

Towards Smarter Small Systems: Vision-Based AI, Edge Computing, and XAI for Next-Generation Industrial Applications

Ivan Ćirić, Zlatan Car and Jelena Milojković

Abstract – The rapid evolution of Industry 5.0 places humans back at the center of intelligent systems, demanding AI technologies that are not only powerful, but also explainable, trustworthy, and deployable in real-time on resource-constrained devices. This talk presents a unified perspective on edge-enabled artificial intelligence covering three emerging domains: sports analytics, collaborative robotics, and smart energy systems. We first discuss the development of vision-based deep learning models for on-device recognition of basketball events, human pose, and gesture dynamics, enabling real-time feedback and laying the foundation for AI-powered coaching assistants. We then transition to collaborative robotics, where robust perception of human motion and hand articulation becomes essential for safe and intuitive human–robot interaction in industrial environments.

Since these tasks inherently unfold over time, we explore sequence modeling and time-series prediction, including LSTM-based and hybrid approaches used for forecasting human motion as well as for large-scale applications such as predictive control in district heating systems. Finally, we emphasize the role of Explainable AI (XAI) in bridging the gap between model performance, decision transparency, and operational trust—particularly in high-impact engineering systems. The talk highlights how edge computing, human-centric design, and XAI jointly shape the next generation of intelligent small-scale systems.

Keywords – Artificial Intelligence, Machine Vision, Edge Computing, eXplainable Artificial Intelligence.

I. FROM INDUSTRY 4.0 TO INDUSTRY 5.0

The transition from Industry 4.0 to Industry 5.0 represents a fundamental shift from technology-driven optimization toward human-centered intelligent ecosystems. While Industry 4.0 is defined by digitalization, automation, cyber-physical systems, and large-scale data analytics, its primary objective is operational efficiency and system autonomy [1]. Industry 5.0 does not negate these technological advances; rather, it reframes them within a broader socio-technical context that emphasizes collaboration, sustainability, resilience, and ethical responsibility.

Ivan Ćirić, University of Niš, Faculty of Mechanical Engineering, ivan.ciric@masfak.ni.ac.rs, Zlatan Car, Faculty of Engineering, Catholic University of Croatia, Zagreb, zlatan.car@unicath.hr, Jelena Milojković, University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 4, Niš, Serbia, jelena.milojkovic@elfak.ni.ac.rs.

In this paradigm, artificial intelligence is no longer conceived merely as an automation instrument, but as an augmentation technology that enhances human capabilities. Central to this transformation is computer vision, which becomes the perceptual backbone of human–machine collaboration. Unlike traditional automation architectures that operate based on predefined rules and sensor thresholds, Industry 5.0 systems require contextual awareness – the ability to interpret complex visual scenes, understand human behavior, and adapt dynamically to unstructured environments.

Computer vision enables machines to transition from passive tools to perceptive agents. Through advanced deep learning architectures, including convolutional neural networks, transformer-based perception models, and multimodal fusion strategies, industrial systems can detect and track human posture, recognize gestures, estimate motion trajectories, and infer intent. Such capabilities are indispensable in collaborative robotic environments, where cobots must operate safely and responsively in close physical proximity to workers [1,2]. Real-time visual tracking ensures that robotic motion planning is continuously adapted to human presence, thereby reducing collision risk and increasing operational flexibility.

Beyond safety, vision-based AI enables semantic understanding of tasks. In intelligent workspaces, visual systems can classify work actions, monitor ergonomic posture, detect anomalies in behavior, and provide real-time decision support. This transforms production lines into adaptive environments where the system responds to human expertise rather than replacing it. In contrast to Industry 4.0’s emphasis on autonomous execution, Industry 5.0 promotes shared autonomy, where perception-driven AI assists and augments human operators (Fig.1).

The expansion of computer vision also underpins the shift from mass production to mass personalization [1,2]. Modern visual inspection systems enable automatic recognition of product variants, verification of customized components, and real-time adjustment of assembly parameters. Vision-guided robotics thus supports highly flexible production without sacrificing efficiency. Digital twins enriched with real-time visual data enable continuous synchronization between physical and digital environments, strengthening predictive maintenance and adaptive planning.

Sustainability objectives further amplify the importance

of computer vision. Intelligent visual systems reduce waste through early defect detection, optimize material consumption by monitoring process deviations, and support automated sorting and recycling operations using classification networks capable of fine-grained material recognition. Vision-based monitoring also contributes to energy efficiency by detecting equipment misuse, overheating, or irregular operational patterns.

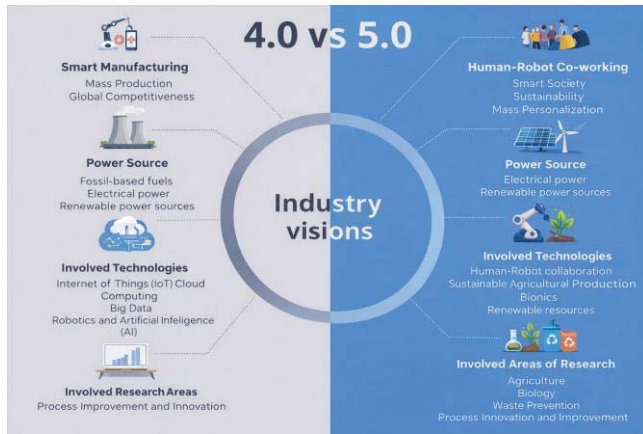


Fig. 1. Key differences between Industry 4.0 and Industry 5.0

However, as computer vision systems assume greater autonomy and responsibility, the necessity of explainability becomes unavoidable. In collaborative settings, AI must justify why a robot altered its trajectory, why a safety stop was activated, or why a predictive recommendation was issued. Explainable computer vision techniques - including saliency mapping, feature attribution, and attention visualization - allow stakeholders to understand which visual cues influenced decision-making. Trust, therefore, arises not solely from accuracy metrics but from transparent reasoning processes.

The evolution from Industry 4.0 to Industry 5.0 signifies a transformation from efficiency-centered automation toward perceptive, adaptive, and trustworthy intelligent systems. Computer vision is no longer an auxiliary sensing modality; it is the cognitive interface between humans and machines. By enabling contextual awareness, safe collaboration, personalization, and sustainability - while remaining interpretable and accountable - vision-based AI systems define the technological core of Industry 5.0.

II. VISION-BASED AI IN INDUSTRY 5.0

Vision-based artificial intelligence constitutes one of the foundational technological pillars of Industry 5.0, not merely as a sensing modality, but as an enabling interface between humans, machines, and complex adaptive systems. Unlike traditional automation frameworks that rely predominantly on predefined logic and scalar measurements, vision-based systems operate on rich

spatial-temporal data streams, enabling contextual interpretation of dynamic environments. In this sense, computer vision provides machines with structured perceptual awareness, transforming raw visual input into semantically meaningful representations that can drive intelligent decision-making (Figure 2).

From a theoretical standpoint, computer vision encompasses mathematical modeling of image formation, geometric transformations, camera calibration, and three-dimensional reconstruction. As elaborated in foundational treatments of machine vision systems [3-6], the transformation from physical scene to actionable information requires a precisely orchestrated pipeline: sensor configuration, illumination control, image acquisition, filtering, feature extraction, segmentation, classification, and integration into higher-level control algorithms. In Industry 5.0, this pipeline extends beyond inspection and quality assurance toward human-aware system design.

A central role of vision-based AI in Industry 5.0 lies in human-state perception. Advanced perception models allow the detection and tracking of full-body pose, limb articulation, gaze direction, and hand gestures. Deep convolutional architectures, pose estimation networks, and transformer-based keypoint detectors enable continuous tracking of multiple individuals under complex environmental conditions. These perception modules feed into temporal filtering or sequential models that encode motion dynamics and behavioral patterns.



Fig. 2. xArm 5 with mounted camera doing manipulation task

What is particularly noteworthy is that identical perceptual mechanisms underlie both industrial and sports applications. In sports analytics, multi-camera vision systems reconstruct player trajectories, compute biomechanical parameters, and estimate kinematic variables with high temporal resolution. These same geometric and statistical principles are directly transferable

to collaborative robotic environments, where the spatial relationship between a human worker and robotic manipulator must be continuously assessed. Object detection, skeletal tracking, and motion prediction algorithms that quantify athletic performance are fundamentally equivalent to those that ensure safe robotic interaction in shared workspaces (Figure 3, Figure 4) [7-10].

The shared methodological foundation includes perspective projection models, intrinsic and extrinsic camera calibration, stereo triangulation, epipolar geometry, and point-cloud reconstruction. In sports contexts, these tools enable high-fidelity reconstruction of movement across large arenas. In industrial mechatronics, they facilitate precise localization of workpieces, dynamic workspace mapping, and robot pose estimation. Thus, sports tracking systems and industrial vision systems converge at the level of geometric modeling and probabilistic estimation.

Another dimension of convergence emerges in temporal reasoning. Both domains require robust motion modeling capable of handling occlusions, abrupt accelerations, and environmental variability. Kalman filtering, particle filtering, and recurrent neural networks are applied to stabilize trajectories and infer hidden states. In basketball analysis, such models reconstruct continuous motion even under partial visibility. In collaborative robotics, similar temporal smoothing ensures reliable estimation of human proximity and intent, directly influencing safety protocols and adaptive control laws.

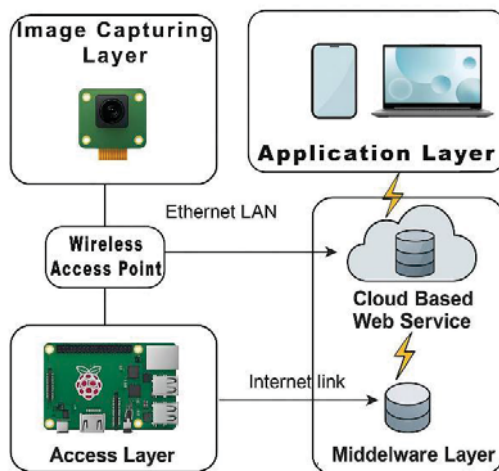


Fig. 3. Smart IoT basketball vision system – KošKo [7]

Industry 5.0 further requires real-time performance under embedded constraints. Modern vision systems are increasingly deployed on edge platforms equipped with specialized accelerators, such as embedded GPUs or AI-specific inference chips. Model compression, quantization, and pruning techniques enable efficient execution of deep learning architectures without compromising perceptual fidelity. This is particularly critical in collaborative

environments, where decision latency must remain within strict millisecond-level thresholds to guarantee safe human-machine interaction.

Robustness in unstructured environments represents another key challenge. Illumination variability, dynamic backgrounds, reflective surfaces, and partial occlusions complicate reliable perception. Consequently, adaptive preprocessing, data augmentation, domain adaptation strategies, and sensor fusion approaches become essential. The integration of RGB cameras with depth sensors, LiDAR systems, inertial measurement units, or force sensors provides complementary information, enhancing stability and resilience in complex operational scenarios [4,6,11].



Fig. 4. Detection of basketball players using YOLOv8

Vision-based AI in Industry 5.0 is not confined to static observation; it actively participates in closed-loop control. In visual servoing architectures, image features directly influence actuator commands, enabling dynamic adjustment of robotic trajectories. Position-based and image-based control strategies translate visual discrepancies into corrective motor signals. In sports analytics, although not traditionally framed as control, visual feedback loops similarly guide performance optimization by providing athletes and coaches with precise kinematic insights.

Furthermore, perceptual systems contribute to personalization and adaptability. In flexible manufacturing systems, visual recognition modules enable automatic configuration changes for different product variants. In sports, individualized biomechanical profiling allows adaptation of training regimens based on data-driven motion analysis. In both cases, visual perception acts as a mediator between heterogeneous physical inputs and adaptive algorithmic responses.

Beyond detection and tracking, Industry 5.0 vision systems increasingly incorporate semantic scene understanding. Scene segmentation models, attention mechanisms, and contextual reasoning networks allow systems to distinguish between task-relevant and irrelevant entities within complex environments. This semantic awareness enables prioritization of salient objects, prediction of interaction zones, and efficient allocation of computational resources.

Finally, the continued advancement of machine learning techniques profoundly influences vision-based AI capabilities. Deep neural networks trained on large-scale datasets provide generalized feature representations that

outperform handcrafted descriptors. Transfer learning and self-supervised learning approaches reduce dependency on extensive labeled data, while graph-based models capture relational structures among detected entities. These advances broaden the applicability of vision systems from structured industrial lines to dynamic, human-populated environments.

III. POSE AND GESTURE ESTIMATION ON EDGE DEVICES

Human pose estimation represents a central problem in modern computer vision, particularly within systems that require understanding of human motion and interaction. As discussed in [1,12], pose estimation aims to determine the spatial configuration of key anatomical landmarks from 2D images or video sequences. These landmarks typically include joints such as shoulders, elbows, wrists, hips, knees, and ankles, forming a skeletal representation of the human body.

In two-dimensional pose estimation, deep neural networks infer the image coordinates of keypoints, often accompanied by confidence measures that quantify detection reliability [13]. Two dominant methodological paradigms are recognized: top-down and bottom-up approaches. Top-down methods first detect the person and then estimate the pose within the detected bounding box, ensuring high precision but potentially increasing computational cost. Bottom-up approaches detect all keypoints in the scene before associating them into full-body skeletons, offering better scalability in multi-person environments.

Recent architectures integrate convolutional neural networks with attention mechanisms, enabling robust localization even under partial occlusion or varying illumination conditions. Systems such as YOLO-Pose and OpenPose demonstrate the feasibility of real-time skeletal detection in complex dynamic scenes [1]. Pose representations commonly rely on structured graphs where nodes correspond to keypoints and edges encode anatomical constraints, preserving biomechanical consistency.



Fig. 5. Real-time skeletal detection using YOLO-Pose [1]

Beyond static pose estimation, motion analysis introduces temporal coherence, allowing the computation of joint angles, angular velocities, and trajectory smoothness. The estimation of joint angles from skeletal keypoints provides quantitative biomechanical descriptors applicable to sports analytics, rehabilitation, and ergonomic assessment [1]. Pose confidence metrics play a crucial role in filtering unreliable detections and ensuring robust downstream inference.

Gesture recognition extends pose estimation by incorporating dynamic temporal patterns. While pose estimation captures instantaneous body configuration, gesture recognition interprets motion sequences across time. Techniques such as recurrent neural networks, temporal convolutional networks, and transformer-based architectures enable modeling of motion dependencies across successive frames. This temporal modeling is essential for distinguishing between similar static poses that differ only in motion trajectory.

In collaborative robotic environments, pose and gesture estimation contribute directly to safety and interaction fluency. Detection of arm extension, stepping motion, or hand signaling allows robotic systems to infer human intent. Gesture vocabularies can be formally defined and mapped to control commands, enabling natural human–robot communication channels.

Robust pose and gesture systems must operate under varying camera viewpoints, scale changes, and illumination variability. Calibration procedures, camera modeling, and perspective geometry remain fundamental for accurate spatial interpretation [1,11]. The integration of stereo vision and depth sensing further enhances pose reliability by introducing three-dimensional consistency constraints.

Moreover, pose estimation frameworks increasingly incorporate multimodal sensing. Fusion of RGB data with depth sensors or inertial measurement units improves resilience in challenging environmental conditions. Such multimodal strategies reduce sensitivity to occlusions and improve motion continuity estimation.

An additional challenge arises in multi-person environments, where skeletal association must prevent identity switching across frames. Tracking-by-detection strategies integrate object detection outputs with motion prediction filters to maintain stable identity assignment (Figure 6).

Overall, pose and gesture estimation constitute a high-level perception module that bridges raw image acquisition and semantic interpretation of human behavior. They enable quantitative modeling of kinematics, support interaction analysis, and serve as the perceptual foundation for adaptive and human-aware intelligent systems.

While high-performance GPU clusters facilitate complex training procedures, practical deployment increasingly requires inference on edge devices with constrained computational resources. Edge deployment

imposes strict requirements on model size, memory footprint, power consumption, and latency. In collaborative robotics and sports applications, inference delay must remain within real-time constraints, often below 20–30 milliseconds per frame (Figure 7).

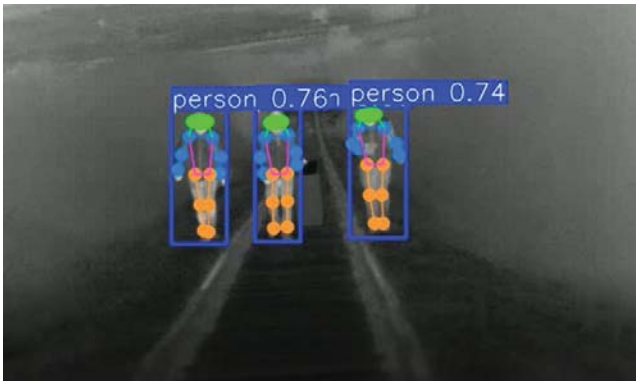


Fig. 6. Real-time skeletal detection of multiple persons, on a railtracks, while image is acquired using thermal camera

Model optimization techniques play a critical role in enabling edge compatibility. Network pruning reduces redundant parameters while preserving structural integrity of learned filters. Quantization compresses weight representations from floating-point to lower-bit precision formats, significantly decreasing memory consumption and computation load. Knowledge distillation enables training of smaller student networks that approximate the performance of larger teacher models.

Efficient backbone architectures such as lightweight CNNs or scaled network families (e.g., EfficientNet variants) provide favorable trade-offs between accuracy and computational cost. Pipeline optimization-including batch normalization fusion, hardware-aware scheduling, and inference acceleration libraries-ensures that edge processors maintain stable frame rates.



Fig. 7. Real-time skeletal detection of basketball player using YOLO-Pose on Koško smart basketball system

In embedded vision systems, hardware selection is equally significant. Devices equipped with CUDA-enabled GPUs or AI accelerators support parallel convolution

operations required for skeletal detection. Careful architectural co-design between neural networks and hardware allows deterministic latency, which is essential in safety-critical environments. Edge deployment also enhances data privacy by limiting external transmission of human-motion data.

The broader relevance of pose and gesture estimation becomes evident when considered in the functional classification of machine vision in mechatronics. As emphasized in the discussion of human-machine interaction applications [1], visual perception enables systems not only to detect objects but also to interpret human presence and behavior. Pose estimation supports collaborative workflows, ergonomic monitoring, and adaptive robotic control.

Furthermore, real-time object detection and embedded deployment, including implementations on various platforms such as NVIDIA Jetson Orin Nano, demonstrate the practical feasibility of executing deep learning models directly within industrial environments [4-6, 14]. The integration of pose estimation algorithms into such embedded systems consolidates visual perception, control logic, and actuator response within a unified mechatronic framework.

In this context, pose and gesture estimation on edge devices represent not merely an academic challenge but a fully integrable subsystem within intelligent, human-aware industrial architectures.

IV. TIME-SERIES PREDICTION IN VISION-BASED APPLICATIONS

Time-series prediction represents a fundamental analytical layer in both energy systems and vision-based intelligent applications. While in district heating systems time-series models are used to forecast thermal load, supply temperature, or energy consumption [15,16], in vision-based systems temporal prediction is applied to motion trajectories, activity patterns, and dynamic scene evolution. Despite the apparent domain differences, both contexts share common methodological challenges: irregular sampling, nonlinear dynamics, regime switching, sensor noise, and delayed system responses [16].

In district heating systems, SCADA data often exhibit missing values, irregular logging intervals, and discontinuities due to sensor failures or communication interruptions [16] (Figure 8). To address such irregular temporal structures, preprocessing strategies integrate interpolation, seasonal decomposition, Kalman filtering, and physically constrained reconstruction of variables. These techniques ensure that reconstructed time series preserve operational realism while enabling reliable machine learning training [17]. A comparable issue arises in vision-based tracking systems, where frame drops, occlusions, or variable frame rates produce temporally inconsistent motion signals (Table I). Similar filtering and smoothing techniques-Kalman filters, moving averages, or

probabilistic trajectory reconstruction-are applied to restore temporal continuity.

From a modeling perspective, ensemble learners such as Random Forest and gradient boosting have demonstrated robustness on irregular, noisy operational data [16]. In district heating forecasting, Random Forest models achieve high explanatory power ($R^2 \approx 0.89$) by combining autoregressive features with environmental inputs such as outdoor temperature and supply temperature. In vision applications, comparable ensemble approaches can be used for short-term trajectory prediction, activity classification based on motion descriptors, or future pose estimation, particularly when computational stability is prioritized over deep sequence modeling complexity.

Sequential neural models-including recurrent neural networks (RNN), Long Short-Term Memory (LSTM), and temporal convolutional architectures-offer enhanced capability to capture long-term dependencies and system inertia [17]. In energy systems, autoregressive lags dominate predictive performance because future consumption is strongly conditioned by recent states [16]. Similarly, in vision-based tracking, future joint positions or object coordinates are tightly correlated with recent velocity and acceleration patterns. This structural parallel underscores that both domains are governed by state-space dynamics with inertia and regime-dependent transitions.

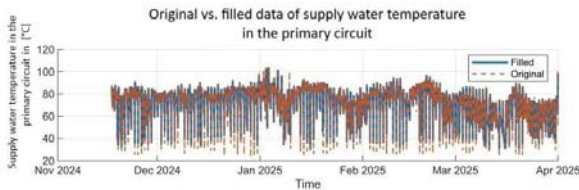


Fig. 8. Time Series Estimation in DHS System

An important consideration emphasized in irregular district heating data is the trade-off between synthetic data augmentation and preservation of realistic operating regimes [18]. Synthetic data generation (e.g., TimeGAN-based augmentation) may improve training robustness but risks introducing artificial bias. In vision-based datasets-such as human action recognition or basketball player tracking-data augmentation through geometric transformations, synthetic motion sequences, or simulation-based rendering provides similar benefits and risks. Consequently, maintaining physical plausibility and biomechanical consistency is equally critical in both domains.

Explainability further bridges energy forecasting and vision-based temporal prediction. In the district heating context, XAI methods such as SHAP and permutation importance identify dominant drivers of load dynamics, often highlighting short-term lags and operational temperatures [16]. In vision-based time-series applications, explainability can reveal which frames, joint trajectories, or velocity spikes most influence a movement classification or

intention prediction. Temporal attention mechanisms, originally applied in heating load forecasting [15,16], can likewise be embedded within sequence models for gesture recognition or activity anticipation to interpret temporal relevance.

Datasets used in vision-based time-series prediction frequently include human pose benchmarks (e.g., multi-person tracking datasets, sports analytics recordings, or human-robot interaction datasets) where annotated sequences provide ground truth for trajectory forecasting. These datasets share structural similarities with SCADA streams: multivariate signals evolving over time, influenced by environmental and operational variables. In both cases, baseline modeling enables deviation detection-whether identifying energy wastage episodes in heating systems or anomalous motion patterns in collaborative robotic spaces.

TABLE I
Change of Characteristic Angles during Basketball Shooting

Frame No.	Right elbow angle (°)	Left elbow angle(°)	Right shoulder angle(°)	Left shoulder angle(°)
1.	32.95	12.95	78.19	132.70
2.	33.03	12.73	77.99	132.25
3.	26.99	1.62	77.31	122.19
4.	75.02	14.22	98.55	65.37
5.	37.36	11.13	77.85	124.75
6.	36.39	6.63	77.16	124.75
7.	32.02	1.14	78.78	113.91
8.	29.88	0.61	76.81	115.35
9.	32.29	0.09	77.81	115.33
10.	31.03	6.45	76.95	113.03
11.	27.35	6.66	76.35	114.94
12.	22.34	12.00	74.38	102.07

Time-series prediction in vision-based applications functions as the predictive extension of perceptual intelligence. Just as district heating models learn expected consumption regimes and detect deviations from baseline [16], vision-based systems learn normal motion evolution and anticipate future states. Integrating temporal modeling with explainability ensures that predictions are not only accurate but interpretable, enabling transparent decision support in both energy and vision-driven intelligent systems.

V. EXPLAINABLE ARTIFICIAL INTELLIGENCE IN VISION-BASED HUMAN TRACKING AND TIME-SERIES PREDICTION SYSTEMS

Explainable AI (XAI) is an emerging research area of considerable importance in sports analytics, collaborative robotics, and smart infrastructure systems. The development of AI in these domains relies on computer vision, data science, machine learning, deep neural networks, and sequential modeling techniques such as

LSTMs and transformer architectures.

Despite their predictive capabilities, these models often lack interpretability. In player tracking applications, deep convolutional networks may detect shooting events or estimate pose without clarifying which biomechanical features drove the decision. In collaborative robotic systems, gesture recognition models may trigger robotic motion without explicitly indicating which features of the human posture or trajectory led to the action. Similarly, time-series forecasting models may predict future system states without revealing the temporal patterns responsible for the forecast.

XAI aims to improve both performance and explainability, thereby facilitating trust, understanding, acceptance, and control of AI systems operating alongside humans (Fig. 9).

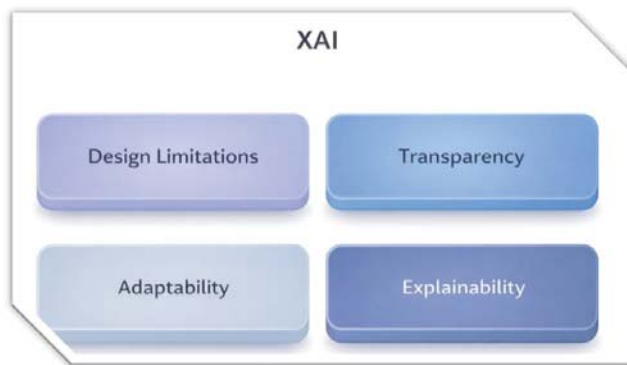


Fig. 9. Main components of Explainable AI (XAI)

Key components of XAI in such environments include design awareness, transparency, explainability, and adaptability. Through improved interpretability of feature extraction in vision models and temporal dependencies in sequence networks, XAI enables detection of bias, spurious correlations, and noise sensitivity in tracking and forecasting tasks. Transparency provides insight into data preprocessing, model structure, and inference logic, ensuring that player-tracking decisions rely on valid biomechanical cues and that robotic safety responses are triggered by meaningful human motion patterns. Explainability helps identify errors within multi-stage pipelines (detection → pose estimation → temporal modeling → action classification), facilitating validation and refinement. Adaptability allows explanations to be tailored to different users, from coaches requiring motion-level insight to engineers seeking feature-level diagnostics.

Recent literature highlights several effective interpretability techniques. Grad-CAM and LRP enable spatial attribution in convolutional models, while attention maps in vision and temporal transformers reveal spatial-temporal dependencies. SHAP and LIME provide model-agnostic feature contribution analysis, particularly useful in time-series forecasting. Complementary statistical tools—such as feature importance ranking and lag analysis—

support interpretation of LSTM-based models. In temporal applications, XAI primarily aims to identify influential time windows and sensor inputs, clarifying whether predictions are driven by short-term fluctuations or long-term dynamics.

Edge deployment introduces additional constraints due to limited computational resources. Consequently, interpretability mechanisms must be lightweight and efficient. Practical solutions include simplified surrogate models, approximate SHAP calculations, attention visualization in compact architectures, and low-resolution saliency mapping. These strategies preserve real-time performance while maintaining transparency.

In sports analytics, collaborative robotics, and predictive control, explainability is essential for trust and operational acceptance. Integrating XAI into perception and prediction pipelines transforms AI systems from opaque black boxes into transparent, human-centered decision-support tools—fully aligned with the collaborative and responsible principles of Industry 5.0.

VI. CONCLUSIONS AND FUTURE OUTLOOK

In this work, we presented a unified perspective on vision-based artificial intelligence as a core enabling technology for next-generation small intelligent systems within the Industry 5.0 paradigm. By integrating computer vision, temporal modeling, edge deployment, and explainable AI, we demonstrated how perceptual intelligence can extend beyond traditional automation toward collaborative, adaptive, and trustworthy systems. The proposed framework bridges three application domains—sports analytics, collaborative robotics, and district heating forecasting, while highlighting their common computational foundations in perception, sequence modeling, and interpretability.

The results indicate that pose and gesture estimation on edge devices provides a viable pathway toward real-time human-machine interaction, while time-series prediction frameworks allow systems to anticipate future states and optimize control strategies. Importantly, the integration of Explainable AI mechanisms ensures that predictions and control actions remain transparent and interpretable to end users. By embedding saliency mapping, SHAP-based feature attribution, temporal attention analysis, and lightweight surrogate modeling into perception and forecasting pipelines, the proposed approach advances AI systems from high-accuracy black boxes toward accountable and trustworthy decision-support tools. In this manner, the synergy of edge computing and XAI directly supports the core principles of Industry 5.0: human-centricity, safety, sustainability, and operational resilience.

Future work will focus on several directions. First, the integration of multimodal perception—combining vision with inertial, force, and environmental sensing—will enhance robustness in unstructured environments. Second, the development of adaptive edge-cloud hybrid

architectures will enable dynamic allocation of computational loads while preserving real-time constraints. Third, further research will address scalable explainability mechanisms tailored for embedded devices, ensuring that interpretability remains feasible even under strict resource limitations. Finally, we envision the integration of digital twins and self-supervised learning frameworks to enable continuous model refinement in operational settings, thereby supporting the long-term deployment of transparent, adaptive, and human-aware intelligent small systems.

ACKNOWLEDGEMENT

This research was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, under the Agreement on Financing the Scientific Research Work of Teaching Staff at the Faculty of Mechanical Engineering, University of Niš - Registration number: 451-03-136/2025-03/200109 and Grant Number: 451-03-136/2025-03/200102.

REFERENCES

- [1] Ćirić, I., Simonović, M., Pavlović, M., Cvetković, S., *Mašinska Vizija u Mehatronici*, 1st ed., 2025, ISBN: 978-86-6055-198-8.
- [2] Chen, N., et al. (2023). Real-time vision-based human hand motion tracking and grasping for a robotic manipulator with a soft hand. *Progress in Canadian Mechanical Engineering*, CSME - CFD-SC2023 May 28-31, 2023, Sherbrooke, QC, Canada
- [3] Ćirić, I., Ivačko, N., Čojbašić, Ž., Pavlović, M., Mitić, D., Lalić, S., Jevtić, D., "Comparison of Simple Object Recognition Algorithms for Robot Vision", *Proc. XVI Int. SAUM Conference on Systems, Automatic Control and Measurements*, Niš, Serbia, Nov. 17–18, 2022.
- [4] Ivačko, N., Ćirić, I., Čojbašić, Ž., Milošević, M., Jevtić, D., "Machine Learning Based Hardware Accelerated Cobot Grasping in the Food Industry", *Proc. 11th Int. Conf. on Electrical, Electronic and Computing Engineering (IcETRAN)*, Niš, Serbia, pp. 1–5, 2024.
- [5] Ivačko, N., Stojiljković, D., Jevtić, D., Simonović, M., Ćirić, I., "AI-Based Robot Vision Algorithm for Fruit Ripeness Classification", *Proc. 17th Int. Conf. on Accomplishments in Mechanical and Industrial Engineering (DEMI 2025)*, Banja Luka, Bosnia and Herzegovina, pp. 441–448, 2025.
- [6] Ivačko, N., Ćirić, I., Stojiljković, D., Čojbašić, Ž., Jevtić, D., "A Cognitive Robotics Approach for Manipulation of Freeform Objects Using CNN-Based Perception and Soft-Gripping", in *Advances in Service and Industrial Robotics: RAAD 2025*, Mechanisms and Machine Science, vol. 190, Cham, Switzerland: Springer, pp. 46–54, 2025, doi: 10.1007/978-3-032-02106-9.
- [7] Ćirić, I.; Ivačko, N.; Milić, M.; Ristić, P.; Krstić, D. Computer Vision-Driven Framework for IoT-Enabled Basketball Score Tracking. *Computers* 2025, 14, 469. <https://doi.org/10.3390/computers14110469>
- [8] Ćirić, I., Milošević, M., Ivačko, N., Pavlović, M., Krstić, D., "Intelligent Computer Vision System for Score Detection in Basketball", *FACTA UNIVERSITATIS, Series: Automatic Control and Robotics*, vol. 22, no. 2, pp. 075–085, 2023.
- [9] Milić, M., Ristić, P., Ćirić, I., Ivačko, N., Krstić, D., "Edge Computing for Sports Analytics: AI-Powered Basketball Score Detection on Raspberry Pi 5", *Proc. IEEE 34th Int. Conf. on Microelectronics (MIEL)*, 2025.
- [10] Hu, M., Hu, Q., "Design of Basketball Game Image Acquisition and Processing System Based on Machine Vision and Image Processor", *Microprocessors and Microsystems*, vol. 82, no. 1, 2021, doi: 10.1016/j.micpro.2021.103904.
- [11] Ćirić, I., Pavlović, M., Ivačko, N., Tomov, P., Nikolić, V., "Autonomous Train Operation Obstacle Detection Based on Night Vision System", *Acta Polytechnica Hungarica*, vol. 23, no. 1, pp. 307–327, Jan. 2026, doi: 10.12700/APH.23.1.2026.1.18.
- [12] Cvetkovic, S., Savic, N., Ćiric, I., "Deep Transfer Learning Approach for Robust Hand Detection", *Intelligent Automation & Soft Computing*, vol. 36, no. 1, pp. 967–979, 2022, doi: 10.32604/iasc.2023.032526.
- [13] Dai, J. Advancements in Deep Learning for Visual Object Tracking, *Appl. Comput. Eng.*, 82, 130–136, 2024.
- [14] Feng, H. et al. Benchmark Analysis of YOLO Performance on Edge Intelligence Devices, *Cryptography*, 6, 16, 2022
- [15] Zdravković, M., Ignjatović, M., Ćirić, I., "Explainable Heat Demand Forecasting for the Novel Control Strategies of District Heating Systems", *Annual Reviews in Control*, vol. 53, pp. 405–413, 2023, doi: 10.1016/j.arcontrol.2022.03.009.
- [16] Ćirić, I., Tasić, M., Ignjatović, M., Stojiljković, D., Ćirić, M., "Machine Learning Approach for Energy Consumption Estimation in District Heating Systems", *Innovative Mechanical Engineering*, vol. 3, pp. 31–47, ISSN 2812-9229.
- [17] Ćirić, M., Predić, B., Stojanović, D., Ćirić, I., "Single and Multiple Separate LSTM Neural Networks for Multiple Output Feature Purchase Prediction", *Electronics*, vol. 12, no. 12, 2616, 2023, doi: 10.3390/electronics12122616.
- [18] Tang, P., Li, Z., Wang, X., Liu, X., & Mou, P. (2025). Time series data augmentation for energy consumption data based on improved TimeGAN. *Sensors*, 25(2), 493. <https://doi.org/10.3390/s25020493>
- [19] Neubauer, A., Brandt, S., and M. Kriegel. 2025. Explainable multi-step heating load forecasting: Using SHAP values and temporal attention mechanisms for enhanced interpretability. *Energy and AI* 20:100480.