Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira, Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria

doi: 10.3850/ESREL2026061419_esrel26-p26221-cd

Achieving Cognitive Safety for Human-Robot Collaboration

Frederik Plahl
Proximity Robotics & Automation GmbH, Pfinztal, Germany.
IAS, University of Stuttgart, Stuttgart, Germany. E-mail: plahl@proximityrobotics.com

Georgios Katranis
IAS, University of Stuttgart, Stuttgart, Germany. E-mail: georgios.katranis@ias.uni-stuttgart.de

Muhammad Atif
University of Florence, Florence, Italy. E-mail: muhammad.atif@unifi.it

Anas Asswad
ResilTech S.R.L., Lecce, Italy. E-mail: anas.asswad@resiltech.com

Tommaso Zoppi
University of Florence, Florence, Italy. E-mail: tommaso.zoppi@unifi.it

Francesco Brancati
ResilTech S.R.L., Pontedera, Italy. E-mail: francesco.brancati@resiltech.com

Andrey Morozov
IAS, University of Stuttgart, Stuttgart, Germany. E-mail: andrey.morozov@ias.uni-stuttgart.de

Ilshat Mamaev
Proximity Robotics & Automation GmbH, Pfinztal, Germany. E-mail: mamaev@proximityrobotics.com

Human-Robot Collaboration introduces safety challenges that traditional 2D sensors or performance-degrading mitigations cannot adequately solve. This paper introduces the concept of CogniSafe3D, a cognitive safety system that utilizes an external 3D LiDAR to monitor the operational environment. The system's methodology is twofold: it first establishes a deterministic foundation for safety by processing LiDAR point clouds with a Signed Distance Field representation to detect safety zone violations. This core functionality is augmented by a cognitive layer that uses Human Pose Estimation and predictive algorithms to anticipate human movements and infer intent. A Risk Monitoring module synthesizes this information to provide real-time, dynamic risk estimates. This enables the robot to proactively mitigate hazards while reducing unnecessary stops, advancing perception-driven safety in compliance with ISO/TS 15066 standards. The methods developed in the CogniSafe3D research project will be implemented in a TRL-6 prototype.

Keywords: Human-Robot Collaboration, Cognitive Safety, Risk Monitoring, Safety Assessment, 3D LiDAR.

1. Introduction

AI integration into robotics has driven significant advancements across manufacturing, healthcare, and logistics, enhancing adaptability and overall performance Shah et al. (2025). Despite these gains, safety remains a fundamental requirement for Human-Robot Collaboration (HRC), and achieving certifiable safety continues to be a major industrial challenge Montini et al. (2024). Current standards emphasize deterministic behavior in collaborative applications to guarantee safety, which conflicts with the non-deterministic nature of deep learning-based systems ISO 10218-2 (2025); ISO/TS 15066 (2016).

Practically, most deployed HRC safety systems still rely on 2D sensors that lack reliable depth information; this forces risk assessments to assume

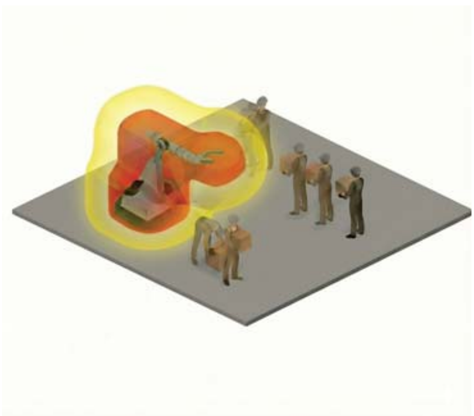

Fig. 1. Concept visualization of the CogniSafe3D system in a collaborative workspace, showing the inner Safety Monitoring zone (red) and outer Cognitive Layer (yellow), monitored by a wall-mounted LiDAR sensor.

worst-case positions for vulnerable body parts and to impose conservative speed/force limits to meet ISO/TS 15066 according to Saleem et al. (2024). Conversely, 3D sensing is typically paired with learning-based pose estimation, which the review finds to be non-deterministic, computationally intensive, and not yet safety-certified for standalone use Saleem et al. (2024).

CogniSafe3D addresses that exact gap. Instead of equipping the robot with costly sensors or constraining task speed, CogniSafe3D uses an external mounted (pole or ceiling), room-scale 3D LiDAR to produce deterministic safety-zone violation detection and real-time Risk Monitoring (RM) from point clouds, augmented by a cognitive layer that provides risk estimates to reduce unnecessary stops while preserving ISO/TS 15066-aligned safety margins. We will implement these methods on a TRL-6 prototype and validate their behavior under realistic environmental variability, sensor noise, and human motion ISO/TS 15066 (2016).

The remainder of this paper is organized as follows. Section 2 introduces the proposed system architecture of the CogniSafe3D system, briefly discusses the relevant methods used, and justifies technical choices with gaps in the current state of the art. Section 3 reports on the requirements on a certification path as well as specific cybersecurity and resilience concerns, letting Section 4 to conclude this paper.

2. Proposed System Architecture

This section details the conceptual architecture and functional design of the CogniSafe3D framework, establishing the methodological groundwork for its implementation in the upcoming TRL-6 prototype.

The CogniSafe3D system follows a methodology, shown in Fig. 2. The system consists of a robot, collaborating with humans, and its controlling computer running Robot Operating System 2 (ROS 2), as well as a 3D LiDAR and an edge computing unit running the main software components. All hardware components use wired Ethernet connections to minimize communication latency. While the 3D LiDAR and the edge computing unit are considered to be the actual CogniSafe3D system, the robot and the computer with ROS 2 are interchangeable and could be replaced by any robot with an external ROS 2 motion planning computer.

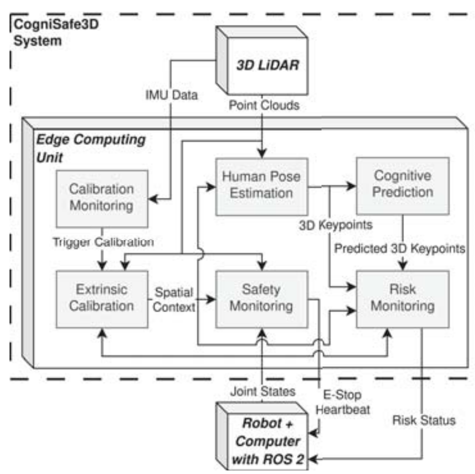

Fig. 2. Proposed system architecture of the CogniSafe3D System.

2.1. *Calibration Monitoring and Extrinsic Calibration*

In order to establish a correlation between the point cloud data and the robot's states, it is necessary to calibrate the two components with respect to each other. The extrinsic calibration is performed by means of a robot-mounted reflector, the movement of which is detected by the LiDAR, according to the values in its intensity channel. The detected positions are then matched with the actual robot positions via the received joint states, by employing a rigid body transformation estimation.

In the context of safety systems, prompt fault reporting and self-assessment are imperative. We rely on the liveliness feature of the ROS 2 Quality of Service settings: each node asserts liveliness on a designated topic, and a missed assertion within the lease duration raises an error. Given that the robot is mounted to a static table and the LiDAR is mounted to a wall, extrinsic calibration can only be misaligned if one of the components shifts through a crash or gradually. The Calibration Monitoring component therefore continuously checks for shifts using static reflective markers; recalibration is triggered after a detected crash, an observed LiDAR shift, or a system reboot.

2.2. *Safety Monitoring*

Detecting violations of a defined safety zone in Human-Robot Interaction (HRI) contexts poses two interlocking challenges. First, the system must respond only to dynamic intruders while ignoring the robot itself, which constantly moves through the protected volume as seen by the static external LiDAR; this requires deterministic, safety-rated masking of the robot's volume that accounts for calibration error and timing jitter. Second, persistent static objects (e.g. the mounting table) must not trigger detection, while legitimate scene changes (added tools, moved objects) must not create unsafe blind spots.

We define the safety zones around the robot as three-dimensional volumes that envelop the robot. The size of the volumes is defined by the robot's dynamic behavior and the characteristics of the sensing system, as shown in ISO/TS 15066 (2016). It is calculated with

$$S_p(t_0) = S_h + S_r + S_s + C + Z_d + Z_r \quad (1)$$

and incorporates multiple factors, like the changes in location from the operator (S_h) or the robot (S_r) or its stopping distance (S_s), as well as position uncertainties ($Z_d + Z_r$) which are typically induced by the sensors. The intrusion distance (C), describing how far a body part could move towards the robot without detection is drastically minimized compared to traditional 2D sensor approaches, since the 3D LiDAR is able to detect the whole body.

Current research on Safety Monitoring (SM) in HRC primarily follows two directions. Learning-based Human Pose Estimation (HPE) explicitly identifies and tracks humans Svarny et al. (2019); Antao et al. (2019); Forlini et al. (2024), allowing the system to disregard the robot and static background, but suffers from a lack of determinism: deep-learning inference pipelines are difficult to validate against safety standards. The second approach uses low-precision representations such as occupancy grids or voxelizations, which are deterministic and easier to validate but introduce significant inaccuracies at zone boundaries and cannot capture fine-grained proximity (Buerkle et al. (2023a); Buerkle et al. (2023b)). The state of the art thus either sacrifices determinism for selectivity (pose estimation) or precision for determinism (grids). To overcome this, we use a Signed Distance Field (SDF) representation with temporal filtering of static and dynamic parts: unlike binary occupancy grids, the SDF encodes a continuous signed distance to the nearest surface, exposing occupancy implicitly via its sign and per-voxel velocity via its temporal evolution, without additional clustering or tracking.

With the LiDAR point cloud at t and the synchronized joint states from the robot. We can exclude the robot based on its URDF model from the point cloud, and we maintain a voxel-indexed SDF. A per-voxel MAD-based background subtraction yields the dynamic voxels, and a violation is declared by the existence of dynamic voxels inside the safety zone, defined based on Eq. 1

and monitored via an AABB membership check. The predicate uses no learned components and is deterministic in the masked cloud and joint state, as required to certify the safety layer.

Upon notice of an intrusion into the defined safety zone or the loss of calibration, the SM module drops an emergency stop heartbeat, signaling the robot to immediately stop its movement. Since these full stops are decreasing the efficiency of the robot, they should be avoided. This is the task of the cognitive layer of the system (comprised of HPE, Cognitive Prediction and RM), eventually calculating risk estimates, pushing the robot to avoid situations in which the SM module would intercept.

2.3. *Human Pose Estimation*

To produce context-aware risk estimates, the system must account for the behavior of human workers collaborating with the robot in the workspace and provide this information to the RM module (discussed further in Section 2.5). HPE plays a central role in this process, by providing pose data, which includes both 3D positional and orientation data of human keypoints. The spatial context provided by the calibration module, helps to transform the detected keypoints from the sensor frame in a global world frame, to put them in context with the position of the robot.

In CogniSafe3D, existing methods for HPE that utilize 3D LiDAR point clouds as input are employed Yang et al. (2025); Ballester et al. (2024). The performance of state-of-the-art HPE algorithms has already been shown to reach a processing time of a few milliseconds Ballester et al. (2024). This level of performance could potentially already meet the requirements for real-time applications. However, further optimization through point cloud filtering techniques, as demonstrated in the work by Gómez et al. (2023), might be a viable strategy to further enhance the performance of these algorithms.

2.4. *Cognitive Prediction*

The Cognitive Prediction module, together with the HPE and RM, forms the cognitive layer of the CogniSafe3D system and predicts future human poses in the shared workspace. Utilizing learning-based techniques to predict future human poses has received considerable attention in recent years, with a steadily growing body of literature Hoffman et al. (2024). Such predictive capabilities have gained traction as they enable robots in collaborative systems to support smoother coordination and more natural interactions Dell'Oca et al. (2025), which are generally preferred by users Pandya et al. (2024).

This module builds on the outputs of HPE. It forecasts full-body human poses up to one second into the future, at predefined discrete timesteps and for a predefined number of samples. The output consists of 6D poses of full-body keypoints. To achieve this, we employ a hierarchical approach, by combining learning-based Human Intention Prediction (HIP) and Human Pose Prediction (HPP), an underexplored strategy in this domain, to more accurately capture future human behavior. The forecasting component uses an encoder–decoder transformer architecture that models spatiotemporal dependencies between observed and future poses. For training and evaluation, we have curated a dataset for this project Plahl et al. (2025). Within the safety system, human intent is inferred with respect to a given task, from scene data, robot data and the past human pose estimates. This allows the subsequent prediction step to focus on a smaller set of plausible future poses corresponding to the inferred human intention. For instance, a human's posture during collaborative tasks differs from one during mere coexistence. By narrowing down the algorithm's search space for the HPP, the short-term future pose prediction is made more efficient and more accurate Gómez-Izquierdo et al. (2025). To mitigate the potential consequences of incorrect intent classification, we plan to integrate safeguards into this module to ensure the system's robustness.

2.5. *Risk Monitoring*

The dynamic nature of HRC makes traditional pre-commissioning risk assessments, as recommended by current standards, difficult to apply Huck et al. (2021). Ensuring risk reduction and maintaining safety during operation is therefore critical. Through this RM module, our safety system, provides robots with real-time risk estimates. Beyond improving safety, equipping robots with risk awareness has been shown to enhance performance in tasks involving decision-making Jiang and Wang (2022).

We address this challenge by leveraging multimodal information from the robot, the human perception pipeline, and the surrounding scene context. The robot provides joint positions, velocities, and torques at high frequency; from these, end-effector poses, velocities, and interaction wrenches are derived via forward kinematics and the robot's dynamic model. The perception pipeline contributes outputs from HPE and HPP, providing both current and short-horizon future poses of the human. These data streams are fused into a unified latent representation of the human–robot system. In our previous work Katranis et al. (2025), we combined heterogeneous inputs using heuristic weighting functions to obtain an aggregated risk index. Here, we generalize this concept through a probabilistic formulation: we model the joint human–robot state as a latent dynamic variable evolving over time. A particle filter propagates and updates this state by integrating the data. This allows the system to handle asynchronous, noisy, and partial observations while maintaining temporal consistency. From the resulting posterior distribution we compute task-relevant risk metrics, such as the expected collision energy, minimum predicted human–robot distance, and relative velocity. These quantities form a multidimensional stochastic risk representation, where each component reflects a distinct hazard dimension that the robot controller can use to proactively mitigate potential hazards.

2.6. *Co-Engineering Cybersecurity and Safety*

To ensure system integrity, a comprehensive Threat Analysis and Risk Assessment (TARA) is applied to the CogniSafe3D system in conjunction with a Hazard Analysis (HARA). The TARA focuses on threats affecting IT/OT infrastructures and cognitive sensing technologies—specifically 3D LiDAR and learning-based methods—to identify mitigation strategies that reduce their impact. Simultaneously, the HARA addresses safety hazards arising from sensor faults, degraded inputs, and machine learning prediction errors.

These processes are conducted in parallel by synchronized teams to identify the minimum necessary mechanisms to mitigate both security and safety hazards without conflict. Key technical mitigations include real-time monitoring, estimation of AI prediction uncertainty, and selective output rejection. Under this framework, safety is only assumed when information is correctly delivered; the absence of data is treated as a potential safety hazard.

2.7. *Preliminary Implementation Results*

To validate the feasibility of the deterministic safety layer at real-time HRC frame rates, we report bench-level timing of a prototype implementation of the dynamic-intruder detection and intrusion check, exclusive of upstream SDF construction. Measurements were taken on commodity hardware (Intel Core i9, NVIDIA GeForce RTX 4070, Ubuntu 22.04).

With an end-to-end latency of 14.1 ± 1.4 ms, it corresponds to ≈ 70 Hz, well above typical 10–30 Hz LiDAR rates; the dominant cost is the GPU statistical filter ($\approx 41\%$). Recent real-time SDF mappers report per-frame updates of 1.9–22.7 ms across 20–10 cm voxel resolutions on comparable hardware Pan et al. (2022), leaving the combined perception-to-violation budget within the LiDAR frame interval and headroom for the asynchronous cognitive layer to run in parallel without contending for the safety-critical timing budget. End-to-end integration on the TRL-6 prototype is the subject of ongoing work.

3. Discussion

The CogniSafe3D system moves beyond traditional, reactive safety towards a proactive framework. However, the integration of cognitive layers into safety-critical environments introduces specific challenges regarding determinism, efficiency, and system integrity.

3.1. *Balancing Determinism and Machine Learning*

A fundamental conflict in modern HRC is the non-deterministic nature of deep learning versus the requirement for certifiable safety. CogniSafe3D addresses this by establishing a twofold methodology. The SM module provides a deterministic foundation by using SDFs to process point clouds for safety zone violations. This ensures a ”fail-safe” baseline.

Simultaneously, the cognitive layer (HPE and HPP) enables the robot to anticipate human intent and movement. By decoupling these, the system can use non-deterministic AI to enhance performance—reducing unnecessary stops—while relying on deterministic SDF-based sensing to guarantee compliance with the applicable standards.

3.2. *Efficiency Gains via Dynamic Risk Assessment*

Traditional HRC safety relies on conservative speed and force limits because 2D sensors lack reliable depth information. This often results in a large protective separation distance Sp, which is calculated in Equation 1.

In this formulation, the intrusion distance C is typically significant for 2D sensors; however, because 3D LiDAR can detect the whole body, C becomes drastically minimized in the CogniSafe3D system. Furthermore, the transition from coarse occupancy grids to high-resolution SDF allows for more precise spatial reasoning. When combined with the RM module’s real-time risk estimates, the robot can proactively mitigate hazards, resulting in a system that is both safer and more efficient than traditional implementations.

3.3. *Strategic Co-Engineering of Safety and Cybersecurity*

Integrating learning-based technologies and 3D LiDAR into industrial environments significantly widens the potential attack surface. A major discussion point in this integration is the inherent conflict between cybersecurity and safety requirements: cybersecurity often demands rapid software patching, whereas safety and resilience require updates to be exhaustively documented and tested before deployment.

CogniSafe3D adopts the philosophy of the IEC 62443 standard, which views safety as an essential function that must be protected from potential attackers. This necessitates an integrated co-engineering approach where TARA and HARA are conducted in parallel. This ensures that mitigations, such as selective output rejection when AI uncertainty is high, serve both safety and security goals without introducing conflicting mechanisms. Ultimately, for cognitive HRC systems to be industrially viable, safety must be assumed only when the underlying sensing and prediction data are verified as both accurate and secure.

4. Conclusion

This paper proposes a novel approach for a safe HRC where an environmental 3D LiDAR senses the scene and continuously gathers point clouds, which are then processed and used to stop robots when safety violations are detected. The execution flow of the proposed CogniSafe3D system spans from perception and preprocessing of point clouds through SM to RM based on HPE and HPP, thereby creating a deterministic and certifiable foundation for safe interaction. Ongoing work is focused on the physical integration of these modules into a TRL-6 environment to validate the efficiency gains within an Eureka-Eurostars EU funded project.

While this research addresses the core challenges of perception-driven safety, future work will focus on enhancing the system’s robustness against occlusions and refining advanced risk mitigation strategies to further optimize collaborative workflows.

Acknowledgement

This work was supported by the CogniSafe3D project (E! 6085) under the Eureka-Eurostars program, with funding from the German Federal Ministry of Research, Technology and Space (BMFTR) and the Italian Ministero dell’Università e della Ricerca (MUR).

References

Antao, L., J. Reis, and G. Goncalves (2019, September). Voxel-based space monitoring in human-robot collaboration environments. In *2019 24th IEEE*

International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 552–559. IEEE.

Ballester, I., O. Peterka, and M. Kampel (2024, December). *SPiKE: 3D Human Pose from Point Cloud Sequences*, pp. 470–486. Springer Nature Switzerland.

Buerkle, C., F. Oboril, and K.-U. Scholl (2023, October). Histodepth - novel depth perception for safe collaborative robots. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9677–9684. IEEE.

Buerkle, C., F. Oboril, O. Zayed, and K.-U. Scholl (2023, September). Histogrid: Robust lidar-based traffic monitoring. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 209–214. IEEE.

Dell'Oca, S., D. Matteri, E. Montini, V. Cutrona, Z. M. Barut, and A. Bettoni (2025). Improving Collaborative Robotics: Insights on the Impact of Human Intention Prediction. In A. Paolillo, A. Giusti, and G. Abbate (Eds.), *Human-Friendly Robotics 2024*, Cham, pp. 1–15. Springer Nature Switzerland.

Forlini, M., F. Neri, M. Ciccarelli, G. Palmieri, and M. Callegari (2024, July). Experimental implementation of skeleton tracking for collision avoidance in collaborative robotics. *The International Journal of Advanced Manufacturing Technology 134*(1–2), 57–73.

Gómez, J., O. Aycard, and J. Baber (2023, May). Efficient detection and tracking of human using 3d lidar sensor. *Sensors 23*(10), 4720.

Gómez-Izquierdo, G., J. Laplaza, A. Sanfeliu, and A. Garrell (2025). Enhancing context-aware human motion prediction for efficient robot handovers.

Hoffman, G., T. Bhattacharjee, and S. Nikolaidis (2024, July). Inferring Human Intent and Predicting Human Action in Human–Robot Collaboration. *Annual Review of Control, Robotics, and Autonomous Systems 7*(Volume 7, 2024), 73–95. Publisher: Annual Reviews.

Huck, T. P., C. Ledermann, and T. Kröger (2021). Testing robot system safety by creating hazardous human worker behavior in simulation. *IEEE Robotics and Automation Letters 7*(2), 770–777.

ISO 10218-2 (2025). Robotics — safety requirements — part 2: Industrial robot applications and robot cells. Technical Report ISO 10218-2:2025, International Organization for Standardization. [Online]. Available: `http://www.iso.org`.

ISO/TS 15066 (2016). Robots and robotic devices - collaborative robots. Technical Report ISO/TS 15066:2016, International Organization for Standardization. [Online]. Available: `http://www.iso.org`.

Jiang, L. and Y. Wang (2022, June). Risk-aware Decision-making in Human-multi-robot Collaborative Search: A Regret Theory Approach. *Journal of Intelligent & Robotic Systems 105*(2), 40.

Katranis, G., F. Plahl, J. Grimstadt, I. Mamaev, S. Vock, and A. Morozov (2025). Dynamic risk assessment for human-robot collaboration using a heuristics-based approach. In *35th European Safety and Reliability Conference (ESREL2025) and the 33rd Society for Risk Analysis Europe Conference (SRA-E 2025)*.

Montini, E., F. Daniele, L. Agbomemewa, M. Confalonieri, V. Cutrona, A. Bettoni, P. Rocco, and A. Ferrario (2024, August). Collaborative robotics: A survey from literature and practitioners perspectives. *Journal of Intelligent & Robotic Systems 110*(3).

Pan, Y., Y. Kompis, L. Bartolomei, R. Mascaro, C. Stachniss, and M. Chli (2022). Voxfield: Non-projective signed distance fields for online planning and 3D reconstruction. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5331–5338. IEEE.

Pandya, R., Z. Wang, Y. Nakahira, and C. Liu (2024, July). Towards Proactive Safe Human-Robot Collaborations via Data-Efficient Conditional Behavior Prediction. arXiv:2311.11893.

Plahl, F., G. Katranis, I. Mamaev, and A. Morozov (2025). LiHRA: A LiDAR-Based HRI Dataset for Automated Risk Monitoring Methods. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Saleem, Z., F. Gustafsson, E. Furey, M. McAfee, and S. Huq (2024, April). A review of external sensors for human detection in a human robot collaborative environment. *Journal of Intelligent Manufacturing 36*(4), 2255–2279.

Shah, R., A. S. A. Doss, and N. Lakshmaiya (2025, September). Advancements in ai-enhanced collaborative robotics: towards safer, smarter, and human-centric industrial automation. *Results in Engineering 27*, 105704.

Svarny, P., M. Tesar, J. K. Behrens, and M. Hoffmann (2019, November). Safe physical hri: Toward a unified treatment of speed and separation monitoring together with power and force limiting. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7580–7587. IEEE.

Yang, J., Y. Liu, J. Li, X. Gu, G.-Z. Yang, and Y. Guo (2025, March). Posesdf++: Point cloud-based 3-d human pose estimation via implicit neural representation. *IEEE Transactions on Industrial Informatics 21*(3), 2689–2698.