-free value function, policy, reward function, state transition function, planning, or exploration [164, 186]. DRL makes autonomous agents take consecutive strategies as per environmental situation, optimized by DL methods cumulatively [16]; which actually started by coping with Atari games in high-level and boomed with winning Go game's champion [197].

Transfer learning (TL) is learning a new task pursuing the related prior earned knowledge of the model by *generalizing* it, like a person who could learn saxophone easier if knew how to play the flute, compared to a person without any musical experience [315]. The TL does the *adaptation* and *generalization* of representations from the source domain, towards connecting it to other homogeneous or heterogeneous target domains, without building a new model from scratch [315, 288]. That makes higher performance when test data distribution distinctly differs from training data, on either conditional or marginal, since domain adaptation corrects differences of both by shifting source and target domains closer together [288]. In DL architectures it is feasible to cut the latter layer(s) off, specifically the last output layer, then compound previous frozen layers' trained parameters weights with new data by replacing it with the NNs' last layer(s) and retrain the added layers.

The procedure of restating an already trained model into a new transferred usage is the *finetuning* and the reuse of a trained model for other tasks is *pretraining*. Pretraining as a side of TL, is to train a model usually as a parameter or initializer for future finetuning by the rise of data and computing power consumption [155]. The *Pre-Trained Model* (PTM) on proper data would do the generalization for the newly transferred model with more data constraints, further to its less training power consumption. A type of finetuning to tune a model by learning it via a dataset of related instructed prompt queries is instruction-tuning, which is relevant in LLMs [312, 175]. If finetuning's aim is to filter the output of the model to not generate some sort of biased, unhelpful, harmful, or hallucinated responses, *alignment-tuning* would be made by methods such as Reinforcement Learning with Human Feedback (RLHF) with human alignment [39, 312, 209].

Multi-Task Learning's (MTL) goal is to achieve multiple related ML tasks *simultaneously* by sharing representations jointly among them in parallel for more generalization [310]. Tasks might have either equal importance or one main task and other auxiliaries, while a model can be a whole with multiple outputs or multiple models with one output. MLT is used mostly when the data has multiple labels, there is a need to get numerous outputs, many corresponding tasks have few amounts of data, or one dataset is sufficient for manifold tasks (data efficiency). For example, training a dataset containing images with multi-label of objects inside them for object detection task should do the recognition at the same time.

All these learning methods are learning algorithms for doing a task (inner learning), but **meta-learning** methods learn how to learn (*learning to learn*) with rectification from a higher stance (outer learning) [114]. Meta-learning's target is to have a better learning way (e.g., transferability, generalization) with more data efficiency on fewer training instances [114]. Representation generalization in meta-learning is directed to learn the learning types and algorithms' metadata as the input primarily [273]. That metadata covers all data about *training* and *algorithm* features, such as hyperparameters, training time, model architecture, model parameters' weights, and evaluation metrics' results [273]. Meta-learning also follows the NNs' reformation in deep meta-learning, based on their metrics, models, or optimizers respectively, further to more conventional algorithm type choosing and hyperparameter optimization in an agile task-agnostic fashion [116].

Federated learning can do the training process without direct reach to further supplementary data for updating or personalization on the main pretrained model, decentralized and asynchronously on plenty of devices collaboratively [263]. Each device trains on its own data using its edge computing capacities locally and if it interacts with the main server, it sends the *updated model parameters* but data [135]. That was an innovation after the growth of data islands affected by stricter legislations on data acquisition caused by *privacy* and *security* issues of users' data accessibility and lower latency in *real-time* decision making [168]. **Self-Supervised Learning** (SSL) does the supervised learning itself only on unlabeled data by previously augmented labels for data, finding inherent similarities of samples to group them by their contrasts, or using both generative and contrastive techniques [176]. Many of these self-supervisions form on data augmentation by making changes in data intentionally and putting the original data as the anchor for posterior labeled transformed data's probability estimation [124]. That change can be a *masked* part of data (e.g., hiding a word in a sentence), adding *noise* (e.g., color changing), image *rotation*, or any other data augmentation or synthetic data generation technique. As SSL is needless to manual annotation, would be practical for pretrained representation learning on lots of unlabeled data.

There are some learning methods only seen in *LLMs* with a level of scale as **In-Context Learning** (ICL) and **reasoning**. ICL is a learning capability in LLMs in which by just feeding the model a *few related samples* as in a *prompt*, the model emergently shows extraordinarily *inference* of prompt's demonstrations and relatively generates outputs [39, 287, 312, 175]. That is needless to any finetuning and parameters' update of the LLM by only *analogy* of contextual demonstration of prompted input. LLMs have reasoning ability to some extent when *step-by-step* guide it by a rationale as multi-stage questions, whether by prompting, finetuning, or pretraining [115]. The majorly used reasoning method is *Chain of Thoughts* (CoT) [286] to make LLMs reason by asking it step-by-step in a sequence of prompts or a prompt made of a series of intermediate requests, to generate both the output and the *processes of reasoning* (rationale). Another method is engineering the rationales to elicit the LLM for reasoning tasks and finetune the LLM on data as of scientific (e.g., Question-Answer, logics, philosophy) and mathematical (e.g., arithmetic, codes) rationales, to modify reasoning power [115]. Another view is whether LLM's reasoning strategy is *single-path* as CoT or *multi-path*, which at each reasoning step multiple probable ways for the next reasoning step will be chosen [280]. A

salient example of a multi-path reasoning strategy is the *Tree of Thoughts* (ToT) [298], as does not consider one specific answer at each step for the further step of thought. ToT like tree branches in a bidimensional sequence of thoughts searches possibilities whether to lead to foresight planning. *Planning* with a determined initial purpose would be naïve by the dynamical perspective of the world, to tackle it further to planning by decomposing the task into sub-tasks, continuous (iterative) *feedback* whether from the environment, human, and the model (e.g., self-refinement), were approached recently [312, 280].

Multimodal learning is learning from multiple modalities into one ML model with fused mapped representations [24]. After representations' fusion and alignment, the multimodal model would extract knowledge by reasoning from relations, intermediates, and logical or causal inferences [165]. Multimodal representations are in two forms of which *joint representations* that amalgamation of multimodalities into one representation simultaneously, while *coordinated representations* are by learning multiple modalities apart from each other but not detached at the end [24]. Lastly, there are **Few-Shot Learning**, **One-Shot Learning**, and **Zero-Shot Learning** methods to have models that can learn with the *least data*. Few-shot learning would be done via some other learning methods, such as *TL*, *meta-learning*, *data augmentation*, and *multimodal learning* [256]. Zero-shot learning tries to learn new *unseen samples* just in time by adapting the semantic information that formerly had learned, such as meta-learning and ICL [221]. LLMs potentially have all the former learning as by finetuning or prompting whether to address, step-by-step reasoning ability as few-shot learning, ICL is caused by one-shot learning, and prompting is zero-shot learning [39, 221, 175].

3.3.2 Model Architectures

The advancements in NNs architectures were cohesively continuous in the last decade, as there were some tipping points. As well as new practical learning ideas in them, a few algorithms to some extent changed the paradigm while were introduced for one specific task, could impact various applications (e.g., Convolution, Transformer). DL techniques are generally categorized into *Discriminative*, *Generative*, and *Hybrid* network models [242, 248]. *Discriminative* models learn the probability of the output's affinity with the subjective data, based on their features and representations (e.g., the picture shows a cat with 80% likelihood). *Generative* models learn the probability of a statistical distribution in the data, which might be regarding or regardless of labels. Discriminative DL detects how probable an input belongs to a particular class, while generative DL generates outputs similar to their trained data. *Hybrid* methods are a combination of different methods to be adopted the competence of each for more development. A Hybrid DL model can be made of multi-Discriminative models or multi-Generative models, a compound of both Discriminative and Generative models, or other types of techniques with each other.

However, based on the anchor of our point of view (as an analyzer), this categorization will differ, as in this paper the focus is on SOTA, it appears that *Transformer architecture* [274] is the tipping point in the last decade of DL architecture progressions. But that does not change the importance of prior architectures or make them useless at all. Some NN architectures in the *history of DL* are more prominent to get AI in its today's eminence, which are: Restricted Boltzmann Machines, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Auto-Encoder, Deep Belief Networks, and Generative Adversarial Networks (GAN). Some of these conventional architectures are covered in this section to get to SOTA DL concisely. Here, traditional ones that are still getting vastly used are mentioned first, and then the SOTA architectures are addressed while in the most abstract way.

Convolutional Neural Networks (CNN), a deep feedforward NN that proved its efficiency in the history of CV on one of the most effective datasets in ML, ImageNet [64]. CNN aims to uncover patterns within related features (e.g., motifs in images) from each latter layer and connect them locally by convolution; the poolings compress features to *dimensionality reduction* towards a fully-connected layer for the reasoning [155, 97]. CNNs learn well on multi-dimensional data arrays including signals, sequences (entirely), and language in 1D, images and audio in 2D, and video in 3D [155]. Some successful convolution-based architectures are *AlexNet* [150], *VGGNet* [251], and *ResNet* [109], in which their citation numbers partly demonstrate how impactful CNNs could be.

Recurrent Neural Networks (RNN) are another effective DL architectures to learn *sequential patterns of arrays* via recurrent cells and remember historical changes along data as NLP, videos, or trajectories through backpropagation through time [155]. That had a major problem with long-term recalling which gradients in lengthy durations of learning (frequent time steps) would continuously get vanished until **Long Short-Term Memory** (LSTM) architecture [112, 113]. In the vanilla (first) LSTM [96], the most famous LSTM type and even mentioned as just the LSTM [272], each block uses three gates of *forget* (Gated Recurrent Units), *input*, and *output*, on the previous layer's *memory cell*. The forget gate calculates to choose unnecessary memories for deletion, the input gate controls which context to add in the cell and previous layer's output, and the output gate concludes the output of the LSTM block [96]. Further to that, *peepholes*' connections make the LSTM cells capable of learning timings better via linking them to gates, thus the LSTM accumulates long-term states to reduce gradient vanishing of consecutive time steps [155, 96, 272].

Generative Adversarial Networks (GAN) is a *game theory* based constructive competition between two NN-based machines in which one of them is a *generator* and the other is a *discriminator* in connected parallel [94]. Their interaction creates a control loop between machines to adjust the cost function; generator tries to generate outputs in a way that discriminator *cannot recognize* while discriminator guides by detecting fake generated ones from the original input data [93]. This confrontation adheres to adjustments to get closer to *Nash's equilibria*, which are local saddle points of both machines' cost functions [101]. Consequently, the generated outputs will be realistic enough to be unrecognizable for any discriminator (maybe humans), whether are fake or not, and capable of learning styles of different modalities in terms of domain adaptation as *synthesis* or *manipulation* [94, 101].

After *seven years* of releasing the first **Transformer** model architecture (2017-2024) [274], it can be stated that Transformers are effectively prominent architectures in AI like CNNs and RNNs as were the basis of myriad further innovations and concepts. The SOTA AI are Transformer-based models since Transformers are called *'transformer'*, their flexible and transposable architectures have worked on almost all *data modalities* (e.g., text, image, video, audio, multimodality) and *AI applications* (NLP, CV, graph learning) [121]. The vanilla

Transformer is a *sequence-to-sequence* model predicated on the *Attention mechanism*, made of an encoder and a decoder; which each involves two layers of multi-head self-attention modules and position-wise feedforward fully connected networks where there are *residual connections* around each [274]. The vanilla Transformer architecture is hinged completely on the *self-attention mechanism* to learn contextual interconnections of sequences, entirely from close converged relationships to track *long-range sequential effects* [171, 121]. That differs from conventional recurrence methods as Transformers try to compute the relations of sequences with each other in parallelized encoder and decoder blocks, while recurrence is just processing the series of sequences recursively.

Whether the vanilla Transformer is not efficient on *'not' large-scale* data, has high complexity in computing very long strings of sequences, and is not adaptable to any downstream task; while further Transformer-based architectures are adjusted in each direction [171]. Beside of these adjustments, Transformers pushed the *TL*'s pretraining paradigm in a new way, as Transformers are in a *neutral prejudgment position* structurally despite CNNs and RNNs [171, 121]. That lack of inductive bias showed to be more appropriate for being in *mainstream Pre-Trained Models* (PTM) on large data, as is the overfitting cause for training on small datasets [314]. Transformers architecturally can be classified by their models' usage namely, *decoder models* (e.g., BERT, LLaMA), *encoder models* (e.g., GPTs), and *encode-decoder models* (e.g., BART, T5) [171].

Foundation models as *Large PTMs* with *huge parameters* trained on *large amounts of data*, provide knowledge generalization and adaptation and make it possible to get higher performance on downstream tasks with limited domain data accessibility [314]. The use of finetuning and prompting on restricted data (few-shot or zero-shot) of downstream tasks has considerable advantages [314, 107]. *Adaptations* by finetuning and prompting on PTMs as mainstream make building a new model from scratch unrequired for other downstream tasks as generalization increases. First, PTM is trained on large-scale data mainly by SSL and that generalization comes from varied big data while the attention mechanism of Transformers learns contextual relations within the dataset (e.g., word embeddings) [121]. Moreover, that makes a large amount of labeled domain-specified data inessential for downstream tasks, which are rare and costly while PTMs can be trained on any data modality and different AI applications such as CV, NLP, and graph learning [314]. Lastly, if the *scale* of the PTM increases enough (e.g., model parameters) as *LLMs*, not only raises downstream tasks' performance and data efficiency [39] but can also cause *emergent abilities* such as ICL and reasoning [287].

Although there are plenty of LLMs currently, here **Generative Pretrained Transformer** (GPT) series as one of the most salient models are discussed which had made LLM-oriented properties (ICL, reasoning) for the first time (GPT-3). *GPT-1* model architecture's core is *decoder-only* Transformer-based for text's long-term dependencies, in a way to have it in *unsupervised SSL PTM*, and *supervised finetuning* toward to be task-agnostic [107, 223, 312]. *GPT-2* [224] is similar to GPT-1 architecturally with adjustments in layer normalization and residual network, besides scaled in model parameters (1.5B) [39]. **GPT-3** as a game changer, the largest scaled PTM of its time at model parameters (175B), revealed emergent abilities such as ICL, reasoning, and following instructions [287]. Compared to the prior architecture, GPT-3's Transformer model used *dense and sparse attention* along with *gradient noise guidance*; while added to the massive model parameters, toward high performance in few-shot learning, and zero-shot learning [287, 312, 201]. Also, GPT-3 was finetuned on *Reinforcement Learning from Human Feedback* (RLHF) [209] for aligning the model to not generate harmful content as *instruction tuning*. Consequently, the *domain adaptation* steps went further to pretraining and finetuning with the addition of instruction tuning and prompting [201, 175].

GPTs got *multimodal* (MLLM) in **GPT-4**, constructed on *images* in addition to *texts* in input training modalities, while it only generates text output with higher performance and reliability than the priors [4]. Afterwards, *GPT-4 Turbo* is with higher knowledge domain and longer context support (tokens) with improvements in multimodality by adding *audio* (speech) modality to image and text [206]. More recently, *GPT-4o* amalgamated more modalities in an end-to-end fashion, which as input has *text*, *audio*, *image*, and *video*, and as output, generates all except videos on one model [317]. While LLMs and MLLMs do reasoning, *planning* is the next step; which, after *decomposition* of the *whole issue into parts* to understand and plan towards future decisions to be made, as in commonsense knowledge [283]. **Agents** can do planning using *feedback* whether from the *environment* (physical or virtual), *humans* (prompts), or their *model* (e.g., GANs, actor-critic, self-refinement) [280]. The *actor planner agents* are not just limited to LLM agents as language-based task planners, but as vision-language multimodality can be on MLLMs [181]. These *robots* can learn via interactions to plan for their reactions based on received (sensed) feedback [280]. Those interactions might have been directed and taken action *control policy* as RL to let the model explore upon it, which the model would be a PTM as MLLMs [68, 181].

A very impactful related example of an *interactive environment with humans* is **ChatGPT**, which finetuned on the GPT model series provides a conversational space of prompts and generations aligned by RLHF based on its proper data of both dialog and alignment samples [314, 312]. As the nature of TL works on three sides of *pretraining*, *finetuning*, and *prompting*, at each section the data is much more effective which in an interactive space if more feedback would be given to the model as data for finetuning, the model gets richer. *Environmental feedback* based on the environmental attributes can be different whether if it is a simulation, physics-based simulation, or dynamic physics-based game environment with the collaboration of multi-agents of humans or programmed features or an AI-based agent [68, 181, 108, 63, 115]. That environment might be in the real world as a robotic arm after being pretrained on simulation, or a digital twin of its physical twin's real system [265].

Visual modality is necessary for environmental feedback while in conversational agents as *GPT-4v* enabled ChatGPT [206], multimodality is getting used ubiquitously these days. However, the **Large Vision Models** (LVM) as Segment Anything Model [144] are like LLMs, while the visual modality is substituted with lingual (text) modality [278]. These CV mainstream models can handle many downstream tasks as the *Segment Anything Model* generalized to do tasks like image analysis, multi object detection (plus segmentation), multi object tracking, and image editing [144, 278]. Properly engineered prompting as an interaction in LVMs is as effective as in LLMs (e.g., CoT, ToT), while in

LVMs the prompt is also an image or a video to lead the model in the aimed direction [278], and with text prompts the model might not exactly get the idea well as for editing [177].

Moreover, some LVMs are made to *surrogate reality* in video physically trying to be at the quality of world simulation, which *Sora* as a text-to-video generator follows [38]. They also can simulate real-world looking renders or imagine unreal scenes close to the physical reality by image synthesis, while cannot always afford reasonable physics-based knock-on effects [177]. In addition to using the *diffusion transformer* model in Sora's architecture, diffusion models are being used in many CV tasks, especially in visual generations [177]. **Diffusion models** similar to LLMs that mainly predict the unseen masked words *add augmented noise* into the data (e.g., image) randomly (e.g., by Gaussian probability distribution) [61]. Afterward, a DL model learns to *denoise* the data to not only detect the original input data from noisy, but it also will generate based on its prediction. As this method is slower than GANs in image generation, it performs with higher *quality* [61]. Although in image generation LVMs are unrecognizable to be fake, they are not still in a quality to generate video physically inclusively which deduces it is an absolutely complex job to do after that much other accomplishments. There are *plenty* of other successful model architectures *not addressed* in this section indeed, but a noteworthy point is here the purpose is to observe just a few architectures *conceptually*, while it is highly recommended to check referred resources. A noteworthy point is that some models are unclear about their architectures such as GPT-4 series.

4 SPACE CONTEXT

The *spatial characteristics* of how the environment is, the details of the *surroundings*' things, and the *physical laws* of space, are about the Space context. The place itself incorporates various information such as where it is by its *geography* (e.g., maps' data), how things are in it by their *positioning*, *physical status* of the environment (e.g., energy usage), what the *applications* of location are (e.g., commercial), the occurring *spatial events* of surroundings (e.g., weather), and the *transportation* conditions (e.g., public transportations). Whether analyzing any spatial attribute inclusively is challenging because of the generic philosophy of this topic, it is tried to consider the more related contexts of Space to AmI. Whether physics is the rules of nature, physical attributes of any space are vulnerable information in Space context. Getting aware of physical conditions as commonsense is what myriad animals have at their young ages while for high-level AI still is challenging. That physics-awareness path is another point of view mentioned in this context. Therefore, Space context is categorized into three sections at *scale*, *location*, and *physics*; among diverse schools of thought about observing the world as a "space".

4.1 Scope

The *scope*, *size*, *complexity*, and *immersion* are some parameters of the scale of an environment that are discussed in this part. Since the IEs would be assumed as *cyber-physical embedded systems*, they can be determined to be any spatial systems networked at multiple scales by computation powers distributed at any level of the system [102]. 3D immersive cyber-physical environments can include both fully *Virtual Realities* (VR) in cyber, and the physical environments (original). Integration of them also as *Augmented Realities* (AR) and *Mixed Reality* (MR) to merge VR into the physical world has common directions with AmI [275, 19]. All types of environments can be three-dimensional (3D), while VR and AR can be in 2D as well. In the 21st century, a *regular* physical environment by the pervasiveness of digital devices (e.g., smart IoTs) potentially is a cyber-physical space.

The *applications* (objectives) of smart cyber-physical spaces (IEs) are various, similar to their scales, which scale of each diverges in quantitative and qualitative factors. For example, *indoor* and *outdoor* is one way to discuss space scale, though an indoor environment might be larger and more complicated than an outdoor one (e.g., hospital vs pickup truck's cargo bed). To start from micro-scale to macro-scale, a place where devices are in it has the capacity to be called an IE. Even though every smart environment needs them as components, one might say any space containing smart devices to be applicable to AmI is potentially a smart environment. A *microscopic* instance might be a smart vehicle that seems to be beyond just a thing, whether if connected to the internet an IoT. Then at the *mesoscopic* level, smart homes, stores, offices, gyms, high streets, schools, and buildings by having a proper infrastructure will further to smart environment towards becoming an IE. *Macroscopic* large-scale ones as are SoS, are malls, hospitals, hotels, museums, universities, industries (e.g., factories, agriculture, grid), airports, ports, and cities.

Any environment, wherever large it is, has boundaries and this delimitation has to be understood, especially for interacting with *the ecosystem*. As everything is a part of the ecosystem, the *ambit* of the IE and *ecosystem* must be distinguished and informed (e.g., causes and effects); where any change in the ecosystem is connected to the people and the environment in *two-way*, such as weather conditions and global warming. In some environments, there is a full enclosure (e.g., hospital) and in some, there is not (e.g., city). Whereas a place like an airport, while has its boundaries, embodies both closed indoor space from the ecosystem (terminal) and in touch with the ecosystem (runway). Any of these places are a separate subject and should be analyzed and studied by its parameters as a wide range point of view, like in *Industry 4.0* [3, 91, 184, 152, 222, 40].

4.2 Location

The *spatial position of humans and things* and what is understandable from that metadata is about location. It mainly is considered as information about vertical and horizontal dimensions, while altitude might be necessary too. The main vision of positioning (i.e., localization) is whether it is *indoor* or *outdoor*. As the most used outdoor positioning method is the Global Navigation Satellite System (GNSS) signals as the Global Positioning System (*GPS*), it is not accurate in an indoor space and does not respond deep indoors [157, 17]. To solve this issue, *Indoor Positioning Systems* (IPS) such as Mobile Node-based Localization (MNL), Reference Node-based Localization (RNL), Inertial Measurement Units (IMU) (e.g., accelerometer, gyroscope, magnetometer), and Proximity-based detection, perform for navigation and

tracking purposes [80]. By the variation of requirements, different communication network methods such as *Wi-Fi*, *RFID*, and *BLE* would be used in IPSs whether omnipresent camera-based approaches work in both indoor and outdoor spaces in a variety of CCTV sizes [80, 211].

In some scenarios, knowing the location of *static objects* is helpful, as in AAL and HAR [53], assuming them in a certain location on a plan layout, besides positioning movable ones. In order to reach location-awareness, location-enabled IoTs are needed pervasively where technologies like Low-Power Wide-Area Network (LPWAN) and 5G are recently ubiquitous [163]. Monitoring location-enabled IoT devices carried by people express multiple contextual information in each context, such as spatial *trajectories* and *paths* (e.g., traffic condition, process management, inventory control). Where the environment itself is placed and how that is, are other qualities of the location. This perspective has the ability of vast inclusion and complexity; as how the *neighborhood* and *geographical* attributes of the environment are even *economically* and *culturally* (e.g., residential or commercial or industrial, economical states, urban or rural, near or far (to another location), weather attributes). Similar to scales' diversities, in this context, inclusiveness needs research on each subject specifically.

4.3 Physics

Physics rules are the *laws of nature* and perceiving them makes humans and animals *aware* of universal realities which can potentially be the exact same direction for machines. The world might physically be a unity of *matter*, *energy*, and *interactions* between the first two. Their conditions (e.g., gaseous substance, liquid substance, or solid matter) and interaction forms (e.g., potential, kinetic, thermal, electrical, chemical) are in a relationship with effective parameters (e.g., temperature, intensity, velocity, frequency). Since we are living in the material world, waves with diverse energy levels are almost everywhere in the cosmos, therefore, any environment has diversified waves through itself, either electromagnetic or mechanical (medium dependent). But as a matter of fact, we humans are *limited* to feeling all wave frequencies unarmed. Whether if it is Radio Frequency or x-rays and gamma rays in electromagnetic, or mechanical infrasonic and ultrasonic in Hertz (20-20000) or Decibel (30-130) out of our sensible domain. However, *our tools* (e.g., IoT, WSN) do sense. As we sense, at construction there are standards for insulation of thermal, moisture, sound, and high energy radiations. With the amazing advancements of AI, now it is feasible for some models to reason step-by-step, solve problems, and plan future actions. Though they are still not learning it the way biological living entities are and have challenges in some simple ones, such as *commonsense*, they can be modeled to be more aware of physics laws. Here, firstly an abstract vision of *sensing* and *monitoring* is overviewed, and then an outlooked side of *AI's physical perception* is covered.

4.3.1 Monitoring

Physical properties of a system should get sensed and measured for documentation, if unfeasible, must be estimated from other measured attributes [265]. *Measuring* application differs depending on its utilizations for the model, whether for its input, update model states (real-time), monitoring, or reasoning. Physical properties in *intuitive physics* are bifurcated into clear observable properties and latent properties [69], which *observable properties* examples include position, velocity, and geometrical shape. *Latent physical properties* require reasoning, such as density, mass, and friction. Furthermore, *monitoring* can be done for many physical subjects as *energy* and *pollution*. The monitoring capabilities of IoT and WSN are extensive as their main applications are for monitoring healthcare and industries [306, 183, 239]. Monitoring the physical environmental conditions by IoT and WSN can be done in all matter substances, whether in air, water, or soil [269]. Smart environments' monitoring has usages in quality control, quality monitoring, and forecasting majorly in air and water quality, which below a level of quality would be *pollution* [269, 152]. Moreover, an important issue is to manage *waste* within the lifecycle as it has an impact on any type of pollution and also energy, for all people by effects of the ecosystem [152]

With IoT-based systems real-time metering of *energy* consumption monitoring is feasibly affordable, even for individuals globally [117]. That can contribute to the decision-making process to raise optimization in effective energy management using ML where the smart grids are the result of this vision [133]. That incorporates the massive power generators and microgrids of the produced electricity by solar panels on the roof of houses towards consumption trends of households [306]. *Pollution* also as another example can be sensed and monitored using WSN and IoT in any three phases as air pollution, water pollution (abnormal pH or toxic), and soil pollution (degraded or salty) [269]. *Air pollution* is a serious problem for everyone to monitor [120], though water and soil pollution is more directly decisive for particular smart environments such as smart agriculture, industries, and smart cities [269]. Air condition of the environment is further than pollution purifying, temperature, lighting, oxygen concentration, and humidity are other parameters that WSN can frequently afford [152]. Beside of these types of pollution, *hazardous radiation* is another challenge, especially in spaces with more variety and ubiquity of wireless electromagnetic devices' exposure [1]. One implication of them is the high specific absorption rate of *electromagnetic fields* affecting humans biologically (e.g., tissues) by their heat whether jointly, as from Radio Frequencies [1].

4.3.2 Perception

Apart from what we cannot sense while machines can, they can do omnipresent management of what physical condition we need or like to have. Whether they have still problems with perception of physics (e.g., commonsense) spatiotemporally as in autonomous driving. To understand the system's physical state and behaviors, the first step is the *physics laws*. *Physical priors*' knowledge as physical laws and rules of the world in a frame of AI relationship would be viewed as *Physics-Informed ML* (PIML) [108]. PIML is categorized into *Partial Differential Equations* (PDE), *symmetry constraints*, and *intuitive physics* [326]. Physics priors' categories from another point of view within *CV* are (1) equations and constraints, (2) data fusion, (3) representations, (4) physical or statistical property, (5) physical variables, and (6) hybridization [25]. The attention here is more on *intuitive physics*, but high inductive biased physical priors as PDEs have plenty of usages in PIML as scientific discoveries [108]. In this section, physics priors' perception is covered in *reasoning* ability and the *modeling* methods for acquiring it are discussed.

4.3.2.1 Modeling

The AmI's *Physics-Awareness* requires linking the *physical state* of the environment to the *digital state* of the model. That connection can be done in three modeling methods if it is completely *physics-based*, *data-driven*, or a *hybrid* amalgamation of both. From less need for data to higher data requirements would be physics-based, hybrid, and data-driven methods [108]. The physics prior knowledge would be a physical awareness enabler in the three ways of modeling, while empirical data are useful in PIML too [108, 25]. Whether physics-awareness is not only dependent on *data* and *properties* parameters based on priors, it (e.g., PDEs) can be derived into the *architecture* as in Physics-Informed Neural Networks (PINN) [226]. If the physical modeling is for CPSs, it is categorized to have modeling based on state machines, rules, and agents [115].

Fully **physics-based** modeling is designed to build a model as in computer-aided engineering, a software based on physical prior like PDEs and geometries would be made and initialized for the model's utilization as in simulation [265]. In abstract exemplar, a physics-based model can be made out of geometrical measurements of objects, while in more high-level modeling, the solid body of an object by its strain can be analyzed. The other types of physics-based analysis and simulation also can be about thermal flow, fluid flow, kinematics, dynamics, and multiphysics [265]. **Data-driven** approaches such as DL by high-level representation learning, in CV for example, can learn relations in video data as multi object recognition and tracking in a way to capture physical priors, while it would not satisfyingly perceive physics. Other ones would be SSL on augmented rotated image data to learn which rotation is a natural image without any knowledge about the gravity of the earth, and data-driven models' capability of learning systems dynamics from sensory data [265]. **Hybrid** methods are nevertheless to be composed of selected good qualities of any modeling approaches. DL and physics hybridization can handle both physics' forward problems (e.g., weather forecasting) and inverse problems (e.g., PDE discovery) [108], as in models like PINNs [226]. For example, in the Human context, for analyzing Humans' behavior and sentiment, DL can learn the physical data or dataset labeled on physical attributes like body skeleton models [27]. While some physical constraints as speed limits can help, physics priors in many scenarios do not work in prediction as in pedestrians' trajectories, which are dependent on psychological state and biophysical constraints beside many other spatiotemporal contexts [134]. Adding biological constraints as prior physical knowledge of human anatomy as in body posture and gesture to the DL model is better learnable than pure data-driven [25].

As physics-based models of CPSs are robust and interpretable, they are not flexibly intelligent, as data-driven and hybrid methods are [225]. Physics-based models are better models of a physical system if they inclusively have all system dynamics measured, get updated, and modeled principally on all levels (micro to macro, inside to outside) of the system's behavior [225]. That is toughly difficult by the costs and the dynamically unpredictable nature of the world to make it much time consuming to be updated. Whether the changes would be noticed and perceived, its complexity puts interpretability and accuracy at risk [225]. On the other side, *model-based* methods (e.g., physics-based models) may need no recorded data for modeling a system except for calibration while data-driven ones (e.g., DL) are data-oriented [25]. If data-driven models have proper data of the task, they can accurately afford it without any other model if there is no need for high-level reasoning while physical perception requires it.

A salient subject to find both physical-digital relationships and a variety of hybridization is twinning. AI-based **digital twins** are the intersection points of real-world systems' data (e.g., from IoTs), physics-based models (e.g., multiphysics simulations), and AI, to not only analyze systems (e.g., counterfactual what-if simulations) but also the finding optimized decisions in each state space [19]. Digital twins follow the goal of being the inclusive twin of a real physical system, no matter which system it is. That covers the relationships too, whether the *physical to virtual* (digital) or *virtual to physical* enabling, to measure and model the physical system's objects and dynamics [265]. Digital twins cover many hybrid modeling approaches as in geometrical modeling (laser scanning, VR, AR), physics modeling (structural analysis, flow analysis, kinematics and dynamics, multiphysics), data-driven modeling (degradation, surrogate, dynamic system identification), and PIML [265].

Hybrid models can be incorporated in either data and architecture or both in cooperation to have both physics-based and data-driven models made [108, 25, 265, 115]. In *data hybridization*, one common approach is to have the output of a physics-based model such as a first-principle multiphysics simulation, as the input of a data-driven model like DL models to learn, because that synthetic data is less limited [108, 63, 115, 265]. It can also be stated that having physics-based simulations as data generators is easier than comprehending the exact system dynamics [115]. The hybrid model can be designed by using the data-driven models to learn datasets that are composed of physical knowledge about a specific context, as physics QA about advanced complex school tests [201, 63]. *Multimodality* and data fusion of multiple domains' heterogeneous contexts as historical, preprocessed (e.g., labeled), and empirical, is another way to cover more physically inclusive data. If one aspect of those is physics-oriented, would lead to being more physically aware [25]. Architectural adjustments to inject physics priors (PDEs) into a data-driven model like DNNs' (DL) loss function can work on both physics' forward and inverse problems, as in PINNs [226, 108]. When these unidimensional physics-awareness enabling methods (data, architecture) get combined, several hybrid methods will be made. One near example is having physical data in PINNs [108], while innovations are more.

Hybridization can be in different levels of architecture such as *model*, *loss function* (objective), *regularization function*, and *optimizer* [108, 134, 225, 25]. Intuitive physics can be utilized in data-driven models' architectures as constraints in kinetic momentums or energy conservation [108]. The recent hybrid techniques of data and architectural inclusiveness can be viewed within learning methods of TL, SSL, prompting, and their amalgamation [265, 108, 25, 292, 175]. TL is mostly based on pretraining on big general data (e.g., in LLMs) or large synthetic datasets (e.g., from physics-based simulations or digital twins), then finetune on real-world physical data (e.g., empirical) with higher access limitation (e.g., in PIML) [108]. The synthetic data might be from the agent's exploration and exploitation with trial and error in the digital environment (simulation or digital twin) through methods like DRL [265]. In SSL large synthetic data would be learned by predicting each time-step, whether if TL gets added, the pretrained SSL model gets finetuned on another limited data. Prompting majorly

specified to foundation models can be done by physical datasets as physics QA and exams or any physics prior knowledge, prompted on large pretrained model as LLMs. Another hybrid way is to have an AI *agent* (e.g., RL) to explore and exploit via action, feedback, and reaction whether in *simulation*, *digital twin*, or the *real physical twin* (e.g., controlling, path planning) [63, 265]. In AI-based digital twins, the interacting feedback might come from the real physical system (physical twin) to learn and react as in controlling and path planning tasks [265].

4.3.2.2 Reasoning

Intuitive physics is about the *commonsense* that we humans and many animals have to reason how our surrounding mechanisms work from understanding systems' behavior to reacting to them [108]. Reasoning physics is a strategy to take physics priors into AI whether as *post-processing* or *inference* [69, 108, 25]. Reasoning with commonsense is about learning physical properties as underlying concepts toward manipulating them [63]. The latent physical properties (e.g., density) have to be inferred after observation [69] enabling the model to physically interpret the scene [134], however, the physical reasoning is not that simple. Physical reasoning tasks in intuitive physics whether in detection, prediction, causality, or inferring, can be categorized into Physical *Properties*, Physical *Interaction Outcome*, Physical *Trajectories*, Physical *Dynamics*, *Visual State*, *Violation of Expectation*, and *counterfactual reasoning* [69]. Reasoning also can be done by hybrid methods, whether by a dataset or benchmark made for physical reasoning, whether in visual (2D or 3D) or text to be learned by *data-driven* models [63]. Physical reasoning can also be made by large models such as *LVMs*, *LLMs*, and *MLLMs* as mentioned in the DL System context.

5 TIME CONTEXT

The philosophy of time and chronological analysis as a physical fact should be viewed by human's perception of time beside of the way machines can read and process it. As vast as modern *philosophy* and *physics* (also in thermodynamics), there is time all across the universe to interpret it through endless standpoints. However, to analyze how IEs can understand time and considering its related issues as a natural interaction, checking it with AmI's systematic settings seems more useful. Though, getting inspiration from humans' and animals' *time perception*, works from philosophy, psychology, neuroscience, cognitive science, and interdisciplinary approaches (e.g., neurophysiology, neuropsychology, and cognitive psychology), would help.

Time is interpretable in terms of the human perception of time via the peripheral signals (stimuli/cues) that our somatic and cognitive (mental) channels sense from space-time and process in neural (nervous) systems. The process of '*becoming'* is defendable in that category [291]. A taxonomy of temporal experiences would be a guiding signal for time perception as the *'elementary time experiences'* taxonomy by Ernst Pöppel [219]. That categorizes time experiences as *subjective* phenomena such as duration estimation, subjective present, temporal organization, temporal processing units (simultaneity, successiveness, and temporal order), and continuum of time [219, 218]. *Duration estimation* is how much time we think had taken for an event in the past subjectively. *Subjective present* is about the short episodic intervals of the present (nearly 3 seconds) which is specious. *Temporal organization* is how we plan and anticipate the possible future actions. *Temporal processing* units are the successive sequences of events to represent the non-simultaneity and the latency of our cognitive procedures to perceive now (present) is near real-time. *Temporal continuity* is about the subjectivity of past to future as the passage of time.

Our elementary experiences apparently have common with the chronological concept that AmI can have. To exemplify, pure real-time even with 5G (even 6G) is *impossible* since it is *near* real-time. As mentioned in the *DL architectures* section, a major objective of the model can just be about temporal processing units (e.g., RNNs). On the other side, assessing neuropsychological research on time perception towards computational models and robotics [29] might be a gleaming peephole in the direction of comprehending how AI may perceive time. Another way to address Time context is *patterns* of time, which are made upon the repeating spatiotemporal occurrences that shaped the lives of all living creatures of the earth.

The world is always changing and the human brain has been trying to extract patterns from Pareidolia to spatiotemporal periodic events, even if they are not iterative as cyclic constituents. Some are by our chronological perception before (past), now (present), and then (future), while some are realities of our life as the daily *routine* of earth rotation at circadian rhythm. These recurring phenomena might revert to one of the previous states or not, which on that basis this part is bifurcated. A pattern might be just algorithmic and not conclude in a complete cycle, but still recur as the succession of future into present and be passed finally, while a past space-time will never be repeated twice. Various phenomena repeat as stationary travelers as the earth's gravity causes recurrence alternate of days (morning, noon, evening, night), moon's rotation, seasons, and years. Here *cyclic* aspect and *flowing* style of temporal patterns are separated as being *one-way* or *two-way* modality of it.

5.1 Flow

Watching time as a flow of spatial spreading might turn our minds towards the *process of continuous events* in order, to change the states of objects. That would lead to observing the world as *dynamic processes* of *'becoming'* transformers, except looking at it as tremendous stable permanent objects. Technologies like AI are not outliers too (with that perspective), by their inclusion with *becoming* objects through time [59]; as the escalation of related aspects such as big data, IoTs, algorithms (e.g., NN architectures), learning methods, and processing capabilities (e.g., hardware and computing layers) are following Moore's law up.

As time passage continues from the past to the future direction, the present is untouchable in pure **real-time,** whether either humans, fauna, or machines. Caused by our limitations, as in sensing, processing, and acting, humans have *latencies* of about several milliseconds to catch the present tense, namely subjective present [219]. That specious illusion of real-time attendance in the present will be magnified in

machines if long queues of latencies are stacked up as lag and jitter. Their slowness of speed would not only be rooted in the transference or hardware specifications but also might be an effect of other causes, like bandwidth or architectural inefficiency. *Delay* issues further to increasing the speed of network or architectural modifications can be handled by the model itself as model compression or distributed local processing like federated learning [313, 168]. Besides, handling *stream* data in near real-time has been becoming a timeliness norm, whether it is listening to a piece of music or monitoring the generated big event log data of a large factory from all processes of the supply chain toward sales. All in all, although we are entities with near real-time reflection, we seem thirsty for *omnipresent real-time* immersive environments with *telepresence* via technologies such as 5G, quantum computing, now-casting, and so on [84].

If *data* has a sequential structure in which each random data point (stochastic process) is related to its prior and posterior records of the dataset, while these orders are indexed on a timestamp, it is called a **time series**. AI's main channel to interact with time is by learning time-series data. As in the System context remarked, an impactful part of ML is focused on learning these types of patterns. Spatial time-series is *spatiotemporal* data, which is a salient joint between Space and Time contexts. A substance of these datatypes is *trajectories* to track the route of an entity (e.g., an agent), timestamped. If the data continuously is generated and sent (for processing) in streaming data and the ordered timestamped, is a time-series data stream [11]. That unbounded data streams in high velocity and scale, toward having minimum latency, unlike *batch* processing needs *online* stream processing.

A vision in the flow of time is analysis's timing perspective, whether *retrospective* or *prospective*; as AI with the power to accomplish the majority of tasks can be interpreted within this angle. The utilization of stream data is usually at real-time processing for tasks that require *rapid reactions* as forecasting a probable dangerous event in alerting systems [11]. The use of DL (and its hybridization with other techniques) in forecasting by learning both weak signals and complex patterns of time series seems to get better performance than other methods [33]. As deep forecasting needs a high-volume sequential data in the time-series, the quantity of *observations* counts in the amount of data [33]. *Anomaly* detection or anomaly prediction are retrospective and prospective tasks respectively, in time-series analysis which identify irregular fluctuations if outlier data points or sequences do not look like the others or fit in their context [11]. Those are more indispensable tasks when there is real-time data streaming, caused by its data poisoning sensitivity [11]. Whether a recognized anomalous signal might be a prediction to prevent a possible catastrophic problem in the future as a consequence of security attacks prospectively [11].

5.2 Cycle

Many phenomena have *periodicity* in cyclic trends with complete oscillations, whether *deterministic* as circular clocks' graduations, or *stochastic* as weather. This type of recurrence includes all information about routines as timepieces, calendars, or any timestamped historical information. AI not only learns deterministic patterns, but can also do pattern mining from stochastic events for time-series analysis tasks such as forecasting (e.g., in action, event, weather), anomaly detection, classification, monitoring, and reasoning. Periodicity in time-series data can be derived by *frequency* transformations to turn the time domain into a frequency domain [300]. That makes complex patterns more perceivable for ML using techniques like convolution theorem and decomposing series to frequency components manifesting representations as seasonality and global dependencies [300].

5.2.1 Timescale

While duration is a one-way factor, its *measuring units* (timing) are repetitive in different timescales which from *micro scale* to *macro* are as microsecond, millisecond, second, minute, hour, day, week, month, year, and decade [29]. The impact of having a range of multiscale periods is noticeable in Context-Awareness (time-awareness) [65, 50] which can be indexed within or further to the above timescales. Those to mention would be *time of the day* (working hours, off-hours, morning, noon, afternoon, evening, night, midnight), *days of week* (working, weekdays), *holidays* (e.g., weekends or occasions), *event time* (e.g., calendar events, visiting an event, attendance duration in an event, event's unveiling time).

5.2.2 Event

Time and event are *separated* phenomena, but as our ancestors distinguished time by *trends of changes*, the relationship between time and event is *bidirectional* as we perceive time by events and event perception comes with its temporal facets [154]. Events as dynamic positional changes are directly interrelated with other contexts too as they happen and influence Systems like the environment (Space) and people (Human) [311]. Beyond timestamping, learning, and reasoning *deterministic* recurrence of events, there is much attention on the recognition and prediction of *stochastic* events. It might be correct that each detection or forecasting task in ML tries to extract patterns out of what betides among the training data representations. Events besides their occurrence probabilities might be sequential whether a procedural time-series trajectory event data (e.g., event logs) has both attributes. An *event* in event sequence data is the most detailed scale, while *sequences of events* represent more inclusive semantics about the entity's journey with temporal and chronological features (e.g., order, successiveness, and duration) [104]. Event sequence analysis is proper beyond the mentioned tasks (detection and prediction) to be used for anomaly detection, recommendation, and causality analysis (e.g., reasoning) [104].

To determine chronological changes, *Change Point Detection* (CPD) works on how to analyze temporal behavior variations inside signals of data with ML techniques [14]. Online CPD focuses only on real-time change recognition as anomaly detection in natural disaster recognition as soon as possible, whether these tasks can be done already by event prediction. Event is a very common term corresponding to its time association, embraces various topics which event prediction covers healthcare, media, transportation, politics, economics, natural disasters, crime, entertainment, and business applications [311].

5.2.3 Memory

Without memorizing, time is not comprehensible as a major subject in time-series analysis is adjusting the *long-term dependencies* learning, to acquire minimum gradient vanishing in DL. *Memory decay* problem and long-range dependencies seem to be modified by using modern DL architectures as *Transformers* but still are not solved enough [300, 289]. Managing plenty of historical profiles about multimodal contexts is another important problem that a variety of methods work to modify, like SSL, which tries to take steps for independence from labeled data even in time-series data [308]. Additionally, memory in real-time streams is not accessible via conventional processing fashions for *working memory* [11] and they should merge *with long-term memory* and *attention* to perceive time better [29].

6 DISCUSSION

The *context categorization taxonomy* indicated how vast AmI is and which capacities are to be expanded because of the diversity in this configuration. Perspectives within contexts have the potential to get deeper into them by *applications* while having independence, are not separated islands, and are *interconnected*. Hence, overlaps show that an improvement in context 'A' may affect context 'B' as the systematic presumption of this survey (DL & IoT are SOTA of AI & ICT). As these sophistications are, they synergize with each other exactly by those *linkages* under the AmI umbrella. Whether the purpose of this paper is to address the point of view of intelligible related contexts of AmI concurrently includes conceptual and technical features, rather than just setting boundaries. Some *interrelationships* in the *Human context* for example are behavioral analysis for gender classification [169], audio recorded by cameras for facial expressions [240], EEG for facial expression [8], BVP for stress recognition [190], and emotional information (e.g., facial expression and gait) as cues for action prediction [148]. Another path to interpret contexts' links is to *contrast dichotomies* for concluding the high valued *priorities* which for example any context can be seen from the Human contextual positions (human-centric); as how any IE system must serve for humanity's sake (humanism). Whether it might be for the System's wise that how contexts might be analyzed as a system's conceptualized role in its architecture design. All in all, while the contexts are distinct, they conceivably are potentially each other's *anchors* for further interpretation, and data about one context is probably *metadata* for another's induction.

Notwithstanding, having the taxonomy of contextual concepts denotes the variety of *requirements* of an IE system, organized to assess the best matching portfolio, to *balance tradeoffs* with specifications. Similar to *resource allocation*, system contexts shall be selected in concept, heterogeneity of varieties, and depth levels, until an engineered Context-Aware system is developed. A very noteworthy factor in this regard is to make a harmonious balance between contexts' tradeoffs in an *orchestration* fashion by myriad differentiations in concept and functionality, as being all-inclusive seems impossible. Another basis of this study, *changes in recent years*, led its references to be mainly review surveys of the last five years. *5G* and low energy-consuming IoT technologies paved the path for more practical real-time infrastructures. It eventually seems if there is a sufficiently proper dataset for a specific task, it can suffice (*software 2.0*) and DL will handle it better than any other method in high-level contexts [290]. DL is also capable of offering *personalized services* via more and more interactions with users nonstop (instruction and prompt tuned), which by preparations can lead to customized smart experiences of IEs.

Out-of-the-box is not necessarily unavailable in the box but might be a ***hybrid*** exploitation of accessible contexts and tools as SoS, to create novel incredible intelligent systems at different scopes. MLLMs (e.g., ChatGPT) as SOTA AI and the closest systems to AGI, are hybrid models with systematically several learning contexts' inclusion. Such foundation models are mainly pretrained (*TL*) in *SSL* fashion on large-scale *multimodal* input data within a large model (e.g., in parameters), which would lead that task-agnostic (*MTL*) generalization to support *few-shot*, one-shot, and *zero-shot learning* whether by finetuning or prompting. LLMs have also emerged the *ICL* (as a form of *meta learners*) and reasoning abilities which in some scenarios might lead to being eventually zero-shot planners [287, 280]. They can have *RL* (e.g., RLHF) to have continuous (online) changing policies or reward (loss) functions to be more and more personalized outputs. These foundation models, architecturally, are also a united compilation of previously presented models' elements [37].

It must have good *reasons* to go through constructing such a powerful entity as a machine, in which a tool like a knife not only can cut autonomously but also decides consecutively on its strategic logic. First of all, in almost all applications that IoT and AI might have, an IE made by those would, whereby an intelligent crane-wise power, homogenized with humans' living area, can help to build *eutopia* as much as a *dystopia*. It presumably seems we humans are condemned to moving there as we are in the midway of the journey, riding hastily. Meanwhile, the hurry is rewarding caused by plenty of economic and social beneficiaries.

The ***Philanthropic*** point of view of IEs seems to be sufficiently effective in varied usages which if only briefly looked at *healthcare* will be: (1) *medical* applications from healthcare systems, smart hospitals in high-level embedded systematic facilities, efficient utilization, or detection of regular disease universally, as people with the highest healthcare limitations can use with least costs (2) *AAL* application of AmI to help people in their routine Activities of Daily Living (ADL) and recognize an emergency event for people with health and medical conditions, as elderly for a longer dependent living (e.g., fall detection) and impairments such as mobility impairments, Autism, Alzheimer, intellectual disabilities (e.g., dawn syndrome), hearing loss, blindness, and other special diseases (e.g., Multiple Sclerosis). (3) *educating* about medical circumstances using recommendation systems as very likely detected illness scenarios or quickly needed reactions in specific emergency occurrences. Another philanthropic aspect, by the pervasiveness and low (possibly zero) *marginal cost* of copying a cyber-physical technology in return for tremendous profits, shall impact seriously the societies' welfare circumstances, likewise the most needful human beings.

As *optimization* and *personalization* are two consequents of context-aware AmI, there are several usages of it which just to name a few would be: education, security, industry 4.0 (e.g., factories, agriculture, grid), infrastructures (e.g., smart cities), entertainment (e.g., immersive environments, gaming), commerce (e.g., marketing, shopping), education (e.g., personalized gamification), and workplace (e.g., make more

convenient synergies). Moreover, the commercial and industrial utilizations are obvious even if not lead to immersive fascinating joyful pleasure for a long time, or move toward optimized agile production with few marginal costs.

6.1 Challenges & Open issues

Majorly works in each context are *application-oriented* (e.g., Industry 4.0, AAL), but the concept they are analyzing may be generalizable while are not expressed that way in broadly covering works. Whether the *interconnections* in the AmI context are highly diversified, by semi-heuristic methodology, sophistication is in a wide range of searching *keywords* at interchangeable parts. On the other side, for the exchange of knowledge and research comprehensively, it is required to consider these diverse aspects of contexts, whether the fields are divergent; it is attempted to cover these heterogeneities under the umbrella terms of *Ambient Intelligence* and *Intelligent Environments*. That is why this research cannot be organized in a systematic literature review format just by overviewing fixed keywords to neither cover SOTA nor comprehensive.

Stability is not a feature of SOTA and purely *all-inclusiveness* does not exist at all, as in this study, each context is extendable. *Human context* might have *additive parts*, for example: expectation (e.g., user taste, QoS), culture, assisting, caring, cognition, ownership, compliance hierarchy (access permission), and safety. The other example to mention for *System context extension* would be further system parts (e.g., Robotics, embedded CPS), design related factors (e.g., requirement-oriented, specification-oriented, system engineering), economic related factors (e.g., financial tradeoffs, supply & demand, marketing), or management related factors (e.g., business, processes, organization, construction).

Tradeoffs between IE contexts are beyond managing the resources, about contexts' orchestration by their conceptual and physical characteristics. That *harmony* depends on the system's metrics' priorities and constraints for allocation and control, which are systematically rooted in the *requirements* and *specifications* of the system's stakeholders. Each tradeoff, including its evaluation metrics by their Key Performance Indicators, should be analyzed including cost, accuracy, precision, reliability, energy consumption, power efficiency, security, privacy, and interoperability. The result is unique for each system to analyze indeed.

Safety and **security** are terrifically serious factors for humanity as one of the biggest concerns in Embedded Systems, such as CPS and IoT [287], where the cyber systems would take actions physically to impact people's lives. It was observed that major security issues are presented in the contexts this article surveyed because of their importance. The security facet of IoT can be analyzed through *confidentiality*, *integrity*, *availability*, *identification*, *authentication*, *privacy*, and *trust* features to be considered in *all IoT layers* [170]. Great technologies and protocols come just to solve security problems such as data and model *decentralization* (e.g., federated learning [135, 263, 168]), high-level *authenticating* [233], encryption [170] (e.g., blockchains), smart *surveillance* (e.g., natural disaster, violence, crime, terrorism) [6, 62, 118, 180, 311], *attack* management [187, 102], or not using vision-based *sensor networks* [279, 190, 53, 51]. The lack of security and privacy puts the whole system at *risk*, whether it is reliable or not, which, if below confidentiality standards, it is better not to continue its duty. The risks seem unlimited as exploiting theft information footprints and malicious attackers might take any harmful action. However, it may have new faces as defying *AGI* to take its beneficiaries. Unethical *apocalyptic* scenarios might be more important than the cognition of a maverick AGI; as exploitation of malicious totalitarian illiberal owners to treat humanity as pigeons of skinner-box just for their own sake.

Fairness problem is more general than these scenarios and exists everywhere caused by its nature in which anyone has her/his own *definition*. A major reason for fairness and **bias** in ML can be in *data*, *algorithm*, and *user interactions* [191, 119], as it looks possible to be in all contexts of AmI. To these concerns and probable futuristic ones, as AmI is potentially much more powerful, it is better to have fundamental standards before it pervasively gets used, regardless of its fairness considerations. A modern IE facing serious consequences must have cyber-physically comprehensive up-to-date legislation in major applications, in parallel with AI's.

As mentioned in the *scope context*, the **scale** of IEs is extensive by size or application, therefore scalability and interoperability should grab attention. *Interoperability* as is a key performance factor in IoT [207] and SoS [126, 203] to cover heterogeneity in functionality, in DL, generalization and adaptation (e.g., in TL) is a significant topic. Though by scaling except growth in *size*, the *complexity* will rise, as in LLMs, emerging unplanned abilities might occur as in the physical phase transition [287, 312, 201].

The degree of freedom of AmI as how much power it can have even in a high-leveled reliability standard might shift it from an independent to maverick's *secretive hypocrisy* is a thread, as that self-governance was potentially autonomy. If not liars, both LLMs and MLLMs are great *hallucinators*, while detecting original from fake would be hard [283]. An entity equipped with the ability to comprehend us and any related context of ours more accurately and agiler than us does not appear in a sci-fi drama in this era. If an intelligent CPS is constructed on physical world scenarios to have meta-learning commonsense physically, can act as a structure *creator* (e.g., using 3D printers and robots). As instructions and manuals (how to) are probably available in such large model's training data, the degree of freedom subject is more salient if that AmI would get or be offered access to a variety of empowering resources (e.g., financial, logistics). Ultimately, a noteworthy point is a system with the above-mentioned specifications is not available now and is only *embodied*. while how much feasible that is or when to achieve it is not clear for sure, though is possible.

6.2 Future works

The represented taxonomy configuration method is semi-heuristic, *any modification* in future works as a customized context taxonomies for specific applications is probable. Because of the rapidly changing nature of technology, it was observed that even philosophical assumptions about technology get adjusted or converted as well as the context taxonomies and computation paradigms. I think there will be further *application-oriented* research in AmI contexts, settled on particular context portfolio settings in theory and practice. Also, as there is an issue in multimodal learning about information fusion, in context-aware AmI, it is noteworthy to research how these contexts will be in an intelligent system, embedded.

What the future of AmI will be is mutually hinged on the *future of its constituents* as AI, ICT, data sciences, and robotics, plus their applications and connections in different fields. For example: (1) if a *new method* for cancer diagnosis tests is discovered, hence *affects* medicine after that data (tests) collected, AI models will learn it; (2) if a robot capable to *lift things* with numerous shapes and weights to *bear* on many surfaces like stairs reliably be invented, by training on DL (e.g., DRL) to do assistive tasks, changes IEs if produced scaled; (3) toward AGI as by progressions in MLLMs that have *high quality multimodal ICL and generalized physics-awareness* (e.g., commonsense), might lead to trustable interactive adaptable agents even in *real world physical tasks*. *Soon*, with more advanced algorithms and architectures in the DL context, *software 2.0* looks bolder whether if the proper data is available, the model can learn it [290]. With the continuum of *Moore's law* in hardware processors, it seems large models get larger and larger, and *inefficient* but highly prone to improve methods like *DRL* and larger foundation models be more *feasible*. Another major trend to be expected is more *systematic transdisciplinary collaborations* to bring new paradigm shifts which might be for example: (1) from Social or Industrial IoT (SIoT, IIoT) toward Social or Industrial AmI (e.g., industry 4.0) (2) from Context-Aware Recommendation Systems to Multimodal Context-Aware Recommendation Systems of IEs (3) from MLLMs to Cyber-Physical Multimodal Large Models (e.g., trained on physical modalities, physics-based twin, robotic enabler). Moreover, there will be more *customized* foundation models for each application with higher fairness and reliability.

Whether *foundation models are SOTA* in AI, there is still a lack of a large multimodal model trained in interactive (multimodal feedback) dynamic physics-supported environment to do *reasoning* about the *physical world physically* [108, 63]. Such models comprehend difficult complex problems differently or might be more inclusive in the *wild*, always changing real-world space for precise interventions [108]; no matter if it is not exactly clear whether LLMs and MLLMs really do reasoning or it is just a property of heuristics [115]. As discussed in the *physics context*, physics-based simulations and digital twins are proper ways to have agents within complex environments to explore and learn with multiple feedback on their policies to get more ready for being used in wild. Learning by multimodal interactions with rewards like nature's gamification as worked in many RL-based works, in order to reach at AGI level might help to be a physical AGI which not necessarily is intimidating.

The future of AmI as *General Ambient Intelligence* if would not be in one large DL model, *systematic thinking* views as *SoS* might be the solution to scale faster by utilizing a variety of systems and contexts as *embedded hybrid intelligent CPSs*' integration. The combination of different services linking to each other is as probable as we experience it in *smartphone apps* to have personalization of our multiple needs so routinely. Whether that customized recommendation is a favorite music, clip, movie, series, or anything we may want to buy, such as foods, clothes, accessories, or even skills we might like to learn. Any operating system supporter IoT (e.g., android) is likely to bring these services into a part of an environment if embedded; like smart kitchens, whether if a necessary ingredient in the refrigerator is finished, it alerts or even orders autonomously.

By surveying SOTA DL contexts, it can be induced that if a system as IEs with *harmonious orchestrated context portfolio* using good qualities of prior works in each context to modify, synchronous with pioneering courage (e.g., scaling in GPT-3), can succeed even in AI. Each significant technology after its emergence paved the way for the next ones by collaborating with other available technologies, playing supportive infrastructure or complementary cooperation whether if impactful enough, regards a paradigm shift. New paradigms in the past vanished in a fashion in which collectors would think those tech products would be counted as antiques or just disappear. While *this time is different*, as the case is the *data* (*contexts*). Data will not disappear because it always is utilizable if the entities in which the dataset recorded them do not exist anymore. Overall, the AGI or Physical AGI seems to be *context-aware complex SoSs* in interactive multimodal feedbacking space with large scale (e.g., in model parameter and data), resulting in transdisciplinary synergetic collaborations. Achieving that is in its best possible form through the history of AI by observing the incredible potentials of SOTA as LLMs, LVMs, and MLLMs.

6.3 Conclusion

In this study, the contexts of AmI were surveyed by proposing a comprehensive taxonomy to configure a conceptual-functional framework through four dimensions: *Human*, *System*, *Space*, and *Time*. The taxonomy framework, by categorizing AmI contexts' heterogeneous nature and mapping interconnections, aims to be used for future hybrid context portfolio allocations in complex intelligent systems, such as IEs. Using a review of over *300* publications in a systematic taxonomy, majorly recent surveys, this study captures the current trends and various facets of SOTA in AmI more comprehensively. The rapid advancements in all contexts' domains and their technological progressions in just a few years indicate the continued influence of Moore's law. By increasing the value of the DL and IoT synergizing utilizations, modern context-aware AmI can potentially become a major technology paradigm prospectively. This framework provides a structured foundation for researchers and practitioners to navigate the complexities of AmI systems and leverage their capabilities in future innovations. Future research could further explore and modify how this taxonomy can be applied in real-world systems while investigating the role of emerging technologies in context-aware AmI will be key to unlocking its full potential. The emergence of unprecedented capabilities is unignorable and unforeseeable, as it was for LLMs, emphasizing the dynamic evolving nature of the field.

ACKNOWLEDGMENTS

No funds or any other support is received. All three Figures infographic were produced by Elahe Kordlou. The images in Figure 2 were generated by ChatGPT 4o. the grammatical check is done by Grammarly and ProWritingAid.

REFERENCES

[1] Mohamed Abdul-Al, Ahmed S I Amar, Issa Elfergani, Richard Littlehales, Naser Ojaroudi Parchin, Yasir Al-Yasir, Chan Hwang See, Dawei Zhou, Zuhairiah Zainal Abidin, and Mohammad Alibakhshikenari. 2022. Wireless electromagnetic radiation assessment based on the specific absorption rate (SAR): A review case study. *Electronics* 11, 4 (2022), 511. https://doi.org/10.3390/electronics11040511

[2] Gregory D Abowd, Anind K Dey, Peter J Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a better understanding of context and context-awareness. In *Handheld and Ubiquitous Computing: First International Symposium, HUC'99 Karlsruhe, Germany, September 27–29, 1999 Proceedings 1*, 1999. Springer, 304–307. https://doi.org/10.1007/3-540-48157-5_29

[3] Giuseppe Aceto, Valerio Persico, and Antonio Pescapé. 2020. Industry 4.0 and health: Internet of things, big data, and cloud computing for healthcare 4.0. *J. Ind. Inf. Integr.* 18, (2020), 100129. https://doi.org/10.1016/j.jii.2020.100129

[4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamal Anadkat. 2023. Gpt-4 technical report. *arXiv Prepr. arXiv2303.08774* (2023). https://doi.org/10.48550/arXiv.2303.08774

[5] Timilehin B Aderinola, Tee Connie, Thian Song Ong, Wei-Chuen Yau, and Andrew Beng Jin Teoh. 2021. Learning age from gait: A survey. *IEEE Access* 9, (2021), 100352–100368. https://doi.org/10.1109/access.2021.3095477

[6] Jake K Aggarwal and Michael S Ryoo. 2011. Human activity analysis: A review. *Acm Comput. Surv.* 43, 3 (2011), 1–43. https://doi.org/10.1145/1922649.1922653

[7] Md Atiqur Rahman Ahad, Thanh Trung Ngo, Anindya Das Antar, Masud Ahmed, Tahera Hossain, Daigo Muramatsu, Yasushi Makihara, Sozo Inoue, and Yasushi Yagi. 2020. Wearable sensor-based gait analysis for age and gender estimation. *Sensors* 20, 8 (2020), 2424. https://doi.org/10.3390/s20082424

[8] Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. 2023. A systematic survey on multimodal emotion recognition using learning algorithms. *Intell. Syst. with Appl.* 17, (2023), 200171. https://doi.org/10.1016/j.iswa.2022.200171

[9] Ian F Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, and Erdal Cayirci. 2002. Wireless sensor networks: a survey. *Comput. networks* 38, 4 (2002), 393–422. https://doi.org/10.1016/s1389-1286(01)00302-4

[10] Omer Ali, Mohamad Khairi Ishak, Muhammad Kamran Liaquat Bhatti, Imran Khan, and Ki-Il Kim. 2022. A comprehensive review of internet of things: Technology stack, middlewares, and fog/edge computing interface. *Sensors* 22, 3 (2022), 995. https://doi.org/10.3390/s22030995

[11] Ana Almeida, Susana Brás, Susana Sargento, and Filipe Cabral Pinto. 2023. Time series big data: a survey on data stream frameworks, analysis and algorithms. *J. Big Data* 10, 1 (2023), 83. https://doi.org/10.1186/s40537-023-00760-1

[12] Md Zahangir Alom, Tarek M Taha, Chris Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Mahmudul Hasan, Brian C Van Essen, Abdul A S Awwal, and Vijayan K Asari. 2019. A state-of-the-art survey on deep learning theory and architectures. *electronics* 8, 3 (2019), 292. https://doi.org/10.3390/electronics8030292

[13] Ahmed A Al-Saedi, Veselka Boeva, Emiliano Casalicchio, and Peter Exner. 2022. Context-Aware Edge-Based AI Models for Wireless Sensor Networks—An Overview. *Sensors* 22, 15 (2022), 5544. https://doi.org/10.3390/s22155544

[14] Samaneh Aminikhanghahi and Diane J Cook. 2017. A survey of methods for time series change point detection. *Knowl. Inf. Syst.* 51, 2 (2017), 339–367. https://doi.org/10.1007/s10115-016-0987-z

[15] Mohd Aquib Ansari and Dushyant Kumar Singh. 2021. Human detection techniques for real time surveillance: a comprehensive survey. *Multimed. Tools Appl.* 80, 6 (2021), 8759–8808. https://doi.org/10.1007/s11042-020-10103-4

[16] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.* 34, 6 (2017), 26–38. https://doi.org/10.1109/msp.2017.2743240

[17] Safar M Asaad and Halgurd S Maghdid. 2022. A comprehensive review of indoor/outdoor localization solutions in IoT era: Research challenges and future perspectives. *Comput. Networks* 212, (2022), 109041. https://doi.org/10.1016/j.comnet.2022.109041

[18] Parvaneh Asghari, Amir Masoud Rahmani, and Hamid Haj Seyyed Javadi. 2019. Internet of Things applications: A systematic review. *Comput. Networks* 148, (2019), 241–261. https://doi.org/10.1016/j.comnet.2018.12.008

[19] Mohsen Attaran and Bilge Gokhan Celik. 2023. Digital Twin: Benefits, use cases, challenges, and opportunities. *Decis. Anal. J.* 6, (2023), 100165. https://doi.org/10.1016/j.dajour.2023.100165

[20] J Augusto, Asier Aztiria, Dean Kramer, and Unai Alegre. 2017. A survey on the evolution of the notion of context-awareness. *Appl. Artif. Intell.* 31, 7–8 (2017), 613–642. https://doi.org/10.1080/08839514.2018.1428490

[21] Juan C Augusto, Vic Callaghan, Diane Cook, Achilles Kameas, and Ichiro Satoh. 2013. Intelligent environments: a manifesto. *Human-centric Comput. Inf. Sci.* 3, (2013), 1–18. https://doi.org/10.1186/2192-1962-3-12

[22] Donald M Baer, Montrose M Wolf, and Todd R Risley. 1987. Some still-current dimensions of applied behavior analysis. *J. Appl. Behav. Anal.* 20, 4 (1987), 313–327. https://doi.org/10.1901/jaba.1987.20-313

[23] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. 2018. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 8340–8348. https://doi.org/10.1109/cvpr.2018.00870

[24] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2 (2018), 423–443. https://doi.org/10.1109/tpami.2018.2798607

[25] Chayan Banerjee, Kien Nguyen, Clinton Fookes, and George Karniadakis. 2023. Physics-informed computer vision: A review and perspectives. *arXiv Prepr. arXiv2305.18035* (2023). https://doi.org/10.1145/3689037

[26] Malcolm Barnard. 2013. *Graphic design as communication*. Routledge. https://doi.org/10.4324/9781315015385

[27] Paola Barra, Carmen Bisogni, Michele Nappi, David Freire-Obregón, and Modesto Castrillón-Santana. 2019. Gender classification on 2D human skeleton. In *2019 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART)*, 2019. IEEE, 1–4. https://doi.org/10.1109/biosmart.2019.8734198

[28] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychol. Sci. public Interes.* 20, 1 (2019), 1–68. https://doi.org/10.1177/1529100619832930

[29] Hamit Basgol, Inci Ayhan, and Emre Ugur. 2021. Time perception: A review on psychological, computational, and robotic models. *IEEE Trans. Cogn. Dev. Syst.* 14, 2 (2021), 301–315. https://doi.org/10.1109/tcds.2021.3059045

[30] Pierfrancesco Bellini, Paolo Nesi, and Gianni Pantaleo. 2022. IoT-enabled smart cities: A review of concepts, frameworks and key technologies. *Appl. Sci.* 12, 3 (2022), 1607. https://doi.org/10.3390/app12031607

[31] Mounir Bendali-Braham, Jonathan Weber, Germain Forestier, Lhassane Idoumghar, and Pierre-Alain Muller. 2021. Recent trends in crowd analysis: A review. *Mach. Learn. with Appl.* 4, (2021), 100023. https://doi.org/10.1016/j.mlwa.2021.100023

[32] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 8 (2013), 1798–1828. https://doi.org/10.1109/tpami.2013.50

[33] Konstantinos Benidis, Syama Sundar Rangapuram, Valentin Flunkert, Yuyang Wang, Danielle Maddix, Caner Turkmen, Jan Gasthaus, Michael Bohlke-Schneider, David Salinas, and Lorenzo Stella. 2022. Deep learning for time series forecasting: Tutorial and literature survey. *ACM Comput. Surv.* 55, 6 (2022), 1–36. https://doi.org/10.1145/3533382

[34] Gunjan Beniwal and Anita Singhrova. 2022. A systematic literature review on IoT gateways. *J. King Saud Univ. Inf. Sci.* 34, 10 (2022), 9541–9563. https://doi.org/10.1016/j.jksuci.2021.11.007

[35] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Syst.* 226, (2021), 107134. https://doi.org/10.1016/j.knosys.2021.107134

[36] Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*. Springer. https://doi.org/10.1007/978-0-387-45528-0

[37] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, and Emma Brunskill. 2021. On the opportunities and risks of foundation models. *arXiv Prepr. arXiv2108.07258* (2021). https://doi.org/10.48550/arXiv.2108.07258

[38] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, and Eric Luhman. 2024. Video generation models as world simulators. 2024. *URL https//openai. com/research/video-generation-models-as-world-simulators* 3, (2024).

[39] Tom B Brown. 2020. Language models are few-shot learners. *arXiv Prepr. ArXiv2005.14165* (2020).

[40] Jamal Bzai, Furqan Alam, Arwa Dhafer, Miroslav Bojović, Saleh M Altowaijri, Imran Khan Niazi, and Rashid Mehmood. 2022. Machine learning-enabled internet of things (iot): Data, applications, and industry perspective. *Electronics* 11, 17 (2022), 2676. https://doi.org/10.3390/electronics11172676

[41] Yang Cai, Angelo Genovese, Vincenzo Piuri, Fabio Scotti, and Mel Siegel. 2019. IoT-based architectures for sensing and local data processing in ambient intelligence: research and industrial trends. In *2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 2019. IEEE, 1–6. https://doi.org/10.1109/i2mtc.2019.8827110

[42] Daniel Canedo and António J R Neves. 2019. Facial expression recognition using computer vision: A systematic review. *Appl. Sci.* 9, 21 (2019), 4678. https://doi.org/10.3390/app9214678

[43] Davide Cannata, Simon M Breil, Bruno Lepri, Mitja D Back, and Denis O'Hora. 2022. Toward an integrative approach to nonverbal personality detection: Connecting psychological and artificial intelligence research. (2022). https://doi.org/10.1037/tmb0000054

[44] Marc-André Carbonneau, Eric Granger, Yazid Attabi, and Ghyslain Gagnon. 2017. Feature learning from spectrograms for assessment of personality traits. *IEEE Trans. Affect. Comput.* 11, 1 (2017), 25–31. https://doi.org/10.1109/taffc.2017.2763132

[45] Francesco Carini, Margherita Mazzola, Chiara Fici, Salvatore Palmeri, Massimo Messina, Provvidenza Damiani, and Giovanni Tomasello. 2017. Posture and posturology, anatomical and physiological profiles: overview and current state of art. *Acta Bio Medica Atenei Parm.* 88, 1 (2017), 11. https://doi.org/10.23750/abm.v88i1.5309

[46] Heather E P Cattell and Alan D Mead. 2008. The sixteen personality factor questionnaire (16PF). *SAGE Handb. Personal. theory Assess.* 2, (2008), 135–159. https://doi.org/10.4135/9781849200479.n7

[47] Alexandros André Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta. 2012. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Syst. Appl.* 39, 12 (2012), 10873–10888. https://doi.org/10.1016/j.eswa.2012.03.005

[48] Samit Chakraborty, Md Saiful Hoque, Naimur Rahman Jeem, Manik Chandra Biswas, Deepayan Bardhan, and Edgar Lobaton. 2021. Fashion recommendation systems, models and methods: A review. In *Informatics*, 2021. MDPI, 49. https://doi.org/10.3390/informatics8030049

[49] Iti Chaturvedi, Erik Cambria, Roy E Welsch, and Francisco Herrera. 2018. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Inf. Fusion* 44, (2018), 65–77. https://doi.org/10.1016/j.inffus.2017.12.006

[50] Guanling Chen and David Kotz. 2000. A survey of context-aware mobile computing research. (2000).

[51] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. 2021. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Comput. Surv.* 54, 4 (2021), 1–40. https://doi.org/10.1145/3447744

[52] Liming Chen, Jesse Hoey, Chris D Nugent, Diane J Cook, and Zhiwen Yu. 2012. Sensor-based activity recognition. *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.* 42, 6 (2012), 790–808. https://doi.org/10.1109/tsmcc.2012.2198883

[53] Liming Chen and Chris D Nugent. 2019. *Human activity recognition and behaviour analysis*. Springer. https://doi.org/10.1007/978-3-030-19408-6

[54] Wen-Huang Cheng, Sijie Song, Chieh-Yun Chen, Shintami Chusnul Hidayati, and Jiaying Liu. 2021. Fashion meets computer vision: A survey. *ACM Comput. Surv.* 54, 4 (2021), 1–41. https://doi.org/10.1145/3447239

[55] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. 2024. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* (2024). https://doi.org/10.1109/tpami.2024.3393571

[56] KR1442 Chowdhary and K R Chowdhary. 2020. Natural language processing. *Fundam. Artif. Intell.* (2020), 603–649. https://doi.org/10.1007/978-81-322-3972-7_19

[57] Kwok Tai Chui, Brij B Gupta, Jiaqi Liu, Varsha Arya, Nadia Nedjah, Ammar Almomani, and Priyanka Chaurasia. 2023. A survey of internet of things and cyber-physical systems: standards, algorithms, applications, security, challenges, and future directions. *Information* 14, 7 (2023), 388. https://doi.org/10.3390/info14070388

[58] Grazia Cicirelli, Roberto Marani, Antonio Petitti, Annalisa Milella, and Tiziana D'Orazio. 2021. Ambient assisted living: a review of technologies, methodologies and future perspectives for healthy aging of population. *Sensors* 21, 10 (2021), 3549. https://doi.org/10.3390/s21103549

[59] Mark Coeckelbergh. 2021. Time machines: Artificial intelligence, process, and narrative. *Philos. Technol.* 34, 4 (2021), 1623–1638. https://doi.org/10.1007/s13347-021-00479-y

[60] Diane J Cook, Juan C Augusto, and Vikramaditya R Jakkula. 2009. Ambient intelligence: Technologies, applications, and opportunities. *Pervasive Mob. Comput.* 5, 4 (2009), 277–298. https://doi.org/10.1016/j.pmcj.2009.04.001

[61] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 9 (2023), 10850–10869. https://doi.org/10.1109/tpami.2023.3261988

[62] L Minh Dang, Kyungbok Min, Hanxiang Wang, Md Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. 2020. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognit.* 108, (2020), 107561. https://doi.org/10.1016/j.patcog.2020.107561

[63] Ernest Davis. 2023. Benchmarks for automated commonsense reasoning: A survey. *ACM Comput. Surv.* 56, 4 (2023), 1–41. https://doi.org/10.1145/3615355

[64] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 2009. Ieee, 248–255. https://doi.org/10.1109/cvprw.2009.5206848

[65] Anind K Dey, Gregory D Abowd, and Daniel Salber. 2001. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human–Computer Interact.* 16, 2–4 (2001), 97–166. https://doi.org/10.1207/s15327051hci16234_02

[66] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv Prepr. arXiv2301.00234* (2022). https://doi.org/10.48550/arXiv.2301.00234

[67] Paul Dourish. 2004. What we talk about when we talk about context. *Pers. ubiquitous Comput.* 8, (2004), 19–30. https://doi.org/10.1007/s00779-003-0253-8

[68] Jiafei Duan, Arijit Dasgupta, Jason Fischer, and Cheston Tan. 2022. A survey on machine learning approaches for modelling intuitive physics. *arXiv Prepr. arXiv2202.06481* (2022). https://doi.org/10.1109/tetci.2022.3141105

[69] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. 2022. A survey of embodied ai: From simulators to research tasks. *IEEE Trans. Emerg. Top. Comput. Intell.* 6, 2 (2022), 230–244. https://doi.org/10.48550/arXiv.2202.06481

[70] Rob Dunne, Tim Morris, and Simon Harper. 2021. A survey of ambient intelligence. *ACM Comput. Surv.* 54, 4 (2021), 1–27. https://doi.org/10.1145/3447242

[71] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. 2020. Human emotion recognition: Review of sensors and methods. *Sensors* 20, 3 (2020), 592. https://doi.org/10.3390/s20030592

[72] Maria Egger, Matthias Ley, and Sten Hanke. 2019. Emotion recognition from physiological signal analysis: A review. *Electron. Notes Theor. Comput. Sci.* 343, (2019), 35–55. https://doi.org/10.1016/j.entcs.2019.04.009

[73] Mahsa Ehsanpour, Alireza Abedin, Fatemeh Saleh, Javen Shi, Ian Reid, and Hamid Rezatofighi. 2020. Joint learning of social groups, individuals action and sub-group activities in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, 2020. Springer, 177–195. https://doi.org/10.1007/978-3-030-58545-7_11

[74] Paul Ekman. 1993. Facial expression and emotion. *Am. Psychol.* 48, 4 (1993), 384. https://doi.org/10.2466/pms.1967.24.3.711

[75] Paul Ekman, Tim Dalgleish, and Mick Power. 1999. Handbook of cognition and emotion. (1999).

[76] Paul Ekman and Wallace V Friesen. 1967. Head and body cues in the judgment of emotion: A reformulation. *Percept. Mot. Skills* 24, 3 (1967), 711–724. https://doi.org/10.1037//0003-066x.48.4.384

[77] Paul Ekman and Wallace V Friesen. 2003. *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk. https://doi.org/10.1002/0470013494.ch3

[78] Hans Jurgen Eysenck. 2012. *A model for personality*. Springer Science & Business Media.

[79] Zizhu Fan, Hong Zhang, Zheng Zhang, Guangming Lu, Yudong Zhang, and Yaowei Wang. 2022. A survey of crowd counting and density estimation based on convolutional neural network. *Neurocomputing* 472, (2022), 224–251. https://doi.org/10.1016/j.neucom.2021.02.103

[80] Pooyan Shams Farahsari, Amirhossein Farahzadi, Javad Rezazadeh, and Alireza Bagheri. 2022. A survey on indoor positioning systems for IoT-based applications. *IEEE Internet Things J.* 9, 10 (2022), 7680–7699. https://doi.org/10.1109/jiot.2022.3149048

[81] Amirhossein Farahzadi, Pooyan Shams, Javad Rezazadeh, and Reza Farahbakhsh. 2018. Middleware technologies for cloud of things: a survey. *Digit. Commun. Networks* 4, 3 (2018), 176–188. https://doi.org/10.1016/j.dcan.2017.04.005

[82] Ling Feng, Peter M G Apers, and Willem Jonker. 2004. Towards context-aware data management for ambient intelligence. In *Database and Expert Systems Applications: 15th International Conference, DEXA 2004, Zaragoza, Spain, August 30-September 3, 2004. Proceedings 15*, 2004. Springer, 422–431. https://doi.org/10.1007/978-3-540-30075-5_41

[83] Wayne W Fisher, Cathleen C Piazza, and Henry S Roane. 2021. *Handbook of applied behavior analysis*. Guilford Publications.

[84] Luciano Floridi. 2021. Digital time: Latency, real-time, and the onlife experience of everyday time. *Philos. Technol.* 34, 3 (2021), 407–412. https://doi.org/10.1007/s13347-021-00472-5

[85] Abdur Rahim Mohammad Forkan, Ibrahim Khalil, Zahir Tari, Sebti Foufou, and Abdelaziz Bouras. 2015. A context-aware approach for long-term behavioural change detection and abnormality prediction in ambient assisted living. *Pattern Recognit.* 48, 3 (2015), 628–641. https://doi.org/10.1016/j.patcog.2014.07.007

[86] Giancarlo Fortino, Claudio Savaglio, Giandomenico Spezzano, and MengChu Zhou. 2020. Internet of things as system of systems: A review of methodologies, frameworks, platforms, and tools. *IEEE Trans. Syst. Man, Cybern. Syst.* 51, 1 (2020), 223–236. https://doi.org/10.1109/tsmc.2020.3042898

[87] Matjaz Gams, Irene Yu-Hua Gu, Aki Härmä, Andrés Muñoz, and Vincent Tam. 2019. Artificial intelligence and ambient intelligence. *J. Ambient Intell. Smart Environ.* 11, 1 (2019), 71–86. https://doi.org/10.3233/ais-180508

[88] Marion Garaus, Udo Wagner, and Ricarda C Rainer. 2021. Emotional targeting using digital signage systems and facial recognition at the point-of-sale. *J. Bus. Res.* 131, (2021), 747–762. https://doi.org/10.1016/j.jbusres.2020.10.065

[89] Mouzhi Ge, Hind Bangui, and Barbora Buhnova. 2018. Big data for internet of things: a survey. *Futur. Gener. Comput. Syst.* 87, (2018), 601–614. https://doi.org/10.1016/j.future.2018.04.053

[90] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. 2019. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019. 5337–5345. https://doi.org/10.1109/cvpr.2019.00548

[91] Morteza Ghobakhloo. 2020. Industry 4.0, digitization, and opportunities for sustainability. *J. Clean. Prod.* 252, (2020), 119869. https://doi.org/10.1016/j.jclepro.2019.119869

[92] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. 2018. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 8359–8367.

[93] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* 27, (2014).

[94] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144. https://doi.org/10.1145/3422622

[95] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, and Moritz Mueller-Freitag. 2017. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, 2017. 5842–5850. https://doi.org/10.1109/iccv.2017.622

[96] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2016. LSTM: A search space odyssey. *IEEE Trans. neural networks Learn. Syst.* 28, 10 (2016), 2222–2232. https://doi.org/10.1109/tnnls.2016.2582924

[97] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, and Jianfei Cai. 2018. Recent advances in convolutional neural networks. *Pattern Recognit.* 77, (2018), 354–377. https://doi.org/10.1016/j.patcog.2017.10.013

[98] Weili Guan, Fangkai Jiao, Xuemeng Song, Haokun Wen, Chung-Hsing Yeh, and Xiaojun Chang. 2022. Personalized fashion compatibility modeling via metapath-guided heterogeneous graph learning. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 2022. 482–491. https://doi.org/10.1145/3477495.3532038

[99] Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, and Marimuthu Palaniswami. 2013. Internet of Things (IoT): A vision, architectural elements, and future directions. *Futur. Gener. Comput. Syst.* 29, 7 (2013), 1645–1660. https://doi.org/10.1016/j.future.2013.01.010

[100] Hajer Guerdelli, Claudio Ferrari, Walid Barhoumi, Haythem Ghazouani, and Stefano Berretti. 2022. Macro-and micro-expressions facial datasets: A survey. *Sensors* 22, 4 (2022), 1524. https://doi.org/10.3390/s22041524

[101] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. 2021. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Trans. Knowl. Data Eng.* 35, 4 (2021), 3313–3332. https://doi.org/10.1109/tkde.2021.3130191

[102] Volkan Gunes, Steffen Peter, Tony Givargis, and Frank Vahid. 2014. A survey on concepts, applications, and challenges in cyber-physical systems. *KSII Trans. Internet Inf. Syst.* 8, 12 (2014), 4242–4268. https://doi.org/10.3837/tiis.2014.12.001

[103] Lin Guo, Zongxing Lu, and Ligang Yao. 2021. Human-machine interaction sensing technology based on hand gesture recognition: A review. *IEEE Trans. Human-Machine Syst.* 51, 4 (2021), 300–309. https://doi.org/10.1109/thms.2021.3086003

[104] Yi Guo, Shunan Guo, Zhuochen Jin, Smiti Kaul, David Gotz, and Nan Cao. 2021. Survey on visual analysis of event sequence data. *IEEE Trans. Vis. Comput. Graph.* 28, 12 (2021), 5091–5112. https://doi.org/10.1109/tvcg.2021.3100413

[105] Jasmin Guth, Uwe Breitenbücher, Michael Falkenthal, Frank Leymann, and Lukas Reinfurt. 2016. Comparison of IoT platform architectures: A field study based on a reference architecture. In *2016 Cloudification of the Internet of Things (CIoT)*, 2016. IEEE, 1–6. https://doi.org/10.1109/ciot.2016.7872918

[106] Yosra Hajjaji, Wadii Boulila, Imed Riadh Farah, Imed Romdhani, and Amir Hussain. 2021. Big data and IoT-based applications in smart environments: A systematic review. *Comput. Sci. Rev.* 39, (2021), 100318. https://doi.org/10.1016/j.cosrev.2020.100318

[107] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, and Liang Zhang. 2021. Pre-trained models: Past, present and future. *AI Open* 2, (2021), 225–250. https://doi.org/10.1016/j.aiopen.2021.08.002

[108] Zhongkai Hao, Songming Liu, Yichi Zhang, Chengyang Ying, Yao Feng, Hang Su, and Jun Zhu. 2022. Physics-informed machine learning: A survey on problems, methods and applications. *arXiv Prepr. arXiv2211.08064* (2022). https://doi.org/10.48550/arXiv.2211.08064

[109] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 770–778. https://doi.org/10.1109/cvpr.2016.90

[110] Karen Henricksen. 2003. A framework for context-aware pervasive computing applications. (2003). https://doi.org/10.14264/106832

[111] Shintami C Hidayati, Chuang-Wen You, Wen-Huang Cheng, and Kai-Lung Hua. 2017. Learning and recognition of clothing genres from full-body images. *IEEE Trans. Cybern.* 48, 5 (2017), 1647–1659. https://doi.org/10.1109/tcyb.2017.2712634

[112] Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.* 6, 02 (1998), 107–116. https://doi.org/10.1162/neco.1997.9.8.1735

[113] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (1997), 1735–1780. https://doi.org/10.1142/s0218488598000094

[114] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2021. Meta-learning in neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 9 (2021), 5149–5169. https://doi.org/10.1109/tpami.2021.3079209

[115] Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv Prepr. arXiv2212.10403* (2022). https://doi.org/10.18653/v1/2023.findings-acl.67

[116] Mike Huisman, Jan N Van Rijn, and Aske Plaat. 2021. A survey of deep meta-learning. *Artif. Intell. Rev.* 54, 6 (2021), 4483–4541. https://doi.org/10.1007/s10462-021-10004-4

[117] Tanveer Hussain, Fath U Min Ullah, Khan Muhammad, Seungmin Rho, Amin Ullah, Eenjun Hwang, Jihoon Moon, and Sung Wook Baik. 2021. Smart and intelligent energy monitoring systems: A comprehensive literature survey and future research guidelines. *Int. J. Energy Res.* 45, 3 (2021), 3590–3614. https://doi.org/10.1002/er.6093

[118] Zawar Hussain, Michael Sheng, and Wei Emma Zhang. 2019. Different approaches for human activity recognition: A survey. *arXiv Prepr. arXiv1906.05074* (2019).

[119] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021. 560–575. https://doi.org/10.1145/3442188.3445918

[120] Zeba Idrees and Lirong Zheng. 2020. Low cost air pollution monitoring systems: A review of protocols and enabling technologies. *J. Ind. Inf. Integr.* 17, (2020), 100123. https://doi.org/10.1016/j.jii.2019.100123

[121] Saidul Islam, Hanae Elmekki, Ahmed Elsebai, Jamal Bentahar, Nagat Drawel, Gaith Rjoub, and Witold Pedrycz. 2023. A comprehensive survey on applications of transformers for deep learning tasks. *Expert Syst. Appl.* (2023), 122666. https://doi.org/10.1016/j.eswa.2023.122666

[122] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 1125–1134. https://doi.org/10.1109/cvpr.2017.632

[123] Rachael E Jack, Oliver G B Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. 2012. Facial expressions of emotion are not culturally universal. *Proc. Natl. Acad. Sci.* 109, 19 (2012), 7241–7244. https://doi.org/10.1073/pnas.1200155109

[124] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies* 9, 1 (2020), 2. https://doi.org/10.3390/technologies9010002

[125] Ahmad Jalal, Yeon-Ho Kim, Yong-Joong Kim, Shaharyar Kamal, and Daijin Kim. 2017. Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recognit.* 61, (2017), 295–308. https://doi.org/10.1016/j.patcog.2016.08.003

[126] Mo Jamshidi. 2017. *Systems of systems engineering: principles and applications*. CRC press. https://doi.org/10.1201/9781420065893

[127] Uthayakumar Jayasankar, Vengattaraman Thirumal, and Dhavachelvan Ponnurangam. 2021. A survey on data compression techniques: From the perspective of data quality, coding schemes, data type and applications. *J. King Saud Univ. Inf. Sci.* 33, 2 (2021), 119–140. https://doi.org/10.1016/j.jksuci.2018.05.006

[128] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. 2020. Vision-based human action recognition: An overview and real world challenges. *Forensic Sci. Int. Digit. Investig.* 32, (2020), 200901. https://doi.org/10.1016/j.fsidi.2019.200901

[129] Min Jiang and Guodong Guo. 2019. Body Weight Analysis from Human Body Images. *IEEE Trans. Inf. Forensics Secur.* 14, 10 (2019), 2676–2688. https://doi.org/10.1109/tifs.2019.2904840

[130] Shuo Jiang, Peiqi Kang, Xinyu Song, Benny P L Lo, and Peter B Shull. 2021. Emerging wearable interfaces and algorithms for hand gesture recognition: A survey. *IEEE Rev. Biomed. Eng.* 15, (2021), 85–102. https://doi.org/10.1109/rbme.2021.3078190

[131] Xiaoyan Jiang, Zuojin Hu, Shuihua Wang, and Yudong Zhang. 2023. A survey on artificial intelligence in posture recognition. *Comput. Model. Eng. Sci. C.* 137, 1 (2023), 35. https://doi.org/10.32604/cmes.2023.027676

[132] Julio C S Jacques Junior, Yağmur Güçlütürk, Marc Pérez, Umut Güçlü, Carlos Andujar, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Marcel A J Van Gerven, and Rob Van Lier. 2019. First impressions: A survey on vision-based apparent personality trait analysis. *IEEE Trans. Affect. Comput.* 13, 1 (2019), 75–95. https://doi.org/10.1109/taffc.2019.2930058

[133] Yasin Kabalci, Ersan Kabalci, Sanjeevikumar Padmanaban, Jens Bo Holm-Nielsen, and Frede Blaabjerg. 2019. Internet of things applications as energy internet in smart grids and smart environments. *Electronics* 8, 9 (2019), 972. https://doi.org/10.3390/electronics8090972

[134] Achuta Kadambi, Celso de Melo, Cho-Jui Hsieh, Mani Srivastava, and Stefano Soatto. 2023. Incorporating physics into data-driven computer vision. *Nat. Mach. Intell.* 5, 6 (2023), 572–580. https://doi.org/10.1038/s42256-023-00662-0

[135] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, and Rachel Cummings. 2021. Advances and open problems in federated learning. *Found. trends® Mach. Learn.* 14, 1–2 (2021), 1–210. http://dx.doi.org/10.1561/2200000083

[136] Dionisis Kandris, Christos Nakas, Dimitrios Vomvas, and Grigorios Koulouras. 2020. Applications of wireless sensor networks: an up-to-date survey. *Appl. Syst. Innov.* 3, 1 (2020), 14. https://doi.org/10.3390/asi3010014

[137] Wafa'a Kassab and Khalid A Darabkh. 2020. A–Z survey of Internet of Things: Architectures, protocols, applications, recent advances, future directions and recommendations. *J. Netw. Comput. Appl.* 163, (2020), 102663. https://doi.org/10.1016/j.jnca.2020.102663

[138] Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press. https://doi.org/10.1017/cbo9780511807572

[139] Rafiullah Khan, Sarmad Ullah Khan, Rifaqat Zaheer, and Shahid Khan. 2012. Future internet: the internet of things architecture, possible applications and key challenges. In *2012 10th international conference on frontiers of information technology*, 2012. IEEE, 257–260. https://doi.org/10.1109/fit.2012.53

[140] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: state of the art, current trends and challenges. *Multimed. Tools Appl.* 82, 3 (2023), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

[141] M Hadi Kiapour, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. 2014. Hipster wars: Discovering elements of fashion styles. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, 2014. Springer, 472–488. https://doi.org/10.1007/978-3-319-10590-1_31

[142] Jonghwa Kim and Elisabeth André. 2008. Emotion recognition based on physiological changes in music listening. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 12 (2008), 2067–2083. https://doi.org/10.1109/tpami.2008.26

[143] Nasser Kimbugwe, Tingrui Pei, and Moses Ntanda Kyebambe. 2021. Application of deep learning for quality of service enhancement in internet of things: A review. *Energies* 14, 19 (2021), 6384. https://doi.org/10.3390/en14196384

[144] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, and Wan-Yen Lo. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 4015–4026. https://doi.org/10.1109/iccv51070.2023.00371

[145] Ahmad F Klaib, Nawaf O Alsrehin, Wasen Y Melhem, Haneen O Bashtawi, and Aws A Magableh. 2021. Eye tracking algorithms, techniques, tools, and applications with an emphasis on machine learning and Internet of Things technologies. *Expert Syst. Appl.* 166, (2021), 114037. https://doi.org/10.1016/j.eswa.2020.114037

[146] Mustafa Kocakulak and Ismail Butun. 2017. An overview of Wireless Sensor Networks towards internet of things. In *2017 IEEE 7th annual computing and communication workshop and conference (CCWC)*, 2017. Ieee, 1–6. https://doi.org/10.1109/ccwc.2017.7868374

[147] Agata Kołakowska, Agnieszka Landowska, Mariusz Szwoch, Wioleta Szwoch, and Michał R Wróbel. 2015. Modeling emotions for affect-aware applications. *Inf. Syst. Dev. Appl.* (2015), 55–69.

[148] Yu Kong and Yun Fu. 2022. Human action recognition and prediction: A survey. *Int. J. Comput. Vis.* 130, 5 (2022), 1366–1401. https://doi.org/10.1007/s11263-022-01594-9

[149] Alexander Kossiakoff, Steven M Biemer, Samuel J Seymour, and David A Flanigan. 2020. *Systems engineering principles and practice*. John Wiley & Sons. https://doi.org/10.1002/9781119516699

[150] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90. https://doi.org/10.1145/3065386

[151] Paula Lago, Claudia Roncancio, and Claudia Jiménez-Guarín. 2019. Learning and managing context enriched behavior patterns in smart homes. *Futur. Gener. Comput. Syst.* 91, (2019), 191–205.https://doi.org/10.1016/j.future.2018.09.004

[152] Suprava Ranjan Laha, Binod Kumar Pattanayak, and Saumendra Pattnaik. 2022. Advancement of environmental monitoring system using IoT and sensor: A comprehensive analysis. *AIMS Environ. Sci.* 9, 6 (2022), 771–800. https://doi.org/10.3934/environsci.2022044

[153] Fanny Larradet, Radoslaw Niewiadomski, Giacinto Barresi, Darwin G Caldwell, and Leonardo S Mattos. 2020. Toward emotion recognition from physiological signals in the wild: approaching the methodological issues in real-life data collection. *Front. Psychol.* 11, (2020), 1111. https://doi.org/10.3389/fpsyg.2020.01111

[154] Robin Le Poidevin. 2000. The experience and perception of time. (2000).

[155] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444. https://doi.org/10.1038/nature14539

[156] Dong-seok Lee, Jong-soo Kim, Seok Chan Jeong, and Soon-kak Kwon. 2020. Human height estimation by color deep learning and depth 3D conversion. *Appl. Sci.* 10, 16 (2020), 5531. https://doi.org/10.3390/app10165531

[157] Alfred Leick, Lev Rapoport, and Dmitry Tatarnikov. 2015. *GPS satellite surveying*. John Wiley & Sons. https://doi.org/10.1002/9781119018612

[158] Athanasios Lentzas and Dimitris Vrakas. 2020. Non-intrusive human activity recognition and abnormal behavior detection on elderly people: A review. *Artif. Intell. Rev.* 53, 3 (2020), 1975–2021. https://doi.org/10.1007/s10462-019-09724-5

[159] Kurt Lewin. 2013. *Principles of topological psychology*. Read Books Ltd. https://doi.org/10.1037/10019-000

[160] Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *Ai Open* 3, (2022), 71–90. https://doi.org/10.1016/j.aiopen.2022.03.001

[161] Meng Li, Aniket Pal, Amirreza Aghakhani, Abdon Pena-Francesch, and Metin Sitti. 2022. Soft actuators for real-world applications. *Nat. Rev. Mater.* 7, 3 (2022), 235–249. https://doi.org/10.1109/taffc.2020.2981446

[162] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* 13, 3 (2020), 1195–1215. https://doi.org/10.1038/s41578-021-00389-7

[163] You Li, Yuan Zhuang, Xin Hu, Zhouzheng Gao, Jia Hu, Long Chen, Zhe He, Ling Pei, Kejie Chen, and Maosong Wang. 2020. Toward location-enabled IoT (LE-IoT): IoT positioning techniques, error sources, and error mitigation. *IEEE Internet Things J.* 8, 6 (2020), 4035–4062. https://doi.org/10.1109/jiot.2020.3019199

[164] Yuxi Li. 2017. Deep reinforcement learning: An overview. *arXiv Prepr. arXiv1701.07274* (2017).

[165] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Comput. Surv.* 56, 10 (2024), 1–42. https://doi.org/10.1145/3656580

[166] Alexander Ligthart, Cagatay Catal, and Bedir Tekinerdogan. 2021. Systematic reviews in sentiment analysis: a tertiary study. *Artif. Intell. Rev.* (2021), 1–57. https://doi.org/10.1007/s10462-021-09973-3

[167] Jia Zheng Lim, James Mountstephens, and Jason Teo. 2020. Emotion recognition using eye-tracking: taxonomy, review and current challenges. *Sensors* 20, 8 (2020), 2384. https://doi.org/10.3390/s20082384

[168] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. 2020. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Commun. Surv. tutorials* 22, 3 (2020), 2031–2063. https://doi.org/10.1109/comst.2020.2986024

[169] Feng Lin, Yingxiao Wu, Yan Zhuang, Xi Long, and Wenyao Xu. 2016. Human gender classification: a review. *Int. J. Biom.* 8, 3–4 (2016), 275–300. https://doi.org/10.1504/ijbm.2016.10003589

[170] Jie Lin, Wei Yu, Nan Zhang, Xinyu Yang, Hanlin Zhang, and Wei Zhao. 2017. A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications. *IEEE internet things J.* 4, 5 (2017), 1125–1142. https://doi.org/10.1109/jiot.2017.2683200

[171] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI open* 3, (2022), 111–132. https://doi.org/10.1016/j.aiopen.2022.10.001

[172] Wenqian Lin and Chao Li. 2023. Review of studies on emotion recognition and judgment based on physiological signals. *Appl. Sci.* 13, 4 (2023), 2573. https://doi.org/10.3390/app13042573

[173] Bing Liu. 2010. Sentiment analysis and subjectivity. *Handb. Nat. Lang. Process.* 2, 2010 (2010), 627–666. https://doi.org/10.1201/9781420085938-36

[174] Bing Liu. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press. https://doi.org/10.1017/9781108639286

[175] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55, 9 (2023), 1–35. https://doi.org/10.1145/3560815

[176] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. *IEEE Trans. Knowl. Data Eng.* 35, 1 (2021), 857–876. https://doi.org/10.1109/tkde.2021.3090866

[177] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, and Jianfeng Gao. 2024. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv Prepr. arXiv2402.17177* (2024). https://doi.org/10.48550/arXiv.2402.17177

[178] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. 2021. Multiple object tracking: A literature review. *Artif. Intell.* 293, (2021), 103448. https://doi.org/10.1016/j.artint.2020.103448

[179] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. 2017. Pose guided person image generation. *Adv. Neural Inf. Process. Syst.* 30, (2017). https://doi.org/10.1109/icip40778.2020.9190773

[180] Nan Ma, Zhixuan Wu, Yiu-ming Cheung, Yuchen Guo, Yue Gao, Jiahong Li, and Beijyan Jiang. 2022. A survey of human action recognition and posture prediction. *Tsinghua Sci. Technol.* 27, 6 (2022), 973–1001. https://doi.org/10.26599/tst.2021.9010068

[181] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2024. A Survey on Vision-Language-Action Models for Embodied AI. *arXiv Prepr. arXiv2405.14093* (2024). https://doi.org/10.48550/arXiv.2405.14093

[182] Mohammad Saeid Mahdavinejad, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi, and Amit P Sheth. 2018. Machine learning for Internet of Things data analysis: A survey. *Digit. Commun. Networks* 4, 3 (2018), 161–175. https://doi.org/10.1016/j.dcan.2017.10.002

[183] M Rezwanul Mahmood, Mohammad Abdul Matin, Panagiotis Sarigiannidis, and Sotirios K Goudos. 2022. A comprehensive review on artificial intelligence/machine learning algorithms for empowering the future IoT toward 6G era. *IEEE Access* 10, (2022), 87535–87562. https://doi.org/10.1109/access.2022.3199689

[184] Mamoona Majid, Shaista Habib, Abdul Rehman Javed, Muhammad Rizwan, Gautam Srivastava, Thippa Reddy Gadekallu, and Jerry Chun-Wei Lin. 2022. Applications of wireless sensor networks and internet of things frameworks in the industry revolution 4.0: A systematic literature review. *Sensors* 22, 6 (2022), 2087. https://doi.org/10.3390/s22062087

[185] Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intell. Syst.* 32, 2 (2017), 74–79. https://doi.org/10.1109/mis.2017.23

[186] Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimed. Tools Appl.* 80, (2021), 9411–9457. https://doi.org/10.1007/s11042-020-10073-7

[187] Peter Marwedel. 2021. *Embedded system design: embedded systems foundations of cyber-physical systems, and the internet of things*. Springer Nature. https://doi.org/10.1007/978-3-030-60910-8

[188] Yutaka Matsuo, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe, and Jun Morimoto. 2022. Deep learning, reinforcement learning, and world models. *Neural Networks* 152, (2022), 267–275. https://doi.org/10.1016/j.neunet.2022.03.037

[189] Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *J. Pers. Soc. Psychol.* 52, 1 (1987), 81. https://doi.org/10.1037//0022-3514.52.1.81

[190] Daniel McDuff. 2023. Camera measurement of physiological vital signs. *ACM Comput. Surv.* 55, 9 (2023), 1–40. https://doi.org/10.1145/3558518

[191] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 54, 6 (2021), 1–35. https://doi.org/10.1145/3457607

[192] Albert Mehrabian. 1974. An approach to environmental psychology. *Massachusetts Inst. Technol.* (1974).

[193] Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. 2020. Recent trends in deep learning based personality detection. *Artif. Intell. Rev.* 53, 4 (2020), 2313–2339. https://doi.org/10.1007/s10462-019-09770-z

[194] Peter Mell. 2011. The NIST Definition of Cloud Computing. *Natl. Inst. Stand. Technol.* (2011). https://doi.org/10.6028/nist.sp.800-145

[195] Walied Merghani, Adrian K Davison, and Moi Hoon Yap. 2018. A review on facial micro-expressions analysis: datasets, features and metrics. *arXiv Prepr. arXiv1805.02397* (2018). https://doi.org/10.48550/arXiv.1805.02397

[196] Seifeddine Messaoud, Abbas Bradai, Syed Hashim Raza Bukhari, Pham Tran Anh Quang, Olfa Ben Ahmed, and Mohamed Atri. 2020. A survey on machine learning in Internet of Things: Algorithms, strategies, and applications. *Internet of Things* 12, (2020), 100314. https://doi.org/10.1016/j.iot.2020.100314

[197] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv Prepr. arXiv1312.5602* (2013). https://doi.org/10.48550/arXiv.1312.5602

[198] Mahdi Momeni-k, Sotirios Ch Diamantas, Fabio Ruggiero, and Bruno Siciliano. 2012. Height estimation from a single camera view. In *International Conference on Computer Vision Theory and Applications*, 2012. SciTePress, 358–364. https://doi.org/10.5220/0003866203580364

[199] I Myers. 1962. The Myers-Briggs Type Indicator. https://doi.org/10.1037/14404-000

[200] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. 2019. Speech recognition using deep neural networks: A systematic review. *IEEE access* 7, (2019), 19143–19165. https://doi.org/10.1109/access.2019.2896880

[201] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv Prepr. arXiv2307.06435* (2023). https://doi.org/10.48550/arXiv.2307.06435

[202] Anne H Ngu, Mario Gutierrez, Vangelis Metsis, Surya Nepal, and Quan Z Sheng. 2016. IoT middleware: A survey on issues and enabling technologies. *IEEE Internet Things J.* 4, 1 (2016), 1–20. https://doi.org/10.1109/jiot.2016.2615180

[203] Claus Ballegaard Nielsen, Peter Gorm Larsen, John Fitzgerald, Jim Woodcock, and Jan Peleska. 2015. Systems of systems engineering: basic concepts, model-based techniques, and research directions. *ACM Comput. Surv.* 48, 2 (2015), 1–41. https://doi.org/10.1145/2794381

[204] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. 2018. Survey on emotional body gesture recognition. *IEEE Trans. Affect. Comput.* 12, 2 (2018), 505–523. https://doi.org/10.1109/taffc.2018.2874986

[205] Mahda Noura, Mohammed Atiquzzaman, and Martin Gaedke. 2019. Interoperability in internet of things: Taxonomies and open challenges. *Mob. networks Appl.* 24, (2019), 796–809. https://doi.org/10.1007/s11036-018-1089-9

[206] G P T OpenAI. 2023. 4V (ision) system card. *preprint* (2023).

[207] Andrew Ortony, Gerald L Clore, and Allan Collins. 2022. *The cognitive structure of emotions*. Cambridge university press. https://doi.org/10.1017/9781108934053

[208] Munir Oudah, Ali Al-Naji, and Javaan Chahl. 2020. Hand gesture recognition based on computer vision: a review of techniques. *J. Imaging* 6, 8 (2020), 73. https://doi.org/10.3390/jimaging6080073

[209] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, and Alex Ray. 2022. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* 35, (2022), 27730–27744.

[210] Indranil Pan, Lachlan R Mason, and Omar K Matar. 2022. Data-centric Engineering: integrating simulation, machine learning and statistics. Challenges and opportunities. *Chem. Eng. Sci.* 249, (2022), 117271. https://doi.org/10.1016/j.ces.2021.117271

[211] Pavel Pascacio, Sven Casteleyn, Joaquin Torres-Sospedra, Elena Simona Lohan, and Jari Nurmi. 2021. Collaborative indoor positioning systems: A systematic review. *Sensors* 21, 3 (2021), 1002. https://doi.org/10.3390/s21031002

[212] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A A Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019. 10975–10985. https://doi.org/10.1109/cvpr.2019.01123

[213] Barbara Pease and Allan Pease. 2008. *The definitive book of body language: The hidden meaning behind people's gestures and expressions*. Bantam.

[214] Charith Perera, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. 2013. Context aware computing for the internet of things: A survey. *IEEE Commun. Surv. tutorials* 16, 1 (2013), 414–454. https://doi.org/10.1109/surv.2013.042313.00197

[215] Le Vy Phan and John F Rauthmann. 2021. Personality computing: New frontiers in personality assessment. *Soc. Personal. Psychol. Compass* 15, 7 (2021), e12624. https://doi.org/10.1111/spc3.12624

[216] Paola Pierleoni, Alberto Belli, Roberto Concetti, Lorenzo Palma, Federica Pinti, Sara Raggiunto, Luisiana Sabbatini, Simone Valenti, and Andrea Monteriù. 2021. Biological age estimation using an eHealth system based on wearable sensors. *J. Ambient Intell. Humaniz. Comput.* 12, (2021), 4449–4460. https://doi.org/10.1007/s12652-019-01593-8

[217] Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Am. Sci.* 89, 4 (2001), 344–350. https://doi.org/10.1511/2001.4.344

[218] Ernst Poppel. 1978. Time perception. *Handb. Sens. Physiol.* 8, (1978), 713–729. https://doi.org/10.1016/s1364-6613(97)01008-5

[219] Ernst Pöppel. 1997. A hierarchical model of temporal perception. *Trends Cogn. Sci.* 1, 2 (1997), 56–61. https://doi.org/10.1007/978-3-642-46354-9_23

[220] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. fusion* 37, (2017), 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

[221] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and Q M Jonathan Wu. 2022. A review of generalized zero-shot learning methods. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 4 (2022), 4051–4070. https://doi.org/10.1109/tpami.2022.3191696

[222] Tie Qiu, Jiancheng Chi, Xiaobo Zhou, Zhaolong Ning, Mohammed Atiquzzaman, and Dapeng Oliver Wu. 2020. Edge computing in industrial internet of things: Architecture, advances and challenges. *IEEE Commun. Surv. Tutorials* 22, 4 (2020), 2462–2488. https://doi.org/10.1109/comst.2020.3009103

[223] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[224] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[225] Rahul Rai and Chandan K Sahu. 2020. Driven by data or derived through physics? a review of hybrid physics guided machine learning techniques with cyber-physical system (cps) focus. *IEEe Access* 8, (2020), 71050–71073. https://doi.org/10.1109/access.2020.2987324

[226] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* 378, (2019), 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

[227] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign language recognition: A deep survey. *Expert Syst. Appl.* 164, (2021), 113794. https://doi.org/10.1016/j.eswa.2020.113794

[228] Abhisek Ray, Maheshkumar H Kolekar, Raman Balasubramanian, and Adel Hafiane. 2023. Transfer learning enhanced vision-based human activity recognition: a decade-long analysis. *Int. J. Inf. Manag. Data Insights* 3, 1 (2023), 100142. https://doi.org/10.1016/j.jjimei.2022.100142

[229] Partha Pratim Ray. 2018. A survey on Internet of Things architectures. *J. King Saud Univ. Inf. Sci.* 30, 3 (2018), 291–319. https://doi.org/10.1016/j.jksuci.2016.10.003

[230] Mohammad Abdur Razzaque, Marija Milojevic-Jevric, Andrei Palade, and Siobhán Clarke. 2015. Middleware for internet of things: a survey. *IEEE Internet things J.* 3, 1 (2015), 70–95. https://doi.org/10.1109/jiot.2015.2498900

[231] Beanbonyka Rim, Nak-Jun Sung, Sedong Min, and Min Hong. 2020. Deep learning in physiological signal data: A survey. *Sensors* 20, 4 (2020), 969. https://doi.org/10.3390/s20040969

[232] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2019. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Trans. Knowl. Data Eng.* 33, 4 (2019), 1328–1347. https://doi.org/10.1109/tkde.2019.2946162

[233] Zhang Rui and Zheng Yan. 2018. A survey on biometric authentication: Toward secure and privacy-preserving identification. *IEEE access* 7, (2018), 5994–6009. https://doi.org/10.1109/access.2018.2889996

[234] James A Russell. 1980. A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 6 (1980), 1161. https://doi.org/10.1037/h0077714

[235] Michael S Ryoo and Jake K Aggarwal. 2009. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *2009 IEEE 12th international conference on computer vision*, 2009. IEEE, 1593–1600. https://doi.org/10.1109/iccv.2009.5459361

[236] Fariba Sadri. 2011. Ambient intelligence: A survey. *ACM Comput. Surv.* 43, 4 (2011), 1–66. https://doi.org/10.1145/1978802.1978815

[237] Stanisław Saganowski, Bartosz Perz, Adam G Polak, and Przemysław Kazienko. 2022. Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review. *IEEE Trans. Affect. Comput.* 14, 3 (2022), 1876–1897. https://doi.org/10.1109/taffc.2022.3176135

[238] Muhammad Sajid, Imtiaz Ahmad Taj, Usama Ijaz Bajwa, and Naeem Iqbal Ratyal. 2018. Facial Asymmetry-Based Age Group Estimation: Role in Recognizing Age-Separated Face Images. *J. Forensic Sci.* 63, 6 (2018), 1727–1749. https://doi.org/10.1111/1556-4029.13798

[239] Tausifa Jan Saleem and Mohammad Ahsan Chishti. 2021. Deep learning for the internet of things: Potential benefits and use-cases. *Digit. Commun. Networks* 7, 4 (2021), 526–542. https://doi.org/10.1016/j.dcan.2020.12.002

[240] Najmeh Samadiani, Guangyan Huang, Borui Cai, Wei Luo, Chi-Hung Chi, Yong Xiang, and Jing He. 2019. A review on automatic facial expression recognition systems assisted by multimodal sensor data. *Sensors* 19, 8 (2019), 1863. https://doi.org/10.3390/s19081863

[241] Francisco Luque Sánchez, Isabelle Hupont, Siham Tabik, and Francisco Herrera. 2020. Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. *Inf. Fusion* 64, (2020), 318–335. https://doi.org/10.1016/j.inffus.2020.07.008

[242] Iqbal H Sarker. 2021. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Comput. Sci.* 2, 6 (2021), 420. https://doi.org/10.1007/s42979-021-00815-1

[243] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proc. ACM Human-Computer Interact.* 3, CSCW (2019), 1–33. https://doi.org/10.1145/3359246

[244] Bill N Schilit and Marvin M Theimer. 1994. Disseminating active map information to mobile hosts. *IEEE Netw.* 8, 5 (1994), 22–32. https://doi.org/10.1109/65.313011

[245] Kim Schouten and Flavius Frasincar. 2015. Survey on aspect-level sentiment analysis. *IEEE Trans. Knowl. Data Eng.* 28, 3 (2015), 813–830. https://doi.org/10.1109/tkde.2015.2485209

[246] Samad Sepasgozar, Reyhaneh Karimi, Leila Farahzadi, Farimah Moezzi, Sara Shirowzhan, Sanee M. Ebrahimzadeh, Felix Hui, and Lu Aye. 2020. A systematic content review of artificial intelligence and the internet of things applications in smart home. *Appl. Sci.* 10, 9 (2020), 3074. https://doi.org/10.3390/app10093074

[247] Pallavi Sethi and Smruti R Sarangi. 2017. Internet of things: architectures, protocols, and applications. *J. Electr. Comput. Eng.* 2017, 1 (2017), 9324035. https://doi.org/10.1155/2017/9324035

[248] Ajay Shrestha and Ausif Mahmood. 2019. Review of deep learning algorithms and architectures. *IEEE access* 7, (2019), 53040–53065. https://doi.org/10.1109/access.2019.2912200

[249] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A review of emotion recognition using physiological signals. *Sensors* 18, 7 (2018), 2074. https://doi.org/10.3390/s18072074

[250] Alexandru-Ionuț Șiean, Cristian Pamparău, Arthur Sluÿters, Radu-Daniel Vatavu, and Jean Vanderdonckt. 2023. Flexible gesture input with radars: systematic literature review and taxonomy of radar sensing integration in ambient intelligence environments. *J. Ambient Intell. Humaniz. Comput.* 14, 6 (2023), 7967–7981. https://doi.org/10.1007/s12652-023-04606-9

[251] Karen Simonyan. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv Prepr. arXiv1409.1556* (2014). https://doi.org/10.48550/arXiv.1409.1556

[252] Vishwanath A Sindagi and Vishal M Patel. 2018. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognit. Lett.* 107, (2018), 3–16. https://doi.org/10.1016/j.patrec.2017.07.007

[253] Vasileios Skaramagkas, Giorgos Giannakakis, Emmanouil Ktistakis, Dimitris Manousos, Ioannis Karatzanis, Nikolaos S Tachos, Evanthia Tripoliti, Kostas Marias, Dimitrios I Fotiadis, and Manolis Tsiknakis. 2021. Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Rev. Biomed. Eng.* 16, (2021), 260–277. https://doi.org/10.1109/rbme.2021.3066072

[254] Sijie Song and Tao Mei. 2018. When multimedia meets fashion. *IEEE Multimed.* 25, 3 (2018), 102–108. https://doi.org/10.1109/mmul.2018.2875860

[255] Xuemeng Song, Xianjing Han, Yunkai Li, Jingyuan Chen, Xin-Shun Xu, and Liqiang Nie. 2019. GP-BPR: Personalized compatibility modeling for clothing matching. In *Proceedings of the 27th ACM international conference on multimedia*, 2019. 320–328. https://doi.org/10.1145/3343031.3350956

[256] Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. 2023. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Comput. Surv.* 55, 13s (2023), 1–40. https://doi.org/10.1145/3582688

[257] Alireza Souri, Aseel Hussien, Mahdi Hoseyninezhad, and Monire Norouzi. 2022. A systematic review of IoT communication strategies for an efficient smart environment. *Trans. Emerg. Telecommun. Technol.* 33, 3 (2022), e3736. https://doi.org/10.1002/ett.3736

[258] Benjamin Stephens-Fripp, Fazel Naghdy, David Stirling, and Golshah Naghdy. 2017. Automatic affect perception based on body gait and posture: A survey. *Int. J. Soc. Robot.* 9, (2017), 617–641. https://doi.org/10.1007/s12369-017-0427-6

[259] Elisa Straulino, Cristina Scarpazza, and Luisa Sartori. 2023. What is missing in the study of emotion expression? *Front. Psychol.* 14, (2023), 1158136. https://doi.org/10.3389/fpsyg.2023.1158136

[260] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe. 2016. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Trans. Affect. Comput.* 9, 2 (2016), 147–160. https://doi.org/10.1109/taffc.2016.2625250

[261] Boštjan Šumak, Saša Brdnik, and Maja Pušnik. 2021. Sensors and artificial intelligence methods and algorithms for human–computer intelligent interaction: A systematic mapping study. *Sensors* 22, 1 (2021), 20. https://doi.org/10.3390/s22010020

[262] Richard S Sutton. 2018. Reinforcement learning: an introduction. *A Bradford B.* (2018).

[263] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. 2018. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, 2018. Springer, 270–279. https://doi.org/10.1109/tnnls.2022.3160699

[264] Divyansh Thakur, Jaspal Kaur Saini, and Srikant Srinivasan. 2023. DeepThink IoT: the strength of deep learning in internet of things. *Artif. Intell. Rev.* 56, 12 (2023), 14663–14730. https://doi.org/10.1007/s10462-023-10513-4

[265] Adam Thelen, Xiaoge Zhang, Olga Fink, Yan Lu, Sayan Ghosh, Byeng D Youn, Michael D Todd, Sankaran Mahadevan, Chao Hu, and Zhen Hu. 2022. A comprehensive review of digital twin—part 1: modeling and twinning enabling technologies. *Struct. Multidiscip. Optim.* 65, 12 (2022), 354. https://doi.org/10.1007/s00158-022-03425-4

[266] Myo Thida, Yoke Leng Yong, Pau Climent-Pérez, How-lung Eng, and Paolo Remagnino. 2013. A literature review on video analytics of crowded scenes. *Intell. Multimed. Surveill. Curr. Trends Res.* (2013), 17–36. https://doi.org/10.1007/978-3-642-41512-8_2

[267] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. 2023. Recovering 3d human mesh from monocular images: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023). https://doi.org/10.1109/tpami.2023.3298850

[268] J Richard Udry. 1994. The nature of gender. *Demography* 31, 4 (1994), 561–573. https://doi.org/10.2307/2061790

[269] Silvia Liberata Ullo and Ganesh Ram Sinha. 2020. Advances in smart environment monitoring systems using IoT and sensors. *Sensors* 20, 11 (2020), 3113. https://doi.org/10.3390/s20113113

[270] Arthur H Van Bunningen, Ling Feng, and Peter M G Apers. 2005. Context for ubiquitous data management. In *International Workshop on Ubiquitous Data Management*, 2005. IEEE, 17–24. https://doi.org/10.1109/udm.2005.7

[271] Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Mach. Learn.* 109, 2 (2020), 373–440. https://doi.org/10.1007/s10994-019-05855-6

[272] Greg Van Houdt, Carlos Mosquera, and Gonzalo Nápoles. 2020. A review on the long short-term memory model. *Artif. Intell. Rev.* 53, 8 (2020), 5929–5955. https://doi.org/10.1007/s10462-020-09838-1

[273] Joaquin Vanschoren. 2018. Meta-learning: A survey. *arXiv Prepr. arXiv1810.03548* (2018). https://doi.org/10.48550/arXiv.1810.03548

[274] Ashish Vaswani. 2017. Attention is all you need. *arXiv Prepr. arXiv1706.03762* (2017).

[275] Radu-Daniel Vatavu. 2022. Are ambient intelligence and augmented reality two sides of the same coin? Implications for human-computer interaction. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022. 1–8. https://doi.org/10.1145/3491101.3519710

[276] Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Trans. Affect. Comput.* 5, 3 (2014), 273–291. https://doi.org/10.1109/taffc.2014.2330816

[277] David D Walden. 2015. Systems engineering handbook: A guide for system life cycle processes and activities. *(No Title)* (2015).

[278] Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, and Songyao Zhang. 2023. Review of large vision models and visual prompt engineering. *Meta-Radiology* (2023), 100047. https://doi.org/10.1016/j.metrad.2023.100047

[279] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognit. Lett.* 119, (2019), 3–11. https://doi.org/10.1016/j.patrec.2018.02.010

[280] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, and Yankai Lin. 2024. A survey on large language model based autonomous agents. *Front. Comput. Sci.* 18, 6 (2024), 186345. https://doi.org/10.1007/s11704-024-40231-1

[281] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, and Wei Zhang. 2022. A systematic review on affective computing: Emotion models, databases, and recent advances. *Inf. Fusion* 83, (2022), 19–52. https://doi.org/10.1016/j.inffus.2022.03.009

[282] Yilun Wang and Michal Kosinski. 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *J. Pers. Soc. Psychol.* 114, 2 (2018), 246. https://doi.org/10.1037/pspa0000098

[283] Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv Prepr. arXiv2401.06805* (2024). https://doi.org/10.48550/arXiv.2401.06805

[284] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. 2021. A comprehensive review of speech emotion recognition systems. *IEEE access* 9, (2021), 47795–47814. https://doi.org/10.1109/access.2021.3068045

[285] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* 55, 7 (2022), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

[286] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, and Donald Metzler. 2022. Emergent abilities of large language models. *arXiv Prepr. arXiv2206.07682* (2022).

[287] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* 35, (2022), 24824–24837. https://doi.org/10.48550/arXiv.2206.07682

[288] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *J. Big data* 3, (2016), 1–40. https://doi.org/10.1186/s40537-016-0043-6

[289] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. 2022. Transformers in time series: A survey. *arXiv Prepr. arXiv2202.07125* (2022). https://doi.org/10.48550/arXiv.2202.07125

[290] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. 2023. Data collection and quality challenges in deep learning: A data-centric ai perspective. *VLDB J.* 32, 4 (2023), 791–813. https://doi.org/10.1007/s00778-022-00775-9

[291] Alfred North Whitehead. 2010. *Process and reality*. Simon and Schuster.

[292] Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. 2022. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Comput. Surv.* 55, 4 (2022), 1–37. https://doi.org/10.1145/3514228

[293] Joshua Wilt and William Revelle. 2015. Affect, behaviour, cognition and desire in the Big Five: An analysis of item content and structure. *Eur. J. Pers.* 29, 4 (2015), 478–497. https://doi.org/10.1002/per.2002

[294] Li-Fang Wu, Qi Wang, Meng Jian, Yu Qiao, and Bo-Xuan Zhao. 2021. A comprehensive review of group activity recognition in videos. *Int. J. Autom. Comput.* 18, 3 (2021), 334–350. https://doi.org/10.1007/s11633-020-1258-8

[295] Shihao Xu, Jing Fang, Xiping Hu, Edith Ngai, Wei Wang, Yi Guo, and Victor C M Leung. 2022. Emotion recognition from gait analyses: Current research and future directions. *IEEE Trans. Comput. Soc. Syst.* 11, 1 (2022), 363–377. https://doi.org/10.1109/tcss.2022.3223251

[296] Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang. 2022. Emotion recognition for multiple context awareness. In *European conference on computer vision*, 2022. Springer, 144–162. https://doi.org/10.1007/978-3-031-19836-6_9

[297] Suorong Yang, Weikang Xiao, Mengchen Zhang, Suhan Guo, Jian Zhao, and Furao Shen. 2022. Image data augmentation for deep learning: A survey. *arXiv Prepr. arXiv2204.08610* (2022). https://doi.org/10.48550/arXiv.2204.08610

[298] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Adv. Neural Inf. Process. Syst.* 36, (2024).

[299] Mais Yasen and Shaidah Jusoh. 2019. A systematic review on hand gesture recognition techniques, challenges and applications. *PeerJ Comput. Sci.* 5, (2019), e218. https://doi.org/10.7717/peerj-cs.218

[300] Kun Yi, Qi Zhang, Longbing Cao, Shoujin Wang, Guodong Long, Liang Hu, Hui He, Zhendong Niu, Wei Fan, and Hui Xiong. 2023. A survey on deep learning based time series analysis with frequency transformation. *arXiv Prepr. arXiv2302.02173* (2023). https://doi.org/10.48550/arXiv.2302.02173

[301] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv Prepr. arXiv2306.13549* (2023). https://doi.org/10.48550/arXiv.2306.13549

[302] Ashkan Yousefpour, Caleb Fung, Tam Nguyen, Krishna Kadiyala, Fatemeh Jalali, Amirreza Niakanlahiji, Jian Kong, and Jason P Jue. 2019. All one needs to know about fog computing and related edge computing paradigms: A complete survey. *J. Syst. Archit.* 98, (2019), 289–330. https://doi.org/10.1016/j.sysarc.2019.02.009

[303] E Zelkha. 1998. The future of information appliances and consumer devices. *Palo Alto Ventur. Palo Alto, Calif.* (1998).

[304] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-centric artificial intelligence: A survey. *arXiv Prepr. arXiv2303.10158* (2023). https://doi.org/10.48550/arXiv.2303.10158

[305] Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. 2019. A comprehensive survey of vision-based human action recognition methods. *Sensors* 19, 5 (2019), 1005. https://doi.org/10.3390/s19051005

[306] Jing Zhang and Dacheng Tao. 2020. Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet Things J.* 8, 10 (2020), 7789–7817. https://doi.org/10.1109/jiot.2020.3039359
[307] Jingbin Zhang, Meng Ma, Ping Wang, and Xiao-dong Sun. 2021. Middleware for the Internet of Things: A survey on requirements, enabling technologies, and solutions. *J. Syst. Archit.* 117, (2021), 102098. https://doi.org/10.1016/j.sysarc.2021.102098
[308] Kexin Zhang, Qingsong Wen, Chaoli Zhang, Rongyao Cai, Ming Jin, Yong Liu, James Y Zhang, Yuxuan Liang, Guansong Pang, and Dongjin Song. 2024. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. *IEEE Trans. Pattern Anal. Mach. Intell.* (2024). https://doi.org/10.1109/tpami.2024.3387317
[309] Shugang Zhang, Zhiqiang Wei, Jie Nie, Lei Huang, Shuang Wang, and Zhen Li. 2017. A review on human activity recognition using vision-based method. *J. Healthc. Eng.* 2017, 1 (2017), 3090343. https://doi.org/10.1155/2017/3090343
[310] Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.* 34, 12 (2021), 5586–5609. https://doi.org/10.1109/tkde.2021.3070203
[311] Liang Zhao. 2021. Event prediction in the big data era: A systematic survey. *ACM Comput. Surv.* 54, 5 (2021), 1–37. https://doi.org/10.1145/3450287
[312] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, and Zican Dong. 2023. A survey of large language models. *arXiv Prepr. arXiv2303.18223* (2023). https://doi.org/10.48550/arXiv.2303.18223
[313] Xiaoming Zhao, Zhiwei Tang, and Shiqing Zhang. 2022. Deep personality trait recognition: a survey. *Front. Psychol.* 13, (2022), 839619. https://doi.org/10.3389/fpsyg.2022.839619
[314] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, and Lifang He. 2023. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv Prepr. arXiv2302.09419* (2023). https://doi.org/10.48550/arXiv.2302.09419
[315] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proc. IEEE* 109, 1 (2020), 43–76. https://doi.org/10.1109/jproc.2020.3004555
[316] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. 2023. Object detection in 20 years: A survey. *Proc. IEEE* 111, 3 (2023), 257–276. https://doi.org/10.1109/jproc.2023.3238524
[317] https://openai.com/index/hello-gpt-4o/