

Design and Development of an AI-Based Security and Safety System for Hydrogen Explosion Hazard Mitigation in Nuclear Fission Power Plants

Mohammad Ali

malixbd@ieee.org

Atlantic International University <https://orcid.org/0009-0001-2184-7978>

Research Article

Keywords: Anomaly detection, artificial intelligence, Bayesian networks, cybersecurity, hydrogen explosion mitigation, ICS/SCADA intrusion detection, long short-term memory, nuclear fission power plant safety, passive autocatalytic recombiner, probabilistic risk assessment

Posted Date: September 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11083681/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Design and Development of an AI-Based Security and Safety System for Hydrogen Explosion Hazard Mitigation in Nuclear Fission Power Plants¹

Mohammad Ali

School of Science & Engineering, Atlantic International University, Honolulu, HI 96813, United States
malixbd@ieee.org

Abstract—This paper presents the design and conceptual development of an artificial-intelligence-based Security and Safety System (AI-S3) for early detection, prediction, and mitigation of hydrogen explosion hazards in nuclear fission power plants (NPPs), together with an integrated cybersecurity layer that protects the safety function itself from cyber-physical attack. The hazard addressed is chemical rather than nuclear-yield: a commercial fission core cannot sustain the fast supercritical chain reaction required for a weapon-type detonation, so the credible explosion hazard is combustible-gas—principally hydrogen—deflagration in auxiliary systems and, under severe-accident conditions, in the containment atmosphere, generated in part by high-temperature zirconium-steam oxidation of fuel cladding. The proposed architecture integrates five functional layers—field sensing, edge acquisition, an AI analytics core, decision and arbitration logic, and mitigation actuation—designed to satisfy the independence and redundancy criteria of IEEE Std 384 while applying long short-term memory networks, autoencoder-based anomaly detection, and Bayesian belief updating to obtain early warning ahead of conventional fixed-setpoint alarms. Mathematical models are developed for hydrogen generation and transport, recombiner reaction kinetics, the anomaly-detection and intrusion-classification sub-models, and a probabilistic risk-reduction formulation linking AI-S3 to established fault-tree and Bayesian-network practice. A qualitative, illustrative-simulation-based evaluation using notional data indicates that AI-based early warning can identify anomalous hydrogen accumulation and anomalous network behavior appreciably earlier than fixed-threshold instrumentation alone, extending operator response margin. The paper closes with the verification and validation requirements, the cybersecurity assurance case, and the experimental and regulatory work needed before such a system could be qualified for plant deployment.

Index Terms—Anomaly detection, artificial intelligence, Bayesian networks, cybersecurity, hydrogen explosion mitigation, ICS/SCADA intrusion detection, long short-term memory, nuclear fission power plant safety, passive autocatalytic recombiner, probabilistic risk assessment.

I. INTRODUCTION

A. Background and Motivation

Nuclear power plants operate under a defense-in-depth safety philosophy that combines physical barriers, redundant safety systems, and administrative controls [1], [2]. A recognized hazard category within this framework is combustible-gas explosion—principally hydrogen—arising in auxiliary systems

such as battery and uninterruptible-power-supply (UPS) rooms, waste-gas systems, and hydrogen-cooled generators during normal operation, and, under severe-accident conditions, within the containment or reactor-building atmosphere itself, as demonstrated at Fukushima Daiichi in 2011 [3], [4]. Conventional protection against this hazard combines hazardous-area electrical classification [5], [6] with passive and active hydrogen-control equipment [4], [7]. Both approaches, however, rely predominantly on fixed instrumentation thresholds and on human operator response, and can therefore be slow relative to the near-exponential concentration growth credible during a severe-accident hydrogen release.

The work reported here investigates whether modern artificial-intelligence (AI) techniques—specifically recurrent neural networks for time-series anomaly detection [8]–[11], probabilistic graphical models for risk estimation [12], and machine-learning-based intrusion detection [13]—can be integrated into a single AI-based Security and Safety System (AI-S3) that (i) detects anomalous precursors to a hydrogen-explosion event earlier than fixed-setpoint alarms, (ii) continuously updates a plant-specific probabilistic risk estimate as new sensor evidence arrives, and (iii) protects the safety system's own digital infrastructure against cyber-physical attack, given the increasing digitalization of instrumentation and control (I&C) systems in modern and next-generation NPPs [14], [15].

B. Scope and Hazard Definition

The explosion hazard addressed in this paper is chemical, not nuclear-yield. Commercial power reactors operating today are fission reactors whose cores cannot undergo the fast supercritical chain reaction required for a weapon-type detonation: low fuel enrichment, typically below 5% U-235, together with core geometry and moderation, precludes it by design. The phrase *nuclear fission explosion*, which appears in parts of the earlier literature and in the author's dissertation from which this work is derived, is therefore avoided here in favor of *hydrogen explosion*, which names the physical event directly and fixes the applicable physics unambiguously.

The hazard class treated throughout is combustible-gas—principally hydrogen—deflagration and, under confinement, deflagration-to-detonation transition, arising from processes

¹Manuscript submitted for review. This work is derived from the author's doctoral research at the School of Science & Engineering, Atlantic International University. The author is with the School of Science & Engineering, Atlantic International University, Honolulu, HI 96813 USA (e-mail: malixbd@ieee.org).

associated with fission-plant operation and severe accidents. The dominant severe-accident source is hydrogen generated by high-temperature oxidation of zirconium fuel cladding by steam during a loss-of-coolant accident (Section III); the lower-consequence normal-operation sources are battery off-gassing, coolant radiolysis, and generator hydrogen-cooling leakage. The distinction is not merely semantic. It fixes the governing physics as the combustion and detonation chemistry of hydrogen–air–steam mixtures rather than nuclear criticality physics, and it therefore fixes the applicable mitigation technologies—recombiners, igniters, engineered venting, and the AI-based early-warning system proposed here—and the applicable standards (Section II-E).

Reactor technologies not yet in commercial operation, in particular magnetic-confinement fusion, present physically distinct energetic hazards such as plasma disruption and magnet quench [16]; a fusion plasma cannot sustain an explosive chain reaction, since loss of confinement extinguishes rather than accelerates the reaction. Such technologies are outside the scope of this work, which is limited throughout to nuclear fission power plants.

C. Problem Statement

Existing hydrogen-explosion protection in NPPs is largely reactive. Fixed-threshold gas detectors trigger discrete alarms and interlocks only after a concentration setpoint has been crossed, and the electrical hazardous-area classification and hydrogen-mitigation disciplines are typically engineered, procured, and maintained as separate, non-integrated systems [17]. Separately, the increasing use of networked digital instrumentation in modern NPP I&C systems introduces a cyber-attack surface that could, in principle, be used to disable or spoof the very sensors and interlocks relied upon for hydrogen-explosion protection [13], [15]. No unified architecture in the current literature addresses AI-based early-warning detection, probabilistic risk estimation, and cybersecurity protection of the safety function as a single, co-designed system.

D. Contributions

The contributions of this paper are as follows.

- A precise physical definition of the explosion-hazard class addressed, and a characterization of the hydrogen-generation mechanisms relevant to NPPs (Section III).
- Mathematical models of hydrogen generation, transport, and catalytic recombination suitable as the physical basis for an AI-based anomaly-detection system (Sections III-B and III-C).
- The design of an AI analytics core combining anomaly detection, predictive risk estimation, and cybersecurity intrusion detection, with an explicit mathematical formulation for each sub-model (Section V).
- An integrated, layered AI-S3 system architecture consistent with the independence criteria of IEEE Std 384 [1] and with IEC 62443 industrial-cybersecurity practice [14] (Section IV).

- An illustrative evaluation using simulated data, together with an explicit statement of the verification, validation, and regulatory-qualification work required before real-plant deployment (Section VI).

E. Organization of the Paper

Section II reviews prior work and the applicable standards and regulatory framework. Section III characterizes the explosion and energetic-hazard sources and develops the hydrogen generation and transport models. Section IV presents the AI-S3 system architecture. Section V develops the mathematical models of the AI analytics core. Section VI presents the illustrative evaluation and discussion, and Section VII concludes and identifies future work.

II. RELATED WORK AND REGULATORY FRAMEWORK

A. Hydrogen Explosion Hazards in Nuclear Facilities

The operating history of hydrogen events in nuclear facilities—battery-room deflagrations, generator hydrogen-cooling leaks, and the Fukushima Daiichi severe-accident explosions—establishes the physical hazard basis for this work [3], [4], [17]. Quantitative severe-accident risk studies [18] and international assessments of hydrogen-management strategies and their supporting computer codes [7] provide the reference framework within which any new detection or mitigation layer must be positioned. Section III summarizes the mechanisms most relevant to the AI-S3 design rather than repeating that operating history in full.

B. AI-Based Anomaly Detection in Industrial and Process Safety Systems

Recurrent neural network architectures, in particular long short-term memory (LSTM) networks [8], and autoencoder- and encoder–decoder-based reconstruction-error methods [9], [10] are established techniques for time-series anomaly detection in industrial process monitoring, chemical-process safety, and rotating-machinery condition monitoring. Comparable methods have been applied directly to cyber-physical process plants, where an LSTM predictor flags deviations between predicted and measured process variables [11]. Their principal advantage over fixed-threshold instrumentation is sensitivity to anomalous temporal patterns—for example, an unusually rapid rate of rise—that precede any fixed setpoint being crossed. The corresponding cost is the need for a training dataset representative of normal operating variability, without which the false-alarm rate becomes unacceptable for a safety-significant application.

C. Probabilistic Risk Assessment and Bayesian Networks

Probabilistic risk assessment (PRA), fault-tree analysis, and Bayesian belief networks are the established quantitative risk-assessment tools in nuclear safety engineering [12], [18]. Research has since explored dynamic, or living, PRA concepts in which a plant's risk estimate is continuously updated from real-time sensor evidence rather than computed once at the design stage. That concept is directly relevant to the predictive

risk engine proposed in Section IV-C and formulated in Section V-B.

D. Cybersecurity of Nuclear Instrumentation and Control Systems

The digitalization of NPP I&C systems has motivated dedicated cybersecurity standards and regulatory guidance [14], [15] and a parallel body of research on machine-learning-based intrusion detection for industrial control system (ICS) and SCADA network traffic [13]. That work forms the technical basis for the cybersecurity sub-model of Section V-D. A recurring finding is that intrusion detection in cyber-physical systems benefits from combining network-traffic features with physical-process consistency checks, which is the approach adopted in the arbitration logic described in Section IV-D.

E. Applicable Standards and Regulatory Framework

Table I summarizes the standards that govern the design. IEEE Std 384 [1] and IEEE Std 603 [2] establish the independence and safety-system criteria for Class 1E equipment; IEEE Std 803 [19] governs unique identification of plant systems; IEEE Std 7000 [20] provides process guidance for the trustworthy design of autonomous and intelligent systems, which is relevant to the AI-S3 decision and arbitration layer (Section IV-D), where AI-generated outputs are gated by deterministic safety interlocks rather than acting autonomously on safety-significant equipment.

IEC 60079 [5] and NFPA 70 Article 500 [6] govern hazardous-area equipment and classification. IEC 61508 [21] and IEC 61513 [22] establish the safety-integrity-level (SIL) framework within which any AI-informed safety function would need to be classified and qualified; Section VI-D discusses the qualification gap between the illustrative AI-S3 concept and a SIL-qualified deployment. IEC 62443 [14] and U.S. NRC Regulatory Guide 5.71 [15] define the applicable cybersecurity framework for the intrusion-detection sub-model of Section V-D, including defense-in-depth network segmentation between safety-related and non-safety-related digital systems. Finally, post-Fukushima regulatory guidance [3] and IAEA Nuclear Energy Series No. NP-T-1.9 [4] govern passive autocatalytic recombiner (PAR) and igniter-based hydrogen-control design, which forms the actuation-layer basis (Section IV-E) that the AI analytics core is intended to trigger earlier and more selectively than fixed-threshold logic alone.

TABLE I
APPLICABLE STANDARDS AND REGULATORY GUIDANCE

Standard	Scope as applied in this work
IEEE Std 384	Independence criteria for Class 1E equipment and circuits
IEEE Std 603	Safety-system criteria for nuclear power generating stations
IEEE Std 803	Unique identification of plant systems and equipment
IEEE Std 7000	Trustworthy process for autonomous / intelligent system design
IEC 60079 series	Equipment and area classification for explosive gas atmospheres

Standard	Scope as applied in this work
IEC 61508 / 61513	Functional safety and nuclear I&C safety-integrity requirements
IEC 62443	Cybersecurity for industrial automation and control systems
NRC RG 5.71	Cybersecurity programs for nuclear facilities (USA)
NFPA 70 Art. 500	Hazardous (classified) location electrical requirements
IAEA NP-T-1.9	Mitigation of hydrogen hazards in water-cooled power reactors

III. EXPLOSION AND ENERGETIC-HAZARD SOURCES

A. Hydrogen Generation Mechanisms

Table II summarizes the hydrogen-generation mechanisms relevant to fission-plant explosion-hazard assessment and identifies the role of each in the proposed AI-S3. The four mechanisms differ by orders of magnitude in rate, which is the principal reason a single fixed detection setpoint serves them poorly: a setpoint tuned to avoid nuisance alarms on continuous, low-rate battery off-gassing is necessarily late for a rapid zirconium–steam transient.

TABLE II
HYDROGEN GENERATION MECHANISMS AND RELEVANCE TO AI-S3

Mechanism	Typical location	Rate regime	AI-S3 relevance
Battery off-gassing	Battery / UPS rooms	Continuous, low rate	Baseline anomaly model
Radiolysis of coolant	Primary system, waste-gas system	Continuous, low to moderate	Baseline anomaly model
Generator H ₂ cooling leakage	Turbine hall, seal-oil system	Intermittent, low rate	Sensor-fusion detection
Zirconium–steam reaction	Reactor core, containment (severe accident)	Rapid, high rate	Primary AI-S3 target

B. Mathematical Model of Hydrogen Concentration Dynamics

The time-dependent hydrogen concentration $C_{H_2}(t)$ (volume fraction) within a well-mixed enclosed volume V , subject to a generation rate $G(t)$ and a ventilation exchange rate Q , is modeled by the lumped-parameter mass balance

$$dC_{H_2} / dt = [G(t) - Q C_{H_2}(t)] / V - R_{PAR}(t) \quad (1)$$

where $R_{PAR}(t)$ is the hydrogen consumption rate due to recombiner action, given by (3). For spatially resolved analysis—required wherever local pocketing near the ceiling of an enclosure is of concern—the governing equation is the advection–diffusion–reaction equation, a generalization of Fick's second law,

$$\partial C / \partial t = D \nabla^2 C - \mathbf{u} \cdot \nabla C + S(x, y, z, t) - kC \quad (2)$$

where D is the turbulent diffusivity, $\mathbf{u}$ is the local ventilation-induced velocity field, S is the volumetric hydrogen source term, and k is a first-order recombination or consumption rate coefficient.

C. Recombiner Reaction Kinetics

The catalytic hydrogen-oxidation rate within a passive autocatalytic recombiner is modeled with an Arrhenius temperature dependence,

$$k(T) = A \exp(-E_a / RT) \quad (3)$$

with the volumetric consumption rate expressed as $R_{\text{PAR}} = k(T)[\text{H}_2]^a[\text{O}_2]^b$, where A is the pre-exponential (frequency) factor, E_a is the activation energy of the catalytic surface reaction, R is the universal gas constant, T is the local gas temperature, and a and b are empirically determined reaction orders.

D. Battery-Room and Turbine Hydrogen Systems

Battery-room and turbine hydrogen-cooling hazards follow the same mass-balance formulation (1), with $G(t)$ parameterized from charging-current-dependent off-gassing rates or from seal-leakage flow rates, respectively. Both are treated as baseline, non-severe-accident training regimes for the anomaly-detection sub-model developed in Section V-A.

E. Scope Relative to Other Reactor Types

The hydrogen-generation models above are specific to fission power plants and are the sole physical basis for the AI-S3 design developed in Sections IV and V. Other reactor technologies not yet in commercial operation, such as magnetic-confinement fusion, present physically distinct energetic hazards [16] that would require different physics models. Extending the AI-S3 sensing and analytics concept to such reactor types is noted only as a possible long-term research direction (Section VII-B) and is not developed further here.

IV. PROPOSED AI-S3 SYSTEM ARCHITECTURE

Fig. 1 shows the proposed five-layer AI-S3 architecture. The layering is deliberate: each layer has a defined interface, a defined failure mode, and a defined qualification route, so that the AI-bearing elements are confined to Layer 3 and are separated from actuation by the deterministic arbitration logic of Layer 4a.

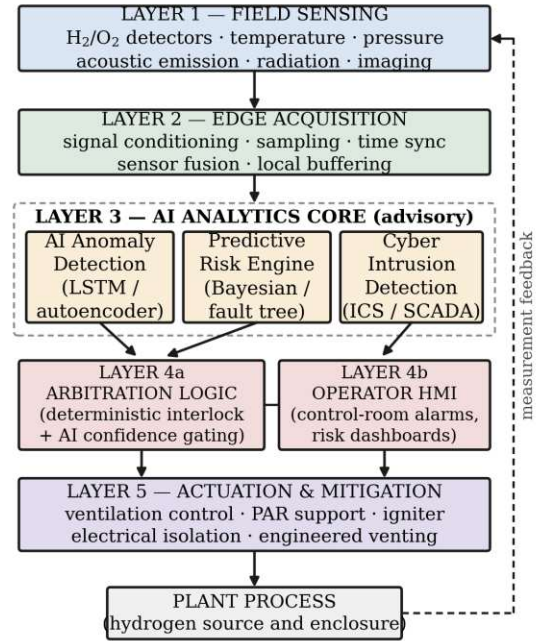


Fig. 1. Layered architecture of the proposed AI-based Security and Safety System (AI-S3). All learned components are confined to Layer 3 and act only as advisory evidence sources; every automatic mitigation command is gated by the deterministic arbitration logic of Layer 4a.

A. Layer 1—Field Sensing

Redundant and diverse sensing includes catalytic-bead and infrared point and open-path hydrogen detectors, oxygen sensors, temperature and pressure transmitters, acoustic-emission sensors for early leak detection, area radiation monitors, and CCTV or thermal imaging, all consistent with the classified-area equipment requirements of Section II-E [5], [6]. Diversity of detection principle is required so that a single failure mechanism—catalyst poisoning, for instance—cannot blind the whole sensing layer.

B. Layer 2—Edge Acquisition and Preprocessing

Field signals are conditioned, sampled, time-synchronized, and locally buffered at the edge before transmission to the AI analytics core, with sensor-fusion pre-alignment of heterogeneous sampling rates. Fig. 2 illustrates the pipeline from raw multi-sensor input through to the alarm decision. Local buffering preserves the pre-event data record needed for post-event analysis even if the upstream network path is lost.

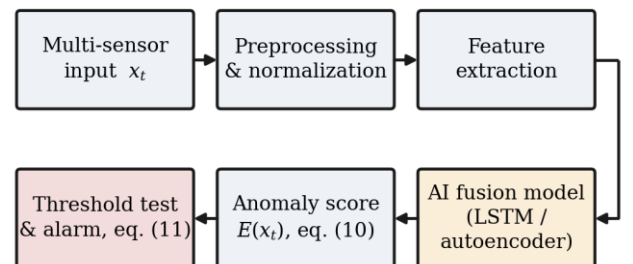


Fig. 2. AI-based multi-sensor fusion and alarm-decision pipeline spanning Layers 2 and 3, from the raw sensor vector x_t to the threshold test of (11).

C. Layer 3—AI Analytics Core

The analytics core comprises three cooperating sub-models, developed mathematically in Section V: an LSTM- and autoencoder-based anomaly-detection model operating on the fused sensor time series; a Bayesian-network predictive risk engine that continuously updates a plant-specific explosion-probability estimate as new evidence arrives; and a machine-learning-based intrusion-detection classifier monitoring ICS and SCADA network traffic for anomalous cyber-physical activity.

D. Layer 4—Decision, Arbitration, and Human–Machine Interface

Consistent with IEEE Std 7000 guidance on trustworthy autonomous and intelligent systems [20] and with nuclear functional-safety practice [22], AI-generated outputs are treated as advisory inputs to a deterministic, rule-based arbitration layer. The AI never directly actuates safety-significant equipment. The arbitration layer combines the AI confidence score with independent hard-wired interlock logic before issuing any automatic mitigation command, and all AI-flagged anomalies are simultaneously presented to control-room operators through the human–machine interface. This separation is what allows the qualification burden of Section VI-D to fall on a deterministic, inspectable layer rather than on the learned model itself.

E. Layer 5—Actuation and Mitigation

Confirmed actuation commands drive ventilation control, PAR support systems, igniter activation, electrical isolation of non-essential loads in the affected area, and, where applicable, engineered emergency venting, consistent with established hydrogen-control practice [4], [7], [17].

F. Independence and Redundancy

Consistent with IEEE Std 384 [1] and IEEE Std 603 [2], the AI-S3 analytics core is implemented on separate, redundant computing channels powered from independent Class 1E divisions, with the arbitration layer requiring agreement, or an independent hard-wired trip, before an automatic mitigation action is issued. Consequently, a single AI-model fault, an episode of training-data drift, or a cyber-intrusion event cannot by itself either cause a spurious actuation or suppress a genuine one.

V. MATHEMATICAL MODELS OF THE AI ANALYTICS CORE

A. Anomaly-Detection Sub-Model: LSTM Network

The anomaly-detection sub-model uses an LSTM recurrent network [8] to forecast the expected next-step sensor vector from recent history. Fig. 3 shows the internal gate structure of a single LSTM cell, and (4)–(9) give the governing equations, in which x_t is the input vector at time t , h_t the hidden state, C_t the cell state, σ the logistic sigmoid function, $\odot$ the Hadamard product, and W and b the trainable weight matrices and bias vectors of each gate.

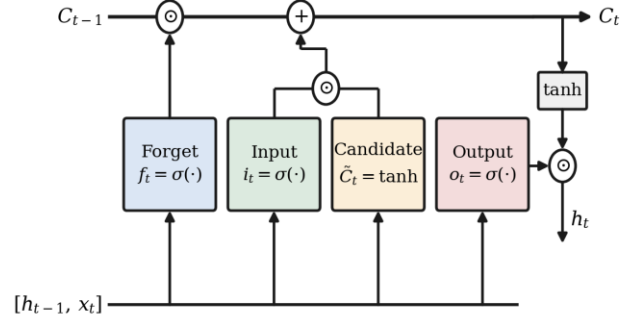


Fig. 3. Simplified LSTM cell used in the anomaly-detection sub-model, corresponding to (4)–(9). The cell-state pathway $C_{t-1} \rightarrow C_t$ carries long-term memory; the hidden state h_t is formed from the output gate and the squashed cell state.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (5)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (6)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (7)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (8)$$

$$h_t = o_t \odot \tanh(C_t) \quad (9)$$

The anomaly score is computed as the squared reconstruction, or prediction, error between the observed and the predicted sensor vector,

$$E(x_t) = \|x_t - \hat{x}_t\|^2 \quad (10)$$

and an alarm is raised when the score exceeds an adaptively estimated threshold $\theta = \mu_E + \alpha \sigma_E$, formed from the rolling mean μ_E and standard deviation σ_E of the reconstruction error over a trailing baseline window with sensitivity factor α (typically 3 to 6),

$$A(t) = 1 \text{ if } E(x_t) > \theta; A(t) = 0 \text{ otherwise} \quad (11)$$

The adaptive threshold, rather than a fixed one, is what allows the sub-model to remain sensitive across the widely differing baseline rate regimes of Table II.

B. Predictive Risk Engine: Bayesian Updating

The predictive risk engine maintains a continuously updated posterior probability of an impending explosion event given accumulated sensor and AI-anomaly evidence ϵ , using Bayes' theorem [12],

$$P(X | \epsilon) = P(\epsilon | X) P(X) / P(\epsilon) \quad (12)$$

where X denotes the explosion (top) event, $P(X)$ is the prior probability drawn from the plant's baseline PRA or fault-tree model [18], and $P(\epsilon | X)$ is the likelihood of the observed evidence pattern conditioned on the developing hazard, learned from the historical output distribution of the anomaly sub-model. For a fault tree with n independent basic events of probability p_i combined through an OR gate—the typical

structure for redundant hydrogen-detection failure paths—the top-event probability is

$$P_{\text{top}} = 1 - \prod_{i=1}^n (1 - p_i) \quad (13)$$

and the risk-reduction factor achieved by the AI-S3 relative to a conventional fixed-threshold baseline is defined as

$$\text{RRF} = P_{\text{top,baseline}} / P_{\text{top,AI-S3}} \quad (14)$$

C. Response-Time Margin Model

The operational benefit of early-warning detection is quantified as the additional response-time margin Δt gained relative to a fixed hard-alarm setpoint C_{alarm} , given the AI early-warning level C_{AI} (with $C_{\text{AI}} < C_{\text{alarm}}$) and the local hydrogen concentration growth rate dC/dt at the time of detection,

$$\Delta t \approx (C_{\text{alarm}} - C_{\text{AI}}) / (dC/dt) \quad (15)$$

Fig. 4 illustrates this concept using an illustrative simulated hydrogen concentration trace. The rate-of-rise test of (11) flags the developing transient while the measured concentration is still far below the hard-alarm setpoint, and the margin Δt of (15), measured between the crossing of C_{AI} and the crossing of C_{alarm} , is the additional interval in which operators may initiate mitigation. Because Δt is inversely proportional to the growth rate, the margin is largest for the slow-developing scenarios and smallest—though still non-zero—for the rapid zirconium–steam transient that dominates severe-accident risk.

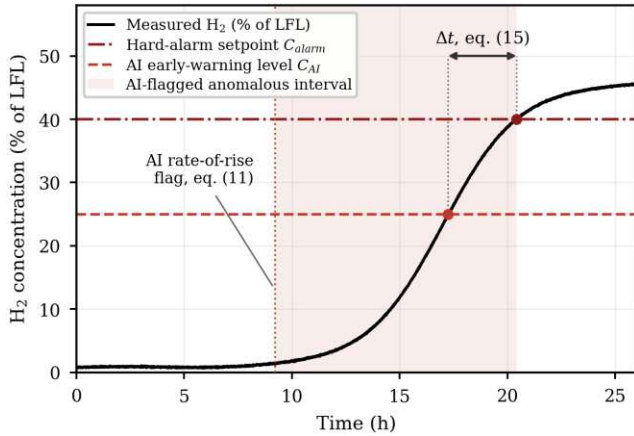


Fig. 4. Illustrative simulated hydrogen concentration trace showing the AI rate-of-rise flag of (11), the early-warning level C_{AI} , the hard-alarm setpoint C_{alarm} , and the resulting response-time margin Δt of (15). Notional data, shown for methodology demonstration only; not a validated plant simulation.

D. Cybersecurity Intrusion-Detection Sub-Model

ICS and SCADA network traffic features z —packet timing, command-sequence statistics, and protocol-conformance indicators—are classified by a supervised model [13] trained to minimize the cross-entropy loss between the predicted intrusion probability $\hat{y}$ and the labeled training class y over a training set of size N ,

$$L = -(1/N) \sum_{j=1}^N [y_j \log \hat{y}_j + (1 - y_j) \log(1 - \hat{y}_j)] \quad (16)$$

A detected intrusion event is treated by the arbitration layer of Section IV-D as a reason to distrust the affected sensor channel's evidence in the Bayesian update of (12), rather than

being used directly to trigger a mitigation actuation. This asymmetry is deliberate: it prevents an adversary from using a spoofed cyber event to force an unwarranted plant response, while still allowing a genuine intrusion to degrade the weight placed on compromised evidence.

E. Illustrative Alarm Prioritization

Fig. 5 shows the qualitative likelihood–consequence risk matrix used by the arbitration layer to rank simultaneous AI-flagged anomalies for operator attention when multiple detection channels report concurrently. The scoring shown is illustrative and would be replaced, in a plant-specific implementation, by the consequence categories of the plant's own PRA.

Consequence	Catastrophic	5	10	15	20	25
	Major	4	8	12	16	20
	Moderate	3	6	9	12	15
	Minor	2	4	6	8	10
	Negligible	1	2	3	4	5
		Likelihood				
		Rare	Unlikely	Possible	Likely	Frequent

Fig. 5. Qualitative likelihood–consequence risk matrix used for AI-prioritized alarm ranking. Illustrative scoring only.

F. Summary of Notation

Table III summarizes the symbols used in Sections III and V.

TABLE III
SUMMARY OF MATHEMATICAL SYMBOLS

Symbol	Definition
$C(t), V, Q$	Hydrogen concentration, enclosure volume, ventilation exchange rate
D, u, k	Turbulent diffusivity, velocity field, first-order consumption coefficient
A, E_a, R, T	Arrhenius pre-exponential factor, activation energy, gas constant, temperature
f_i, i_t, o_t, C_t, h_t	LSTM forget, input, and output gates; cell state; hidden state at time t
$E(x), \theta, \alpha$	Anomaly reconstruction error, adaptive alarm threshold, sensitivity factor
$P(X \epsilon), \text{RRF}$	Posterior explosion probability given evidence; risk-reduction factor
$\Delta t, C_{\text{alarm}}, C_{\text{AI}}$	Response-time margin; hard-alarm setpoint; AI early-warning level
$L, y, \hat{y}$	Cross-entropy loss, true label, predicted intrusion probability

VI. ILLUSTRATIVE EVALUATION AND DISCUSSION

A. Basis of Evaluation

Because no physical AI-S3 prototype was deployed at an operating facility as part of this work, the evaluation presented here is illustrative. It uses simulated and notional sensor data (Fig. 4) to demonstrate how the mathematical framework of Section V would behave, and it reports representative characteristics drawn from the anomaly-detection and intrusion-detection literature reviewed in Section II [9]–[11], [13] rather than plant-specific empirical results. Table IV should be read accordingly, as a methodology-demonstration summary and not as a validated qualification result.

B. Illustrative Performance Summary

TABLE IV
ILLUSTRATIVE COMPARISON OF BASELINE AND PROPOSED AI-S3
CHARACTERISTICS (METHODOLOGY DEMONSTRATION ONLY)

Metric	Fixed-threshold baseline	AI-S3 (illustrative)
Detection lead time Δt , (15)	Zero by definition	Minutes to hours, scenario-dependent
False-alarm sensitivity	Low (fixed setpoint)	Requires threshold tuning per (11)
Cyber-intrusion awareness	None	Present, Section V-D
Basis of figures shown	Not applicable	Notional / simulated, not field-validated

C. Risk-Reduction Discussion

Within the fault-tree formulation of (13), the AI-S3 acts to reduce the effective basic-event probabilities p_i associated with late or missed detection; it does not alter the physical hydrogen-generation source terms of (1) and (2). This is consistent with defense-in-depth practice: the AI-S3 augments, but does not replace, the physical PAR and igniter mitigation layer or the electrical hazardous-area classification layer described in Section III and in [17]. It follows that the risk-reduction factor of (14) is bounded by the fraction of the top-event probability attributable to detection failure, and that a plant whose residual risk is dominated by mitigation-equipment unavailability rather than by late detection would benefit correspondingly less.

D. Qualification Gap and Verification Requirements

Deployment of any AI-informed function in a safety-significant role would require qualification against the safety-integrity-level requirements of IEC 61508 and IEC 61513 [21], [22], formal verification and validation of the trained model's behavior across the plant's credible operating envelope, and a documented cybersecurity assurance case under IEC 62443 and NRC RG 5.71 [14], [15]. None of that qualification work has been performed within the scope of this study; it is identified as essential future work in Section VII-B. The architectural separation of Section IV-D is intended to make that qualification tractable, by confining the safety-significant decision to a deterministic layer that can be exhaustively analyzed, with the learned models acting only as advisory evidence sources.

VII. CONCLUSION AND FUTURE WORK

A. Conclusion

This paper has proposed AI-S3, a layered architecture integrating AI-based anomaly detection, Bayesian predictive risk estimation, and cybersecurity intrusion detection with conventional hydrogen-explosion mitigation equipment for nuclear fission power plants, together with the underlying mathematical models for hydrogen transport, recombiner kinetics, the LSTM anomaly-detection sub-model, Bayesian risk updating, and ICS intrusion classification. The scope was fixed at the outset in Section I-B: the hazard addressed is the chemical hydrogen explosion credible in today's commercial fission fleet, not a nuclear-yield detonation, and the models, architecture, and evaluation presented here are limited throughout to that hazard class. The central design decision is the confinement of all learned components to an advisory analytics layer whose outputs are gated by deterministic arbitration logic, which preserves the independence and single-failure properties required of a Class 1E-associated function while still allowing the detection margin of (15) to be realized.

B. Future Work

Five lines of work are required before the concept could be advanced toward deployment. First, development and training of the LSTM, autoencoder, and intrusion-detection models against representative plant-simulator or test-loop data, replacing the notional data used in Section VI. Second, formal safety-integrity-level qualification and verification and validation of the arbitration-layer logic under IEC 61508 and IEC 61513. Third, quantitative, plant-specific PRA integration of the risk-reduction factor of (14). Fourth, independent cybersecurity penetration testing of the proposed network-segmentation and intrusion-detection design under IEC 62443. Fifth, and as a long-term research direction only, exploratory extension of the AI-S3 sensing and analytics concept to the physically distinct energetic hazards of reactor types not yet in commercial operation, as noted in Section III-E.

REFERENCES

- [1] *IEEE Standard Criteria for Independence of Class 1E Equipment and Circuits*, IEEE Std 384-2008, IEEE, New York, NY, USA, 2008.
- [2] *IEEE Standard Criteria for Safety Systems for Nuclear Power Generating Stations*, IEEE Std 603-2018, IEEE, New York, NY, USA, 2018.
- [3] U.S. Nuclear Regulatory Commission, "Recommendations for enhancing reactor safety in the 21st century: The near-term task force review of insights from the Fukushima Dai-ichi accident," NRC, Washington, DC, USA, 2011.
- [4] International Atomic Energy Agency, "Mitigation of hydrogen hazards in water cooled power reactors," IAEA Nuclear Energy Series No. NP-T-1.9, IAEA, Vienna, Austria, 2011.
- [5] *Explosive Atmospheres*, IEC 60079 series, International Electrotechnical Commission, Geneva, Switzerland.
- [6] *National Electrical Code, Article 500—Hazardous (Classified) Locations*, NFPA 70, National Fire Protection Association, Quincy, MA, USA.
- [7] OECD Nuclear Energy Agency, "Status report on hydrogen management and related computer codes," NEA/CSNI/R(2014)8, OECD/NEA, Paris, France, 2014.
- [8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [9] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in *Proc. Eur. Symp. Artif. Neural Netw. (ESANN)*, Bruges, Belgium, 2015, pp. 89–94.

- [10] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTM-based encoder-decoder for multi-sensor anomaly detection," arXiv:1607.00148, 2016.
- [11] J. Goh, S. Adepur, M. Tan, and Z. S. Lee, "Anomaly detection in cyber physical systems using recurrent neural networks," in *Proc. IEEE 18th Int. Symp. High Assurance Syst. Eng. (HASE)*, Singapore, 2017, pp. 140–145.
- [12] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann, 1988.
- [13] R. Mitchell and I.-R. Chen, "A survey of intrusion detection techniques for cyber-physical systems," *ACM Comput. Surv.*, vol. 46, no. 4, art. 55, Mar. 2014.
- [14] *Security for Industrial Automation and Control Systems*, IEC 62443 series, International Electrotechnical Commission, Geneva, Switzerland.
- [15] U.S. Nuclear Regulatory Commission, "Cyber security programs for nuclear facilities," Regulatory Guide 5.71, NRC, Washington, DC, USA.
- [16] ITER Organization, "ITER research plan and safety basis," ITER Organization, Saint-Paul-lez-Durance, France.
- [17] M. Ali, "Explosion-proof electrical design for nuclear power plant hazardous areas," thesis, School of Science & Engineering, Atlantic International University, Honolulu, HI, USA, 2026. *[Author to confirm exact title, degree, and year before submission.]*
- [18] U.S. Nuclear Regulatory Commission, "Severe accident risks: An assessment for five U.S. nuclear power plants," NUREG-1150, NRC, Washington, DC, USA, 1990.
- [19] *IEEE Recommended Practice for Unique Identification in Power Plants and Related Facilities*, IEEE Std 803-1983 (R2002), IEEE, New York, NY, USA, 2002.
- [20] *IEEE Standard Model Process for Addressing Ethical Concerns During System Design*, IEEE Std 7000-2021, IEEE, New York, NY, USA, 2021.
- [21] *Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems*, IEC 61508, International Electrotechnical Commission, Geneva, Switzerland.
- [22] *Nuclear Power Plants—Instrumentation and Control Important to Safety—General Requirements for Systems*, IEC 61513, International Electrotechnical Commission, Geneva, Switzerland.