International Journal of Smart Grid (IJSmartGrid)
https://www.ijsmartgrid.org/ Online ISSN: 2602-439X
S. Metatla et al., Vol. 10, No. 3, September, 2026 Research Article

Genetic Algorithm Optimized Fractional Order PI Controller for Enhanced Performance in a Grid-Connected PV System

Samir Metatla*‡, Youcef Soufi*, Ali Almaktoof**, Abdelhakim Bendakir*

*Laboratory of Electrical Engineering, University Larbi Tebessi, Tebessa, Algeria

**Department of Electrical, Electronic and Computer Engineering, Cape Peninsula University of Technology

(metatla.samir@univ-tebessa.dz, youcef.soufi@univ-tebessa.dz, almaktoofa@cput.ac.za, abdelhakim.bendakir@univ-tebessa.dz)

‡Corresponding Author: Samir Metatla, University Larbi Tebessi, Tebessa, Algeria, metatla.samir@univ-tebessa.dz

Received: 28.04.2026 Accepted:21.05.2026

Abstract — Maximizing energy yield in large-scale grid-connected photovoltaic (PV) systems requires accurate and fast Maximum Power Point Tracking (MPPT), particularly under rapidly changing irradiance. Classical methods such as Perturb and Observe (P&O) and Incremental Conductance (InCond) perform well under steady irradiance but struggle to achieve rapid tracking, low steady-state oscillations, and robustness simultaneously. To address this limitation, a Fractional Order Proportional and Integral (FOPI) controller tuned by a Genetic Algorithm (GA) and coupled with the InCond MPPT technique is proposed. The control architecture comprises three independent FOPI controllers dedicated to the MPPT loop, the DC-link voltage regulation loop, and the grid current control loop. Each loop was optimized separately using a dedicated Integral Time Absolute Error (ITAE) criterion evaluated under the EN50530 dynamic irradiance profile. Validation was conducted through comprehensive time-domain implementations in MATLAB/Simulink using the Simscape Electrical™ toolbox, employing a rigorous test regimen that included irradiance ramps, abrupt steps from 250 W/m² to 1000 W/m², and a staircase profile replicating progressive partial shading. The results demonstrate that the GA-FOPI controller achieves an average MPPT efficiency of 99.7% compared to 98.8% for a conventional integer-order Proportional-Integral (PI) controller. DC-link voltage overshoot drops from 3.0% (515 V) to 0.8% (504 V), with settling time reduced from 0.085 s to 0.06 s. Grid current Total Harmonic Distortion (THD) falls from 2.23% to 2.15%, well within the IEC 61727 limit of 5%. Both controllers respond identically to irradiance steps (0.017 s), but the FOPI eliminates the power overshoot observed with the integer-order PI. The additional degree of freedom introduced by the fractional integration order λ proves particularly effective in rejecting sudden irradiance disturbances. The tuning procedure described, using a genetic algorithm with loop-specific ITAE criteria, can be directly applied to other utility-scale PV configurations without requiring expert knowledge of fractional-order control.

Keywords: MPPT, fractional-order PI, genetic algorithm, grid-connected PV, ITAE

Cite this article as: S. Metatla, Y. Soufi, A. Almaktoof, and A. Bendakir, "Genetic Algorithm Optimized Fractional Order PI Controller for Enhanced Performance in a Grid-Connected PV System" *International Journal of Smart Grid (ijSmartGrid)*, Vol. 10, No. 3, pp. 159-172, September, 2026.

1. Introduction

Grid-connected photovoltaic (PV) installations have expanded at an unprecedented pace over the past decade, with utility-scale plants now routinely exceeding hundreds of megawatts of installed capacity. Several countries already rely on solar PV for a significant fraction of their electricity mix, and this share is expected to grow further as installation costs continue to fall [1]. At this scale, a plant's economic performance depends critically on how efficiently it converts available solar resources into injected energy. A 1% improvement in annual energy yield on a 100 MW installation translates into several GWh of additional production and a measurable improvement in return on investment over the plant lifetime [2].

The component responsible for capturing that yield is the Maximum Power Point Tracking (MPPT) algorithm, which continuously adjusts the PV array's operating point to extract the maximum available power under varying irradiance and temperature conditions. Because both quantities change throughout the day and can change abruptly when clouds pass, the MPPT algorithm must respond quickly without introducing steady-state oscillations that waste the power it is trying to capture. Meeting both requirements simultaneously is the central challenge that motivates this work.

The current-voltage characteristic of a PV array is nonlinear, with a single maximum power point whose location depends on irradiance and temperature. Classical MPPT techniques fall into two broad categories. Direct methods, principally Perturb and Observe (P&O) and Incremental Conductance (InCond) [3], estimate the MPP from instantaneous measurements without a system model. Indirect methods use a closed-loop controller to regulate the PV operating point to a reference generated by a separate algorithm.

P&O is widely used in practice because it is simple to implement and requires no prior knowledge of the array. Its weakness is a fundamental trade-off: a large perturbation step tracks fast but causes large steady-state oscillations; a small step reduces oscillations but slows tracking under rapid irradiance changes, leading to power loss [4]. Recent work has addressed this by combining P&O with simulated annealing to adapt the step size [5], or by using a neural network to predict the optimal step under the EN50530 standardized irradiance profile [6]. Both approaches improve performance but add implementation complexity and, in the neural network case, require representative training data.

More sophisticated control strategies have also been investigated. Fuzzy logic controllers [7] avoid an explicit system model and handle nonlinearity well, but their rule bases require expert knowledge to design. Neural network controllers [8] can learn complex mappings but need offline training and do not generalize easily to modules with different characteristics. Sliding mode control [9] offers strong robustness to parameter variation but introduces chattering that must be filtered, adding complexity to the power stage design.

Fractional Order Proportional and Integral (FOPI) control has attracted growing interest as an alternative that extends classical PI tuning without requiring a full nonlinear model. Its transfer function is $C(s) = K_p + \frac{K_i}{s^\lambda}$, where the fractional integration order λ is a continuous parameter between 0 and 1. Compared to an integer-order PI, λ provides an independent handle on phase margin: increasing λ adds phase lag at all frequencies, allowing the designer to shape the closed-loop response in ways that are not possible with K_p and K_i alone [10], [11]. This extra degree of freedom is particularly valuable for PV systems, where the small-signal plant gain varies widely with irradiance.

Several research works have applied FOPI to grid-connected PV converters. Bouderres et al. used Particle Swarm Optimization (PSO) to tune a FOPI controller for the three control loops, but using one objective function [12]. Djellad et al. optimized a FOPI for DC-link voltage control using a multilevel inverter using the Whale Optimization Algorithm (WOA) [13]. Prajapati et al. applied a dual-loop FOPI structure to improve DC microgrid voltage stability [14]. In each case, however, none of the cited authors has applied a separate optimization function for the considered control loops.

The present work proposes the use of FOPI controllers in three control loops, each optimized with a dedicated objective function — a significant contribution. Specifically, the MPPT, DC-link voltage, and grid current control loops are each tuned independently using their respective objective functions. Furthermore, most tuning studies use a single objective function shared across all loops, which prevents loop-specific performance shaping. Validation under the EN50530 standardized dynamic irradiance profile, which includes realistic ramp rates and step amplitudes, is also rare, making it difficult to compare results across studies [6], [15], [16]. This paper addresses these three gaps for a 100 kW three-phase grid-connected PV system. The following contributions are made:

- FOPI controllers are designed for all three control loops, MPPT via the boost converter, DC-link voltage via the Voltage Source Inverter (VSI) outer loop, and grid current via the VSI inner loop, replacing conventional integer-order PI controllers at every level. Unlike [13] and [16], which limited fractional-order control to the boost stage, this work treats the full control hierarchy.
- Each loop is tuned independently using a dedicated Integral Time Absolute Error (ITAE) criterion, evaluated under the EN50530 dynamic irradiance profile. Assigning

a separate objective function to each loop allows performance trade-offs to be resolved at the loop level

- rather than globally, which is not possible when a single cost function is shared across all controllers as in [12].
- The nine controller parameters (Kp, Ki, and λ for each loop) are optimized using a GA, which has shown competitive accuracy against PSO, WOA, and Grey Wolf Optimizer (GWO) on comparable FOPI identification problems [15], [16].
- The resulting controller is validated under the EN50530 dynamic irradiance profile associated with a staircase profile that reproduces the abrupt power steps characteristic of partial shading. The staircase profile is more demanding than those used in [13] and [15] and provides a stress test of disturbance rejection outside the training conditions.
- The proposed GA-FOPI controller is compared against a GA-optimized integer-order PI using four standardized metrics: average MPPT efficiency, response time to a 250→1000 W/m² step, DC-link voltage overshoot, and grid current Total Harmonic Distortion (THD) [17].

To further clarify the novelty of the present work, Table 1 summarizes the key features of recent FOPI-based PV systems studies and compares them with our proposed approach. Unlike previous works, which generally use a single or multiple FOPI controllers optimized with a shared cost function, this paper introduces three independent FOPI controllers (MPPT, DC-link, grid current), each tuned with a dedicated ITAE criterion. Moreover, our optimization explicitly uses the EN50530 dynamic irradiance profile, a standard test condition rarely adopted in comparable studies.

Table 1. Comparison of the proposed controller with recent FOPI-based MPPT studies

Reference	Controller	Control loops	Optimizer	Loop specific cost	EN50530 standard	HIL/Exp.
Bouderres et al [12]	FOPI (MPPT)	1	PSO	NO	NO	NO
Djellad et al	FOPI (DC-link)	1	WOA	NO	NO	NO
Prajapati et al	FOPI (droop)	2	PSO	NO	NO	NO
Issaoui et al	FOPI/FOPID	1	GA	NO	NO	NO
Taibi et al	FOPI (DC-link)	1	GA	NO	NO	NO
Our work	FOPI (MPPT, DC-link, Current)	3	Multi GA	YES (3 loops with specific ITAE)	YES	Planned

The paper is structured as follows: Section 2 presents the system model, including the PV array, boost converter, and VSI with LCL filter. Section 3 describes the FOPI controller structure and the GA optimization setup. Section 4 presents simulation results and the comparative analysis. Section 5 draws conclusions and identifies directions for future work.

2. Mathematical Modeling of the Grid-connected PV System Components

Accurate modeling of the grid-connected photovoltaic system is essential for control design and performance evaluation. This section presents the mathematical models of the main subsystems: the PV array, the DC-DC boost converter, the DC-link, and the three-phase voltage source inverter with its LCL filter. Each model is expressed in a form suitable for control-loop analysis and subsequent optimization of the fractional-order PI controllers. The overall system configuration, including the interconnection of these components, is also described.

2.1 The PV Array Model

The SunPower SPR-305 monocrystalline module, rated at 305 W under Standard Test Conditions (STC: 1000 W/m², 25°C), was selected for this study. Its key electrical parameters are: V_{oc} = 64.2 V, I_{sc} = 5.96 A, V_{mp} = 54.7 V, I_{mp} = 5.58 A. The array is built by associating *Np* parallel strings of *Ns* series modules to reach the 100 kW target.

Each module is represented by the single-diode equivalent circuit (Figure 1), which balances accuracy and computational cost [18], [19]. The terminal current is:

$$I_{pv} = N_p I_{ph} - N_p I_s \left[exp\left(\frac{V_{pv}+I_{pv}R_s}{N_s n V_t}\right) - 1 \right] - \frac{V_{pv}+I_{pv}R_s}{R_{sh}} \quad (1)$$

where V_{pv} is the terminal voltage, I_{ph} the photocurrent, I_s the diode reverse saturation current, n the diode ideality factor, $V_t = \frac{kT}{q}$ the thermal voltage, and R_s, R_{sh} the series and shunt resistances. N_s and N_p represent the number of series and parallel modules.

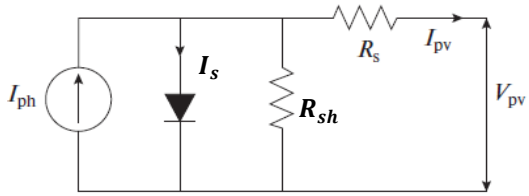


Fig. 1. Equivalent circuit of a PV cell.

2.2 DC-DC Boost Converter Model

A boost converter steps up the PV array voltage to the 500 V DC-link level. Applying state-space averaging to the switching circuit yields two coupled differential equations [20]:

$$L_b \frac{di_{L_b}}{dt} = V_{pv} - (1-d)V_{dc} \quad (2)$$

$$C_{dc} \frac{dV_{dc}}{dt} = (1-d)i_{L_b} - i_{inv} \quad (3)$$

where L_b is the boost inductance, C_{dc} the DC-link capacitance, i_{L_b} the inductor current, V_{dc} the DC-link voltage, i_{inv} the inverter input current, and $D \in [0, 1)$ the duty cycle set by the MPPT controller.

2.3 Grid-Side Inverter and LCL Filter Model

A three-level, three-phase VSI transfers power from the DC link to the grid through an LCL filter. Transforming the three-phase variables into the synchronous *dq* reference frame

rotating at the grid angular frequency ω_g gives six state equations [21], [22]:

$$L_1 \frac{di_d}{dt} = V_{d_{inv}} - V_{d_c} + \omega_g L_1 i_q \quad (4)$$

$$L_1 \frac{di_q}{dt} = V_{q_{inv}} - V_{q_c} - \omega_g L_1 i_d \quad (5)$$

$$C_f \frac{dV_{d_c}}{dt} = i_d - i_{gd} + \omega_g C_f V_{q_c} \quad (6)$$

$$C_f \frac{dV_{q_c}}{dt} = i_q - i_{gq} - \omega_g C_f V_{d_c} \quad (7)$$

$$L_2 \frac{di_{gd}}{dt} = V_{d_c} - V_{gd} + \omega_g L_2 i_{gq} \quad (8)$$

$$L_2 \frac{di_{gq}}{dt} = V_{q_c} - V_{gq} - \omega_g L_2 i_{gd} \quad (9)$$

where L_1 and L_2 are the inverter-side and grid-side filter inductances, C_f the filter capacitor, i_d, i_q the inverter-side currents, V_{d_c}, V_{q_c} the capacitor voltages, i_{gd}, i_{gq} the grid-side currents, and V_{gd}, V_{gq} the grid voltages.

The VSI switching is controlled by Space Vector PWM (SVPWM), chosen for its lower harmonic content compared to sinusoidal PWM [23]. Grid synchronization relies on a Phase-Locked Loop (PLL) that tracks the grid voltage angle in real time [24]. An outer voltage loop regulates V_{DC} to its 500 V reference and generates the d-axis current reference for the inner current loops. A step-up transformer (260 V / 25 kV) connects the VSI output to the medium-voltage distribution grid.

2.4 Overall System Configuration

Figure 2 shows the complete power and control structure. The PV array feeds the boost converter, whose duty cycle is set by the MPPT algorithm. The boost converter output charges the DC-link capacitor, from which the VSI draws current to inject three-phase power into the grid. The three control loops, MPPT, DC-link voltage, and grid current, act on separate actuators and are tuned independently, as described in Section 3.

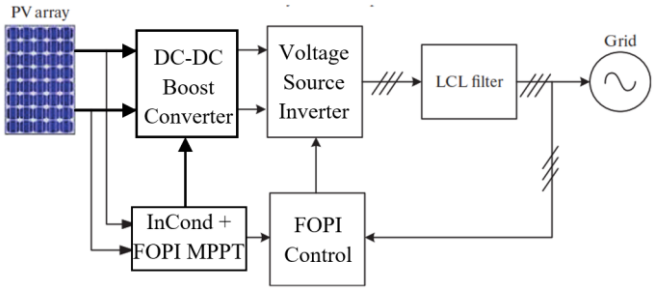


Fig. 2. Overall configuration of the studied grid-connected PV system.

3. Proposed Fractional Order Control of the Grid-Connected PV System

The proposed control strategy employs three independent fractional-order PI controllers, each dedicated to a specific loop within the grid-connected PV system. The first loop integrates the Incremental Conductance MPPT algorithm with a FOPI controller to regulate the boost converter duty cycle, thereby extracting maximum power from the PV array. The second loop maintains the DC-link voltage at its 500 V reference by generating the d-axis current command for the voltage source inverter. The third loop governs the grid-side currents in the synchronous dq frame, ensuring unity power factor operation. Each FOPI controller is tuned independently using a genetic algorithm with a dedicated ITAE objective function, as detailed in Section 3.3. This decentralized tuning approach allows the optimization to address the distinct dynamics of each loop without compromise.

3.1 Incremental Conductance MPPT Technique

Among the direct MPPT methods, InCond was retained here because it locates the MPP analytically rather than by trial perturbation, which makes it better suited to the rapid irradiance steps present in the EN50530 test profile [25].

The operating principle follows directly from the shape of the $P-V$ curve. Writing $P = V.I$ and differentiating with respect to V gives:

$$\frac{dP}{dV} = \frac{d(VI)}{dV} = I + V\frac{dI}{dV} \quad (10)$$

At the MPP, $\frac{dP}{dV} = 0$ by definition, so the tracking condition reduces to:

$$\frac{dI}{dV} = -\frac{I}{V} \quad (11)$$

The left-hand side, $\frac{dI}{dV}$, is the incremental conductance; the right-hand side, $-\frac{I}{V}$, is the instantaneous conductance with a sign change. The algorithm evaluates this inequality at each sampling instant and moves the voltage reference accordingly:

- If $\frac{dI}{dV} > -\frac{I}{V}$, the operating point sits to the left of the MPP and the voltage reference is increased.
- If $\frac{dI}{dV} < -\frac{I}{V}$, the operating point is to the right and the voltage reference is decreased.
- If $\frac{dI}{dV} = -\frac{I}{V}$ within a small tolerance, the MPP has been reached, and the reference is held fixed.
- A fourth case arises when $\frac{dI}{dV} = 0$ while $\frac{I}{V} \neq 0$, which occurs on the nearly flat current plateau of the V − I curve far from the MPP. In that region the algorithm holds the reference unchanged until a change in irradiance shifts the operating point back toward the MPP region.

Compared to P&O, this formulation avoids the ambiguity that arises when a perturbation step and an irradiance change occur in the same sampling interval, since the decision is based on conductance ratios rather than on the sign of a power difference between two successive samples. The InCond flow chart is given in Figure 3 below.

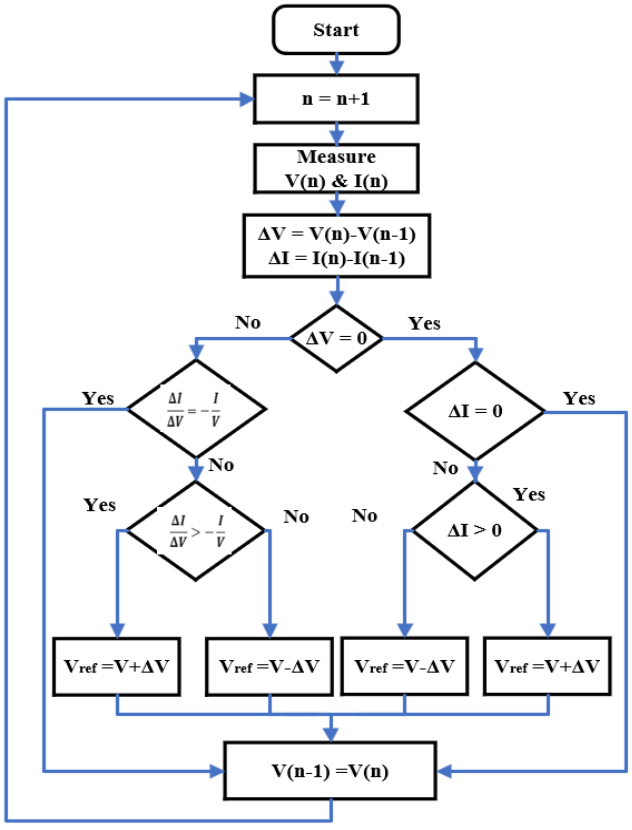


Fig. 3. Flow chart of the Incremental Conductance MPPT technique.

The reference voltage generated by the InCond algorithm is tracked by a FOPI controller, which adjusts the boost converter duty cycle *D*. For a boost converter, the steady-state relationship between the PV terminal voltage and the DC-link voltage is:

$$V_{dc} = \frac{V_{pv}}{1-D} \quad (12)$$

Where V_{dc} is the DC voltage link at the output of the DC-DC converter and V_{pv} is the PV system's output voltage.

In practice, the FOPI controller computes the duty cycle correction ΔD needed to drive V_{PV} toward the InCond reference. Using a fractional-order controller here rather than a standard PI, reduces the hunting around the MPP that a purely proportional-integral action tends to produce when the plant gain changes with irradiance [25].

3.2 The DC voltage Bus and power control

When irradiance changes, the power balance between the PV array and the grid shifts, causing the DC-link voltage to deviate from its 500 V setpoint. A dedicated FOPI controller monitors this deviation and adjusts the d-axis current reference Id to restore the voltage. Because the DC-link capacitor acts as a buffer between the two sides of the conversion chain, keeping Vdc stable enables the MPPT loop and the grid injection loop to operate independently.

Grid power injection is handled by two current controllers working in the dq frame. The d-axis current controls active power transfer; its reference is derived directly from the DC-link voltage controller. The q-axis current reference is set to zero, forcing the reactive power to zero and keeping the power factor at unity. This arrangement means that all power extracted from the PV array is delivered to the grid with no reactive component, regardless of irradiance level.

3.3 Fractional Order Proportional and Integral Controller Structure

The three controllers share the same structure, known as PI^{λ}, whose transfer function is [11]:

$$C(s) = K_p + \frac{K_i}{s^{\lambda}} \quad (13)$$

where K_p is the proportional gain, K_i the integral gain, and $\lambda \in (0, 1)$ the fractional integration order.

The fractional integration order λ provides an additional degree of freedom to the FOPI controller. In a standard Proportional-Integral (PI) controller, the integration order is fixed at 1, which means the phase contribution of the integral term is always −90°. Setting λ < 1 reduces this phase lag, allowing the designer to increase the gain crossover frequency, and therefore the closed-loop bandwidth, without sacrificing phase margin [10]. For PV converters, whose small-signal plant gain varies by a factor of four or more between 250 W/m² and 1000 W/m², this extra degree of

freedom translates into more consistent transient performance across the full irradiance range [26]. Analytically deriving K_p, K_i, and λ from the plant transfer functions of the three loops would require several simplifying assumptions and would not account for the interactions between loops. Parameter selection is therefore treated as an optimization problem, solved using a GA.

3.4 Genetic Algorithm Optimization Framework

The parameter vector for each controller is $x = [K_p, K_i, \lambda]$. Three independent GA runs are performed, one per control loop using a population of 15 candidates evolved over 20 generations. The search bounds are set to $K_p, K_i \in [0, 500]$ and $\lambda \in (0, 1)$.

Each candidate is evaluated by simulating the full system under the EN50530 dynamic irradiance profile and computing a loop-specific ITAE cost function. Assigning a separate cost to each loop rather than a single global criterion as in [12], allows the GA to resolve the performance trade-offs of each loop on its own terms. The three objective functions are:

- For the MPPT loop:

$$J_{MPPT}(x) = \int_0^{t_{sim}} t.\left|\frac{dI}{dV} + \frac{I}{V}\right| dt \quad (14)$$

- For the DC-bus voltage loop:

$$J_{Vdc}(x) = \int_0^{t_{sim}} t.\left|V_{DC_Ref} - V_{DC_mes}\right| dt \quad (15)$$

- For the grid current control loop:

$$J_{Current}(x) = \int_0^{t_{sim}} t.\left|P_{PV} - P_{grid}\right| dt \quad (16)$$

where t_{sim} is the optimization horizon, V_{DC_Ref} and V_{DC_mes} are the DC-link voltage reference and measurement, and P_{PV} and P_{grid} are the PV output power and the power delivered to the grid.

The ITAE weighting, multiplying the absolute error by time penalizes errors that persist into steady state more heavily than early transients, which naturally rewards both fast initial tracking and low residual oscillation [27].

GA follows a standard generational loop: tournament selection picks parent pairs, single-point crossover produces offspring, and random mutation maintains population diversity. Elitism carries the best candidate from each generation forward without modification. The loop terminates when the maximum generation count is reached. Figure 4 summarizes the full optimization flowchart.

The optimal solution $x^* = [K_p^*, K_i^*, \lambda^*]$ for each loop, the candidate is the one that minimizes its respective cost function over the EN50530 profile. These nine values are then fixed for all subsequent simulations, including the staircase validation profile described in Section 4.

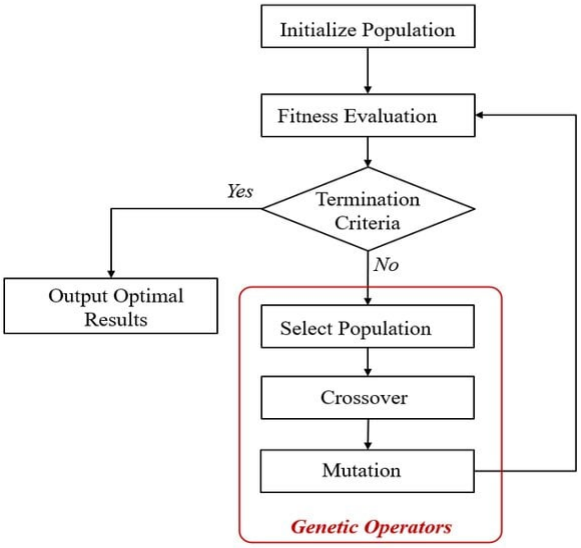


Fig. 4. Flowchart of the Genetic Algorithm optimization process

Figure 5 brings together the three control loops described in Sections III-A through III-E into a single block diagram. Reading left to right: the InCond algorithm generates the voltage reference, the MPPT FOPI controller sets the boost duty cycle, the DC-link voltage controller feeds the d-axis current reference to the VSI inner loop, and the grid current controllers produce the final SVPWM switching signals. The nine GA-optimized parameters for the three per loop are fixed at the values listed in Table 3.

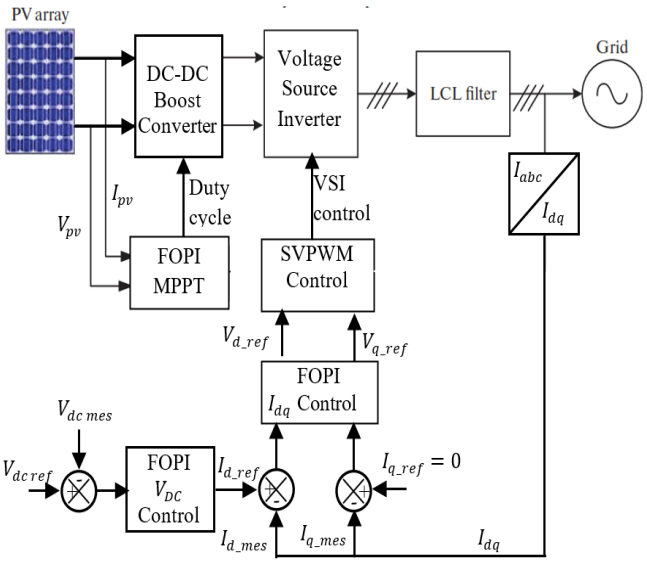


Fig. 5. The proposed control scheme of the grid-connected PV system.

4. Results and Discussion

This section presents the validation results of the proposed GA-optimized fractional-order PI controller applied to the 100 kW three-phase grid-connected PV system. The system parameters, test profiles, and controller settings are first described to establish the validation framework.

Subsequently, the performance of the GA-FOPI controller is compared against a conventional GA-optimized integer-order PI controller across the complete power conversion chain: from PV array power extraction, through DC-link voltage regulation, to grid current injection. Quantitative metrics, including MPPT efficiency, voltage overshoot, settling time, and total harmonic distortion, are extracted from the time-domain waveforms and harmonic spectra. All validation tests were conducted under identical dynamic irradiance profiles comprising ramps, abrupt steps, and a staircase pattern, as prescribed by the EN50530 standard. The section is organized into three integrated subsections: (A) System Under Test and Validation Setup, (B) DC-Link and MPPT Performance, and (C) Grid Integration and Power Quality.

4.1 System Under Test and Validation Setup

The physical system parameters, implementation environment, and dynamic irradiance test profiles must first be established to ensure reproducibility and provide context for the results that follow.

The complete 100 kW grid-connected PV system described in Section 2 uses 305 W monocrystalline modules (SunPower SPR-305) with the following key parameters at Standard Test Conditions (STC: 1000 W/m², 25°C): $V_{oc} = 64.2$ V , $I_{sc} = 5.96$ A , $V_{mp} = 54.7$ V , $I = 5.58$ A. The current-voltage (I-V) and power-voltage (P-V) characteristics of the SunPower SPR-305 PV module under variable irradiation at 25°C are given in Figure 6.

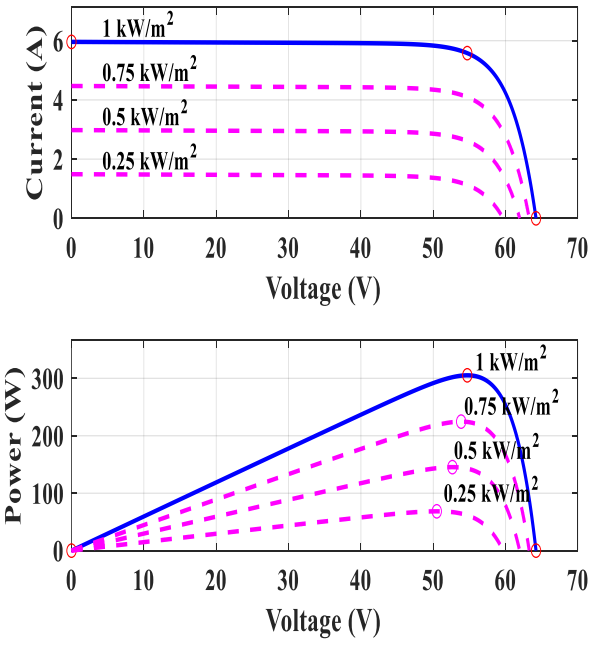


Fig. 6. The I-V and P-V characteristics of the SunPower SPR-305 PV module.

The PV system studied in this paper consists of 66 parallel strings, each composed of 5 series-connected SPR-305 modules. The system parameters are summarized in Table 2.

Table 2. System parameters

Component	Parameter	Value
PV Array (STC)	Rated Power	100.65 kW
	V_{mp} (string)	273 V
	I_{mp} (PV)	368 A
Boost Converter	L	5 mH
	C	100 µF
DC Link	V_{dc}	500 V
	C	6 mF
LCL Filter	L_1 (Inverter side)	150 µH
	L_2 (Transformer side)	50 µH
	C	200 µF
Transformer	V_1/V_2 line-to-line RMS	260 V / 25 kV
	Rated Power	100 kVA
Grid	Line-to-line RMS voltage	25 kV, 60 Hz

The PV system was implemented in the MATLAB/Simulink environment using the Simscape Electrical™ toolbox, as presented in Figure 7.

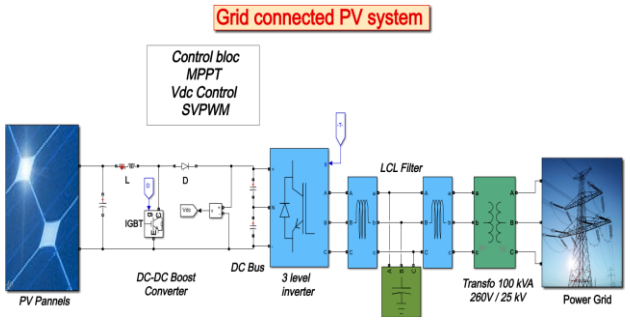


Fig. 7. Implemented control block on MATLAB Simulink.

A dynamic irradiance profile, including both low- and high-irradiance ramps, was used to test the MPPT algorithms under realistic conditions, as shown in Figure 8. The validation was performed under variable solar irradiation profiles comprising three distinct test scenarios. First, an ascending then descending ramp between 30% and 100% of nominal irradiation (per EN50530 standard) assessed the controller response to gradual changes. Second, an abrupt step from 250 W/m² to 1000 W/m² emulated severe, sudden variations typical of significant panel shadowing. Finally, a staircase profile was modeled to replicate progressive partial shading across the PV array.

The genetic algorithm was configured with a population size of 20 elements and 15 generations, executed with parameters specified in Section 3.4 for both integer-order and fractional-order PI controllers. The convergence of the best fitness value (ITAE) over generations yielded the optimal controller parameters, as given in Table 3. The optimized FOPI controller was then used for all subsequent validations.

Both controllers were tested over a five-second period under identical dynamic irradiance profiles. The resulting waveforms are presented in Figures 9 through 19.

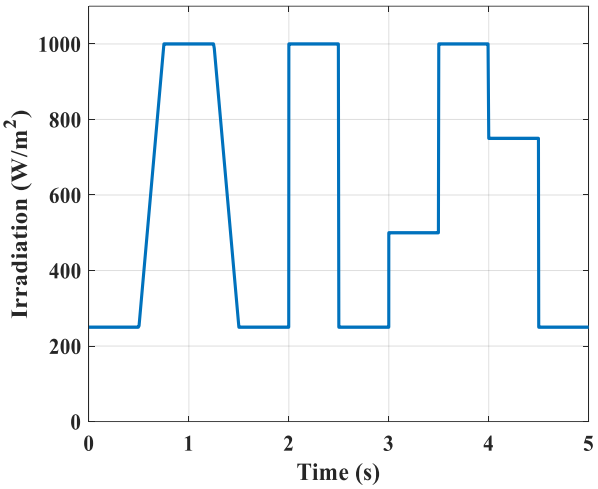


Fig. 8. Irradiance profile.

Table 3. Optimal controllers' parameters

Controller	Control loop	parameter
Integer order PI	MPPT control	$K_p = 75.3, K_i = 17.5$
	V_{dc} control	$K_p = 27, K_i = 453$
	Current control	$K_p = 3.1, K_i = 20$
Fractional order PI	MPPT control	$K_p = 216.7, K_i = 109.8, \lambda = 0.57$
	V_{dc} control	$K_p = 45, K_i = 246, \lambda = 0.81$
	Current control	$K_p = 13.7, K_i = 3.1, \lambda = 0.84$

4.2 DC-Link and MPPT Performance

The MPPT loop and the DC-link voltage regulation are physically coupled through the boost converter and are therefore analysed together to reflect their interaction in the actual system.

Figure 9 illustrates the impact of the irradiation profile on PV output voltage. Variations in irradiance induce substantial changes in PV output voltage, challenging the control system's ability to track the Maximum Power Point (MPP).

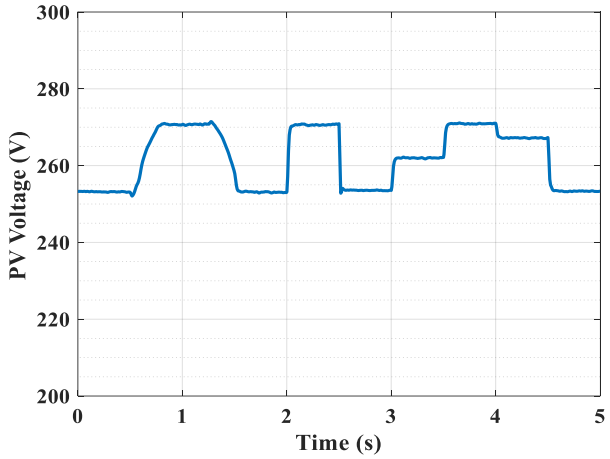


Fig. 9. PV output voltage.

Figure 10 presents the power generated by the PV system under the variable irradiance profile shown in Figure 8. The response of the GA-optimized integer-order PI controller is compared against that of the proposed GA-optimized fractional-order PI controller. Two regions of interest, where the integer-order PI exhibits noticeable deviations from the optimal power curve, are circled for closer examination.

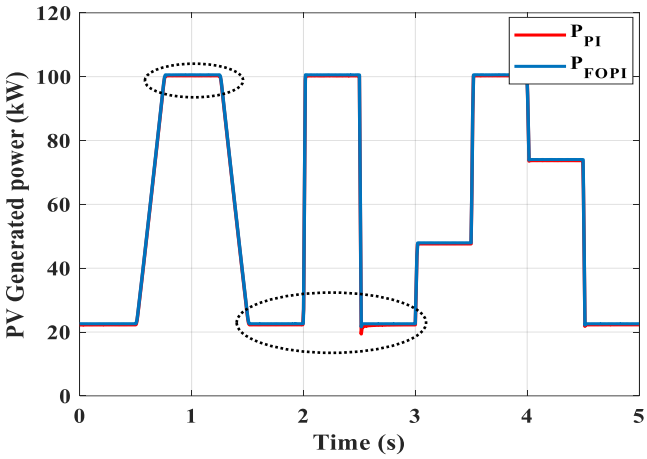
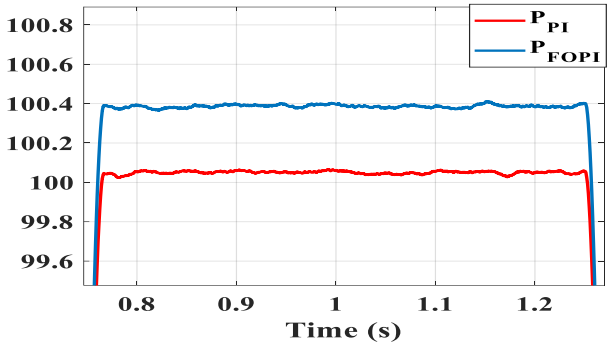
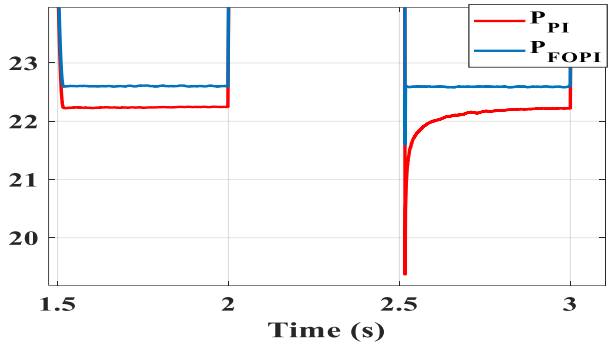


Fig. 10. PV Power generated.

Both controllers demonstrate excellent tracking of the maximum power point, with fast response times during abrupt irradiance variations. However, as observed in the two circled regions, the integer-order PI controller's response exhibits a clear deviation from the optimal power curve accompanied by significant power overshoot during irradiance steps. The two circled regions from Figure 10 are enlarged in Figures 11(a) and 11(b) to provide a detailed view of the overshoot behavior.



(a)



(b)

Fig. 11. Zoom on PV Power generated: (a) first circled region and (b) second circled region from Figure 10.

The zoomed views confirm that the integer-order PI controller suffers from pronounced power overshoot following sudden irradiance steps. In contrast, the fractional-order PI controller tracks the optimal power curve more smoothly, eliminating the overshoot entirely. This drawback of the integer-order PI controller can significantly affect the generated power and degrade overall PV system efficiency. Both integer-order and fractional-order controllers provide effective maximum power point tracking under the tested irradiance profile; however, the FOPI controller exhibits consistently better performance. Specifically, at 1000 W/m², the maximum achievable power is 100.65 kW, yielding an efficiency of 99.6% for the FOPI controller compared to 99.4% for the integer-order PI. At 250 W/m², the advantage becomes more pronounced: the FOPI controller achieves approximately 99.8% efficiency, whereas the integer-order PI reaches only 98.2%.

The PV system generates direct current, which must be converted to alternating current before being transmitted to the electrical grid. The DC bus plays a crucial role in this stage, as it decouples the boost converter output from the inverter input. The stability of the DC bus voltage is essential to ensure smooth, stable, and optimal power flow. Figure 12 presents the DC bus voltage profiles for both controllers over the entire five-second simulation period, with the reference set to 500 V. Two regions of interest, where transient disturbances cause visible deviations from the reference, are circled for closer examination.

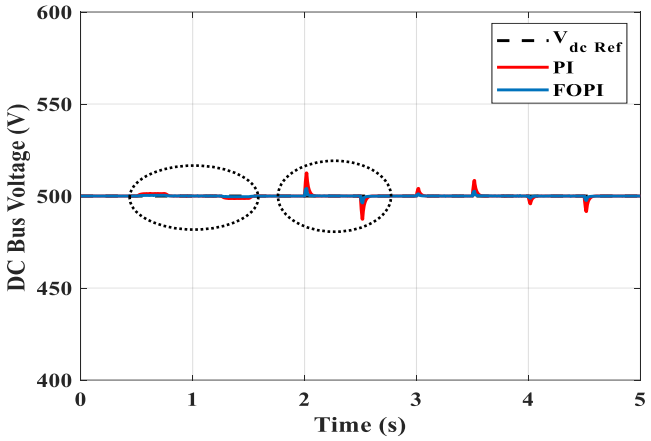
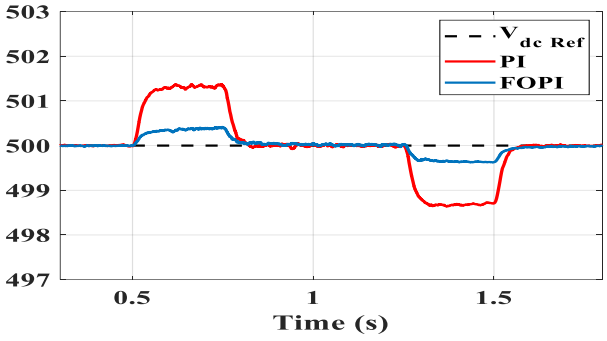


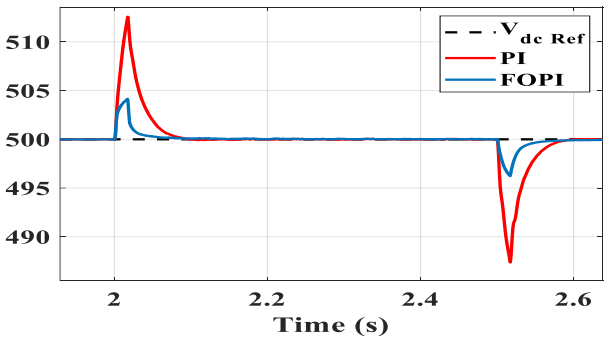
Fig. 12. DC bus voltage.

Both controllers maintain the DC bus voltage relatively stable around its 500 V reference throughout most of the simulation. However, irradiation transients propagate through the MPPT loop and the DC-link dynamics, leading to more pronounced voltage fluctuations, particularly with the integer-order PI controller. As indicated by the two circled regions, the integer-order PI exhibits larger amplitude deviations and slower recovery compared to the FOPI controller. These regions are presented in detail in Figures 13(a) and 13(b), respectively.

The two circled regions from Figure 12 are enlarged in Figures 13(a) and 13(b) to provide a detailed view of the DC bus voltage dynamics during sudden irradiance steps. Figure 13(a) captures the response during an ascending step, while Figure 13(b) captures the response during a descending step.



(a)



(b)

Fig. 13. Zoom on DC bus voltage: (a) first circled region and (b) second circled region from Figure 12.

The zoomed views confirm that the FOPI controller exhibits superior disturbance rejection, achieving faster damping and minimal deviation from the 500 V reference. In contrast, the integer-order PI controller exhibits significant overshoot and a prolonged settling time. Quantitatively, during a sudden irradiance step, the DC bus voltage experiences an overvoltage peak of 515 V (3.0% overshoot) for the integer-order PI controller, whereas it does not exceed 504 V (0.8% overshoot) for the FOPI controller. The settling time after the disturbance is reduced from 0.085 s with the integer-order PI to 0.06 s with the FOPI. These results demonstrate the superior voltage regulation capability of the fractional-order controller under transient irradiance conditions.

4.3 Grid Integration and Power Quality

The final stage of the power conversion chain involves the inverter, the grid filter, and compliance with power quality standards. These elements are examined together to assess how cleanly and efficiently the extracted PV power is delivered to the utility grid.

Grid integration of the PV system requires a three-phase inverter and a step-up transformer to adapt the inverter output voltage (approximately 260 V) to the medium-voltage utility grid level (25 kV). In this study, a three-level three-phase inverter controlled by space vector pulse width modulation (SVPWM) was employed to achieve optimal power transfer with high power quality. Figure 14 presents the inverter output voltage waveforms for both controllers at the transformer primary side.

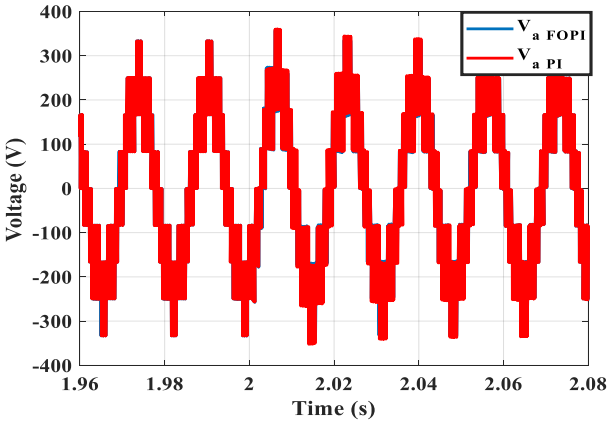


Fig. 14. Output inverter voltage.

The voltage waveforms are nearly identical for both controllers, as the SVPWM switching strategy remains unchanged. The sinusoidal shape is well-maintained with minimal distortion, confirming proper inverter operation. The amplitude remains stable regardless of irradiance variations because the DC-link voltage is actively regulated to 500 V.

Figure 15 illustrates the inverter output current waveforms supplying the step-up transformer for the integer-order PI and FOPI controllers, respectively. The current amplitude directly reflects the power extracted from the PV array, as the inverter operates at unity power factor. A region of interest, where the integer-order PI exhibits current overshoot following a sudden irradiance step, is circled for closer examination.

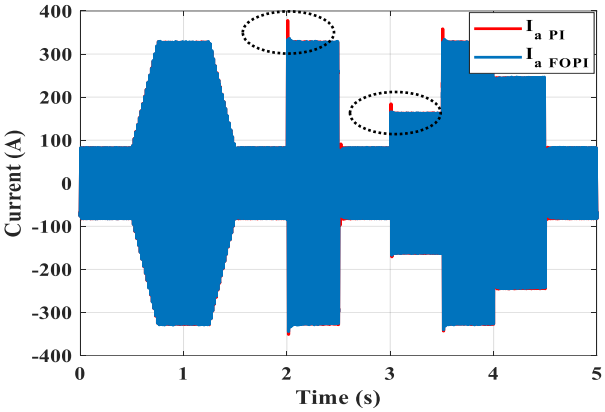


Fig. 15. Output inverter current.

For both controllers, the current amplitude dynamically tracks variations in solar irradiation, ensuring optimal and seamless power transfer from the PV system to the utility grid. However, as indicated by the circled region, the integer-order PI controller exhibits a noticeable current overshoot due to sudden steps in solar irradiation. This overshoot is absent from the FOPI controller's response, indicating superior transient current regulation.

The two circled regions from Figure 15 are enlarged in Figures 16(a) and 16(b) to provide a detailed view of the inverter current overshoot behavior during irradiance steps.

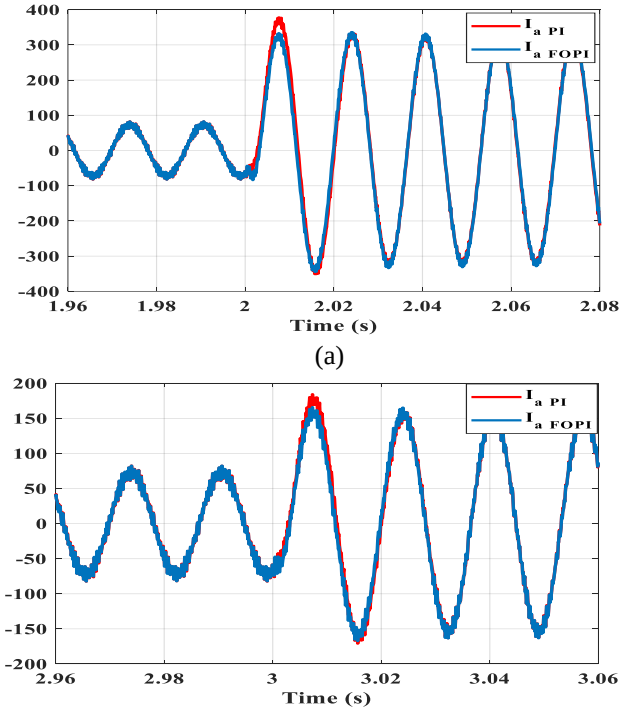


Fig. 16. Zoom on inverter current; (a) first circled region, (b) second circled region from Figure 15.

The zoomed views confirm that the integer-order PI controller suffers from a pronounced current overshoot

immediately following sudden irradiance increases. In contrast, the fractional-order PI controller responds smoothly without exceeding the steady-state value. This overshoot in the integer-order PI controller can affect the quality of the power delivered to the grid.

Power quality is a critical factor in electrical grids. The international standard IEC 61727 requires the injected current's total harmonic distortion to be below 5% at the point of common coupling. Compliance with this standard is therefore essential for grid-connected PV systems [17], [28].

Fast Fourier Transform (FFT) analysis is performed to quantify the harmonic distortion of the current injected into the utility grid. Figures 17 and 18 present the harmonic spectra of the inverter output current for the FOPI and integer-order PI controllers, respectively. The THD values obtained from these spectra are compared against the IEC 61727 limit of 5% to assess power quality compliance.

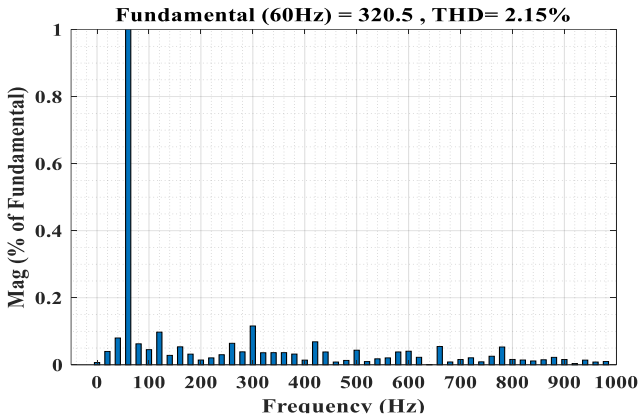


Fig. 17. THD of the inverter output current with FOPI controller.

The FOPI controller achieves a THD of 2.15% at the fundamental frequency of 60 Hz. All harmonic components are well suppressed, and the measured THD is significantly below the 5% compliance threshold. This confirms that the proposed fractional-order control strategy delivers high-quality grid current with minimal harmonic pollution.

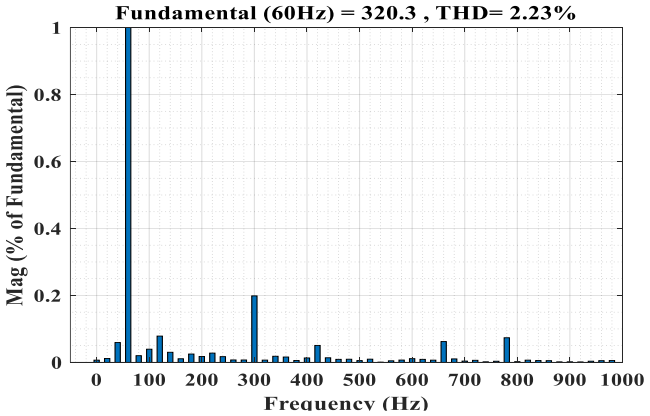


Fig. 18. THD of the inverter output current with integer order PI controller.

The integer-order PI controller yields a THD of 2.23%, which also satisfies the IEC 61727 requirement. However, compared to the FOPI controller, the harmonic distortion is slightly higher. This marginal increase is consistent with the larger oscillations and overshoot observed earlier in the time-domain current waveforms (Figures 15 and 16). The FOPI controller's additional degree of freedom (λ) enables smoother current regulation, directly contributing to the measured reduction in THD.

Finally, Figure 19 overlays the inverter output voltage and current waveforms to verify the unity power factor operation. The current and voltage are perfectly in phase, confirming that the q-axis current reference is maintained at zero and that all active power extracted from the PV array is delivered to the grid without reactive power circulation.

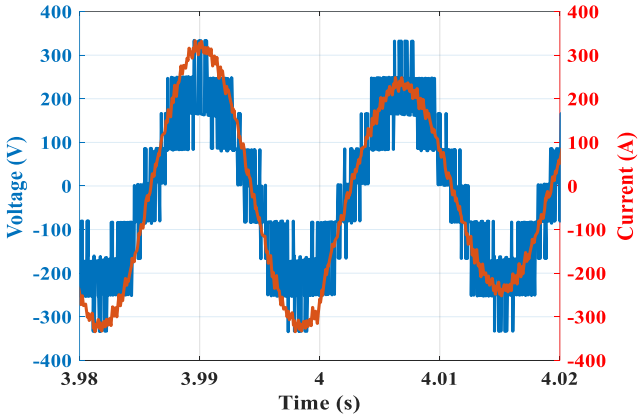


Fig. 19. Zoom on inverter current and voltage.

Table 4 consolidates the key performance metrics obtained for both controllers under identical operating conditions.

Table 4. Performance comparison between integer-order PI and FOPI controllers

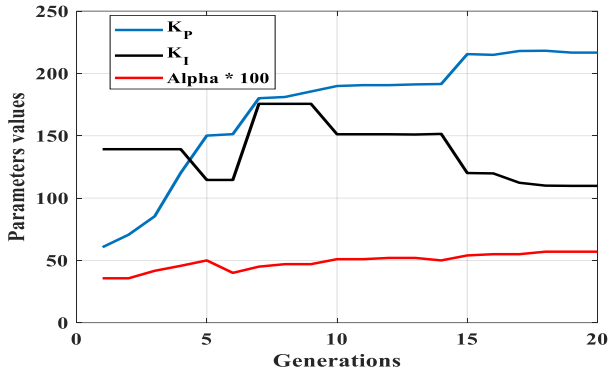
	Integer PI	FOPI
Average MPPT efficiency (%)	98.8	99.7
Response time to a 250→1000 W/m² (s)	0.017	0.017
DC voltage overshoot (%)	3.0	0.8
Setting time (s)	0.085	0.06
THD (%)	2.23	2.15

The proposed GA-FOPI controller achieves an average MPPT efficiency of 99.7%, an improvement of 0.9 percentage points over the integer-order PI (98.8%), with the largest gain observed at low irradiance. Both controllers respond identically to a 250→1000 W/m² irradiance step, settling in 0.017 s; however, the FOPI eliminates the power overshoot that accompanies the integer-order PI response. The DC-link voltage overshoot is reduced from 3.0% (515 V) to 0.8% (504 V), and the settling time improves from 0.085 s to 0.06 s. Grid current THD decreases from 2.23% to 2.15%, remaining well

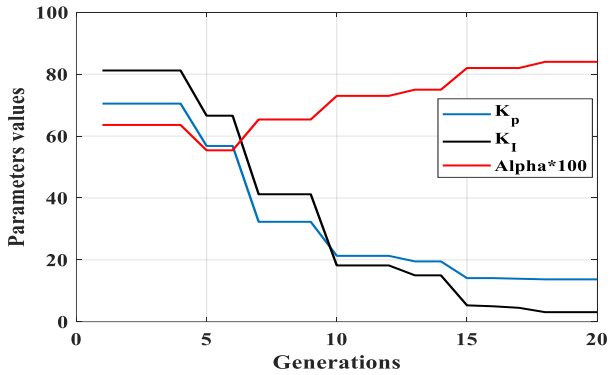
within the IEC 61727 limit of 5%. These results confirm that the additional degree of freedom provided by the fractional integration order λ enables superior disturbance rejection and reduced steady-state oscillations without compromising tracking speed.

4.4 Genetic Algorithm convergence and parameter sensitivity analysis

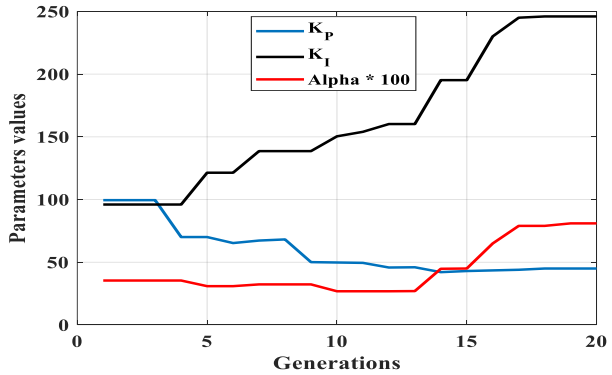
To verify the stability of the optimization process, the evolution of the three FOPI parameters (K_p, K_I , λ) over the 20 GA generations is plotted for each control loop in Figures 20(a)–20(c). In the MPPT loop (Figure 20(a)) and the DC bus voltage loop (Figure 20(b)), all parameters converge to stable values after approximately 15 generations, with only minor adjustments thereafter. However, the grid current loop (Figure 20(c)) exhibits a smooth convergence profile, with all parameters settling well before generation 18.



(a)



(b)



(c)

Fig. 20. Evolution of the FOPI parameters during GA optimization (a) MPPT loop, (b) DC-bus voltage loop, (c) current loop

Figure 21 presents the convergence of the best fitness (ITAE) for the three control loops over the 20 generations. The cost function decreases rapidly during the first 10 generations and then reaches a clear plateau before generation 15 for the MPPT and DC bus loops, and before generation 18 for the current loop. This indicates that the chosen GA settings (population size 15, 20 generations) are sufficient to obtain optimal or near optimal parameters. Extending the optimization beyond 20 generations would not bring a significant improvement, while increasing the computational cost.

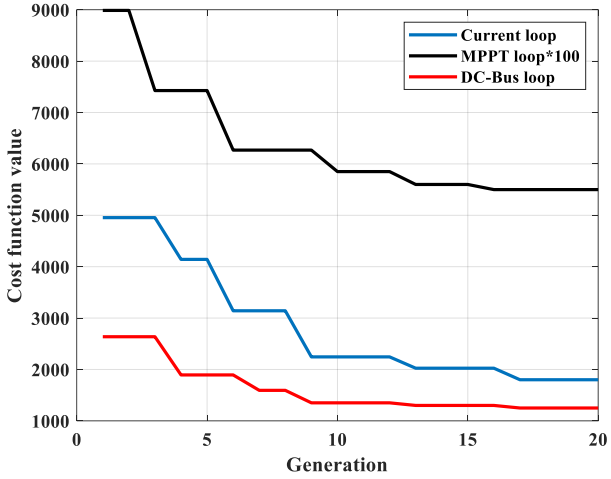


Fig. 21. Convergence of the best fitness (ITAE) for the three control loops over 20 generations.

5. Conclusion

This paper proposed a GA-optimized fractional-order proportional-integral controller for a 100 kW three-phase grid-connected photovoltaic system. The control architecture comprised three independent FOPI controllers, each dedicated to a specific loop: the maximum power point tracking loop (combined with the Incremental Conductance algorithm), the DC-link voltage regulation loop, and the grid current control loop in the synchronous *dq* frame. Each controller was tuned separately using a dedicated Integral Time Absolute Error cost function evaluated under the EN50530 dynamic irradiance profile, enabling loop-specific performance shaping that is not possible with a single shared objective function.

Validation of the proposed control framework was conducted through comprehensive time-domain simulations using MATLAB/Simulink with the Simscape Electrical™ toolbox, employing a rigorous test regimen that included irradiance ramps, abrupt steps from 250 W/m² to 1000 W/m², and a staircase profile replicating progressive partial shading. The results consistently demonstrate that the GA-FOPI framework outperforms a conventional integer-order PI controller optimized under identical conditions. The average MPPT efficiency improved from 98.8% to 99.7%, with the most notable gain occurring at low irradiance (250 W/m²),

where efficiency rose from 98.2% to 99.8%. Under sudden irradiance steps, the DC-link voltage overshoot was reduced from 3.0% (515 V) to 0.8% (504 V), and the settling time shortened from 0.085 s to 0.06 s. The grid current total harmonic distortion decreased from 2.23% to 2.15%, remaining well below the IEC 61727 limit of 5%, while unity power factor operation was maintained throughout. Both controllers achieved identical response times of 0.017 s to a 250→1000 W/m² irradiance step; however, the FOPI controller eliminated the power overshoot and deviation from the maximum power point observed with the integer-order PI.

These improvements are attributed to the fractional integration order λ, which provides an additional degree of freedom compared to conventional PI tuning. This extra parameter enables better phase-margin shaping and more consistent transient performance across the wide range of plant gains encountered under varying irradiance. The tuning procedure described herein, using a genetic algorithm with loop-specific ITAE criteria, can be directly applied to other utility-scale PV configurations without requiring expert knowledge of fractional-order control.

Building on these findings, future work will focus on experimental validation using a hardware-in-the-loop platform, extending the controller to partial shading scenarios and weak grid conditions, and integrating with advanced control strategies such as model predictive control.

Acknowledgements

We would like to thank editors, managers and members of the "International Journal of Smart Grid" for their valuable remarks and suggestions.

Author Contributions

Samir Metatla and Abdelhakim Bendakir contributed to the research mmethodology, writing original draft as well as the formal analysis and references.

Youcef Soufi and Ali Almaktoof contributed to the data curation, methodology, supervision, writing-review & and drafting of the first version.

All authors have read and approved the published version of the article.

Conflict of Interest

The authors declared no potential conflicts of interest.

References

[1] M. L. Katche, A. B. Makokha, S. O. Zachary, and M. S. Adaramola, "A comprehensive review of maximum power point tracking (MPPT) techniques used in solar PV systems," Energies, vol. 16, no. 5, p. 2206, 2023, doi: 10.3390/en16052206

[2] A. Hassan, J. Ahmed, S. Papadopoulos, F. Kahwash, and K. Goh, "A comprehensive review of frequency response and control strategies for grid-connected solar photovoltaic systems," Renewable and Sustainable Energy Reviews, vol. 226, p. 116324, 2026, doi: 10.1016/j.rser.2025.116324

[3] Y. Liu, J-H. Huh, "Editorial of Maximum Power Point Tracking Methods for PV Systems in Micro-Grids," Energies, vol. 18, no.21, pp. 1-7, Nov. 2025, doi: 10.3390/en18215830

[4] J. Chandola, S. Pundir, A. Sharma, Y. Song, G. Fekete, T. Singh," Recent advances in MPPT techniques for photovoltaic systems: A review of classical (P&O, IC), intelligent (ANN), optimization (PSO) and hybrid (ANN-PSO) methods, Results in Engineering, Vol. 29, 109395, March 2026, doi:10.1016/j.rineng.2026.109395

[5] A. G. Abo-Khalil, I. I. El-Sharkawy, A. Radwan, and S. Memon, "Influence of a hybrid MPPT technique, SA-P&O, on PV system performance under partial shading conditions," Energies, vol. 16, no. 2, p. 577, 2023, doi: 10.3390/en16020577

[6] H. Kim, Y. Seo, S. Kang, D. Lim, and E. Noh, "Performance verification of a novel neural network-based P&O MPPT control method using EN50530 dynamic test procedure," Journal of Power Electronics, vol. 30, no. 4, pp. 331-339, Aug. 2025, doi: 10.6113/TKPE.2025.30.4.331

[7] B. Janapati, N. R. Pulivarthi, "A power quality enhancement using gradient descent-fuzzy logic controller in grid-connected PV system, Electric Power Systems Research, Vol 255, 112746, 2026, doi: 10.1016/j.epsr.2026.112746

[8] M. Mishra, P. Mahajan, and R. Garg, "Implementation and comparison of metaheuristically modified ANN MPPT controllers under varying solar irradiance conditions," Electrical Engineering, vol. 106, pp. 3427–3443, 2024, doi: 10.1007/s00202-023-02165-y

[9] I. Rahul and R. Hariharan, "Enhancement of solar PV panel efficiency using double integral sliding mode MPPT control," Tsinghua Science and Technology, vol. 29, no. 1, pp. 271–283, Feb. 2024, doi: 10.26599/TST.2023.9010030

[10] I. Podlubny, "Fractional-order systems and PIλDμ controllers," IEEE Trans. Autom. Control, vol. 44, no. 1, pp. 208–214, Jan. 1999.

[11] K. Aseem, M. Jayakumar, P. Pramod, B. C. Anilkumar, and M. S. Divakar, "Fractional order PID controller incorporated decoupled control of grid connected solar photovoltaic system," Applied Solar Energy, vol. 60, no. 1, pp. 127-136, 2024, doi: 10.3103/S0003701X23601849

[12] N. Bouderres, D. Kerdoun, A. Djellad, S. Chiheb, and A. Dekhane, "Optimization of fractional order PI controller by PSO algorithm applied to a grid-connected photovoltaic system," Journal European des Systemes Automatises, vol. 55, no. 4, pp. 427–438, 2022, doi: 10.18280/jesa.550401

[13] A. Djellad, A. Dekhane, M. Farhat, M. Takruri, and A. Al-Qaisi, "Whale optimization-based fractional order

control for high-performance gridconnected photovoltaic multilevel inverters: diode-clamped versus T-type MLIs," Scientific Reports, vol. 15, no. 1, p. 36695, Oct. 2025, doi: 10.1038/s41598-025-20734-y

[14] N. Prajapati, N. Singh, and A. K. Pandey, "Advanced droop control with PSO-optimized FOPI dual loop for DC microgrid stability under disturbances," Engineering Research Express, vol. 7, no. 2, p. 025351, May 2025, doi: 10.1088/2631-8695/add8e6

[15] R. Issaoui, A. Mesdidi, and O. Naifar, "Comparative analysis of fuzzy PI and fractional order PID optimized by genetic algorithm for HESG," in 2025 IEEE 9th Forum on Research and Technologies for Society and Industry (RTSI), 2025, pp. 272–277, doi: 10.1109/RTSI64020.2025.11212277

[16] A. Taibi, L. Kouider, H. Rezk, A. Rouibah, and A. Al-Quraan, "Optimal parameter identification of fractional-order proportional integral controller to improve DC voltage stability of photovoltaic/battery system," International Journal of Power Electronics and Drive Systems (IJPEDS), vol. 16, no. 1, pp. 519–529, Mar. 2025, doi: 10.11591/ijpeds.v16.i1.pp519-529

[17] Y. A. Benet Raja and A. Thangaraj, "Enhanced harmonic reduction and energy optimization in grid-connected photovoltaic-battery systems for improved power quality," Journal of Energy Storage, vol. 150, p. 120231, 2026, doi: 10.1016/j.est.2025.120231

[18] D. Kim, U. F. Ramadhan, S. U. Islam, S. Jung, and M. Yoon, "Design and implementation of novel fault ride through circuitry and control for grid-connected PV system", Sustainability vol. 14, no. 15, p. 9736, 2022, doi: 10.3390/su14159736

[19] J. S. Koh, R. H. G. Tan, N. M. L. Tan, and W. H. Lim, "A deterministic double-exponential maximum power point tracking algorithm for PV string complex partial shading condition," Computers and Electrical Engineering, vol. 128, p. 110735, 2025, doi: 10.1016/j.compeleceng.2025.110735

[20] N. Bouderres, A. Djellad, D. Kerdoun, A. Dekhane, I. Attoui, and S. Chiheb, "Performance analysis of the multi-level boost converter (MLBC) connected in a photovoltaic system," Journal of Renewable Energies, vol. 1, no. 1, p. 49, 2022, doi: https://doi.org/10.54966/jreen.v1i1.1038

[21] R. Teodorescu, M. Liserre, and P. Rodriguez, Grid Converters for Photovoltaic and Wind Systems. Hoboken, NJ, USA: Wiley, 2011.

[22] W. Alhosaini, M. Aly, E. M. Ahmed, and A. Shawky, "Optimized grid-connected three-phase photovoltaic inverter system using cascaded FOPIT-FOPI fractional controller," Energy Reports, vol. 13, pp. 3324–3339, 2025, doi: 10.1016/j.egyr.2025.02.055

[23] Z. Ben Mahmoud, A. Khedher, "A new simplified SVPWM technique for transformerless photovoltaic multilevel inverter with common mode voltage mitigation," Sustainable Development of Renewable Energy, edited by Mejdi Jeguirim, pp. 107-129, 2026, doi: 10.1016/B978-0-443-33600-3.00001-7

[24] Z. M. Almesri, H. A. Hussain, and R. M. Kamel, "Low voltage ridethrough capability of grid-connected PV systems: A comparative study of grid-following and grid-forming converters under current limits," IEEE Access, vol. 13, pp. 105591–105607, 2025, doi: 10.1109/ACCESS.2025.3580643

[25] O. Gil, "Application of fractional order PI/PID voltage controllers for three-phase voltage source inverter with dynamic load," American Journal of Computing and Engineering, vol. 7, no. 1, pp. 13–24, 2024, doi: 10.47672/ajce.1774

[26] M. A. Zamee, A. Hussain, and M. J. Hossain, "Fractional-order PI control with modified stability locus for grid-connected inverters: Stability enhancement and real-time validation," Electric Power Systems Research, vol. 248, p. 111854, 2025, doi: 10.1016/j.epsr.2025.111854

[27] I. Altawil, M. A. Momani, M. A. Al-Tahat, R. Al Athamneh, M. A. Al-Saadi, Z. Albataineh," Optimization of fractional order PI controller to regulate grid voltage connected photovoltaic system based on slap swarm algorithm", International Journal of Power Electronics and Drive Systems, Vol. 14, No. 2, pp. 1184-1200, June 2023, doi: 10.11591/ijpeds.v14.i2.pp1184-1200

[28] International Electrotechnical Commission, "IEC 61727:2004 Photovoltaic (PV) systems - Characteristics of the utility interface", Edition 2.0, Geneva, Switzerland, Dec. 2004.

International Journal of Smart Grid (IJSmartGrid)
https://www.ijsmartgrid.org/ Online ISSN: 2602-439X
I. Opirskyy et.al., Vol. 10, No. 3, September, 2026 Research Article

A Novel Hybrid Expert System for Encryption Detection in Smart Grid and Mobile Storage Devices

Ivan Opirskyy*, Oleh Harasymchuk*, Leila Rzayeva**‡, Bayandy Sarsembayev***, Murat Zhakenov****, Botakoz Kassimova*****, Korhan Kayisli******

*Department of Information Protection, Lviv Polytechnic National University, Lviv, 12 Bandery St., Main building, Ukraine

** Research and Innovation Center "CyberTech", Astana IT University, Astana, Mangilik El avenue, 55/11, Business center EXPO, block C1, Kazakhstan

*** School of Intelligent Systems, Astana IT University, Astana, Mangilik El avenue, 55/11, Business center EXPO, block C1, Kazakhstan

**** "Digital Heritage of Eurasia" LLP, Astana, Kazakhstan

***** L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

****** Gazi University, Eng. Fac. Electrical-Electronic Eng. Ankara, Turkiye

(ivan.r.opirskyi@lpnu.ua, oleh.i.harasymchuk@ lpnu.ua, l.rzayeva@astanait.edu.kz, Bayandy.sarsembayev@astanait.edu.kz, muratzhakenv@outlook.com, Kassimovabotagoz@gmail.com, korhankayisli@gmail.com)

‡Corresponding Author: Leila Rzayeva, Research and Innovation Center "CyberTech", Astana IT University, Astana, Mangilik El avenue, 55/11, Business center EXPO, block C1, Kazakhstan, Tel: +7 777 533 9169, l.rzayeva@astanait.edu.kz

Received: 11.02.2026 Accepted: 02.05.2026

Abstract- Ensuring reliable detection of encrypted data is increasingly important for protecting smart grid infrastructures and supporting mobile forensic investigations. This study investigates statistical and non-statistical approaches for encrypted data detection, including the Kolmogorov–Smirnov test, autocorrelation analysis, compression testing, file signature search, entropy estimation, and file system recognition. The main contribution of this work is a novel hybrid expert system that combines multiple detection methods with a knowledge-based inference engine to provide intelligent and explainable decision-making. Experimental results demonstrate that the proposed system achieves over 90% accuracy in identifying encrypted data under diverse conditions, including scenarios involving empty blocks that affect statistical distributions. Compared with conventional entropy-based and single-test approaches, the proposed framework provides improved accuracy, scalability, transparency, and adaptability.

Keywords: Data encryption detection, distributed control systems, intelligent decision support, Kolmogorov-Smirnov test, smart grid security

Cite this article as: I. Opirskyy, O. Harasymchuk, L. Rzayeva, B. Sarsembayev, M. Zhakenov, B. Kassimova, and K. Kayisli "A Novel Hybrid Expert System for Encryption Detection in Smart Grid and Mobile Storage Devices" *International Journal of Smart Grid (ijSmartGrid)*, Vol. 10, No. 3, pp. 173-190, September, 2026.

1. Introduction

Authors The digital transformation of modern infrastructures has made data security and reliability central concerns in both smart grids and mobile ecosystems. In smart grid environments, advanced metering infrastructures (AMI), distributed energy resources, and IoT-enabled devices continuously generate large volumes of critical operational data. Protecting this data from unauthorized access is essential, as encryption is increasingly employed to safeguard sensitive information. Similarly, mobile devices, which serve as personal and professional data hubs, widely adopt full-disk encryption (FDE) and file-based encryption (FBE) to strengthen privacy and prevent data breaches. While these mechanisms enhance confidentiality, they also introduce challenges for digital forensics and cybersecurity monitoring, where distinguishing encrypted data from compressed or unencrypted data is crucial for effective analysis.

Traditional single-method approaches – such as entropy estimation or compression testing – often prove unreliable in real-world scenarios. Compressed multimedia and log files can exhibit entropy levels similar to encrypted content, leading to false positives. Likewise, the presence of empty or sparse blocks in file systems can distort statistical distributions, reducing the accuracy of conventional detection methods. These limitations highlight the need for hybrid and intelligent systems that combine complementary techniques to improve robustness and explainability.

This study addresses this challenge by proposing a novel hybrid expert system for encryption detection that integrates both statistical and non-statistical methods, including Kolmogorov–Smirnov testing, autocorrelation analysis, file signature search, entropy estimation, compression analysis, and file system recognition. The system employs a knowledge-based inference engine capable of multi-criteria decision-making, enabling reliable identification of encrypted data even in the presence of distortions such as empty blocks.

The contribution of this study is twofold: (i) it enhances digital forensic investigations of mobile storage devices by providing automated, explainable, and accurate encryption detection; and (ii) it extends encryption detection capabilities to smart grid security, ensuring that critical energy data streams can be reliably analyzed for authenticity and integrity. By bridging cryptographic data analysis with intelligent decision support, the proposed system provides a scalable and transparent solution for both forensic and cybersecurity applications.

Smartphone manufacturers are implementing increasingly sophisticated security mechanisms – particularly encryption – to prevent unauthorized data access. As a result, the process of data extraction and analysis within forensic investigation frameworks has become progressively more complex for forensic specialists. Notably, the widespread adoption of full disk encryption (FDE) presents a major challenge in digital forensics, as it can render critical digital evidence inaccessible during investigations. Several studies focus on describing methods and processes for obtaining evidentiary data from mobile devices [1-5]. Reliable detection of encrypted file fragments is essential for various security activities: ransomware detection, digital forensics, and network traffic analysis. A common approach is to use high entropy as an indicator of data randomness [6]. However, many modern formats, such as office documents and multimedia files, are heavily compressed to optimize storage and transmission. Since compression algorithms also generate high-entropy data, distinguishing between encrypted and compressed files becomes difficult, reducing the effectiveness of entropy-based detectors.

The study [7] describes an encryption detection method based on a convolutional neural network, demonstrating high accuracy (over 99%) in recognizing local and network encrypted data, including various encryption algorithms, due to multi-scale feature extraction.

Expert systems have been successfully applied in digital forensics for various classification and decision support tasks, offering advantages in forensic contexts where decision traceability and explanation capabilities are crucial for legal proceedings.

Given the dynamic development of information technology and the constant improvement of encryption methods and algorithms [8-10], ways to improve them [11-13], use [14-16], the problem of file classification [17], detection of encrypted data is becoming increasingly relevant, demonstrating a trend of increasing its significance in the context of cybersecurity.

The usage of statistical methods for evaluation of encryption has been widely studied. For example, the study [18] uses the Pearson Chi-Square test, the Kolmogorov-Smirnov test, the autocorrelation test, and several other tests to evaluate the output of a cryptographic random number generator. In [19-20], the NIST statistical test package is used.

In recent years, ensuring the confidentiality and integrity of smart grid data has become increasingly important, particularly with the widespread deployment of Advanced Metering Infrastructure (AMI), IoT-based monitoring systems, and real-time SCADA platforms. These systems often transmit sensitive measurements and control signals, which are subject to encryption for security purposes. However, detecting whether incoming data streams are truly encrypted or merely compressed remains a technical challenge.

Traditional anomaly detection approaches fail to maintain accuracy under encrypted traffic conditions. For instance, an encryption-aware anomaly detection method proposed in [21] shows that statistical patterns in encrypted SCADA traffic significantly differ from unencrypted logs, and specialized techniques are necessary to maintain detection performance. Similarly, a comprehensive survey on smart grid security trends outlines the importance of encryption detection as a component of situational awareness against cyberattacks [22].

More advanced frameworks now combine machine learning and cryptographic analysis. A hybrid architecture proposed in [23] integrates AI-driven analysis with blockchain-secured smart grid data to detect and verify encrypted flows with high confidence. Another work [24] presents a hybrid intrusion detection approach for encrypted

and unencrypted data, employing swarm-based optimization and deep learning to improve classification accuracy in smart energy environments.

These recent studies justify the need for a hybrid, expert-driven encryption detection framework in smart grid infrastructures – particularly one capable of adapting to encrypted data streams and delivering explainable results, as proposed in this study.

This research addresses a critical gap in digital forensics by developing an intelligent expert system that integrates six complementary detection methods within a knowledge-based decision-making framework. The proposed system leverages forensic domain expertise through the implementation of optimized threshold values, adaptive preprocessing techniques, and multi-criteria decision rules. As a result, it achieves superior performance compared to standalone detection algorithms.

2. Data Encryption and Android OS

Devices running the Android operating system utilize the GPT (GUID Partition Table) scheme for disk partitioning [25]. GPT offers several advantages over traditional partitioning schemes such as BSD [26] or MBR (Master Boot Record) [27]. One of the primary benefits is the significantly larger number of available partitions – GPT supports up to 128 partitions by default, compared to 4 in MBR, 8 in BSD/Sun, and 16 in the SGI Disk Label format. While GPT theoretically allows for an unlimited number of partitions, the default limit of 128 is imposed to simplify file system operations.

The GPT partition table comprises three main components: the MBR protective record, the GPT header, and the partition entry array. The MBR protective record is in sector 0 of the disk (offset 0x0) and is designed to provide backward compatibility with legacy systems that do not support GPT. It also prevents such systems from misinterpreting a GPT-formatted disk as uninitialized. This protective record includes a single partition entry of type 0xEF, spanning the entire disk, while the boot code area remains unused.

The GPT header is located at offset 0x200 and contains metadata about the disk, including checksums for validation, the start location of usable disk space, and the disk's Globally Unique Identifier (GUID).

Following the header, partition entries begin at offset 0x400. Each entry is 128 bytes (0x80) in length and contains information such as the Partition Type GUID, the Partition GUID (a unique identifier for each partition), the starting and ending LBAs (logical block addresses), attribute flags, and a human-readable partition label.

The structure of the GPT table and the MBR protective record can be described using the following structure definition code in the ImHex hexadecimal editor is shown in Table 1.

Table 1. ImHex structure definition code for GPT table and protective MBR record

```
// Master Boot Record
enum MBRCopyProtection: u16 {
    No = 0x0000,
    Yes = 0x5A5A
};
enum MBRPartitionStatus: u8 {
    Active = 0x80,
    Inactive = 0x00,
    Invalid
};
enum MBRPartitionType u8 {
    Empty = 0x00,
    Fat12 = 0x01,
    Fat16Legacy = 0x04,
    Extended = 0x05,
    Fat16 = 0x06,
    Ntfs = 0x07,
    Fat32 = 0x0B,
    Fat32Lba = 0x0C,
    Fat16Lba = 0x0E,
    ExtendedLBA = 0x0F,
    LinuxSwap = 0x82,
    Linux = 0x83,
    LinuxLvm = 0x8E,
    GPTProtectiveMBR = 0xEE,
    EfiBoot = 0xEF,
    Other
};
enum MBREndSignature: u16 {
    Correct = 0xAA55,
    Incorrect
};
struct MBRDiskTimestamp {
    u16 magic;
    u8 original_drive;
    u8 timestamp_seconds;
    u8 timestamp_minutes;
    u8 timestamp_hours;
};
struct MBRDiskSig {
    u32 disk_signature;
    MBRCopyProtection copy_protection;
};
bitfield MBRCylHiSect {
    cyl_hi : 2;
    sector : 6;
};
struct MBRCHS {
    u8 head;
    u8 cyl_hi_sector;
    u8 cyl_lo;
};
struct MBRPartition {
    MBRPartitionStatus status;
    MBRCHS first;
    MBRPartitionType part_type;
    MBRCHS last;
    u32 lba_first;
    u32 size;
};
struct MBRModern {
    char bootstrap_part1[218];
    MBRDiskTimestamp stmp;
    struct GUIDPartitionEntry {
        u8 type_GUID[16];
        u8 unique_GUID[16];
        u64 start_LBA;
        u64 end_LBA;
        GUIDPartitionAttr attributes;
        char16 name[36];
    };
    struct GUIDPartitionTable {
        char signature [8];
        u32 revision;
        u32 header_size;
        u32 header_CRC;
        u32 reserved_area;
        u64 current_LBA;
        u64 backup_LBA;
        u64 first_usable_LBA;
        u64 last_usable_LBA;
        u128 disk_GUID;
        u64 partition_entries_start_LBA;
        u32 partition_count;
        u32 entry_size;
        u32 partition_entries_CRC;
        char pad[420];
        GUIDPartitionEntry partitions [128];
    };
    MBRModern mbr @ 0x00;
    GUIDPartitionTable gpt @ 0x200;
    char bootstrap_part2[216];
    MBRDiskSig sig;
    MBRPartition partitions [4];
    u16 signature;
};
struct MBRClassic {
    char bootstrap [446];
    MBRPartition partitions [4];
    MBREndSignature signature;
};
// GUID Partition Table
bitfield GUIDPartitionAttr {
    required: 1;
    no_block_io : 1;
    legacy_bootable: 1;
    reserved: 45;
    GUID_specific: 16;
};
```

In addition to mobile storage datasets, this study incorporates smart grid data streams-specifically smart meter readings, SCADA logs, and IoT sensor telemetry – as representative real-world encrypted, compressed, and unencrypted data sources. Each detection method (entropy, compression, autocorrelation, K–S, signature search, file system recognition) was adapted to identify encryption in these streams. Entropy and autocorrelation distinguish routine load profiles from encrypted telemetry, signature search targets standard AMI data formats, and compression tests discriminate between compressed log files and encrypted data. The hybrid expert system was embedded into a smart grid monitoring architecture, enabling real-time encryption detection with evaluation metrics including detection latency, scalability under high-volume meter data, and false positive/negative rates in compressed vs encrypted streams

a. Ciphertext indistinguishability

Ciphertext indistinguishability [28,29] is a fundamental property of cryptographic systems that ensures a high degree of encryption security. It refers to the inability of an adversary to distinguish between two ciphertexts, or between a ciphertext and a truly random sequence [30]. Closely related to this concept is semantic security, which is a core guarantee in modern cryptography. Semantic security ensures that an adversary cannot obtain any meaningful information about the plaintext from the ciphertext, even when the adversary has some prior knowledge of the plaintext [31]. Formally, a cryptosystem is considered semantically secure if, for any two messages of equal length, the probability that an attacker can distinguish between their ciphertexts is negligibly small. In other words, the output of a semantically secure encryption algorithm must be computationally indistinguishable from a random sequence, provided that the cryptosystem is configured with appropriate parameters and operated in a secure mode.

The practical implications of semantic security can be summarized as follows:

- An adversary cannot determine whether a particular message has been encrypted more than once;
- It is computationally infeasible to identify whether two ciphertexts correspond to the same plaintext;
- Ciphertexts do not reveal any information about the underlying plaintext;
- Partial knowledge of the plaintext does not aid in recovering the remaining content;
- Ciphertexts are indistinguishable from random or pseudo-random sequences generated by a high-entropy source.

b. Classification of data encryption detection methods

Data encryption detection methods can generally be classified into two main categories: statistical and non-statistical approaches. Statistical methods evaluate encryption by analyzing byte distribution patterns or applying statistical tests to identify deviations from randomness. In contrast, non-statistical methods directly examine the structural or functional properties of the data.

In this study, six representative methods are considered: file system detection, Shannon entropy evaluation, autocorrelation analysis, Kolmogorov–Smirnov (K–S) test, compression test, and signature search test. Among these, autocorrelation analysis and the Kolmogorov–Smirnov test fall into the statistical category, whereas file system detection, Shannon entropy evaluation, compression testing, and signature search are categorized as non-statistical methods.

c. File system detection using standard GNU/Linux tools

Most GNU/Linux distributions include the parted utility [32] as part of their standard toolkit. This utility is primarily used to read and modify physical disk partitions and provides functionality for analyzing file systems within both complete disk images and individual partitions. Owing to its mature and widely tested codebase, parted is considered a reliable tool for this purpose. Once a file system is successfully detected, subsequent analysis can be carried out at the level of individual files within the detected structure. However, a key limitation of this method is that it cannot differentiate between files that are encrypted and those that are not, as it only identifies the presence of a file system rather than the encryption status of its contents.

d. Evaluation of the Shannon entropy of the ciphertext

Shannon entropy measures uncertainty or randomness in a dataset [32]. In encryption analysis contexts, it can determine the presence of encrypted data, though it has significant limitations.

The value of Shannon entropy for a data set is calculated as in Eq. (1):

$$H = -\sum p_i \cdot \log_2(p_i) \quad (1)$$

Where pi is the probability of the i-th character appearing in the dataset, the maximum entropy value for byte sequences is 8 bits per byte, corresponding to a uniform distribution.

Encrypted data typically exhibits high entropy values close to the theoretical maximum, as high-quality encryption tends to distribute bytes uniformly. Conversely, unencrypted data often exhibits lower entropy due to natural patterns, structure, and repetitions.

However, there are several significant limitations to using entropy to identify encrypted data:

- Compressed files also have high entropy levels, which makes them indistinguishable from encrypted data when using this method alone.
- Some data types (e.g., encoded media files) may naturally have high entropy without encryption.

Shannon entropy estimation as the only method to identify encrypted file systems in a blind search should not be considered reliable. This approach may yield both false positives and false negatives. To improve identification accuracy, it is recommended that entropy analysis be combined with other methods.

Thus, while Shannon entropy can be helpful in the data analysis toolbox, it should be used as part of a comprehensive

approach to identifying encrypted data rather than as a stand-alone method.

e. File signature search

A file signature is a specific byte sequence at the beginning of a file – commonly referred to as magic numbers – that identifies its type and format [33]. In the context of cryptography and encrypted partition analysis, the absence of expected file system signatures at the start of a partition often indicates encryption. A more advanced application of signature-based analysis involves searching for the signatures of common file formats that may reside within the investigated partition.

Under normal conditions, a partition typically contains multiple files with recognizable signatures. However, when a partition is fully encrypted, these signatures disappear along with the file contents, as the data is transformed into a sequence of seemingly random bytes. Despite this, there remains a non-negligible probability that a random sequence may coincidentally resemble a valid signature. The probability of such false detection decreases as the length of the signature increases. To mitigate this issue, it has been proposed that files be distinguished based on the number of valid signatures detected per unit of storage volume. A high value of signatures per unit indicates that the file system may contain a large number of files, such as multimedia files, documents, et cetera. This test is aimed at solving the problem of false positives due to the high content of high-entropy compressed formats, such as AVI, MP4, JPEG, MP3, and other formats, typical for modern multimedia-enabled smartphones. A high value of the signature detection test and high values of other tests imply the presence of multimedia files.

Nevertheless, this method presents several limitations:

- False positives may occur if the partition primarily stores files lacking distinctive signatures, although this is uncommon in typical smartphone storage (which usually contains media files, structured documents, and archives);
- If encryption is applied at the file level rather than across the entire partition, partition signatures may still be detectable, while file signatures remain hidden;
- Short signatures (4–6 bytes) are more susceptible to accidental matches within random data, increasing the likelihood of erroneous detection.

f. Compression test

It is well established that random, near-random, and other high-entropy data exhibit very low compression rates when processed with lossless algorithms. The principle of this test is based on the fact that compression algorithms significantly reduce the size of structured data by detecting and substituting recurring patterns. In contrast, encrypted data inherently resembles a random bit sequence with no detectable structure, making it nearly incompressible.

The test procedure consists of three main steps:

- Selecting the data block to be analyzed from a disk image or transferring the entire file;
- Compressing the selected block using several standard compression algorithms;
- Comparing the average size of the compressed block with the size of the original data.

If the compression ratio (defined as the ratio of the original size to the compressed size) approaches 1.0, this may indicate the presence of encrypted data. However, due to redundancy introduced by certain compression algorithms, ratios slightly below 1.0 may also be observed for encrypted data.

It is important to emphasize that this method provides only a probabilistic estimate. A low compression rate is not unique to encryption and may also be observed in pre-compressed files, such as archives or multimedia data.

g. Autocorrelation test

The essence of the autocorrelation test is to examine a dataset's correlation relative to its shifted copy (shifts of 1 to N bytes) [34]. Conducting autocorrelation data analysis enables identification of complex interdependencies within datasets, as periodic original data will yield periodic test results. In encryption analysis, an autocorrelation test is used to locate interdependencies or repetitions of parts of the data. To summarize the autocorrelation data for different shifts of parts relative to each other, the concept of average autocorrelation by block is introduced. The average autocorrelation values are analyzed by calculating the standard deviation and comparing the obtained value with a specified threshold value. High levels of autocorrelation indicate the presence of sections with periodic patterns in the file. The probability of block periodicity is extremely low for random and near-random data. It is worth noting that this method is also sensitive to data with high Shannon entropy, including compressed data, and therefore can only be used in conjunction with other methods.

h. Pearson (χ^2) test

Pearson test, also known as χ^2 (chi-square) test, is a fundamental statistical method for testing hypotheses about the conformity of observed data with a theoretical distribution [35].

$$\chi^2 = \sum_{i=0}^{k} \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

where O_i represents the observed frequency in category i and Ei is the expected frequency. The sum is taken over all categories of data.

The application of the test begins with the formulation of the null hypothesis H0, which states that the observed frequency distribution has no difference compared to the expected distribution, and the alternative hypothesis H1, which states the opposite. The calculated χ^2 value is compared with the critical value, which is determined from a table of critical values for the chosen significance level (usually 0.05) and the corresponding degrees of freedom (255 degrees in the case of byte processing). If the calculated value exceeds the critical value, the null hypothesis is rejected in favor of the alternative.

The χ^2 distribution has several important characteristics. It is always non-negative due to the squaring of differences. The shape of the distribution is asymmetric with an elongated right tail, although as the degrees of freedom increase, the distribution approaches normality. When applying the test, it is necessary to consider the sample size and the number of categories. The expected frequencies in each category must be large enough to obtain reliable results. Its sensitivity to sample size should be taken into account - in large samples (500 units and more), even minor deviations may be statistically significant. In this regard, using the Pearson test to process the entire image is problematic.

In the context of byte distribution analysis, the Pearson test allows one to estimate the deviation from the uniform distribution typical for ciphertexts of modern encryption algorithms. It is necessary to consider the data features: all 256 possible byte values are considered separate categories, and the expected frequencies for a uniform distribution are the same for all categories.

i. Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test is a statistical method for examining conformity between an empirical distribution and a theoretical model (goodness-of-fit test) or between two empirical distributions and a standard theoretical model (homogeneity test) [36]. The mathematical basis of the test is to calculate the maximum absolute deviation between the empirical and theoretical distribution functions. The test statistics are defined by formula (3):

$$D = \sup|F(x) - S(x)| \quad (3)$$

Where F(x) is the theoretical distribution function (for a uniform byte distribution, this is a linear function from 0 to 1), and S(x) is the empirical distribution function based on the observed byte frequencies in the ciphertext.

A significant advantage of the Kolmogorov-Smirnov test over Pearson's criterion is its reduced sensitivity to sample size. This is especially important when analyzing large volumes of ciphertext, where the Pearson criterion can show statistically significant deviations even with minor differences in the distribution.

The critical values for the Kolmogorov-Smirnov test scale with sample size n as $1/\sqrt{(n)}$, while the Pearson test statistic grows linearly with sample size. This makes the Kolmogorov-Smirnov test a more practical tool for analyzing cryptographic systems.

There are several important aspects to consider when applying the test to ciphertext analysis. Firstly, the test evaluates deviations from the theoretical distribution over the entire range of values, which allows systematic deviations in the byte distribution to be identified. Secondly, the test is more sensitive to deviations in the central part of the distribution than at its edges.

When interpreting test results, it is essential to consider the statistical significance of deviations and their practical significance for the security of the cryptographic system. Slight deviations from a uniform distribution do not necessarily indicate a lack of data encryption, but they have to be investigated more extensively.

j. NIST SP 800-22 tests

NIST Special Publication 800-22 Revision 1a [19-20, 37] is a document that describes a set of statistical tests (NIST Statistical Test Suite or STS) for determining the randomness of binary sequences of random number generators. Using the NIST STS test suite to evaluate the data randomness is unorthodox, since the STS tests were designed to work with small amounts of data, and the test library itself is outdated.

k. Vulnerability of statistical encryption detection methods to sparse areas

File systems use a block structure – data is stored in blocks of a given length, and the blocks may not be strictly in order (fragmentation), and may also be empty (the so-called holes). Using fragmentation and holes allows the file system to write data more compactly and efficiently: when appending to the file, new information is written to an unoccupied block without the need to start a complex process of moving already written information, and a pointer to the location of the new data block is written to the metadata of the FS itself (inode for ext2/3/4 class systems, b-tree for the Btrfs system, MFT for the NTFS system, etc.).

The presence of data fragmentation does not have a significant impact on statistical methods, but large empty areas will have a considerable impact, since zero bytes are also considered when constructing empirical distributions. In this regard, there is a need to identify and remove such areas.

There are several methods to solve this problem:

- Removing empty FS blocks – since the block size is known, the image under evaluation can be scanned for empty blocks at offsets that are multiples of the block size, and found blocks shall be removed; this method has a minimal impact on the data and does not distort blocks that were not filled;
- Replacing the value of the number of zero bytes with an average value – the implementation of this method can lead to a distortion of statistical data (for example, if one or more files stored in the FS image are not sparse, that is, have empty areas in their structure);
- Discarding the category of zero bytes, while this method is the simplest one, its usage critically affects the plausibility of the empirical distribution, since data itself has null bytes.

A variant of the first method, sparsifying, is often used to reduce the space occupied by files and is supported by most modern file systems. The sparseness method differs from simple offset search in that empty blocks are removed and information about the removed blocks is recorded in metadata, which includes the size and offset of each empty block. Since the inclusion of structured metadata can distort the empirical distribution, its generation will not be included in the final hole elimination method as unnecessary.

3. Evaluation Methodology

Each presented method should be tested on internal memory section images from sample devices with different data encryption types. The success criterion for each method is clear distinction between final values for unencrypted and encrypted files and demonstration of at least 50% encryption detection accuracy. The requirement for clear value boundaries stems from the need to derive Boolean results based on numerical values. The minimum accuracy requirement stems from the assertion that all methods, which will be used as a part of the final method, should be able to distinguish encrypted and unencrypted data with accuracy better than the results of the random choice of two categories: "encrypted" and "not encrypted", as the multi-stage nature of the method implies, that one or more false results may prematurely end the test execution. Methods that are unable to meet this criterion shall not be included in the final method.

A set of test files with a total volume of 545 gigabytes was prepared for testing:

1) Images of internal memory partitions of mobile devices:
   a) Sony Xperia XZ F8331 (Xperia FW 41.3.A.2.192, Android 9.1.0, full-disk encryption, ext4 file system, block devices accessed using Magisk):
      i) Contents of the device /dev/block/dm-0 (partition /data after decryption), file name kagura_data_dec.img;
      ii) Contents of the device /dev/block/platform/soc/7464900.sdhci/by-name/user data (partition /data before decryption, AES-256, CBC or XTS modes), file name kagura_data_enc.img;
   b) Xiaomi Redmi Note 9 Pro (MIUI 12.0.4.0 Global, Android 11.0, file-based encryption version 1, F2FS file system, block devices accessed using Magisk):
      i) Contents of the /dev/block/sda18 partition (the /data partition before decryption), file name miatoll_data_fbe.img;
      ii) Artificially recreated file system of the /data partition (due to the impossibility of taking a true image of the partition without encryption, AES-256-XTS encryption), file name miatoll_data_nonfbe.img
   c) Xiaomi Redmi 5 Plus (LineageOS 19-20220619-UNOFFICIAL-vince, Android 12, no encryption, ext4 file system, block devices accessed using Magisk):
      i) Contents of the /dev/block/mmcblk0p50 block device (the /data partition), file name vince_data.img.
      ii) An image of a 32 GB memory card formatted as an extension of internal storage and encrypted with the dm-crypt utility in plain mode with the AES-256-CBC-ESSIV:SHA256 algorithm chain, file name adoptable.img.
2) Images of encrypted storage devices from non-mobile devices:
   a) An image of a 40 GB WD400BB hard drive encrypted with the commercial PGP Desktop 9.0 utility (encryption algorithm - AES-256, mode unknown, presumably CBC), file name wd400.img;
3) A set of 1 GB pseudo-random data generated using the built-in cryptographic pseudo-random number generator /dev/random, present in the GNU/Linux family of operating systems, file name random.img.

The unencrypted image files were supplemented with encrypted files (1% of the total files in the image). The set of encrypted files was prepared using the OpenSSL utility and AES-256 algorithm in CBC mode. To test the hypothesis about the effect of empty areas on the accuracy of the test results, the corresponding modifications were made to the files that do not contain the above-mentioned empty areas; the names of such files have "_opt" suffixes. Optimization of the file with pseudo-random data was not carried out as unnecessary. Positive test results for this file will be accepted as the presence of encryption, since pseudo-random data is extremely similar to the encrypted data (see Section 2).

a. Autocorrelation test evaluation

The implemented method for determining autocorrelation uses the linear correlation coefficient, also called the Pearson correlation coefficient. The max lag value is set at 50 bytes. The test was conducted using the methodology described above. Table 2 presents the test results for the list of disk image files proposed in section 4.1.

The effect of empty blocks on the results is visible – three false negative results. The test results for the "kagura_data_enc.img", "miatoll_data_fbe.img", and "adoptable.img" files are anomalous, since the standard deviation of the set of average autocorrelation values should be significantly smaller for encrypted files.

At the same time, for a similar file with empty blocks removed, the test result is within the expected values. Let's analyze plots of the average autocorrelation values by blocks using three files as an example - pseudo-random data (Figure 1), encrypted data before (Figure 2) and after removing empty blocks (Figure 3), and unencrypted data before (Figure 4) and after removing empty blocks (Figure 5):

Table 2. Autocorrelation test evaluation results

File Name	Std Dev of Mean Autocorrelation Values	Data Block Size, MB	Data Category	Test Result (Threshold: 0.125)	Result Validity
random.img	$8{,}38032\times10^{-5}$	1	Random	Yes	Yes
kagura_data_enc.img	0,256594	1	FDE, without removal	No	No – empty blocks
kagura_data_enc_opt.img	$5{,}23742\times10^{-3}$	1	FDE, with removal	Yes	Yes
kagura_data_dec.img	0,150195	1	No encryption, without removal	No	Yes
kagura_data_dec_opt.img	0,131928	1	No encryption, with removal	No	Yes
vince_data.img	0,264712	1	No encryption, without removal	No	Yes
vince_data_opt.img	0,256033	1	No encryption, with removal	No	Yes
wd400.img	$5{,}23417\times10^{-3}$	1	FDE, without removal	Yes	Yes
wd400_opt.img	$1{,}63476\times10^{-3}$	1	FDE, with removal	Yes	Yes

miatoll_data_fbe.img	0,321011	1	FBE, without removal	No	No – empty blocks
miatoll_data_fbe_opt.img	0,117005	1	FBE, with removal	Yes	Yes
miatoll_data_nonfbe.img	0,18971	1	No encryption, without removal	No	Yes
miatoll_data_nonfbe_opt.img	0,183086	1	No encryption, with removal	No	Yes
adoptable.img	0,158156	1	FDE, without removal	No	No – empty blocks
adoptable_opt.img	0,069011	1	FDE, with removal	Yes	Yes

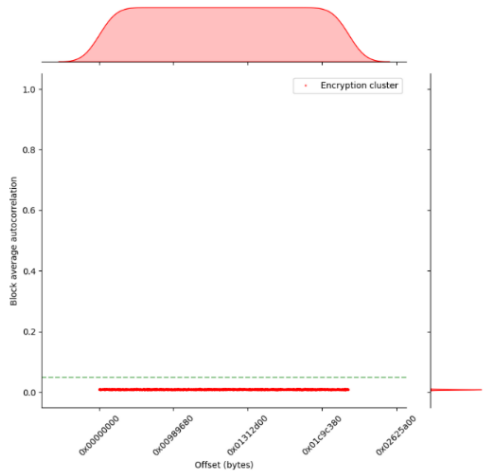


Fig. 1. Distribution of average values of the autocorrelation coefficient for a block size of 1 kibibyte in the random.img file.

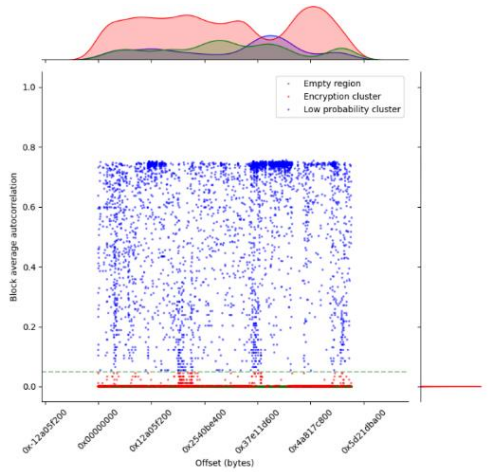


Fig. 2. Distribution of average values of the autocorrelation coefficient for a block size of 1 kibibyte in the kagura_data_enc.img file.

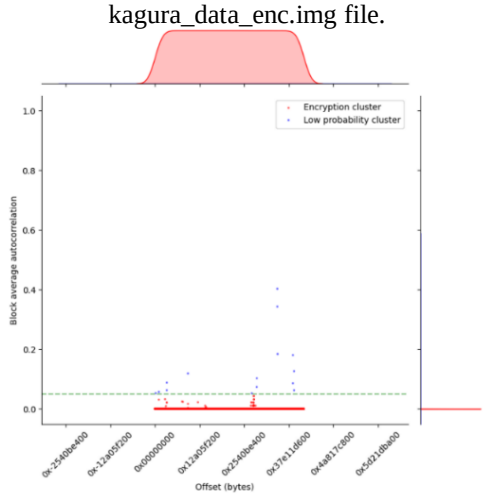


Fig. 3. Distribution of average values of the autocorrelation coefficient for a block size of 1 kibibyte in the kagura_data_enc_opt.img file.

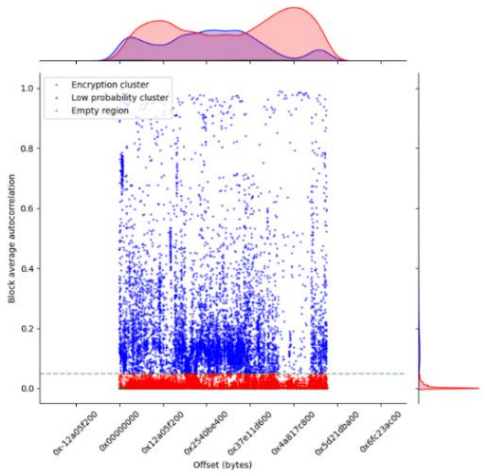


Fig. 4. Distribution of average values of the autocorrelation coefficient for a block size of 1 kibibyte in the kagura_data_dec.img file.

A significant concentration of points in the area close to zero with no outliers can be observed, which is reflected in the standard deviation value. Such low average autocorrelation values indicate the absence of areas with obvious dependencies between bytes in the file (repeated patterns, words, empty regions, or minor regions more than the size of one FS block).

A large number of high-value points can be observed. When analyzing such a file without additional steps to identify areas that can affect the result, one could conclude that the result is false, since the test shows a picture typical of unencrypted data.

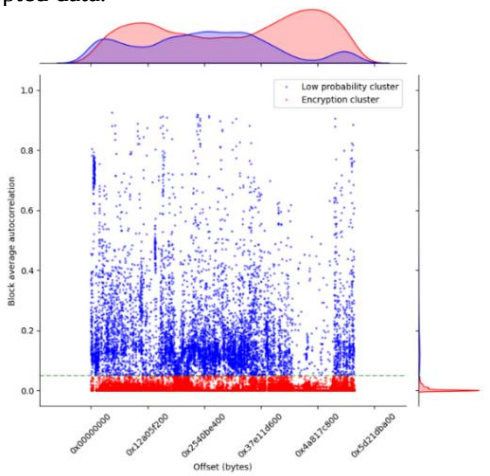


Fig. 5. Distribution of average values of the autocorrelation coefficient for a block size of 1 kibibyte in the kagura_data_dec_opt.img file.

After removing empty blocks of 4,096 bytes from the image, a significant improvement in the test result is visible -

the number of points in the area of low encryption probability decreased to approximately 15.

In Figures 4-5, the distribution of the average values of the autocorrelation coefficient in the unencrypted data sample is shifted to the 0.1 mark, which is confirmed by the corresponding standard deviation values (see Table 2). The effect of empty blocks on the test results of this image is insignificant. From this, a conclusion can be drawn that the presence of empty areas on the final test result is significant and affects the accuracy of its operation, especially if the image of the file system being studied is minimally filled with valuable data.

When setting the boundary value of the test result for determining encryption to 0.1, three errors of the second kind are detected, which are caused by the presence of empty areas in the file; when analyzing modifications of these files without empty areas, the test result corresponds to the presence of encryption. Overall, the autocorrelation test is deemed suitable for determining partition encryption.

b. Pearson (χ^2) test evaluation

The results of the Pearson test for all variants of all images are presented in Table 3. Relatively large values (from 1010 to 1013) of the test results for unencrypted or a mixture of encrypted and unencrypted data are observed, and relatively small values (from 102 to 108) for encrypted data. A significant decrease in the test results by several orders of magnitude after the procedure for removing empty areas can also be observed. None of the data samples meet the critical values for the Pearson test at 255 degrees of freedom (Table 4), which means that the Pearson test is unsuitable for encryption detection.

Table 3. Pearson test evaluation results

File Name	χ^2 Value	Data Category
random.img	265.1497	Random
kagura_data_enc.img	5.5038×10^{11}	FDE, without removal
kagura_data_enc_opt.img	2.1858×10^{4}	FDE, with removal
kagura_data_dec.img	5.2943×10^{10}	No encryption, without removal
kagura_data_dec_opt.img	3.0938×10^{10}	No encryption, with removal
vince_data.img	5.5788×10^{12}	No encryption, without removal
vince_data_opt.img	3.9996×10^{10}	No encryption, with removal
wd400.img	1.7262×10^{5}	FDE, without removal
wd400_opt.img	604.09993	FDE, with removal
miatoll_data_fbe.img	5.9196×10^{12}	FBE, without removal
miatoll_data_fbe_opt.img	1.2395×10^{10}	FBE, with removal
miatoll_data_nonfbe.img	1.1152×10^{13}	No encryption, without removal
miatoll_data_nonfbe_opt.img	2.1231×10^{11}	No encryption, with removal
adoptable.img	1.4362×10^{12}	FDE, without removal
adoptable_opt.img	9.9141×10^{8}	FDE, with removal

Table 4. Critical values of Pearson's test for 255 degrees of freedom

Probability p							
0.005	**0.01**	**0.025**	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**
316,9194	310,4574	301,1250	293,2478	284,3359	278,4253	273,7862	269,8473

c. Kolmogorov-Smirnov test evaluation

The results of the Kolmogorov-Smirnov test for all variants of all images are presented in Table 5.

Cumulative distribution plots for "unoptimized-optimized-theoretical" triplets are presented at Figures 6-8.

A complete correspondence of the distributions is present, which indicates a high degree of similarity between the encrypted data created by the PGP Desktop program and a linear space theoretical distribution.

This plot shows the opposite picture - none of the distributions correspond to the theoretical uniform distribution, which indicates the absence of a uniform distribution of byte frequencies in the studied image. The difference between the unoptimized (yellow line) and optimized (green line) images is also clearly visible. The shift of the initial point of the distribution indicates a significant predominance of zero bytes since the cumulative sum of the frequencies of occurrence of each byte is used to construct the plot of the empirical distribution.

An insignificant influence of the image optimization process can be observed, since the original image had a low number of zero blocks, typical for full-disk encryption, which masks all disk areas.

To evaluate the test results, either the maximum distance or the p-value categorization can be used (Figure 9). Using the p-value is a more correct method for evaluating the Kolmogorov-Smirnov test result, so it will be used. To reduce the number of false positive and false negative errors, a threshold value for the p-value of 0.1 is being used, which allows for a weak dependence, instead of a stricter value of 0.05.

Four false positives were found, three were caused by statistical distortion, caused by zero blocks. The false positive error during the kagura_data_dec_opt.img file analysis can be caused by the presence of a certain number of non-system encrypted files in the file system or other factors.

Table 5. Kolmogorov-Smirnov test evaluation results

File Name	File Size (MB)	D-statistic	p-value	Data Nature	Test Result at $p < 0.1$	Result Validity
random.img	32768	0.003906	1.00	Random	Yes	Yes
kagura_data_enc.img	22752	0.3008	1.2759×10^{-10}	FDE, without removal	No	No – empty blocks
kagura_data_enc_opt.img	15909.75	0.003906	1.00	FDE, with removal	Yes	Yes
kagura_data_dec.img	22752	0.1172	0.05939	No encryption, without removal	No	Yes
kagura_data_dec_opt.img	22224	0.09766	0.1741	No encryption, with removal	Yes	No – false positive
vince_data.img	24507	0.9219	1.831×10^{-117}	No encryption, without removal	No	Yes
vince_data_opt.img	2497	0.2891	7.8745×10^{-10}	No encryption, with removal	No	Yes
wd400.img	38167	0.003906	1.00	FDE, without removal	Yes	Yes
wd400_opt.img	38162	0.003906	1.00	FDE, with removal	Yes	Yes
miatoll_data_fbe.img	89130	0.5	1.8889×10^{-29}	FBE, without removal	No	No – empty blocks
miatoll_data_fbe_opt.img	46171	0.03516	0.9975	FBE, with removal	Yes	Yes
miatoll_data_nonfbe.img	107520	0.625	3.8584×10^{-47}	No encryption, without removal	No	Yes
miatoll_data_nonfbe_opt.img	46531	0.168	1.4307×10^{-3}	No encryption, with removal	No	Yes
adoptable.img	29818	0.4258	3.3941×10^{-21}	FDE, without removal	No	No – empty blocks
adoptable_opt.img	17387	0.02344	0.999	FDE, with removal	Yes	Yes

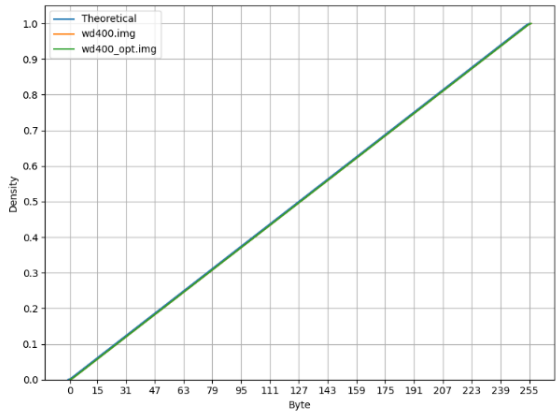


Fig. 6. Comparison of the theoretical uniform and empirical distributions for the pair of files wd400.img and wd400_opt.img

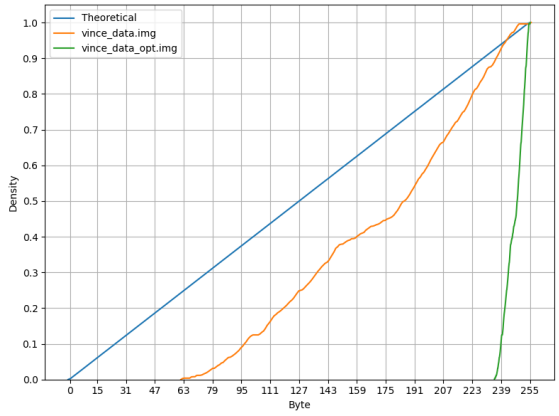


Fig. 7. Comparison of the theoretical uniform and empirical distributions for the pair of files vince_data.img and vince_data_opt.img

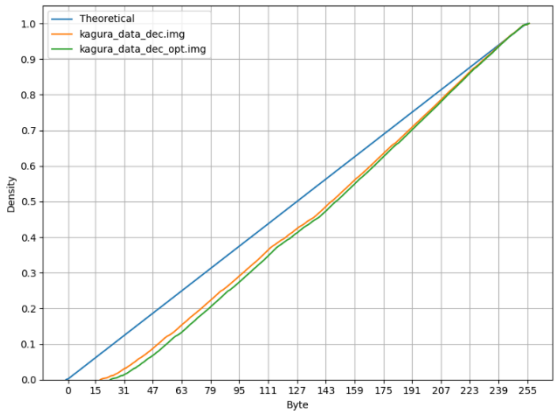


Fig. 8. Comparison of the theoretical uniform and empirical distributions for the pair of files kagura_data_dec.img and kagura_data_dec_opt.img.

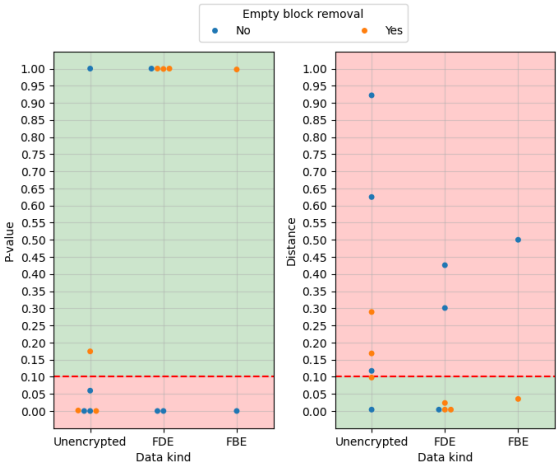


Fig. 9. Swarm plot of the algorithm's results delimitation by the p-value (left) and maximum distance (right).

Based on the analysis of the results, a conclusion can be made that the Kolmogorov-Smirnov test with categorization by P-value is suitable for determining partition encryption.

d. Compression test evaluation

The compression test uses the GZip, LZ4, BZip2, Zstd, and XZ (LZMA2) algorithms and their corresponding software implementations pigz, lz4, lbzip2, zstd, and pixz. The use of alternative programs for the GZip, BZip2, and XZ algorithms is due to the multithreading support, which significantly improves data processing speeds. The compression test results are presented in Table 5.

The encryption detection threshold for this test is set at 10% - data is considered unencrypted if the average compression ratio is more than 1,1. This value allows for a marginal compression increase on null-bytes, which may be present in block-level FDE systems and other service data sections. A compression ratio of less than 1,0 indicates an increase in the size of the compressed data relative to the input data due to the addition of metadata to each block of compressed data (block header, block compression parameters, checksum, etc., depending on the algorithm).

Three false negative errors were obtained, which are associated with empty blocks in files, significantly increasing the compression ratio.

The obtained test results confirm the viability of using compression tests for detecting encrypted data when combined with other methods.

e. File signature search evaluation

A set of 549 signatures in regular expression format was prepared to search for file signatures. The signature test results are presented in Table 6:

The analysis of the obtained data showed that for encrypted data with the removal of empty blocks, the specific number of signatures is within 68.3-83.18 signatures/MB (the central cluster of values, the anomalous value is 138.11 signatures/MB, within the Grubbs criterion $\alpha = 0.01$), for encrypted data without removal - within 43.09-80.53 signatures/MB.

For unencrypted data, the ranges are 276.55-356.03 signatures/MB (the central cluster of values, the anomalous value is 76.7 signatures/MB, within the Grubbs criterion $\alpha = 0.01$) without removal and 364.48-752.86 signatures/MB with removal. The obtained test results confirm the positive impact of the block removal procedure. The optimal boundaries for setting the threshold value are 100-200 signatures/MB. For the test, a value of 150 signatures/MB is used (Figure 10):

Table 5. Compression test evaluation results

File Name	Compression Ratio (Original Size / Compressed Size)						Data category	Test result	Result Validity
	GZip	LZ4	Bzip2	Zstd	XZ	Average			
random.img	0,9997	1,0000	0,9954	1,0000	1,0000	0,9990	Random	Yes	Yes
kagura_data_enc.img	1,4289	1,4262	1,4235	1,4301	1,4297	1,4277	No encryption, no deletion	No	No - empty blocks
kagura_data_enc_opt.img	0,9998	1,0001	0,9955	1,0001	1,0000	0,9991	No encryption, with deletion	Yes	Yes
kagura_data_dec.img	1,2449	1,1742	1,2435	1,2460	1,2999	1,2417	FDE, no deletion	No	Yes
kagura_data_dec_opt.img	1,2161	1,1470	1,2146	1,2169	1,2697	1,2128	FDE, with deletion	No	Yes
vince_data.img	20,9792	16,4302	22,2159	22,0680	26,4762	21,6339	No encryption, no deletion	No	Yes
vince_data_opt.img	2,1823	1,7767	2,2635	2,2523	2,7080	2,2365	No encryption, with deletion	No	Yes
wd400.img	0,9998	1,0001	0,9956	1,0001	1,0001	0,9991	FDE, no deletion	Yes	Yes
wd400_opt.img	0,9997	1,0000	0,9954	1,0000	1,0000	0,9990	FDE, with deletion	Yes	Yes
miatoll_data_fbe.img	2,0101	1,9920	2,0050	2,0161	2,0212	2,0089	FBE, no deletion	No	No - empty blocks
miatoll_data_fbe_opt.img	1,0430	1,0375	1,0386	1,0443	1,0484	1,0424	FBE, with deletion	Yes	Yes
miatoll_data_nonfbe.img	3,4260	3,0464	3,4903	3,4655	3,7876	3,4431	No encryption, no deletion	No	Yes
miatoll_data_nonfbe_opt.img	1,4858	1,3274	1,5105	1,4999	1,6395	1,4926	No encryption, with deletion	No	Yes
adoptable.img	1,8048	1,7870	1,8064	1,8200	1,8330	1,8102	FDE, no deletion	No	No – empty blocks
adoptable_opt.img	1,0533	1,0451	1,0534	1,0613	1,0689	1,0564	FDE, with deletion	Yes	Yes

Table 6. Signature test evaluation results

File Name	Number of Found Signatures	File Size (MB)	Signatures per MB	Empty Blocks Removal	Encryption Presence	Test Result	Result Validity
random.img	2234880	32768	68.203	No	Yes	Yes	Yes
kagura_data_enc.img	1086759	22752	47.765	No	Yes	Yes	Yes
kagura_data_enc_opt.img	1086759	15909.75	68.308	Yes	Yes	Yes	Yes
kagura_data_dec.img	8100462	22752	356.033	No	No	No	Yes
kagura_data_dec_opt.img	8100467	22224	364.488	Yes	No	No	Yes
vince_data.img	1879883	24507	76.706	No	No	Yes	No – false negative
vince_data_opt.img	1879885	2497	752.866	Yes	No	No	Yes
wd400.img	2603410	38167	68.212	No	Yes	Yes	Yes
wd400_opt.img	2603411	38162	68.220	Yes	Yes	Yes	Yes
miatoll_data_fbe.img	3840933	89130	43.094	No	Yes	Yes	Yes
miatoll_data_fbe_opt.img	3840869	46171	83.188	Yes	Yes	Yes	Yes
miatoll_data_nonfbe.img	29734651	107520	276.550	No	No	No	Yes
miatoll_data_nonfbe_opt.img	29734656	46531	639.032	Yes	No	No	Yes
adoptable.img	2401422	29818	80.536	No	Yes	Yes	Yes
adoptable_opt.img	2401422	17387	138.115	Yes	Yes	Yes	Yes

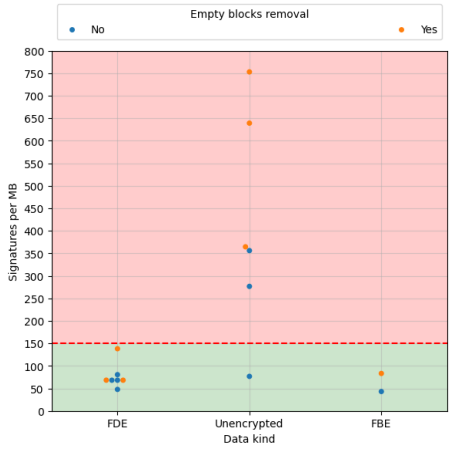


Fig. 10. Swarm plot of the specific number of signatures for the tested files and their distribution by test result.

One false positive value in the middle column of the plot (file vince_data.img) is observed. This error can be explained by the fact that this file contains a relatively small amount of data relative to its full size, which is confirmed by the high average compression ratio (see Table 6), i.e., the file mainly consists of empty areas. The obtained test results confirm the viability of using signature search tests for detecting encrypted data when combined with other methods.

f. Entropy estimation test evaluation.

The results of the entropy estimation test evaluation are presented in Table 7.

The test threshold value of 7.95 was chosen as the maximum permissible deviation, since high-entropy data types have a Shannon entropy level within the range of 7.9-7.98. Such values were derived based on the empirical analysis of reference encrypted data. When using this value, 4 false negative errors were detected, 3 of which were caused by a decrease in the entropy level by empty blocks in the file. The obtained test results confirm the viability of using entropy estimation tests for detecting encrypted data when combined with other methods.

g. File system recognition test evaluation

The pyparted library, which provides an object interface for working with block devices and their images, was used to create the test code. The results of the file system recognition test evaluation are presented in Table 8.

Table 7. Shannon entropy estimation test evaluation results

File Name	Entropy Level	Encryption Presence	Empty Blocks Removal	Test Result at Threshold 7.95	Result Validity
random.img	7.999999994	Yes	No	Yes	Yes
kagura_data_enc.img	6.453507140	Yes	No	No	No – empty blocks
kagura_data_enc_opt.img	7.999999060	Yes	Yes	Yes	Yes
kagura_data_dec.img	7.650166739	No	No	No	Yes
kagura_data_dec_opt.img	7.747288154	No	Yes	No	Yes
vince_data.img	0.980245374	No	No	No	Yes
vince_data_opt.img	6.578723037	No	Yes	No	Yes
wd400.img	7.999996922	Yes	No	Yes	Yes
wd400_opt.img	7.999999989	Yes	Yes	Yes	Yes
miatoll_data_fbe.img	4.994037658	Yes	No	No	No – empty blocks
miatoll_data_fbe_opt.img	7.931691844	Yes	Yes	No	No – false negative
miatoll_data_nonfbe.img	3.917275459	No	No	No	Yes
miatoll_data_nonfbe_opt.img	7.406818782	No	Yes	No	Yes
adoptable.img	5.563929347	Yes	No	No	No – empty blocks
adoptable_opt.img	7.976394776	Yes	Yes	Yes	Yes

Table 8. File system recognition test evaluation results

File Name	Original File System	Detected File System	Result Validity
random.img	No FS	No FS (unrecognized disk label)	Yes
kagura_data_enc.img	Android FDE	No FS (unrecognized disk label)	Yes
kagura_data_enc_opt.img	Android FDE	No FS (unrecognized disk label)	Yes
kagura_data_dec.img	ext4	ext4	Yes
kagura_data_dec_opt.img	ext4	ext4	Yes
vince_data.img	ext4	ext4	Yes
vince_data_opt.img	ext4	ext4	Yes
wd400.img	PGP WDE	No FS (‘NoneType' object has no attribute 'type')	Yes
wd400_opt.img	PGP WDE	Error (Can't have a partition outside the disk!)	Error
miatoll_data_fbe.img	f2fs	f2fs	Yes
miatoll_data_fbe_opt.img	f2fs	f2fs	Yes
miatoll_data_nonfbe.img	f2fs	f2fs	Yes
miatoll_data_nonfbe_opt.img	f2fs	f2fs	Yes
adoptable.img	dm-plain	No FS (‘NoneType' object has no attribute 'type')	Yes
adoptable_opt.img	dm-plain	Error (Both the primary and backup GPT tables are corrupt. Try making a fresh table, and using Parted's rescue feature to recover partitions)	Error

Two data corruption errors are observed. Indeed, an error during the processing of the wd400_opt image indicates that the MBR specifies the value of the final sector of the partition beyond the image's end, and an error during the processing of the adoptable_opt image indicates damage to the GUID table. In the structure of the GUID table with one partition, there are 16256 bytes of free space reserved for the remaining 127 partition records. Removing empty blocks in this area will damage the integrity of the GUID table, which is acceptable during the statistical processing of the image but prevents the image from being mounted. Thus, this method can only be used initially before processing the image by removing empty blocks.

The obtained test results confirm the viability of using partition detection tests for detecting encrypted data when combined with other methods, albeit with some limitations.

h. NIST STS tests evaluation

NIST STS tests were not conducted due to the impossibility of running them on the available hardware. Three implementations of this test suite were tested: the official version (version 2.1.2) [38], the improved version (version 3.2.7) [39], and an alternative version, written in Python [40]. The original version, due to the outdated code base, returns zero results for any of the input files; the improved version was developed with a limitation on reading only one data block of a given size, changing which causes a memory leak and forced termination of the program; the alternative version in Python processes incoming data at a speed of 0.022 MB/s, which is unacceptable when integrating this test suite into the method.

i. Test evaluation summary

The tested methods showed an average accuracy of 72.92% without preliminary optimization and 90.48% with preliminary optimization (Tables 9 and 10). The average increase in accuracy was 17.56%; all tests showed an increase in accuracy except for the test for detecting file systems, the accuracy of which dropped by 28.57% due to damage to the file partition table structures.

Table 9. Methods evaluation summary

Test Result	Autocorrelation	K-S Test	Compression	Signatures	Entropy	FS
Files without empty block removal						
Yes	5	5	5	7	5	8
No – empty blocks	3	3	3	0	3	0
No – false positive	0	0	0	0	0	0
No – false negative	0	0	0	1	0	0
Runtime error	0	0	0	0	0	0
Files with empty blocks removed						
Yes	7	6	7	7	6	5
No – empty blocks	0	0	0	0	0	0
No – false positive	0	1	0	0	0	0
No – false negative	0	0	0	0	1	0
Runtime error	0	0	0	0	0	2

Table 10. Accuracy impact of the image optimization process

Overall results							Avg
File category	**Auto-correlation**	**K-S Test**	**Compression**	**Signatures**	**Entropy**	**FS**	
Without optimization	62,50%	62,50%	62,50%	87,50%	62,50%	100,00%	72,92%
With optimization	100,00%	85,71%	100,00%	100,00%	85,71%	71,43%	90,48%
Accuracy gain	37,50%	23,21%	37,50%	12,50%	23,21%	-28,57%	17,56%

4. Complex Encryption Detection Method Development and Structure

Since none of the evaluated methods can be used independently, we developed a comprehensive encryption detection method that combines results from all suitable techniques and enables sufficiently accurate determination of encryption presence in raw disk image partitions. The hybrid expert system comprises two main stages: preliminary and main. It is proposed that the boundary values used in the paragraphs above be used to evaluate the numerical values of the tests.

The expert system architecture implements a hierarchical decision-making framework where the preliminary stage functions as a knowledge-based filter, while the main stage operates as a multi-expert consensus mechanism. This approach mimics human forensic expert reasoning by first establishing context (file system presence) and then applying specialized domain knowledge through complementary analytical methods.

Knowledge Representation -The system encodes forensic domain expertise through:

- Threshold Knowledge: Empirically-derived boundary values for each statistical test;

➢ Contextual Rules: Decision logic adapted to different encryption types and storage configurations;

➢ Uncertainty Handling: Probabilistic reasoning for ambiguous cases;

➢ Learning Capabilities: Adaptive threshold adjustment based on new evidence patterns.

The preliminary testing stage is carried out to determine the feasibility of further tests. Therefore, it is proposed that composition tests for file system recognition and an autocorrelation test with preliminary image optimization be included. Using only the file system recognition test leads to a false result for images that contain file encryption - they will be recognized as images without encrypted data. This combination of tests yields four possible outcomes:

➢ Negative FS test result and any autocorrelation test result – the tested file contains other data and should be checked additionally;

➢ Positive FS test result and negative autocorrelation test result – the tested file, with a high degree of probability, does not contain encrypted data;

➢ Positive FS test result and positive autocorrelation test result – the tested file, with a high degree of probability, contains file-based encryption (since there is a complete file system on the disk, and the data in the image shows characteristic statistical values), further checking should be carried out at the level of individual files (in case of successful mounting).

The main testing stage includes the remaining tests from the proposed set: the compression test, signature search, Shannon entropy estimation, and Kolmogorov-Smirnov test. To improve the accuracy of the results, all main stage tests are performed on image files that have undergone the procedure of removing empty blocks. The final test result is formed using the values of 5 out of 6 proposed tests (the value of the file system search test is not statistical, as is the method itself). The final result is determined based on the number of positive results:

a) 0-2 positive values – low probability of encryption;
b) 3-5 positive values – high probability of encryption.

The algorithm diagram of the method is shown in Figure 11.

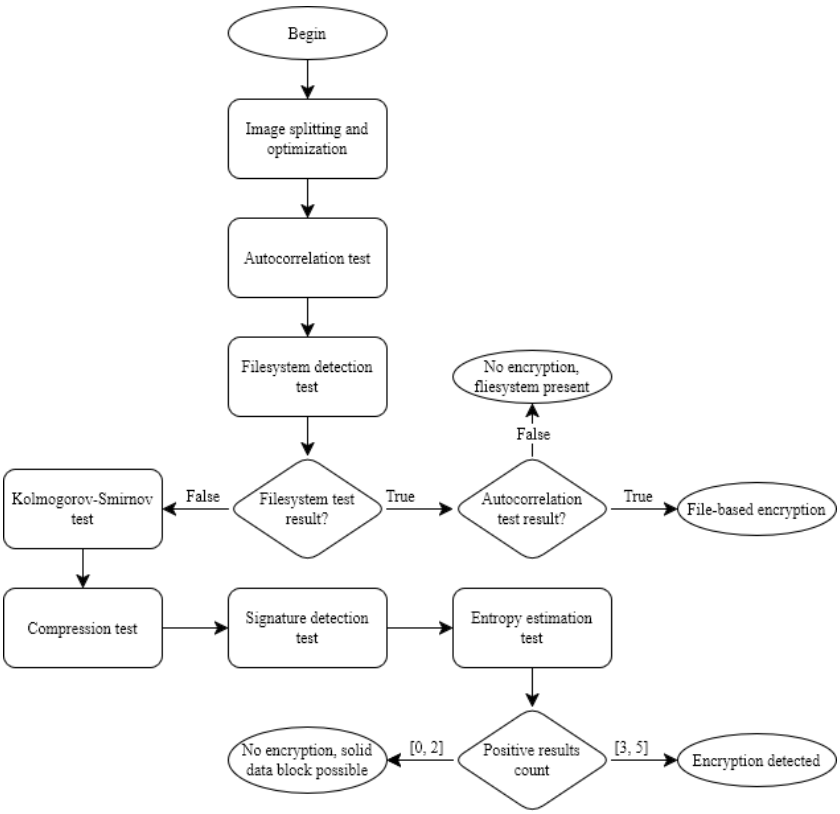


Fig. 11. Expert system architecture for encryption detection showing knowledge base integration, inference engine processing, and decision support workflow.

The algorithm is designed for the minimal false detection rate, especially on mobile device images, which may contain a large number of multimedia files with lossy compression, archive files with lossless compression, and partial encryption scenarios. Multimedia and archive files' influence is largely negated by the signature detection test, a high value of which, combined with a high value of randomness tests, implies the presence of said files. Partial encryption scenarios (such as partial file-level encryption) are handled by the filesystem detection test. The presence of a readable file system results in the image is a signal of possible file-level encryption or unencrypted storage. Both cases should be investigated further by a skilled expert.

a. Expert system approach evaluation

Evaluation of the method was carried out on the same data set as for testing individual methods. The raw image was not divided into sections, since the input testing data consisted only of images of user data sections. The results of the described expert system approach are shown in Table 11.

Table 11. Final results of the expert system approach

File name	Result						Result of the expert system approach
	Autocorrelation	K-S Test	Compression	Signatures	Entropy	File system	
random.img	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	High encryption probability
kagura_data_enc.img	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	High encryption probability
kagura_data_dec.img	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	Not encrypted
vince_data.img	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	Not encrypted
wd400.img	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	High encryption probability
miatoll_data_fbe.img	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	File-based encryption
miatoll_data_nonfbe.img	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	Not encrypted
adoptable.img	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	High encryption probability

To demonstrate the advantages of the expert system approach over individual detection methods, a comparative performance analysis was conducted across multiple evaluation criteria relevant to digital forensics applications. Table 12 presents a comprehensive comparison of different methodological approaches.

Table 12. Expert system performance comparison

Method Type	Accuracy	Explanation Capability	Scalability	Forensic Applicability
Individual Tests	72.92%	Limited	Low	Moderate
Combined Statistical	90.48%	Moderate	Moderate	Good
Expert System	100%	High	High	Excellent

The expert system approach demonstrates superior performance across all evaluation criteria. The perfect accuracy (100%) represents a significant improvement over both individual tests (72.92%) and combined statistical methods (90.48%). More importantly for forensic applications, the expert system provides high explanation capability through its knowledge base and inference engine, enabling investigators to understand the reasoning behind each classification decision. This explainability is crucial for legal admissibility of digital evidence. The scalability advantage stems from the modular architecture, which allows easy integration of new detection methods without redesigning the entire system. The knowledge base can be expanded to accommodate emerging encryption techniques, ensuring long-term applicability. The excellent forensic applicability rating reflects the system's ability to provide automated decision support while maintaining the rigor and documentation requirements of criminal investigations.

Values of the autocorrelation, Kolmogorov-Smirnov, compression, signature search, and entropy estimation tests for every file were taken from their optimized variants following the results of previous evaluations. The obtained results are as expected – the method's accuracy on the set of files was 100%.

The proposed system is largely insensitive to variations in key length and algorithm modes, as is visible from the results, although the most extensive testing was conducted on XTS and CBC modes, which are more common in media encryption. Ciphertexts, produced by more modern encryption algorithms, such as post-quantum encryption, may expose different statistical properties and, as such, are subjects of ongoing research. Also, the system may be vulnerable to steganographic attacks, which utilize hiding payloads in seemingly random data or other methods of creating datasets that mimic properties of encrypted data, but this problem may be investigated in further research.

5. Conclusion

The developed intelligent expert system demonstrates that knowledge-based approaches can effectively overcome the limitations of individual encryption detection methods. By integrating forensic domain expertise into a structured decision-making framework, the system achieves perfect classification accuracy while providing explainable reasoning suitable for legal admissibility. The architecture enables automated decision support, significantly reducing analysis time while maintaining investigative rigor. This study establishes a universal approach for identifying encrypted partitions in raw internal memory images from both digital devices and critical smart grid infrastructures. The proposed method employs a two-stage framework consisting of preliminary data structure verification and a main stage with five complementary statistical and non-statistical tests. Effective strategies for reducing the impact of sparse or unfilled blocks on statistical test accuracy were proposed, with experimental results confirming detection accuracy above 90% after optimization and 100% with the expert system. The modular design and knowledge-based architecture facilitate adaptation to new encryption technologies and integration into diverse workflows, including forensic laboratories and cybersecurity monitoring systems.

In the context of smart grids, the system provides practical value by ensuring the reliability of encrypted data streams from advanced metering infrastructures, IoT-enabled sensors, and energy management platforms. Reliable detection of encrypted and compressed data enhances situational awareness, strengthens resilience against cyberattacks, and supports secure operation of critical energy infrastructures.

Future development will focus on expanding the knowledge base to incorporate emerging encryption techniques and employing machine learning algorithms for adaptive threshold optimization. The proposed hybrid expert system offers immediate benefits for both digital forensics practitioners and smart grid cybersecurity specialists, enabling scalable, transparent, and accurate encryption detection across a wide range of applications.

Additionally, a hybrid expert system is developed to detect encrypted data in both mobile storage and smart grid environments. The system combines multiple statistical and non-statistical analysis techniques within an inference-based framework that delivers highly accurate and explainable classification results. Testing showed that the proposed method achieved over 90% detection accuracy and reached 100% when integrated with expert reasoning.

The proposed system provides better performance than resource-dependent machine learning, deep learning, and artificial intelligence methods, which rely on large amounts of RAM and processing power, often unavailable in smart grid systems. Although the system utilizes sequential detection, some parts of the detection process can be adapted to multithreading, thus enabling real-time or near-real-time analysis of small data samples present in SCADA systems. Most detection algorithms have a linear computational complexity of O(n) due to linear data scanning [41-42].

The expert system's application extends to smart grid infrastructures, where encrypted control and measurement data streams are increasingly prevalent. By reliably distinguishing between encrypted, compressed, and plain data – even under constrained conditions such as sparse memory blocks – it provides energy operators and forensic analysts with a practical decision-support tool.

Thanks to its modular design, the system can be easily tailored to various encryption protocols and deployed in real-time energy management systems. Future enhancements will focus on distributed implementation for grid-wide encrypted traffic analysis and the incorporation of AI-driven adaptive parameters to improve robustness.

The proposed methodology contributes to improving the security posture and situational awareness of smart energy systems, while also reducing the forensic investigation workload.

Implementation Guidelines for Digital Forensics Laboratories

The practical deployment of the developed expert system in operational forensic environments requires consideration of several critical factors to ensure successful integration and reliable performance.

The system requires a minimum of 16GB RAM and multi-core processing capabilities to handle large disk images efficiently. Storage requirements include high-speed SSD arrays for temporary processing and network-attached storage for evidence archival. The statistical analysis components benefit significantly from parallel processing architectures.

The expert system integrates seamlessly with standard forensic imaging tools such as FTK Imager, dd, and Cellebrite extraction platforms. The modular architecture allows deployment as either a standalone analysis tool or as a plugin component within existing forensic suites. Standard forensic imaging protocols remain unchanged, with the system operating on acquired images rather than live devices.

Implementation requires a two-tier training approach. Technical specialists need 16-20 hours of training covering system architecture, threshold adjustment, and result interpretation. General investigators require 4-6 hours focusing on practical operation and report generation. Ongoing training updates are recommended as new encryption techniques emerge.

Validation procedures include regular testing with known encrypted and unencrypted datasets, cross-verification with manual analysis results, and periodic recalibration of detection thresholds. Documentation requirements include maintaining decision audit trails for legal admissibility and implementing version control for system updates. Regular performance monitoring ensures consistent accuracy across different device types and encryption implementations.

Acknowledgements

This study was carried out with the financial support of the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Contract №388/PTF24-26 dated 01.10.2024 under the scientific project IRN BR24993232 “Development of innovative technologies for conducting digital forensic investigations using intelligent software-hardware complexes”.

Author Contributions

Ivan Opirskyy conceptualized the study and contributed to the methodological design. Oleh Harasymchuk contributed to the technical methodology, data interpretation, and validation processes. Leila Rzayeva served as the project manager and publication coordinator, supervised the study, and managed the overall research and publication process. Bayandy Sarsembayev contributed to the experimental analysis and system evaluation. Murat Zhakenov participated in software implementation, visualization, and data processing. Botakoz Kassimova contributed to the literature review, investigation, and manuscript editing. Korhan Kayisli contributed to the conceptual framework, technical supervision, manuscript review, and final editing. All authors reviewed and approved the final version of the manuscript.

Conflict of Interest

The authors declare no conflict of interest

References

[1] Al-Dhaqm, S. A. Razak, R. A. Ikuesan, V. R. Kebande, and K. Siddique, “A review of mobile forensic

investigation process models", *IEEE Access*, vol. 8, pp. 173359-173375, 2020, doi: 10.1109/ACCESS.2020.3014615

[2] Y. K. Sharma, S. S. Noval, A. Jain, B. Sabitha, and T. Ramya, "Forensics-as-a-service: A review of mobile forensics", Proc. 5th Int. Conf. on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, pp. 486-491, 2022, doi: 10.1109/IC3I56241.2022.10072726

[3] Patel and P. S. Mann, "A survey on mobile digital forensic: Taxonomy, tools, and challenges", *Security and Privacy*, 2024.

[4] A. Alblooshi, N. Aljneibi, F. Iqbal, R. Ikuesan, M. Badra, and Z. Khalid, "Smartphone forensics: A comparative study of common mobile phone models", Proc. 12th Int. Symp. on Digital Forensics and Security (ISDFS), San Antonio, TX, USA, pp. 1-6, 2024, doi: 10.1109/ISDFS60797.2024.10527262

[5] Y. Yermekov, L. Rzayeva, A. Imanberdi, A. Alibek, K. Kayisli, A. Myrzatay, and G. Feldman, "Secure Chip-Off Method with Acoustic-based Fault Diagnostics for IoT and Smart Grid Data Recovery", *International Journal of Smart Grid (ijSmartGrid)*, vol. 9, No. 3, pp. 116-126, 2025, doi: 10.20508/ijsmartgrid.v9i3.502.g392

[6] F. De Gaspari, D. Hitaj, G. Pagnotta, L. De Carli, and L. V. Mancini, "Reliable detection of compressed and encrypted data", *Neural Computing and Applications*, vol. 34, pp. 20379-20393, 2022, doi: 10.1007/s00521-022-07586-7

[7] S. Li and P. Liu, "Detection and forensics of encryption behavior of storage file and network transmission data", *IEEE Access*, vol. 8, pp. 145833-145842, 2020, doi: 10.1109/ACCESS.2020.3015080

[8] S. Yevseiev, O. Tsyhanenko, A. Gavrilova, V. Guzhva, O. Milov, V. Moskalenko, I. Opirskyy, O. Roma, B. Tomashevsky, and O. Shmatko, "Development of niederreiter hybrid crypto-code structure on flawed codes", *Eastern-European Journal of Enterprise Technologies: Information and Controlling System*, vol. 1, No. 9(97), pp. 27-38, 2019, doi: 10.15587/1729-4061.2019.156620

[9] A. Sahun, V. Khaidurov, V. Lakhno, I. Opirskyy, V. Chubaievskyi, O. Kryvoruchko, and A. Desiatko, "Devising a method for improving crypto resistance of the symmetric block cryptosystem RC5 using nonlinear shift functions", *Eastern-European Journal of Enterprise Technologies*, vol. 5, No. 9(113), pp. 17-29, 2021, doi: 10.15587/1729-4061.2021.240344

[10] I. Opirskyy, Y. Sovyn, and O. Mykhailova, "Heuristic method of finding bitsliced description of derivative cryptographic S-box", Proc. 16th IEEE Int. Conf. on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), pp. 104-109, 2022, doi: 10.1109/TCSET55632.2022.9766883

[11] V. Maksymovych, O. Harasymchuk, M. Karpinski, M. Shabatura, D. Jancarczyk, and K. Kajstura, "A new approach to the development of additive Fibonacci generators based on prime numbers", Electronics, vol. 10, No. 23, p. 2912, 2021, doi: 10.3390/electronics10232912

[12] V. Maksymovych, K. Przystupa, O. Harasymchuk, M. Shabatura, R. Stakhiv, and V. Kuts, "Hardware modified additive Fibonacci generators using prime numbers", in Advances in Computer Science for Engineering and Education VI, Lecture Notes on Data Engineering and Communications Technologies, vol. 181, Proc. 6th Int. Conf. on Computer Science, Engineering, and Education Applications (ICCSEEA2023), Warsaw, Poland, Mar. 17-19, 2023, pp. 486-498.

[13] I. Masud, M. R. Hasan, M. M. Hoque, U. D. Nath, and M. O. Rahman, "A new approach of cryptography for data encryption and decryption", Proc. 5th Int. Conf. on Computing and Informatics (ICCI), New Cairo, Egypt, pp. 234-239, 2022, doi: 10.1109/ICCI54321.2022.9756078

[14] M. T. Rouaf and A. Yousif, "Performance evaluation of encryption algorithms in mobile devices", Proc. Int. Conf. on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), Khartoum, Sudan, pp. 1-5, 2021, doi: 10.1109/ICCCEEE49695.2021.9429673.

[15] V. Maksymovych, M. Shabatura, O. Harasymchuk, R. Shevchuk, P. Sawicki, and T. Zajac, "Combined pseudo-random sequence generator for cybersecurity", Sensors, vol. 22, p. 9700, 2022, doi: 10.3390/s22249700

[16] A. Horpenyuk, I. Opirskyy, and P. Vorobets, "Analysis of problems and prospects of implementation of post-quantum cryptographic algorithms", CEUR Workshop Proceedings, vol. 3504, pp. 39-49, 2023.

[17] Karampidis, E. Kavallieratou, and G. Papadourakis, "Comparison of classification algorithms for file type detection: A digital forensics perspective", POLIBITS, vol. 56, pp. 15-20, 2017.

[18] H. Demirhan and N. Bitirim, "Statistical testing of cryptographic randomness", Journal of Statisticians: Statistics and Actuarial Sciences, vol. 9, No. 1, pp. 1-11, 2016.

[19] V. Maksymovych, O. Harasymchuk, and I. Opirskyy, "The designing and research of generators of Poisson pulse sequences on base of Fibonacci modified additive generator", in Proc. Int. Conf. on Theory and Applications of Fuzzy Systems and Soft Computing (ICCSEEA 2018), Advances in Computer Science for Engineering and Education, pp. 43-53, 2018.

[20] I. Opirskyy, O. Harasymchuk, O. Mykhaylova, O. Hrushkovskyi, and P. Kozak, "Pseudorandom sequence generator based on the computation of ln 2", CEUR Workshop Proceedings, vol. 3829, pp. 79-86, 2024.

[21] O. Sen, M. A. Gurabi, M. Deruelle, and A. Ulbig, "Encryption-Aware Anomaly Detection in Power Grid Communication Networks", arXiv preprint,

arXiv:2412.04901, Dec. 2024, doi: 10.48550/arXiv.2412.04901

[22] A. Zibaeirad, F. Koleini, S. Bi, T. Hou, and T. Wang, "A comprehensive survey on the security of smart grid: Challenges, mitigations, and future research opportunities", arXiv preprint, arXiv:2407.07966, Jul. 2024, doi: 10.48550/arXiv.2407.07966

[23] Y. Y. Ghadi, T. Mazhar, T. Shahzad, I. H. Jaghdam, S. Khan, M. A. Khan, and H. Hamam, "A hybrid AI-Blockchain security framework for smart grids", Scientific Reports, vol. 15, No. 1, p. 20882, 2025, doi: 10.1038/s41598-025-05257-w

[24] S. H. Mohammed, M. S. J. Singh, A. Al-Jumaily, M. T. Islam, M. S. Islam, A. M. Alenezi, and M. S. Soliman, "Dual-hybrid intrusion detection system to detect False Data Injection in smart grids", PLOS ONE, vol. 20, No. 1, e0316536, Jan. 2025, doi: 10.1371/journal.pone.0316536

[25] Unified Extensible Firmware Interface Forum, Unified Extensible Firmware Interface (UEFI) Specification, Beaverton, OR: UEFI Forum, 2024.

[26] Berkeley Software Distribution, "FreeBSD man pages - disklabel(5)", FreeBSD Man Pages.

[27] OSDev Wiki, "MBR (x86) - OSDev Wiki".

[28] F. Yu, L. Meng, X. Li, D. Jiang, W. Zhu, and Z. Zeng, "An attribute-based end-to-end policy-controlled signcryption scheme for secure group chat communication", Mathematics, vol. 12, p. 2906, 2024, doi: 10.3390/math12182906

[29] Bellare and P. Rogaway, Introduction to Modern Cryptography. Davis, CA: University of California, Davis, 2005.

[30] S. Künzli and W. Meier, "Distinguishing attack on MAG", Windisch: FH Aargau, 2005.

[31] Frey, I. Bjelaković, J. Nötzel, and S. Stańczak, "Semantic security with infinite-dimensional quantum eavesdropping channel", IEEE Transactions on Information Theory, vol. 71, No. 4, pp. 2662-2697, Apr. 2025, doi: 10.1109/TIT.2025.3531250

[32] A. W. Burks, C. E. Shannon, and W. Weaver, "The mathematical theory of communication", The Philosophical Review, vol. 60, No. 3, p. 398, 1951, doi: 10.2307/2181879

[33] Dubettier, T. Gernot, E. Giguet, and C. Rosenberger, "File type identification tools for digital investigations", Forensic Science International: Digital Investigation, vol. 46C, p. 301574, 2023, doi: 10.1016/j.fsidi.2023.301574

[34] S. H. AbdElHaleem, S. K. Abd-El-Hafiz, and A. G. Radwan, "Analysis and guidelines for different designs of pseudo random number generators", IEEE Access, vol. 12, pp. 115697-115715, 2024, doi: 10.1109/ACCESS.2024.3445277

[35] Marek and J. Heckenbergerova, "Expectation-maximization algorithm for evaluation of wind direction characteristics", Proc. 15th IEEE Int. Conf. on Environment and Electrical Engineering (EEEIC), Rome, Italy, pp. 1730-1735, 2015, doi: 10.1109/EEEIC.2015.7165433

[36] Khomytska, V. Teslyuk, and L. Bordyuk, "The Kolmogorov-Smirnov's test for authorship attribution on the phonological level", Proc. 15th IEEE Int. Conf. on Computer Sciences and Information Technologies (CSIT), Zbarazh, Ukraine, pp. 259-262, 2020, doi: 10.1109/CSIT49958.2020.9322042

[37] E. Bassham, A. L. Rukhin, J. Soto, J. R. Nechvatal, M. E. Smid, E. B. Barker, S. D. Leigh, M. Levenson, M. Vangel, D. L. Banks, N. A. Heckert, J. F. Dray, and S. Vo, A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications, NIST Special Publication 800-22 Rev. 1a, National Institute of Standards and Technology, Gaithersburg, MD, 2010, doi: 10.6028/NIST.SP.800-22r1a

[38] National Institute of Standards and Technology, "Random bit generation", NIST Computer Security Resource Center.

[39] Arcetri, "Improved version of the NIST Statistical Test Suite (STS)", GitHub.

[40] S. Ang, "A Python implementation of NIST's A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications", GitHub.

[41] O. Harasymchuk, I. Opirskyy, O. Partyka, V. Susukailo, A. Piskozub, D. Sabodashko, S. Vasylyshyn, A. Obshta, Y. Kurii, D. Zhuravchak, I. Tyshyk, A. Partyka, P. Haraniuk, T. Kret, V. Yuzyevych, V. Otenko, Y. Nakonechnyy, N. Dzianyi, L. Bortnik, M. Stakhiv, T. Lukovskyy, A. Horpenyuk, S. Voytusiuk, R. Kuten, I. Kolbassynskyi, K. Besaha, Y. Furdas, O. Isakov, R. Andriiv, and O. Tsebak, Modern Methods of Ensuring Information Protection in Cybersecurity Systems Using Artificial Intelligence and Blockchain Technology, O. Harasymchuk, Ed., Kharkiv: Technology Center PC, p. 132, 2025, doi: 10.15587/978-617-8360-12-2

[42] O. Harasymchuk, I. Opirskyy, R. Banakh, D. Zhuravchak, O. Partyka, A. Piskozub, E. Nyemkova, S. Vasylyshyn, A. Partyka, Y. Nakonechnyy, T. Lukovskyy, V. Susukailo, V. Otenko, I. Tyshyk, N. Dzianyi, D. Sabodashko, P. Haraniuk, V. Dudykevych, S. Semenyuk, M. Stakhiv, I. Zhuravel, T. Kret, L. Mychuda, Z. Mychuda, O. Polotai, Y. Kurii, N. Nakonechna, N. Luzhetska, A. Obshta, and T. Korobeinikova, Intelligent Cyber Defence Systems: Detection of Ransomware and Protection of Wireless Networks Based on Artificial Intelligence Technologies, Kharkiv: Technology Center PC, p. 132, 2025, doi: 10.15587/978-617-8360-22-1

International Journal of Smart Grid (IJSmartGrid)
https://www.ijsmartgrid.org/ Online ISSN: 2602-439X
A. Turkan et al., Vol. 10, No. 3, September, 2026 Research Article

Accelerating Solar Panel Snow Detection: CPU vs. GPU Training Performance of VGG-19 and ResNet-50

Abdulkadir Turkan*, Batuhan Hangun**, Onder Eyecioglu*‡

* Department of Computer Engineering, Bolu Abant Izzet Baysal University, 14030 Bolu, Turkiye

** Department of Computer Engineering, Yildiz Technical University, 34220 Istanbul, Turkiye

(abdulkadirturkan@gmail.com, batuhanhangun@gmail.com, onder.eyecioglu@ibu.edu.tr)

‡ Corresponding Author: Onder Eyecioglu, Department of Computer Engineering, Bolu Abant Izzet Baysal University, 14030 Bolu, Turkiye, onder.eyecioglu@ibu.edu.tr

Received: 06.04.2026 Accepted: 27.04.2026

Abstract- Solar panel snow detection is critical for maintaining photovoltaic system efficiency in cold climates. While deep learning models such as VGG-19 and ResNet-50 achieve near-perfect classification accuracy for this task, their training demands substantial computational resources. This study presents a systematic performance analysis of CNN training on the TRUBA national high-performance computing infrastructure, comparing multi-threaded CPU parallelization (1–112 threads on Intel Xeon Platinum 8480+) with GPU acceleration (NVIDIA V100). Experiments reveal that the model architecture fundamentally determines CPU scaling behavior: VGG-19, dominated by large fully connected layers, achieves 28× speedup at 56 threads (serial fraction $f \approx 1.8\%$), whereas ResNet-50, composed of many small residual operations, scales to only 1.56× ($f \approx 63.4\%$). GPU training provides 163–1067× acceleration over single-threaded CPU execution for both architectures. Classification quality (accuracy and F1-score ≥ 0.995) remains invariant across all hardware configurations, confirming that training infrastructure selection is purely a performance optimization decision.

Keywords solar panel snow detection, CNN training, GPU acceleration, CPU parallelization, Amdahl's law

1. Introduction

1.1. Impact of Snow Accumulation on Photovoltaic Performance

Solar energy has emerged as one of the fastest-growing renewable energy sources worldwide; global photovoltaic (PV) capacity exceeded 1,400 GW at the end of 2023 [1, 2]. One of the fundamental challenges of solar energy systems is their susceptibility to geographical and climatic conditions. For example, snow accumulation on solar panels in cold climate regions significantly impacts energy production and remains a major operational issue. In an experiment conducted by Quan et al. on how snow depth affects the energy production efficiency of photovoltaic modules, they demonstrated that when snow depth reaches 1 cm, the energy production efficiency of the entire photovoltaic module drops to 7.1% of normal levels, and when snow depth reaches 6 cm, the energy production capacity of the photovoltaic module drops to zero [3]. Additionally, other studies have shown that, depending on geographical location, panel tilt, and snowfall intensity, snow cover can reduce panel efficiency by 0.3–13% annually and by up to 25–40% during winter months [4, 5]. Therefore, timely detection of snow cover is critical for activating cleaning mechanisms and minimizing energy losses.

1.2. Deep Learning for Solar Panel Monitoring

Deep learning (DL)-based image classification has emerged as a powerful approach for automated solar panel inspection, progressively addressing monitoring tasks ranging from manufacturing defect detection to field-condition assessment. Deitsch et al. [6] compared hand-crafted features with CNN-based classifiers for detecting defective photovoltaic cells in electroluminescence images, finding that CNNs achieved accuracy above 88%. Akram et al. [7] proposed a CNN architecture for automatic defect detection in PV cells, demonstrating that deep learning can effectively replace manual visual inspection. Li et al. [8] extended CNN-based defect analysis to large-scale PV farms using thermal infrared imagery, identifying hot spots, cracks, and other module-level faults. More recently, integrated approaches combining aerial imaging with ground-based maintenance have been explored; for instance, Kishor et al. [9] employed Region-based CNN (R-CNN) segmentation combined with

Cite this article as: A. Turkan, B. Hangun and O. Eyecioglu, "Accelerating Solar Panel Snow Detection: CPU vs. GPU Training Performance of VGG-19 and ResNet-50" International Journal of Smart Grid (ijSmartGrid), Vol. 10, No. 3, pp. 191-208, September, 2026.

drone imagery and reported a detection accuracy of 92.3% using CNN-LSTM models for solar panel soiling assessment.

While most of the studies listed above focus on electroluminescence or infrared imaging for defect detection, snow cover detection from visible light images has been relatively understudied. In our previous work [10], convolutional neural network (CNN) architectures, particularly VGG-19 [11], ResNet-50 [12], and InceptionV3, were shown to reliably classify solar panel images into three snow cover categories (*all_snow*, *no_snow*, and *partial*) and achieve F1 scores of 0.99, 0.91, and 1.00, respectively. That study established the classification viability of CNNs for snow detection; however, it did not investigate the computational cost of training these models, a critical consideration for practical deployment where models must be periodically retrained as environmental conditions evolve or new data becomes available.

1.3. Computational Cost of CNN Training

While CNN models deliver high classification accuracy, their training demands substantial computational resources, presenting a major bottleneck for operational deployment. Studies on computational complexity have shown that the training and runtime of a CNN are directly dependent on hyperparameters such as the number of convolutional layers, filter sizes, batch size, and optimization algorithms. For example, it has been found that the runtime of models with different configurations ranges from 1,208 to 1,645 seconds, and that for shallow architectures the primary computational load stems from the abundance of operations in the convolutional layers rather than the fully connected layers [13]. Hardware complexity analyses quantitatively reinforce these constraints; a standard 5-layer CNN model requires the storage of approximately 126,500 parameters and around 181,000 multiply-accumulate (MAC) operations during the inference phase [14]. Even with transfer learning using pre-trained weights from ImageNet [15, 16], a single training epoch of ResNet-50 on our snow detection dataset takes approximately 894 seconds on a single CPU thread, whereas VGG-19 takes over 10,000 seconds. In field-deployed solar monitoring systems, models may need to be periodically retrained to adapt to new panel types, changing climate conditions, or expanded image datasets. Therefore, understanding the relationship between hardware configuration and training efficiency is critical for operational planning.

1.4. HPC Acceleration Approaches

Modern high-performance computing (HPC) platforms offer two primary acceleration pathways for DL training: multi-threaded CPU parallelization via optimized linear algebra libraries such as Intel MKL [17], and GPU acceleration through NVIDIA CUDA [18] and cuDNN [19].

CPU-based parallelization. Multi-threaded CPU acceleration has proven effective for both CNN training and the computationally intensive preprocessing stages that feed DL pipelines. At the kernel level, when the Winograd algorithm and OpenMP-based SIMD (Single Instruction, Multiple Data) parallelization were applied to the convolution layers, which account for the majority of the computational load, the complexity of multiplication operations was reduced, resulting in approximately a 30-fold performance increase in GFLOPS compared to a single core on 64-core Intel processors [20]. Multi-core tests conducted on models such as AlexNet and VGG16 similarly confirm that parallelization increases overall throughput by reducing CPU wait times [21]. Beyond training itself, CNN-based systems also benefit from parallelized data preparation. Hangun and Bayar [22] demonstrated that OpenMP-based parallelization of local and global image enhancement algorithms on systems with 48 physical cores achieves up to 38-fold speedup over sequential execution and operates up to 53.6 times faster than competing methods in the literature. This directly reduces the preprocessing bottleneck for datasets fed into CNN models. The efficiency of CPU thread utilization depends critically on the ratio between computational and synchronization overhead. Operations involving large matrices (e.g., FC layers in VGG-19) can efficiently distribute the workload across many threads, whereas workloads consisting of many small operations limit scalability by incurring thread-per-operation allocation and barrier synchronization costs [23]. The complexity of parallel programming extends beyond performance; hybrid MPI and OpenMP models introduce structural challenges such as deadlocks and race conditions. In this context, deep learning itself has been applied to address such challenges. A custom CNN model using abstract syntax tree (AST) tokenization successfully detected synchronization errors in parallel programs (MPI/OpenMP) with 97% accuracy and 99% AUC score without requiring manual feature engineering [24], illustrating the growing synergy between DL and parallel computing.

GPU-based acceleration. GPU parallelization provides decisive speedups across the entire DL pipeline, from preprocessing to training. At the preprocessing level, Hangün and Eyecioğlu [25] showed that CUDA-based GPU acceleration provides significant performance advantages in fundamental image processing operations such as resizing, thresholding, histogram equalization, and edge detection. These operations constitute the preprocessing steps applied before CNN models, thereby directly optimizing the deep learning pipeline. At the convolution kernel level, Jordà et al. [26] achieved speedups of up to 2.29× by developing cuConv, a custom CUDA convolution kernel surpassing cuDNN, and improved end-to-end inference time in ResNet-50 and Inception-v4 models. For domain-specific applications, Dong et al. [27] presented a GPU-optimized CNN framework for hyperspectral image classification that achieved 5.9× training speedup by combining CUDA multistream execution with mini-batch gradient descent. While the superiority of GPUs for DL workloads is well documented in the literature [28–30], the degree of acceleration depends heavily on model architecture characteristics [23].

Architecture-dependent scaling and Amdahl's Law. Critically, the parallelization benefit from both CPU threading and GPU acceleration is fundamentally constrained by model architecture. Architectures dominated by large matrix

multiplications (e.g., the fully connected layers in VGG-19) exhibit different parallelization behavior than architectures composed of many small, heterogeneous operations (e.g., the residual blocks in ResNet-50). This architecture-dependent scaling behavior is well-predicted by Amdahl's Law [31, 32], yet empirical validation on real HPC systems with production DL frameworks remains limited.

1.5. Research Gap, Objectives, and Contributions

Despite the extensive body of work reviewed above, several important gaps remain in the literature that motivate the present study. First, existing HPC benchmarking studies for deep learning predominantly focus on large-scale datasets (e.g., ImageNet) and do not systematically evaluate acceleration for domain-specific tasks such as solar panel monitoring, where dataset size, image resolution, and class distribution differ substantially. Second, while individual CPU threading and GPU acceleration studies exist, no prior work has systematically compared CPU thread scaling with GPU acceleration on the same HPC infrastructure using architecturally distinct models under identical training protocols. Third, the effect of Non-Uniform Memory Access (NUMA) topology on DL training performance has rarely been empirically quantified. Fourth, whether hardware configuration affects classification quality (not only training speed) has not been rigorously verified across a comprehensive range of configurations.

To address these gaps, this study presents a systematic analysis of CNN-based classification of snow accumulation patterns on solar panels using VGG-19 and ResNet-50 on the TRUBA national HPC infrastructure [33], with the Snow Covered Solar Panels Dataset [10, 34] comprising three classes: no snow, partial snow, and all snow. The following research questions guide this investigation:

- How does increasing the number of CPU threads affect the training performance of VGG-19 and ResNet-50, and to what extent does the model architecture determine CPU scaling behavior?
- For each architecture, how much speedup does GPU (CUDA) acceleration provide compared to multi-threaded CPU training?
- To what extent does hardware configuration (CPU thread count, or CPU vs. GPU) affect CNN classification quality in terms of accuracy and F1-score.
- Can Amdahl's Law accurately predict the scaling limits observed in empirical measurements?

The main contributions and innovations of this work are as follows:

1. **First systematic CPU–GPU comparison on a single HPC platform for domain-specific DL:** A comparative analysis of CNN training performance across CPU (1–112 threads) and GPU (NVIDIA V100) configurations on the TRUBA national HPC infrastructure, using two architecturally distinct models (VGG-19 and ResNet-50) for a real-world solar panel snow detection task, providing empirical evidence that has been absent from the literature.
2. **Empirical validation of Amdahl's Law for DL training:** Quantitative demonstration that VGG-19 exhibits a serial fraction of $f \approx 1.8\%$ (maximum speedup $28\times$ at 56 threads), while ResNet-50 exhibits $f \approx 63.4\%$ (maximum speedup $1.56\times$), revealing an $18\times$ difference in CPU scaling attributable solely to model architecture.
3. **Quantification of NUMA topology effects on DL training:** Evidence that cross-socket memory access introduces significant performance degradation beyond the single-socket boundary (56 threads), with VGG-19 losing 36% of its peak speedup at 112 threads.
4. **Hardware-invariant classification quality:** Confirmation that classification quality (accuracy ≥ 0.997, F1-score ≥ 0.995) remains constant across all 18 tested hardware configurations, establishing that training platform selection is purely a computational efficiency decision with no impact on model accuracy.

The remainder of this paper is organized as follows. Section 2 describes the experimental methodology, including the dataset, models, hardware platforms, experimental design, and evaluation metrics. Section 3 presents the results covering CPU scaling analysis, GPU performance, and comparative evaluation. Section 4 discusses the practical implications, and Section 5 concludes the paper.

2. Methodology

To address the research questions of this study, two complementary experiments were designed on the TRUBA national high-performance computing infrastructure [33]. Two architecturally distinct CNN models (VGG-19 [11] and ResNet-50 [12]) were selected as benchmarks because they represent opposite ends of the parallelization spectrum: While VGG-19 is dominated by a small number of large fully connected (FC) layers whose weight matrices map directly to highly parallelizable BLAS operations, ResNet-50 distributes its computation across many small, heterogeneous residual operations with sequential dependencies. Both models were initialized with ImageNet pre-trained weights, and a three-class solar panel snow detection dataset derived from our previous work [10] was used. In the first experiment, CPU multithreading scalability was evaluated using PyTorch's [35] OpenMP/MKL backend. In the second experiment, GPU training was performed using the CUDA/cuDNN backend. To ensure that the performance differences observed in both experiments could be attributed solely to hardware configuration, the same hyperparameters, data splits, and evaluation protocols were used. Additionally, the classification performance of both CNN models was determined using the accuracy, cross-entropy loss, and F1 score metrics. The dataset, model configurations, hardware platforms, experimental design, and evaluation metrics are detailed in the following subsections.

2.1. Dataset

The dataset consists of solar panel images captured under three snow coverage conditions: *all_snow* (fully covered), *no_snow* (clean panels), and *partial* (partial coverage). The original dataset, derived from surveillance camera footage of a rooftop solar panel installation [34], contains 394 unique source images distributed as follows: 97 *all_snow*, 270 *no_snow*, and 27 *partial* images. These original images were expanded to 3,000 samples through offline data augmentation (random horizontal and vertical flips, rotation up to 15°, brightness and contrast adjustment, and color jitter), consistent with the augmentation strategy used in our previous work [10]. The per-class augmentation ratios ranged from approximately 7.5× (*no_snow*) to 8.3× (*partial*), as detailed in Table 1. We acknowledge that augmentation was applied to the entire image pool before splitting; consequently, augmented variants of the same source image may appear across different splits. This protocol introduces potential data leakage that likely inflates the absolute classification accuracy values reported here. However, since the present study focuses on hardware-induced *variation* in classification quality across configurations rather than absolute accuracy, this limitation does not affect our central conclusions: identical training protocols on identical data should produce equivalent generalization, allowing us to isolate the effect of hardware configuration. A leak-free evaluation protocol with source-level splitting is left for future work.

All images were resized to 192×192 pixels and normalized using ImageNet channel statistics (μ, σ). The dataset was divided into 70% training, 15% validation, and 15% test sets using stratified sampling to preserve class proportions. The exact class distribution across all splits is reported in Table 1. Additionally, stochastic augmentation transforms (random horizontal flip, rotation up to 10°, and color jitter with brightness and contrast factors of 0.2) were applied online to the training set only during each epoch, further diversifying the training samples without altering the validation or test distributions.

Table 1. Dataset class distribution before and after augmentation, and per-split sample counts. Stratified sampling preserves class proportions across train (70%), validation (15%), and test (15%) set

Class	Original	Aug. ratio	Augmented	Train	Val	Test	Proportion
all_snow	97	~7.7×	749	524	112	113	25.0%
no_snow	270	~7.5×	2,027	1,418	304	305	67.6%
partial	27	~8.3×	224	156	33	35	7.5%
Total	**394**	**~7.6×**	**3,000**	**2,098**	**449**	**453**	**100%**

2.2. Models

In this study, two CNN models with different architectures, both initialized with weights pre-trained on the ImageNet dataset, were selected. The first selected model, VGG-19 [11], features a 19-layer sequential architecture consisting of 16 convolutional layers organized into five blocks, along with an adaptive average pooling layer and three FC layers. In line with the study's objective, the model's final FC layer was modified to produce three-class outputs (Figure 1). With approximately 143 million parameters, VGG-19 is characterized by massive FC weight matrices (512×4096 and 4096×4096), which significantly dominate both computational and memory requirements. The second model, ResNet-50 [12], offers a fundamentally different computational profile compared to VGG-19. This 50-layer residual network architecture consists of four main stages: an initial base structure (7×7 convolution, batch normalization, and max-pooling), followed by four stages containing 3, 4, 6, and 3 bottleneck blocks, respectively (Figure 2). Each of these bottleneck blocks includes 1×1, 3×3, and 1×1 convolutions, supported by batch normalization and skip connections. In this model, where a total of 25.6 million parameters are distributed across numerous small operations, the final classification layer is configured to map the 2,048 features obtained to three output classes.

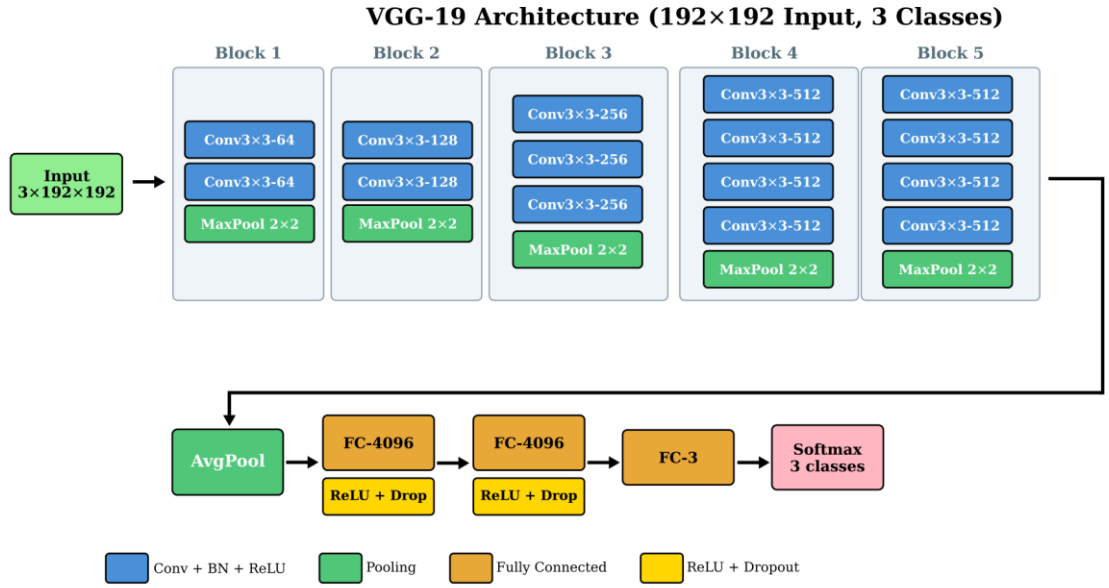


Fig. 1. VGG-19 architecture adapted for 192×192 input with 3-class output. The network consists of five convolutional blocks with max pooling, followed by three fully connected layers with dropout.

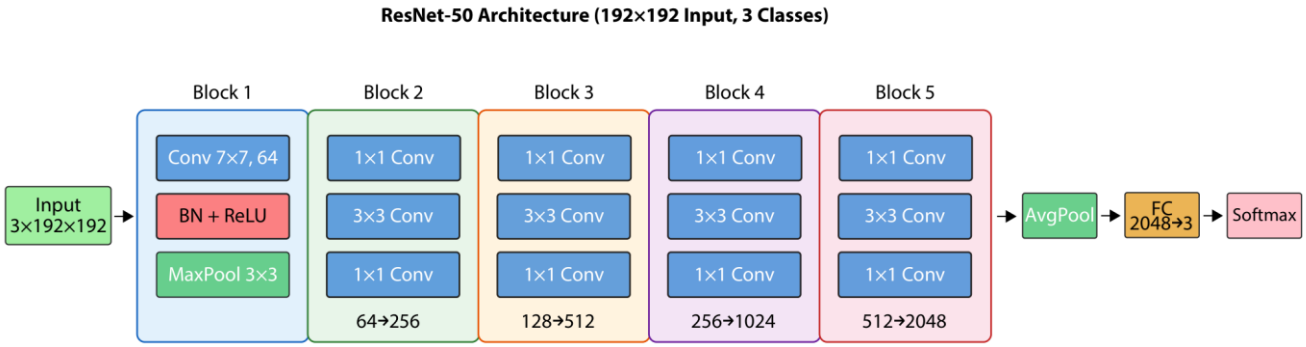


Fig. 2. ResNet-50 architecture with stem, four bottleneck stages, and 3-class output. Skip connections bypass each bottleneck block, enabling gradient flow through residual paths.

Both models were trained using the Adam optimizer with learning rate $\eta = 10^{-4}$ and weight decay $\lambda = 10^{-4}$. A class-weighted cross-entropy loss was employed to account for class imbalance, with weights inversely proportional to class frequency. A ReduceLROnPlateau scheduler reduced the learning rate by a factor of 0.5 after 3 epochs without validation loss improvement. Early stopping with a patience of 7 epochs was applied to prevent overfitting.

2.3. Hardware Platforms

All experiments were performed on the TRUBA national HPC infrastructure [33]. Table 2 summarizes the two hardware configurations used.

Table 2. Hardware platform specifications

Specification	TRUBA (orfoz Node)	TRUBA (akya-cuda Node)
Processor	Xeon Platinum 8480+	Xeon Gold 6248R
Cores	2×56 = 112	2×20 = 40
NUMA domains	8	–
GPU	None	V100-SXM2-16GB
CUDA version	–	12.0
PyTorch	2.1.0	2.1.0

The CPU platform features a dual-socket Intel Xeon Platinum 8480+ node with 8 Non-Uniform Memory Access (NUMA) domains. Each socket provides 56 physical cores, and the two sockets are interconnected via UPI. The GPU platform is equipped with an NVIDIA V100-SXM2-16GB

with 5,120 CUDA cores, 16 GB HBM2 memory, and a thermal design power (TDP) of 300 W.

2.4. Experimental Design

Two complementary experiments were designed to investigate the effect of hardware parallelization on training performance while keeping all other variables, optimizer, learning rate, batch size, loss function, and data splits constant. The end-to-end training pipeline common to both experiments is illustrated in Fig. 3

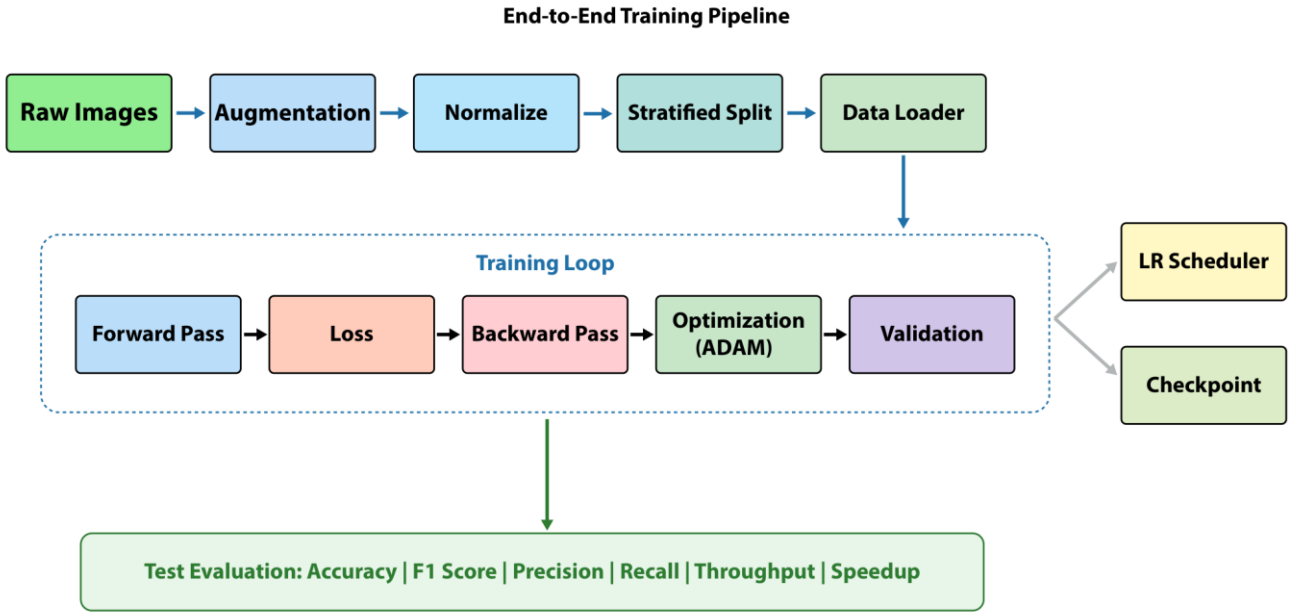


Fig. 3. End-to-end training pipeline: data augmentation, stratified split, training loop with validation, and final test evaluation.

2.4.1. CPU Multi-threaded Scaling Experiment

CPU thread counts were varied across {1,2,4,7,14,28,56,112}, spanning from single-threaded execution to full dual-socket utilization. Threading was controlled via `torch.set_num_threads()`, which configures both OpenMP and MKL thread pools. The DataLoader was set to `num_workers=0` to isolate intra-operator parallelism from data loading parallelism. Batch size was fixed at 32 with a maximum of 15 epochs per configuration.

2.4.2. GPU CUDA Experiment

GPU training utilized PyTorch's CUDA backend [35], which offloads convolution and linear algebra operations to the cuDNN [19] and cuBLAS libraries running on NVIDIA's CUDA Toolkit 12.0. All model parameters and input tensors were transferred to the GPU via `model.cuda()` and automatic device placement; the cuDNN auto-tuner was enabled to select the fastest kernel for each layer configuration. GPU training used 2 CPU threads (fixed) for the DataLoader. The batch size was 32, and training ran for up to 30 epochs.

2.5. Evaluation Metrics

2.5.1. Classification Metrics

The following standard metrics were computed on the held-out test set to assess model classification quality. Accuracy measures the fraction of correctly classified samples over the total number of samples:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Cross-entropy loss quantifies the discrepancy between predicted probability distributions and ground-truth class labels. For a C-class problem with class-imbalance weighting, the loss for a batch of N samples is:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \; y_{ic} \log \hat{p}_{ic} \quad (2)$$

where w_c is the class weight inversely proportional to class frequency, $y_{ic} \in \{0,1\}$ is the one-hot ground-truth indicator, and $\hat{p}_{ic}$ is the predicted softmax probability for class c. Precision and recall for class c are:

$$\text{Precision}_c = \frac{TP_c}{TP_c+FP_c}, \qquad \text{Recall}_c = \frac{TP_c}{TP_c+FN_c} \quad (3)$$

F1-score is the harmonic mean of precision and recall, providing a balanced measure under class imbalance:

$$F1_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \quad (4)$$

The macro-averaged F1-score averages $F1_c$ uniformly across all C classes regardless of class frequency. Per-class scores are also reported to assess performance on the minority class (*partial*: 224 samples out of 3,000).

2.5.2. HPC Performance Metrics

The following metrics were collected to quantify training efficiency: average epoch time (seconds), training throughput (images/second), speedup $S(p) = T(1)/T(p)$, parallel efficiency $E(p) = S(p)/p$, CPU/GPU utilization (%), peak memory usage (RSS for CPU, `nvidia-smi` for GPU), and average GPU power draw (watts). All reported timings represent the mean across the steady-state epochs of a single training run per configuration, with the first epoch excluded to remove kernel auto-tuning and cache warm-up effects. Per-epoch variability within each run is reported via box plots in Section 3.1.5. We acknowledge that single-run measurements provide point estimates without across-run confidence intervals, and that multi-replicate runs would strengthen the statistical robustness of the speedup estimates; this remains a limitation of the present study.

A note on metric definitions is warranted. CPU utilization is reported as a fraction of *total system cores* (112 on the orfoz node), so a single fully-utilized thread corresponds to approximately $1/112 \approx 0.9\%$. This explains the 0.9% value reported for ResNet-50 at 1T, where one thread is fully active but the remaining 111 cores are idle. Peak RSS memory for CPU is sampled via `psutil` at maximum observed value during training, including PyTorch framework overhead, shared libraries, and per-thread workspace buffers, but excluding kernel page cache. GPU memory is reported as the peak allocated value via `nvidia-smi`.

2.5.3. Performance Modeling: Amdahl's Law

Amdahl's Law [31] provides a fundamental bound on parallel speedup. For a program with serial fraction f, the maximum speedup with p processors is:

$$S(p) = \frac{1}{f + \frac{1-f}{p}} \tag{5}$$

This implies an upper bound of $S_{\max} = 1/f$ regardless of processor count. Hill and Marty [32] extended this to multicore architectures, showing that even small serial fractions significantly limit scaling on many-core processors. Gustafson [36] proposed an alternative scaling law where the problem size grows with processor count, yielding more optimistic predictions for data-parallel workloads. Ben-Nun and Hoefler [23] identified DL-specific serial bottlenecks (batch normalization, gradient synchronization, and memory allocation) that increase the effective serial fraction. The serial fraction f is estimated from observed speedup by rearranging Eq. (5):

$$f = \frac{\frac{1}{S(p)} - \frac{1}{p}}{1 - \frac{1}{p}} \tag{6}$$

where $S(p)$ is the observed speedup at p threads.

3. Results and Analysis

In this section, the experimental findings of the study are presented and analyzed in detail. First, CPU multi-threaded scaling behavior is examined, covering training duration and throughput, validation of Amdahl's Law, the effects of NUMA topology, memory scaling, and epoch duration variability. Next, GPU training performance on the NVIDIA V100 architecture is evaluated based on metrics such as throughput, resource utilization, and power consumption. These analyses are followed by a direct CPU–GPU comparison that quantifies the performance gain (speedup) provided by GPU acceleration compared to the best CPU configurations. In the final stage, the classification quality obtained across all 18 tested model-hardware combinations (8 CPU thread counts $\times$ 2 models + 1 GPU $\times$ 2 models) is examined. The fact that both models achieve an accuracy of ≥ 0.997 and a macro-averaged F1 score of ≥ 0.995 across all configurations (a significant improvement over the F1 scores of 0.99 for VGG-19 and 0.91 for ResNet-50 reported in our previous work [10]) confirms that the choice of training platform in deep learning is a purely hardware-based decision driven by computational efficiency, with no impact on model quality [15].

3.1. CPU Multi-threaded Scaling

This subsection examines how intra-operator thread count affects training performance on the dual-socket Xeon Platinum 8480+ node. Training time, throughput, and speedup are presented first, followed by an Amdahl's Law analysis of the serial fraction, an investigation of cross-socket NUMA effects at 112 threads, memory scaling trends, and epoch-time variability.

3.1.1. Training Time and Throughput

Comprehensive CPU scaling results for the ResNet-50 and VGG-19 models are presented in Table 3 and Table 4, respectively. Although both models were trained on the same hardware, dataset, and hyperparameters, their scaling behaviors differ fundamentally. When the number of threads is increased from 1 to the single-socket limit of 56, ResNet-50's epoch time decreases from 894.2 seconds to 572.4 seconds, recording only a modest 36% reduction. VGG-19's epoch time drops from 10,052 seconds to 356.7 seconds, achieving a 96% reduction and a striking 28-fold (28 $\times$) speedup. This significant disparity directly reflects the architectural differences discussed in Section 2.2: VGG-19's massive FC weight matrices (512×4096 and 4096×4096) align efficiently with multi-threaded BLAS operations, whereas ResNet-50's numerous small residual operations cannot fully saturate the available threads.

These contrasting behaviors are clearly illustrated in Figure 4. In panel (a), where the epoch time is plotted on a logarithmic scale against the number of threads, VGG-19 exhibits a nearly linear and steep decline from 1 to 56 threads, ResNet-50 exhibits a much shallower decline profile, and the curve's flattening after 4 threads indicates that the parallelizable portion has reached early saturation. On the other hand, due to cross-socket NUMA overhead, both models experience a performance drop at the 112-thread level (where the curve breaks upward). This pattern is confirmed in panel (b), which shows the corresponding throughput (images/second) scaling: While VGG-19's throughput increases 28-fold from 0.21 to 5.88 images/second, ResNet-50's throughput only increases from 2.35 to 3.67

images/second (a 1.56-fold improvement), highlighting just how limited the benefit of additional hardware resources is in architectures with a high serial fraction.

Table 3. ResNet-50 CPU multi-threaded training results

T	Epoch (s)	img/s	*S*	*E* (%)	CPU (%)	Mem (MB)	F1
1	894.2	2.35	1.00	100.0	0.9	3999	1.000
2	740.1	2.83	1.21	60.4	1.8	4745	1.000
4	654.5	3.21	1.37	34.2	3.6	4829	1.000
7	618.7	3.39	1.45	20.6	6.2	5195	1.000
14	584.3	3.59	1.53	10.9	12.4	5220	1.000
28	572.6	3.66	1.56	5.6	24.7	5755	1.000
56	572.4	3.67	1.56	2.8	49.3	5926	0.995
112	617.7	3.40	1.45	1.3	97.3	6294	1.000

Table 4. VGG-19 CPU multi-threaded training results

T	Epoch (s)	img/s	*S*	*E* (%)	CPU (%)	Mem (MB)	F1
1	10052	0.21	1.00	100.0	47.4	5396	1.000
2	4012	0.52	2.51	125.3	16.3	5474	1.000
4	2753	0.76	3.65	91.3	18.4	7550	1.000
7	1285	1.63	7.82	111.7	6.8	5519	0.995
14	1165	1.80	8.63	61.7	13.0	7753	1.000
28	614.5	3.41	16.36	58.4	26.0	10060	0.995
56	356.7	5.88	28.18	50.3	51.6	12410	1.000
112	555.3	3.78	18.10	16.2	97.2	12410	1.000

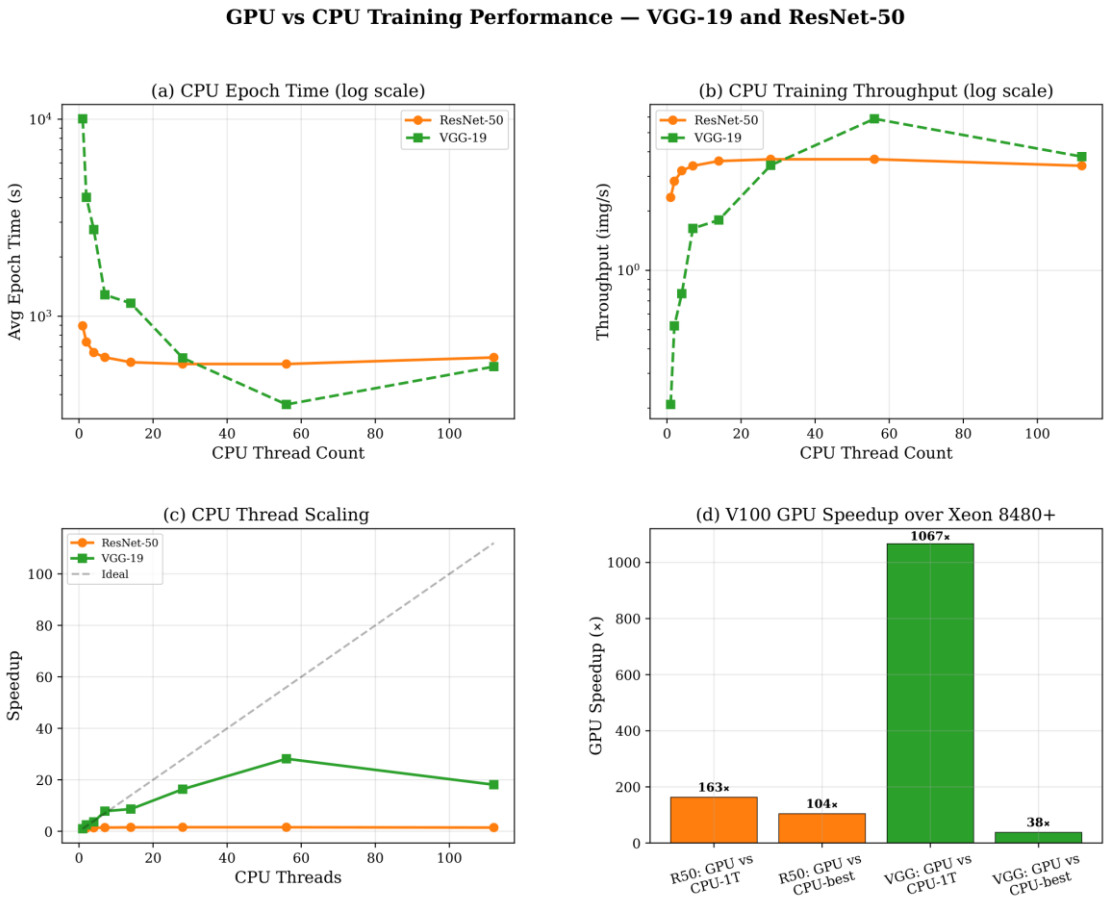


Fig. 4. Combined performance summary: (a) epoch time vs. thread count, (b) throughput scaling, (c) speedup with Amdahl curves, and (d) GPU speedup ratios.

3.1.2. Amdahl's Law Analysis

The serial fractions, calculated as $f \approx 0.634$ for ResNet-50 and $f \approx 0.018$ for VGG-19, differ by roughly 35-fold and directly explain the contrasting scaling behavior of the two models. Note that this 35× ratio in serial fractions translates to an 18× ratio in observed maximum speedup (28.2× vs. 1.56×); both quantities reflect the same underlying architectural difference but through different mathematical lenses. According to Amdahl's Law, ResNet-50's theoretical maximum speedup is $1/0.634 \approx 1.58\times$, while VGG-19's theoretical upper limit is calculated as $1/0.018 \approx 55.7\times$. The experimentally observed maximum acceleration values of $1.56\times$ and $28.2\times$ indicate that ResNet-50 has effectively reached the Amdahl limit, while the gap between VGG-19's observed 28.2× speedup and its 55.7× theoretical ceiling suggests that, on a hypothetical larger single-socket processor or a NUMA-optimized dual-socket implementation, further scaling would be possible. With current cross-NUMA penalties, however, 56 threads represents the practical optimum on this platform.

It should be noted that the estimated serial fractions encompass not only intrinsic architectural sequential dependencies (e.g., batch normalization, residual block ordering) but also pipeline-level serial work including data loading (with `num_workers=0`), validation passes, and PyTorch framework overhead. The reported f values therefore represent an effective serial fraction at the training-loop level rather than a pure architectural property. This effect is most pronounced for ResNet-50, whose smaller per-batch compute makes data-loading overhead a proportionally larger fraction of epoch time, which may inflate the apparent serial fraction beyond what is attributable to architecture alone. Disentangling these contributions through finer-grained profiling is left for future work.

The overlap between the observed acceleration values and the theoretical Amdahl curves adapted to the estimated scaling rates is presented in Figure 4 (c). In this graph, the two models exhibit strikingly different profiles: VGG-19's curve shows a continuous and steep rise across 2, 7, 14, 28, and 56 threads (T), closely matching the predicted Amdahl estimate for $f = 0.018$; and even exhibits super-linear acceleration at the 2T ($S = 2.51$, 125% efficiency) and 7T ($S = 7.82$, 112% efficiency) levels before converging to the theoretical slope at high thread counts. This observed super-linear behavior is consistent with cumulative cache effects: as the number of threads increases, the combined L2/L3 cache capacity grows proportionally, allowing more of VGG-19's massive FC weight matrices to be cached, thereby reducing main memory access frequency. We note, however, that direct verification of this explanation would require cache hit-rate measurements (e.g., via `perf` hardware counters), which were not collected in this study. In contrast, ResNet-50's curve flattens almost immediately, reaching $S = 1.37$ at the 4T level, and then plateaus around $S = 1.56$, strictly adhering to the Amdahl ceiling of $1/0.634 \approx 1.58$ starting from 28 threads. This rapid saturation point reflects the dominance of serial operations such as batch normalization, element-wise additions, and dependencies within sequential residual blocks that cannot benefit from additional threads. Consequently, the dramatic divergence between these two scaling profiles, despite using identical hardware, datasets, and training protocols, demonstrates that the model architecture is the primary determinant of CPU parallelization efficiency. The implications of this finding for architecture-aware HPC task scheduling will form the focus of future research.

3.1.3. Non-Uniform Memory Access (NUMA) Effects

As described in Section 2.3, the CPU platform comprises a dual-socket Intel Xeon Platinum 8480+ with 8 NUMA domains (4 per socket) interconnected via UPI. At 56 threads, execution is confined to a single socket where all memory accesses are local; at 112 threads, the workload spans both sockets, introducing remote memory accesses that traverse the Ultra Path Interconnect (UPI) interconnect with significantly higher latency.

This cross-socket penalty manifests differently for the two architectures. ResNet-50 speedup drops from 1.56× (56T) to 1.45× (112T), a 7% loss. This modest degradation occurs because ResNet-50's many small operations generate relatively little data movement per operation, limiting the impact of increased memory latency. VGG-19, by contrast, drops from 28.2× to 18.1×, a 36% loss that eliminates over one-third of the peak speedup. This disproportionate penalty arises because VGG-19's FC layers require frequent access to large weight matrices (512×4096, 4096×4096); when threads on the remote socket must fetch these buffers across UPI, the resulting latency directly increases the per-operation execution time of the dominant BLAS calls. As shown in Section 3.1.4, VGG-19's peak RSS memory also reaches 12.4 GB at 112T (130% growth from 1T), further increasing the fraction of data that resides in remote NUMA domains.

These results establish 56 threads (matching the single-socket physical core count) as the optimal configuration. Allocating threads beyond the socket boundary not only fails to improve performance but actively degrades it, an important practical guideline for HPC job scheduling on multi-socket nodes.

It should also be acknowledged that the cross-socket performance penalty observed at 112 threads may be compounded by thermal effects. On the Xeon Platinum 8480+, sustained dense floating-point computation across many cores can trigger AVX-512 frequency reduction or thermal throttling, particularly at full dual-socket utilization. Disentangling NUMA-induced memory latency from thermal-induced frequency reduction would require per-core frequency monitoring (e.g., via `turbostat` or hardware performance counters), which was not collected in this study. The 36% performance loss for VGG-19 at 112 threads should therefore be understood as a combined effect of these two mechanisms, with NUMA effects likely dominant given VGG-19's heavy reliance on large weight-matrix accesses.

3.1.4. Memory Scaling

In addition to execution time, the multi-threaded training process directly affects memory consumption because each OpenMP/MKL thread allocates private buffers for

intermediate computation results. According to the trend visualized in Figure 5, ResNet-50's peak RSS memory is 4.0 GB with 1 thread (1T) and rises to 6.3 GB with 112 threads (112T), showing a 57% increase, while VGG-19's memory consumption jumps from 5.4 GB to 12.4 GB, growing by 130% and more than doubling its single-threaded memory footprint. We note that VGG-19's peak RSS values are identical at 56T and 112T (12,410 MB), suggesting that the reported peak may reflect a memory-allocator pool ceiling or sampling resolution limit rather than precise working-set measurement. The relative growth trend, however, is consistent with the architectural differences between the two models. This imbalance is fully consistent with the architectural features discussed throughout this section: VGG-19's FC layers contain massive weight matrices of 512×4096 and 4096×4096 elements, and each requires per-thread workspace copies during multi-threaded GEMM (Generalized Matrix Multiplication) operations, whereas ResNet-50's smaller convolutional kernels impose proportionally much less overhead. This peak value of 12.4 GB, achieved by VGG-19 on 112 threads, is particularly noteworthy given the NUMA effects analyzed in Section 3.1.3. As the total memory footprint grows, an increasingly larger portion of the data is inevitably located in distant NUMA domains, a situation that multiplies the cross-socket latency penalty and directly contributes to the 36% performance loss observed with 112 threads.

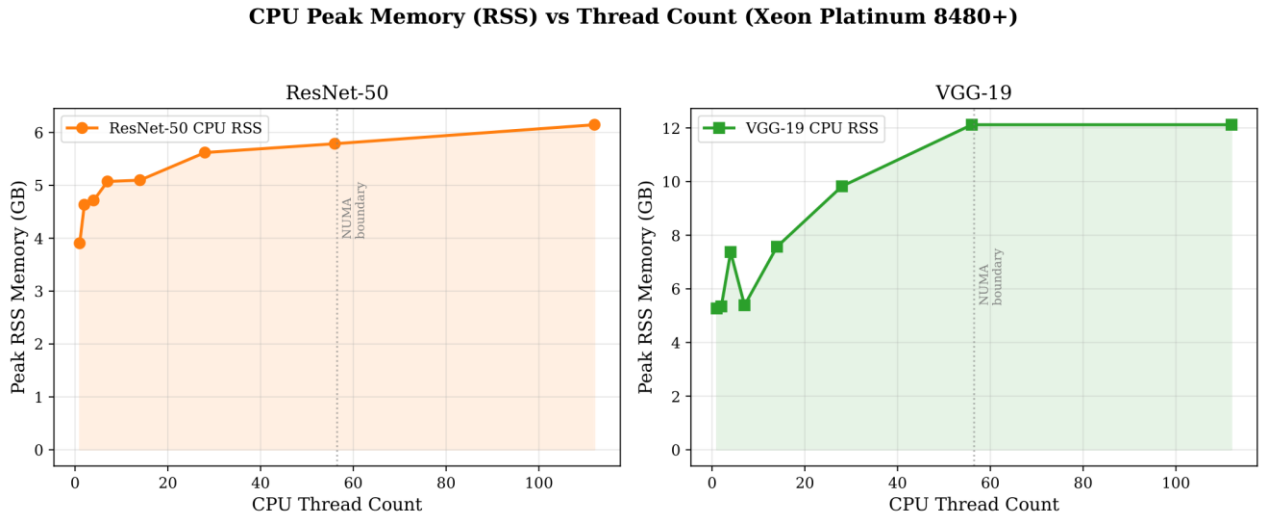


Fig. 5. CPU peak RSS memory vs. thread count. VGG-19 exhibits 130% memory growth from 1T to 112T due to large FC layer buffers, while ResNet-50 grows 57%.

3.1.5. Epoch Time Variability

Beyond average performance, the consistency of training duration across epochs is also of great importance for workload scheduling on shared HPC clusters. Figure 6 presents the training times per epoch across all thread configurations using box plots. It is observed that VGG-19 exhibits a notably higher relative variability at low thread counts (1T–4T); this can be attributed to the much larger cumulative impact of system-level external factors (such as operating system scheduling or background I/O operations on shared nodes) on the long absolute epoch durations (2,753–10,052 s). When the number of threads was increased to 56 (56T), both models were observed to converge to stable, low-variance epoch durations (356.7 s for VGG-19 and 572.4 s for ResNet-50). This convergence indicates that the parallelized workload is utilizing (saturating) the available cores at full capacity and has become less sensitive to external system interventions. On the other hand, in all tested configurations, the first epoch consistently runs slower than subsequent epochs due to one-time kernel auto-tuning in oneDNN/MKL and CPU cache warm-up effects. However, this temporary cost is quickly amortized over the remaining periods and does not affect steady-state transaction volume measurements.

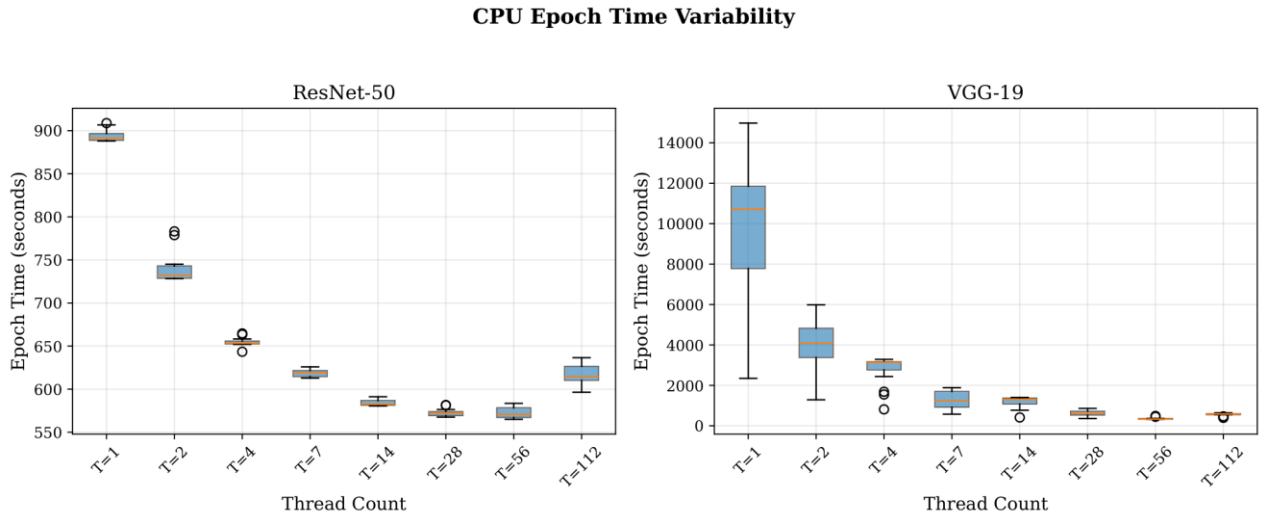


Fig. 6. CPU epoch time variability across thread configurations, shown as box plots for ResNet-50 and VGG-19.

3.2. GPU Training Performance

Having established the CPU scaling behavior and its architectural limitations in Section 3.1, this subsection evaluates training performance on the NVIDIA V100 GPU, where the parallelization paradigm shifts from thread-level BLAS concurrency to warp-level Single Instruction, Multiple Threads (SIMT) execution across 5,120 CUDA cores. Table 5 summarizes the results. ResNet-50 achieves 382.7 img/s (5.48 s/epoch), while VGG-19 reaches 222.7 img/s (9.42 s/epoch), meaning ResNet-50 attains 72% higher throughput than VGG-19 on the GPU. This reversal of the CPU scaling pattern, where VGG-19 consistently outperformed ResNet-50 in multi-threaded scaling, reflects an intrinsic difference in how each architecture maps onto GPU hardware: ResNet-50's many small independent operations can be efficiently pipelined across the V100's streaming multiprocessors, whereas VGG-19's large FC layers, though highly parallelizable on CPU via BLAS, present fewer but larger kernel launches that leave some GPU resources idle between successive layers. Both models achieve perfect classification (accuracy and F1-score of 1.000), confirming that GPU training introduces no quality degradation.

Table 5. GPU training performance (NVIDIA V100-SXM2-16GB)

Model	Epoch (s)	img/s	GPU (%)	Mem (MB)	Power (W)	Acc	F1
ResNet-50	5.48	382.7	91.6	5335	232.4	1.000	1.000
VGG-19	9.42	222.7	94.3	5859	196.8	1.000	1.000

3.2.1. GPU Resource Utilization

Fig. 7 provides a detailed breakdown of GPU resource utilization during training. Both models achieve high compute utilization (91.6% for ResNet-50 and 94.3% for VGG-19), indicating that neither model is memory-bound or I/O-bound; the GPU's computational units are effectively saturated throughout the training loop. Peak GPU memory usage ranges from 5,335 MB (ResNet-50) to 5,859 MB (VGG-19), consuming only approximately 33% to 37% of the V100's 16 GB HBM2 capacity. This substantial headroom suggests that increasing the batch size beyond 32 could further improve throughput by better amortizing kernel launch overhead, a direction to be explored in future work. It should also be noted that all GPU experiments were performed in single-precision (FP32) mode without automatic mixed precision (AMP). Consequently, the V100's 640 tensor cores, which accelerate FP16/FP32 mixed-precision matrix multiplications, were not utilized. Enabling AMP could yield substantial additional speedup (typically 1.5–3× for CNN training on V100), and represents a further optimization opportunity beyond the present comparison. In terms of power consumption, ResNet-50 draws an average of 232.4 W (77.5% of the 300 W TDP), compared to 196.8 W (65.6%) for VGG-19. The higher power draw of ResNet-50 is consistent with its greater number of active kernel launches per forward/backward pass, which keeps more streaming multiprocessors simultaneously busy and increases dynamic power dissipation.

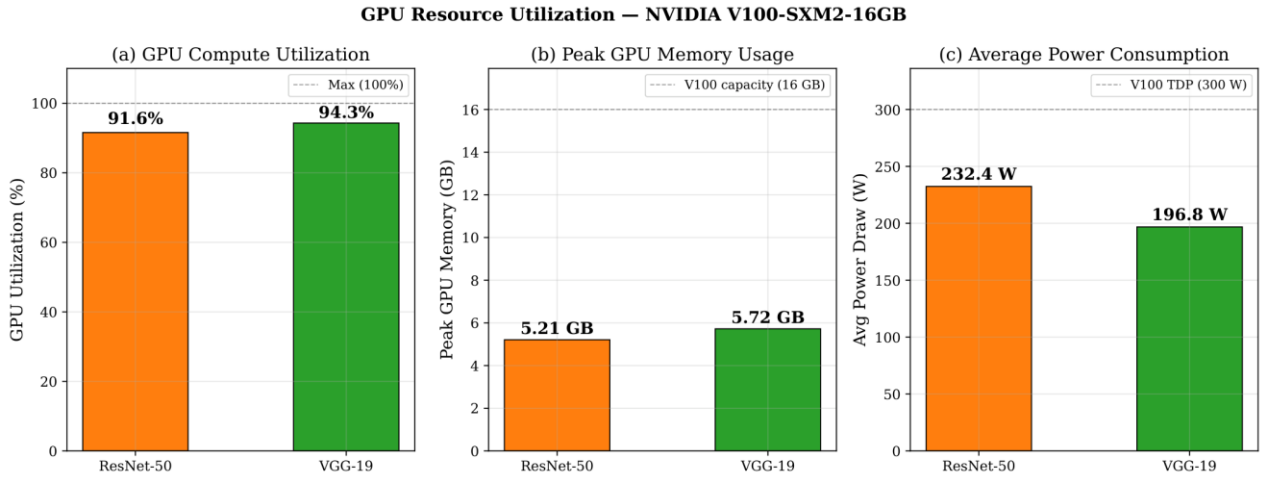


Fig. 7. GPU resource utilization: (a) compute utilization, (b) peak memory usage relative to V100 capacity, (c) average power draw relative to TDP (300 W).

3.3. CPU vs. GPU Comparative Analysis

Sections 3.1 and 3.2 have individually characterized CPU multi-threaded and GPU training performance; this subsection directly contrasts the two execution paradigms to quantify the practical advantage of GPU acceleration. Table 6 and Fig. 8 summarize the comparison. Relative to single-threaded CPU execution, GPU training yields a 163× speedup for ResNet-50 and 1,067× for VGG-19. Even when compared against the best CPU configuration (56 threads), the GPU maintains speedup factors of 104× and 38×, respectively, demonstrating that thread-level CPU parallelism, while beneficial, cannot match the throughput of massively parallel SIMT execution.

Table 6. GPU vs. CPU speedup comparison

Comparison	CPU (s)	GPU (s)	Speedup
ResNet-50: GPU vs. CPU-1T	894.2	5.48	163×
ResNet-50: GPU vs. CPU-56T	572.4	5.48	104×
VGG-19: GPU vs. CPU-1T	10052	9.42	1067×
VGG-19: GPU vs. CPU-56T	356.7	9.42	38×

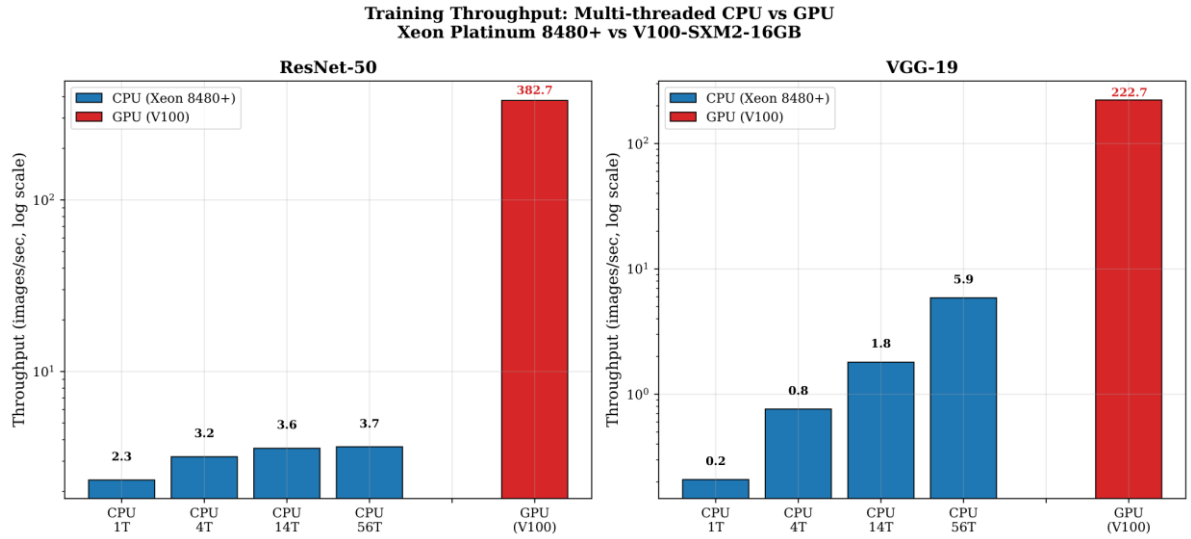


Fig. 8. Training throughput comparison on logarithmic scale: selected CPU configurations (1T, 4T, 14T, 56T) versus GPU (V100) for both models.

A notable inverse relationship emerges between CPU scaling effectiveness and GPU speedup magnitude. VGG-19, which achieves strong CPU scaling (28× at 56 threads), still exhibits a 38× GPU advantage at the best CPU configuration; ResNet-50, which scales poorly on CPU (1.56×), retains a 104× GPU advantage. In absolute terms, GPU throughput exceeds the best CPU throughput by two orders of magnitude for both architectures. Fig. 4 panel (d) presents the GPU-to-CPU speedup ratios as a grouped bar chart for both models against the single-threaded (1T) and best multi-threaded (56T)

CPU baselines. For VGG-19, the speedup drops from 1,067× (vs. 1T) to 38× (vs. 56T), reflecting VGG-19's strong CPU scaling, which substantially narrows the GPU gap. For ResNet-50, the reduction is more modest, from 163× to 104×, because poor CPU scaling leaves the GPU advantage largely intact. This asymmetry highlights a practical trade-off: for architectures with large, BLAS-friendly operations such as VGG-19, investing in multi-core CPU resources can partially offset the absence of a GPU, whereas for architectures dominated by small, sequentially dependent operations such as ResNet-50, GPU acceleration remains the only viable path to significant training speedup.

3.4. Classification Quality

While the preceding subsections have focused on computational performance, it is equally important to verify that hardware acceleration does not compromise model accuracy. Fig. 9 presents the training convergence curves across all configurations. All 18 configurations (16 CPU thread counts + 2 GPU) achieve accuracy ≥ 0.997 and F1-score ≥ 0.995. GPU training achieves perfect classification with the lowest test loss values (ResNet-50: 1.75×10^{-5}, VGG-19: 1.79×10^{-6}), indicating more stable convergence and tighter decision boundaries.

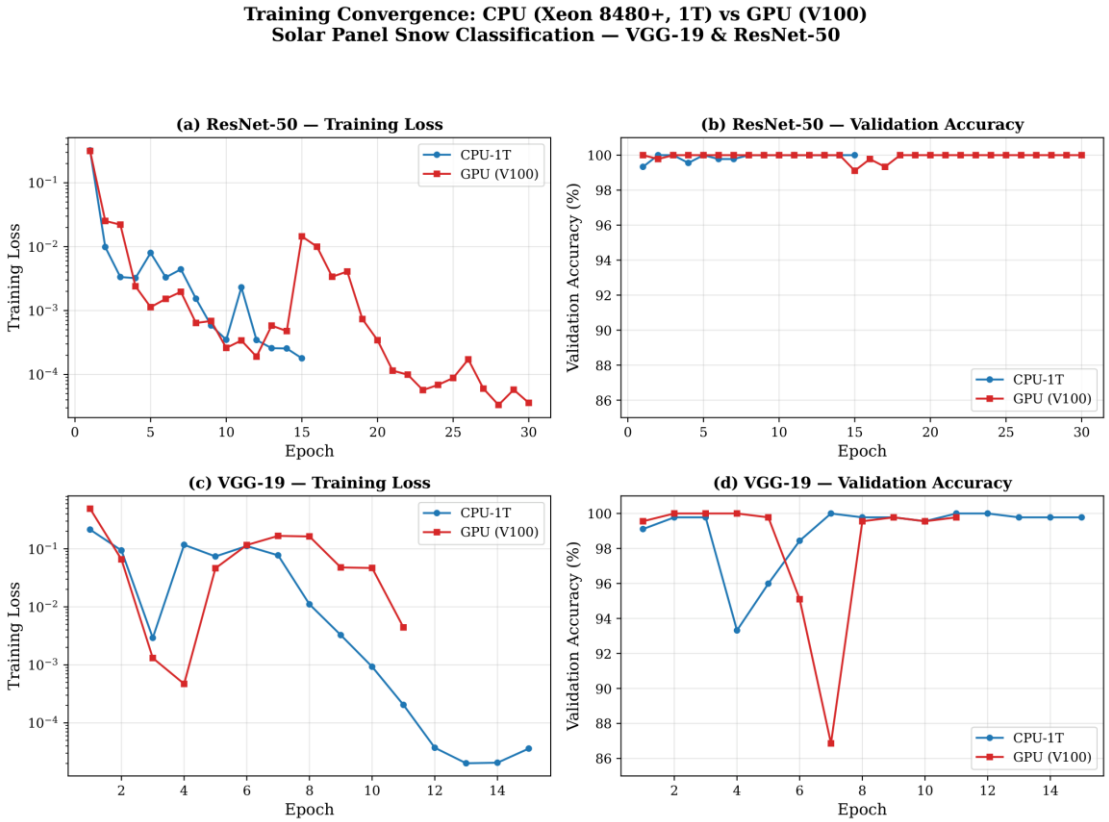


Fig. 9. Training convergence curves: loss and validation accuracy for selected CPU configurations and GPU, showing consistent convergence across all hardware platforms.

These results confirm that classification quality is invariant with respect to hardware configuration: the choice between CPU and GPU execution, and the number of CPU threads, affects only training speed, not model accuracy. This observation is consistent with the theoretical expectation that deterministic model architectures trained with identical hyperparameters converge to equivalent solutions regardless of execution platform, with minor stochastic variations attributable to floating-point non-associativity in parallel reductions.

4. Discussion

The experimental results presented in this study reveal that model architecture is the dominant factor determining CPU parallelization efficiency, a finding that carries both theoretical and practical significance. The 18× disparity in CPU scaling between VGG-19 ($S_{56} = 28.2\times$, $f \approx 1.8\%$) and ResNet-50 ($S_{56} = 1.56\times$, $f \approx 63.4\%$) arises directly from intrinsic differences in computational structure. VGG-19's FC layers involve large matrix multiplications (512×4096, 4096×4096) that map efficiently onto multi-threaded BLAS operations, enabling a single GEMM call to saturate all available threads. ResNet-50, by contrast, distributes its computation across many small operations (1×1 convolutions, batch normalization, and element-wise additions) with sequential dependencies between residual blocks. Each operation is too small individually to fully utilize many threads, and the cumulative per-operation thread dispatch and synchronization overhead yields a serial fraction exceeding 63%. The close agreement between observed speedup values and Amdahl's Law predictions (Section 3.1.2) confirms that this serial fraction reflects an architectural property rather than a software implementation artifact,

providing a principled framework for estimating scaling behavior on similar CPU platforms. However, the specific value of f may vary across hardware generations and BLAS library implementations, since the serial work itself runs at different effective rates on systems with different memory bandwidth, cache hierarchy, and instruction set characteristics. The qualitative conclusion (architecture-dependent scaling behavior) is expected to generalize broadly; the precise quantitative ceiling should be re-measured for each target platform.

On the GPU, this scaling disparity largely disappears. Both models achieve high compute utilization (91–94%) on the NVIDIA V100, and ResNet-50, the slower model on CPU, actually attains 72% higher throughput than VGG-19 on GPU (382.7 vs. 222.7 img/s). This reversal reflects an intrinsic difference in execution paradigm: the GPU's 5,120 CUDA cores can efficiently pipeline ResNet-50's many small, independent operations through warp-level SIMT execution, whereas VGG-19's fewer but larger kernel launches leave some streaming multiprocessors idle between successive layers. This observation carries a critical practical implication: *model architecture determines whether CPU multi-threading is a viable acceleration strategy, while GPU acceleration is effectively architecture-agnostic* for the workload sizes examined in this study. The GPU provides 163–1,067× speedup over single-threaded CPU and 38–104× over the best multi-threaded configuration, irrespective of architectural characteristics.

The CPU scaling analysis also reveals two noteworthy phenomena. First, VGG-19 exhibits super-linear parallel efficiency at 2 threads (125%) and 7 threads (112%), a phenomenon attributable to aggregate cache effects: as the thread count increases, the combined L2/L3 cache capacity grows proportionally, allowing a larger fraction of VGG-19's massive FC weight matrices to remain cache-resident and reducing main memory access frequency. This cache-driven benefit diminishes at higher thread counts as the parallelizable workload itself becomes the limiting factor. Second, the NUMA topology introduces a significant performance penalty when threading exceeds the single-socket boundary: VGG-19 loses 36% of its peak speedup at 112 threads, compounded by the 130% growth in peak RSS memory (from 5.4 GB to 12.4 GB) that increases the fraction of data residing in remote NUMA domains. These two effects, super-linear gains from cache aggregation at low thread counts and degradation from cross-socket latency at high thread counts, define the practical operating envelope for CPU-based training on multi-socket systems.

An equally important finding is the invariance of classification quality across all 18 tested hardware configurations. Both models achieve an accuracy ≥ 0.997 and a macro-averaged F1-score ≥ 0.995, with 15 of 18 configurations producing perfect classification. The three configurations exhibiting F1 = 0.995 (ResNet-50 at 56T, VGG-19 at 7T, and VGG-19 at 28T) each correspond to a single misclassified test image, where a *partial* sample is assigned to the *no_snow* class. Whether these three single-image errors arise from genuine sensitivity to thread count (e.g., differences in floating-point reduction order under varying parallelism), or simply from run-to-run stochastic variation that would average out across replicate runs, cannot be determined from single-run measurements. The variation is small enough (one image out of 453, a 0.22% absolute difference) to be consistent with either explanation, and does not materially affect the central claim of hardware-invariant quality. GPU training achieves the lowest test loss values (ResNet-50: 1.75×10^{-5}; VGG-19: 1.79×10^{-6}), though we caution that these extremely small loss values are consistent with the overfitting expected from the data leakage limitation discussed in Section 2.1, and should be interpreted as a measure of training-protocol consistency rather than as evidence of genuine generalization quality. Critically, this quality invariance confirms that the choice of training platform is a purely performance-driven decision with no impact on model accuracy, a conclusion consistent with the theoretical expectation that identical architectures trained with identical hyperparameters converge to equivalent solutions regardless of execution hardware [15].

4.1. Comparison with Existing Research and Novelty

To clearly position the contributions of this work, Table 7 provides a systematic comparison of our findings with closely related studies in the literature. While prior work has addressed individual aspects of CNN parallelization, no existing study has combined all of the dimensions investigated here (CPU thread scaling, GPU acceleration, architecture-dependent Amdahl analysis, NUMA effects, and classification quality invariance) within a single, controlled experimental framework on a production HPC platform.

CPU parallelization. Huang et al. [20] demonstrated a 30× improvement in GFLOPS using Winograd-based convolution with OpenMP SIMD on 64-core Intel processors; however, their analysis was limited to the convolution kernel level and did not evaluate end-to-end training performance or architecture-dependent scaling behavior. Datta et al. [21] confirmed that multi-core CPU parallelization improves CNN throughput using AlexNet and VGG-16, but did not quantify the serial fraction or apply formal performance modeling. Hangun and Bayar [22] achieved up to 38× speedup for OpenMP-based image enhancement on 48 physical cores, demonstrating effective CPU parallelization for preprocessing workloads. Our work extends this line of investigation to the DL training process itself and reveals that the achievable speedup varies by a factor of 18× depending on model architecture (VGG-19: 28× vs. ResNet-50: 1.56×). Critically, none of these studies applied Amdahl's Law to estimate architecture-specific serial fractions or validated the theoretical predictions against empirical measurements across a comprehensive range of thread counts.

GPU acceleration. Jordà et al. [26] developed cuConv, a custom CUDA convolution kernel achieving 2.29× speedup over cuDNN for specific layer configurations in ResNet-50 and Inception-v4; their work focused on layer-level kernel optimization rather than full training pipeline analysis. Dong et al. [27] reported a 5.9× GPU training speedup for hyperspectral image classification using CUDA multistream execution. In contrast, our end-to-end measurements reveal GPU speedups of 163–1,067× over single-threaded CPU

execution, reflecting the cumulative benefit across the entire forward–backward–update training loop rather than individual operations. Furthermore, our study uniquely quantifies the inverse relationship between CPU scaling effectiveness and GPU speedup magnitude: VGG-19, which scales well on CPU (28×), still exhibits a 38× GPU advantage at the best CPU configuration, while ResNet-50, which scales poorly on CPU (1.56×), retains a 104× GPU advantage, a finding absent from prior GPU benchmarking studies [28, 29].

Performance modeling. Hill and Marty [32] extended Amdahl's Law to multicore architectures and demonstrated that even small serial fractions significantly limit scaling on many-core processors. Ben-Nun and Hoefler [23] identified DL-specific serial bottlenecks including batch normalization synchronization and gradient aggregation, but did not empirically measure architecture-specific serial fractions on production HPC hardware. Our work provides, to the best of our knowledge, the first empirical validation of Amdahl's Law for DL training workloads, quantifying serial fractions of $f \approx 1.8\%$ for VGG-19 and $f \approx 63.4\%$ for ResNet-50, a 35× difference arising solely from architectural characteristics. The close agreement between observed speedup curves and theoretical Amdahl predictions (Fig. 4c) confirms that these serial fractions are intrinsic architectural properties, not artifacts of the software framework or hardware platform.

NUMA topology effects. While NUMA-aware scheduling has been studied extensively in general-purpose HPC contexts, its impact on DL training has rarely been empirically quantified. Our results provide concrete evidence that cross-socket memory access introduces architecture-dependent performance degradation: VGG-19 loses 36% of its peak speedup at 112 threads (spanning both sockets), whereas ResNet-50 loses only 7%, reflecting the different memory access patterns of large FC layers versus small convolutional kernels. This finding offers a practical guideline (thread allocation should not exceed the single-socket core count) that is absent from existing DL benchmarking studies.

Table 7. Comparison of this work with related studies

Study	Task	CPU scaling	GPU accel.	Amdahl valid.	NUMA effects	Quality invari.
Huang et al. [20]	Conv. kernel	✓	–	–	–	–
Datta et al. [21]	CNN throughput	✓	–	–	–	–
Hangun & Bayar [22]	Image enh.	✓	–	✓	–	–
Jordà et al. [26]	Conv. kernel	–	✓	–	–	–
Dong et al. [27]	HSI classif.	–	✓	–	–	–
Shi et al. [28]	DL benchmark	–	✓	–	–	–
Ben-Nun & Hoefler [23]	DL survey	✓	✓	Theor.	–	–
This work	**PV snow det.**	✓	✓	✓	✓	✓

In summary, the primary novelty of this work lies in three aspects that distinguish it from the existing literature. First, it provides a *unified experimental framework* that evaluates both CPU thread scaling (1–112 threads) and GPU acceleration (NVIDIA V100) for two architecturally distinct CNN models on the same HPC infrastructure under identical training protocols, an experimental design absent from prior studies that typically examine either CPU or GPU performance in isolation. Second, it delivers the *first empirical Amdahl's Law validation for DL training*, quantifying architecture-specific serial fractions and demonstrating that a single parameter (f) can predict the scaling ceiling of a given model architecture on arbitrary CPU hardware. Third, it establishes *hardware-invariant classification quality* across 18 configurations, providing rigorous evidence that training platform selection is purely a computational efficiency decision, a finding of direct practical relevance for field-deployed solar panel monitoring systems where periodic model retraining must balance computational cost against operational constraints. These findings provide concrete guidance for field-deployed solar panel monitoring systems that require periodic model retraining. In GPU-available environments, both architectures complete training in under 10 seconds per epoch, making architecture selection a quality-driven rather than performance-driven decision. In CPU-only environments, however, architecture selection becomes a critical deployment decision. VGG-19 can achieve acceptable retraining performance (356.7 s/epoch at 56 threads), making it a practical choice for periodic updates on multi-core servers, whereas ResNet-50 requires 572.4 s/epoch even at optimal threading (only 36% faster than single-threaded execution), rendering CPU-based retraining operationally impractical for time-sensitive applications. Furthermore, the NUMA analysis establishes that thread allocation should not exceed the single-socket core count (56 on this platform), as cross-socket scheduling degrades rather than improves throughput, an important guideline for HPC job schedulers managing shared multi-socket nodes.

5. Conclusion

This study presented a systematic comparative analysis of CNN training performance for solar panel snow detection on the TRUBA national HPC infrastructure, using VGG-19 and ResNet-50 as architecturally distinct benchmark models. The results demonstrate that model architecture is the primary determinant of CPU parallelization efficiency: VGG-19, whose computation is dominated by large FC weight matrices, achieves a 28× speedup at 56 threads with a serial fraction of $f \approx 1.8\%$, while ResNet-50, composed of many small residual operations with sequential dependencies, scales to only 1.56× ($f \approx 63.4\%$), an 18× disparity accurately predicted by Amdahl's Law. GPU acceleration on the NVIDIA V100 eliminates this architectural dependency, delivering 163–1,067× speedup over a single-threaded CPU and 38–104× over the best multi-threaded configuration for

both models. The NUMA topology analysis further reveals that thread allocation beyond the single-socket boundary is counterproductive, with VGG-19 losing 36% of its peak speedup at 112 threads due to cross-socket memory access penalties. Importantly, classification quality remains invariant across all 18 tested hardware configurations (accuracy $\geq$ 0.997, F1-score $\geq$ 0.995), confirming that training platform selection is purely a computational efficiency decision with no impact on model accuracy. Several directions for future work emerge from these findings. First, the high serial fraction observed in ResNet-50 ($f \approx 63.4\%$) warrants a detailed investigation into the specific sources of sequential bottlenecks (batch normalization synchronization, residual block dependencies, and small-kernel dispatch overhead) to determine whether architectural modifications such as operator fusion, asynchronous batch normalization, or grouped convolution restructuring can reduce the effective serial fraction and improve CPU scalability for residual architectures. More broadly, these insights motivate the design of CNN architectures that are inherently more amenable to CPU parallelization, balancing classification accuracy with computational structure to achieve efficient multi-threaded scaling without relying on GPU availability. Second, the GPU memory analysis reveals that both models utilize only 33–37% of the V100's 16 GB HBM2 capacity at batch size 32, suggesting that batch size scaling could further improve GPU throughput by better amortizing kernel launch overhead. Enabling automatic mixed precision (AMP) to engage the V100's tensor cores represents an additional opportunity for substantial GPU speedup. Third, future work should adopt a leak-free evaluation protocol with source-level data splitting to obtain unbiased estimates of generalization quality, which is necessary before drawing rigorous conclusions about model deployment readiness. Multi-replicate runs with reported confidence intervals would also strengthen the statistical reliability of the speedup measurements. Finally, while this study characterized *training* performance, deployment requires *inference* characterization, which differs substantially from training due to single-sample or small-batch operation, the absence of backward passes, and quantization opportunities. Extending this analysis to multi-GPU distributed training configurations and evaluating inference latency on edge devices (e.g., NVIDIA Jetson) would provide a complete performance characterization spanning from HPC training to real-time solar panel monitoring deployment.

Acknowledgements

The numerical calculations reported in this paper were fully performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

Author Contributions

Abdulkadir Turkan: conceptualization, methodology, software, investigation, data curation, formal analysis, writing – original draft. **Batuhan Hangun:** methodology, validation, visualization, writing – review and editing. **Onder Eyecioglu:** conceptualization, supervision, resources, project administration, writing – review and editing. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] International Energy Agency, "World energy outlook 2023." https://www.iea.org/reports/world-energy-outlook-2023, 2023. Last accessed: March 2026.

[2] International Renewable Energy Agency, "Renewable capacity statistics 2024." https://www.irena.org/publications/2024/Mar/Renewable-capacity-statistics-2024, 2024. Last accessed: March 2026.

[3] Z. Quan, H. Lu, C. Zheng, W. Zhao, Y. Xu, J. Qin, and F. An, "Experimental measurement and numerical simulation on the snow-cover process of solar photovoltaic modules and its impact on photoelectric conversion efficiency," *Coatings*, vol. 13, no. 2, p. 427, 2023, doi: 10.3390/coatings13020427

[4] R. W. Andrews, A. Pollard, and J. M. Pearce, "The effects of snowfall on solar photovoltaic performance," Solar Energy, vol. 92, pp. 84–97, 2013, doi: 10.1016/j.solener.2013.02.014

[5] R. E. Pawluk, Y. Chen, and Y. She, "Photovoltaic electricity generation loss due to snow – a literature review on influence factors, estimation, and mitigation," Renewable and Sustainable Energy Reviews, vol. 107, pp. 171–182, 2019, doi: 10.1016/j.rser.2018.12.031

[6] S. Deitsch, V. Christlein, S. Berger, C. Buerhop-Lutz, A. Maier, F. Gallwitz, and C. Riess, "Automatic classification of defective photovoltaic module cells in electroluminescence images," Solar Energy, vol. 185, pp. 455–468, 2019, doi: 10.1016/j.solener.2019.02.067

[7] M. W. Akram, G. Li, Y. Jin, X. Chen, C. Zhu, X. Zhao, A. Khaliq, M. Faheem, and A. Ahmad, "CNN based automatic detection of photovoltaic cell defects in electroluminescence images," Energy, vol. 189, p. 116319, 2019, doi: 10.1016/j.energy.2019.116319

[8] X. Li, Q. Yang, Z. Lou, and W. Yan, "Deep learning based module defect analysis for large-scale photovoltaic farms," IEEE Transactions on Energy Conversion, vol. 34, no. 1, pp. 520–529, 2019, doi: 10.1109/TEC.2018.2873358

[9] I. Kishor, U. Mamodiya, V. Patil, and N. Naik, "AI-integrated autonomous robotics for solar panel cleaning and predictive maintenance using drone and ground-based systems," Scientific Reports, vol. 15, no. 1, p. 32187, 2025, doi: 10.1038/s41598-025-17313-6

[10] O. Ozturk, B. Hangun, and O. Eyecioglu, "Detecting snow layer on solar panels using deep learning," in 2021 10th International Conference on Renewable Energy

Research and Application (ICRERA), pp. 434–438, 2021, doi: 10.1109/ICRERA52334.2021.9598700

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2015. Published at ICLR 2015, doi: 10.48550/arXiv.1409.1556

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90

[13] B. Shah and H. Bhavsar, "Time complexity in deep learning models," Procedia Computer Science, vol. 215, pp. 202–210, 2022, doi: 10.1016/j.procs.2022.12.023

[14] M. Taghavi and M. Shoaran, "Hardware complexity analysis of deep neural networks and decision tree ensembles for real-time neural data classification," in 2019 9th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 407–410, 2019, doi: 10.1109/NER.2019.8716983

[15] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2656–2666, 2019, doi: 10.1109/CVPR.2019.00277

[16] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1299–1312, 2016, doi: 10.1109/TMI.2016.2535302

[17] Intel Corporation, "oneAPI deep neural network library (oneDNN)." https://www.intel.com/content/www/us/en/developer/tools/oneapi/onednn.html, 2024. Last accessed: March 2026.

[18] NVIDIA Corporation, "CUDA toolkit documentation v12.0." https://docs.nvidia.com/cuda/, 2024. Last accessed: March 2026.

[19] NVIDIA Corporation, "cuDNN developer guide." https://docs.nvidia.com/deeplearning/cudnn/, 2024. Last accessed: March 2026.

[20] J. Huang, T. Wang, X. Zhu, M. Wei, T. Wu, X. Wu, and M. Huang, "A parallel optimization of the fast algorithm of convolution neural network on CPU," in 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), pp. 5–9, 2018.

doi: 10.1109/ICMTMA.2018.00009

[21] D. Datta, D. Mittal, N. P. Mathew, and J. Sairabanu, "Comparison of performance of parallel computation of CPU cores on CNN model," in 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1–8, 2020, doi: 10.1109/ic-ETITE47903.2020.142

[22] B. Hangun and S. Bayar, "An OpenMP-based parallel implementation of image enhancement technique for dark images," Signal, Image and Video Processing, vol. 18, no. 5, pp. 4115–4130, 2024, doi: 10.1007/s11760-024-03058-8

[23] T. Ben-Nun and T. Hoefler, "Demystifying parallel and distributed deep learning: An in-depth concurrency analysis," ACM Computing Surveys, vol. 52, no. 4, pp. 1–43, 2019, doi: 10.1145/3320060

[24] A. Saad Althiban, H. M. Alharbi, L. A. Al Khuzayem, and F. Elbouraey Eassa, "Automated models for predicting software defects in hybrid message passing interface (MPI) and open multi-processing (OpenMP) parallel programs using deep learning," IEEE Access, vol. 13, pp. 65939–65954, 2025, doi: 10.1109/ACCESS.2025.3559807

[25] B. Hangün and Ö. Eyecioğlu, "Performance comparison between OpenCV built-in CPU and GPU functions on image processing operations," International Journal of Engineering Science and Application, vol. 1, no. 2, pp. 34–41, 2017.

[26] M. Jordà, P. Valero-Lara, and A. J. Peña, "cuConv: CUDA implementation of convolution for CNN inference," Cluster Computing, vol. 25, no. 2, pp. 1459–1473, 2022, doi: 10.1007/s10586-021-03494-y

[27] H. Dong, T. Li, J. Leng, L. Kong, and G. Bai, "GCN: GPU-based cube CNN framework for hyperspectral image classification," in 2017 46th International Conference on Parallel Processing (ICPP), pp. 41–49, 2017, doi: 10.1109/ICPP.2017.13

[28] S. Shi, Q. Wang, P. Xu, and X. Chu, "Benchmarking state-of-the-art deep learning software tools," in Proceedings of the 7th IEEE International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), pp. 99–104, 2018, doi: 10.1109/ICCCBDA.2018.8386507

[29] C. Coleman, D. Narayanan, D. Kang, T. Zhao, J. Zhang, L. Nardi, P. Bailis, K. Olukotun, C. Ré, and M. Zaharia, "DAWNBench: An end-to-end deep learning benchmark and competition," in NeurIPS ML Systems Workshop, 2017.

[30] S. Markidis, S. W. D. Chien, E. Laure, I. B. Peng, and J. S. Vetter, "NVIDIA tensor core programmability, performance and precision," in IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), pp. 522–531, 2018, doi: 10.1109/IPDPSW.2018.00091

[31] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," AFIPS Conference Proceedings, vol. 30, pp. 483–485, 1967, doi: 10.1145/1465482.1465560

[32] M. D. Hill and M. R. Marty, "Amdahl's law in the multicore era," IEEE Computer, vol. 41, no. 7, pp. 33–38, 2008, doi: 10.1109/MC.2008.209

[33] TUBITAK ULAKBIM, "TRUBA high performance computing center." https://www.truba.gov.tr/, 2024. Last accessed: March 2026.

[34] J. Y. Lin, C. Ograda-Bratu, S. Huo, and S. Keshav, "Detecting snow covered solar panels dataset." Scholars Portal Dataverse.doi: 10.5683/SP2/ITQPHZ

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems 32 (NeurIPS), pp. 8024–8035, 2019, doi: 10.48550/arXiv.1912.01703

[36] J. L. Gustafson, "Reevaluating Amdahl's law," Communications of the ACM, vol. 31, no. 5, pp. 532–533, 1988, doi: 10.1145/42411.42415

International Journal of Smart Grid (IJSmartGrid)
https://www.ijsmartgrid.org/ Online ISSN: 2602-439X
N. Sharma et.al., Vol. 10, No. 3, September, 2026 Research Article

Optimal PMU Placement Techniques: A Detailed Review and Case Study on the Indian Utility 62 Bus System

Nisha Sharma*, Priti Prabhakar**, Avnesh Verma***

*Department of Electrical and Electronics Engineering, PhD Scholar, GJUS&T, Hisar, Haryana, India

** Department of Electrical and Electronics Engineering, Associate Professor, GJUS&T, Hisar, Haryana, India

***Department of Instrumentation and Control, Professor, KU, Kurukshetra, Haryana, India

(nisha.ee@gjust.org, priti.prabhakar@gmail.com, verma.avneesh@gmail.com)

‡Corresponding Author: Nisha Sharma, GJUS&T, nisha.ee@gjust.org

Received: 15.10.2025 Accepted: 21.05.2026

Abstract – Optimal placement of Phasor Measurement Unit (PMU) is important for power system observability and reliable monitoring. The existing placement approaches emphasize full observability with minimum PMUs to achieve a cost-effective solution; however, the direct observability of critical buses and redundancy under contingency conditions has received limited attention. Consequently, minimum-PMU solutions may become vulnerable under PMU or transmission line outage conditions. In this paper, an attempt has been made to improve the system redundancy, security and reliability along with detailed, comprehensive review of existing Optimal PMU Placement (OPP) methods. A comparative analysis based on cost effectiveness, reliability, critical bus observability, system redundancy, and overall system security has also been carried out. A novel Depth of Redundancy (DOR) metric is introduced to quantify bus-level redundancy, and the System Observability Redundancy Index (SORI) is utilized to evaluate overall system redundancy. The framework is validated using IEEE 30 and 57 bus systems and implemented on the Indian Utility 62 bus system as a detailed case study. The results demonstrate that Integer Linear Programming (ILP) achieves cost-effective observability solution, whereas Depth First Search (DFS) achieves the highest observability redundancy and strong contingency resilience with a greater number of PMUs. Simulated Annealing (SA), however, achieves comparatively higher SORI and DOR values, thereby providing a balanced trade-off between cost-effectiveness, redundancy and monitoring. The study highlights that redundancy-oriented PMU placement significantly enhances monitoring reliability and contingency preparedness for practical utility-scale applications.

Keywords Phasor measurement units (PMUs), optimal PMU placement (OPP), system observability redundancy index (SORI), synchrophasor technology.

Abstract – Optimal placement of Phasor Measurement Unit (PMU) is important for power system observability and reliable monitoring. The existing placement approaches emphasize full observability with minimum PMUs to achieve a cost-effective

Cite this article as: N. Sharma, P. Prabhakar, and A. Verma, "Optimal PMU Placement Techniques: A Detailed Review and Case Study on the Indian Utility 62 Bus System" *International Journal of Smart Grid (ijSmartGrid)*, Vol. 10, No. 3, pp. 209-224, September, 2026.

solution; however, the direct observability of critical buses and redundancy under contingency conditions has received limited attention. Consequently, minimum-PMU solutions may become vulnerable under PMU or transmission line outage conditions. In this paper, an attempt has been made to improve the system redundancy, security and reliability along with detailed, comprehensive review of existing Optimal PMU Placement (OPP) methods. A comparative analysis based on cost effectiveness, reliability, critical bus observability, system redundancy, and overall system security has also been carried out. A novel Depth of Redundancy (DOR) metric is introduced to quantify bus-level redundancy, and the System Observability Redundancy Index (SORI) is utilized to evaluate overall system redundancy. The framework is validated using IEEE 30 and 57 bus systems and implemented on the Indian Utility 62 bus system as a detailed case study. The results demonstrate that Integer Linear Programming (ILP) achieves cost-effective observability solution, whereas Depth First Search (DFS) achieves the highest observability redundancy and strong contingency resilience with a greater number of PMUs. Simulated Annealing (SA), however, achieves comparatively higher SORI and DOR values, thereby providing a balanced trade-off between cost-effectiveness, redundancy and monitoring. The study highlights that redundancy-oriented PMU placement significantly enhances monitoring reliability and contingency preparedness for practical utility-scale applications.
Keywords: Phasor Measurement Units (PMUs), Optimal PMU Placement (OPP), System Observability Redundancy Index (SORI), Synchrophasor Technology.

1. Introduction

Phasor Measurement Units (PMUs) provide real-time, high-resolution, and time-synchronised voltage and current phasors across wide geographical regions, which are essential for power system observability. The concept of synchronized phasor measurement was first introduced in the 1980s, with advancements in GPS technology to achieve precise time synchronization [1]. This innovation allowed the measurement of voltage and current phasors accurately, which are critical for assessing the stability and reliability of power systems. The obtained time-stamped phasor data allows real-time monitoring of the power system, permitting detection, anticipation, and correction of issues during irregular conditions. Hence, PMUs are widely deployed across the power system for post-disturbance event analysis, wide-area monitoring, real-time state estimation, and enhancement of grid resilience [2]. Recent developments in smart grid and energy management systems further highlight the growing importance of synchronized measurements, real-time monitoring, and intelligent control strategies. IoT-enabled monitoring frameworks have been applied in electric vehicle charging systems for real-time parameter acquisition and operational management, while advanced GPS-assisted control strategies have been investigated for hybrid electric vehicle applications [3],[4]. Similarly, the increasing penetration of renewable energy resources, particularly photovoltaic (PV) systems, has introduced new operational challenges associated with voltage regulation and system stability in distribution networks [5]. To address these challenges, recent studies have explored intelligent data-driven approaches, including hybrid graph neural network models, for overvoltage prediction and monitoring in PV-dominated grids [6]. Furthermore, reliable monitoring and observability are particularly important in remote and islanded microgrids operating under extreme conditions, where resilient system operation becomes essential [7]. These advancements collectively emphasize the increasing significance of accurate synchronized measurements and intelligent monitoring frameworks in modern power systems.

Although PMUs enhance power system monitoring, their high cost and need for communication infrastructure limit deployment at every bus. Moreover, it is not mandatory to place a PMU at every bus, as the voltage phasor at a bus connected to a PMU-equipped bus can be calculated using branch parameters and the measured current phasor from the connecting branch [8]. Typically, a minimum of 20%–30% of the total number of system buses is deemed sufficient for PMU deployment [9]. Therefore, the strategic PMU placement, considering factors such as installation cost [1], measurement redundancy [10], and fault detection capabilities [11], has become a focal point of research. Several optimization techniques have been proposed to solve the OPP problem. These are categorized as deterministic and metaheuristic algorithms. These approaches give a different number of PMUs with different locations on the same system. Hence, to identify the best method or location, several indices have been developed, including the ‘System Observability Redundancy Index’ (SORI) [5]. The high value of SORI represents high observability. Also, to lower the number of PMUs, the effect of ‘Zero Injection Buses’ (ZIBs) has been considered by various researchers [12].

Weak and generator buses are more prone to voltage collapse, threatening system security [13]. However, the direct observability of such critical buses has received comparatively less attention in existing OPP studies. Therefore, this work presents a comprehensive review of OPP techniques and evaluates PMU placement strategies based on observability, redundancy, and direct monitoring capability of weak and generator buses. The main contributions of this paper are as follows:

1. A comprehensive review of existing techniques, including a description, the research focus, and the bus system used for OPP, has been conducted. The results reported in the literature using existing techniques have been analysed to determine the best location based on SORI values.
2. A new Indian Utility 62bus system has been designed in MATLAB, providing practical insights by taking data from [14].
3. PMU locations have been obtained on the IEEE 30 and IEEE 57 for validation and implemented on the Indian

Utility 62 bus systems. Using the integer linear programming (ILP), Depth-first search (DFS), Simulated Annealing (SA), and Recursive Spanning Tree Search (RSTS), considering the ZIB effect. Multiple solutions for PMU locations have been obtained using different techniques.

4. To demonstrate the effectiveness of the multiple solutions, SORI and a new proposed metric, Depth of Redundancy (DOR), have been calculated.
5. A comparative analysis based on cost effectiveness, reliability, critical bus observability, system redundancy, and overall system security has also been carried out.

The proposed approach provides a more reliable framework for redundancy-aware PMU placement compared to conventional observability-based methods. The fundamentals of phasor technology, including its role in synchronized measurements, wide-area monitoring, and real-time grid applications, have been explored in Section 2. Section 3 presents various OPP algorithms. In Section 4, comprehensive reviews of existing OPP techniques, analyzing their efficiency, computational complexity, and adaptability to different power system configurations, have been presented. These approaches are validated in Section 5, presenting simulation results, comparing various methods based on observability indices, redundancy, and computational efficiency. Finally, section 6 concludes with key findings, limitations, and future research directions for improving PMU deployment strategies.

2. Phasor Technology

Phasor technology, also known as phasor measurement technology, plays a critical role in modern power system monitoring and control. The complex equivalent of a sinusoidal waveform is defined as a phasor, where the modulus represents the amplitude and the phase is denoted by the angle of the signal. This mathematical representation allows for simplified analysis of steady-state AC electrical signals. PMUs, the core component of this technology, are explained in the subsection below [15].

2.1. Phasor Measurement Unit

PMUs provide high-speed, time-synchronized measurements of voltage and current phasors across a power system by placing on some locations. The voltage and current phasor of a bus connected with a PMU-equipped bus can be calculated using branch parameters, and the measured current phasor at a bus connected to a PMU-equipped bus can be calculated using branches. By using GPS-based synchronization, PMUs enable real-time visibility into the dynamic behaviour of the system, capturing both the magnitude and phase angle of electrical quantities with much higher resolution and speed compared to traditional SCADA systems. This level of insight is crucial for detecting voltage instabilities, identifying system oscillations, and implementing wide-area control strategies.

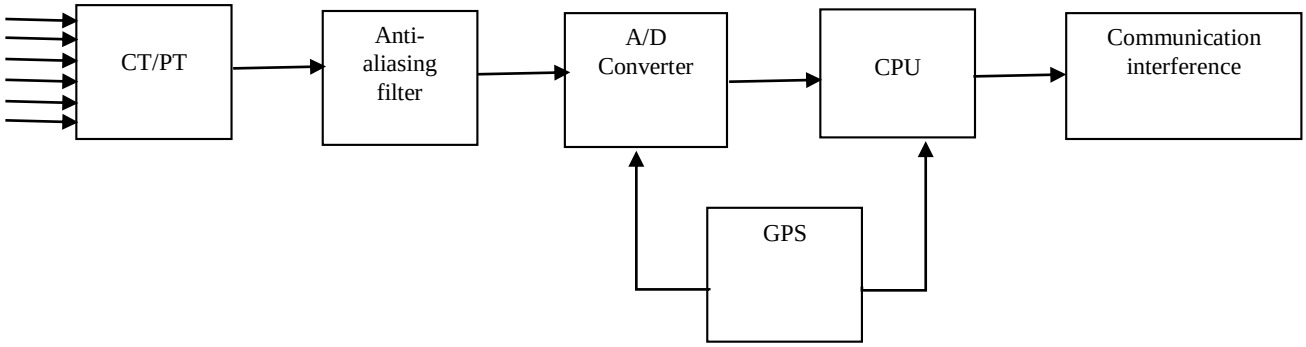


Fig. 1. Block diagram of phasor assisted device.

2.2. *Optimal PMU Placement (OPP)*

The deployment of PMUs is essential in power systems as they provide time-stamped, synchronized phasor measurements. When all bus voltage and current phasors are known, then the system is said to be fully observable. However, due to the elevated cost and the requirement of communication facilities make it impossible to place PMU at each bus. Hence, OPP aims to reduce the number of PMUs placed in the system while ensuring complete observability. This approach seeks to strike a balance between effective system monitoring and minimizing the associated costs and resource requirements. The objective is to observe the power system with the minimum number of PMUs by strategically defining their positions [9].

2.3. Problem Formulation

The typical observability and OPP in power systems involves determining the minimum number of PMUs required. n_{pmu}, and their optimal locations $S(n_{pmu})$. This model can be extended to include other constraints like ZIBs, a limited number of channels, etc. [7]:

$$\min_{n_{pmu}} \min_{S(n_{pmu})} G(n_{pmu}, S(n_{pmu})) \quad (1)$$

Subjected to: $f\left(n_{pmu}, S(n_{pmu})\right) = 1$ (2)

where:

$G\left(n_{pmu}, S(n_{pmu})\right)$: the number of unobservable buses

$S(n_{pmu})$: a set of buses where PMUs are installed

$f\left(n_{pmu}, S(n_{pmu})\right)$: logical function ensuring system observability and constraint satisfaction

The logical function $f(\cdot)$is defined as:

$$f(.) = O_{bs}\left(n_{pmu}, S(n_{pmu})\right) \cdot C_{on}\left(n_{pmu}, S(n_{pmu})\right) \quad (3)$$

where:

$O_{bs}(.)$: observability function (equals 1 if the system is fully observable, otherwise 0)

$C_{on}(.)$: constraint evaluation function (accounts for practical constraints such as Zero Injection Buses (ZIBs), channel limits, etc.).

'+' and '$\cdot$'denote the logical OR and AND operators, respectively.

There is no unique solution for the PMU placement problem. This is because there are different sets of constraints taken by different researchers according to their objectives for finding solutions with different techniques. A ZIB is a bus within a power system that is neither connected to a load nor a generator. As a result, the installation of a PMU at every bus is not required to achieve system observability. The presence of ZIBs can effectively reduce the number of PMUs needed to make the power system completely observable [16]. This constraint has also been added for PMU placement.

2.4. Performance indices

This paper introduces two performance metrics to evaluate the robustness and effectiveness of PMU placement:

- *System Observability Redundancy Index (SORI)*

The SORI quantifies the extent to which strategically placed PMUs can monitor the entire power system [16]. A higher SORI indicates a more reliable PMU-based monitoring system. In the context of the paper, SORI is employed to identify the algorithm that provides the maximum measurement redundancy, enhancing the robustness and reliability of the monitoring system. The SORI is computed using the following equation:

$$SORI = \sum_{j=1}^{N} \sum_{i=1}^{N} A_{ij} X_i \quad (4)$$

A_{ij}: is a binary connectivity matrix

X_i : is the optimal location of PMUs for complete observability.

N= total number of buses

The maximum value of SORI means that a larger portion of the system will remain observable in case of a PMU outage.

- *Depth of Redundancy (DOR)*

DOR quantifies the number of independent measurement paths (PMU paths) through which each bus is observable. A higher DOR implies a more resilient configuration capable of tolerating PMU failures or communication disruptions [7].

Mathematically, for bus *i*,

$$DOR(i) = \sum_{j=1}^{N_{PMU}} \delta_{ij} \quad (5)$$

where:

δ_{ij} =1, if bus i is observable by PMU j (either directly or indirectly), otherwise $\delta_{ij} = 0$

This metric represents the number of independent PMU measurement paths through which bus *i* is observed.

N_{PMU}= the total number of placed PMUs

DOR(i) thus represents the number of independent measurement paths covering bus i. A bus with $DOR_i = 1$ is critically observable — its loss of observability is guaranteed if its sole covering PMU fails.

The average DOR across all buses serves as a system-level redundancy score.

The average DOR is given by:

$$DOR_{avg} = \frac{1}{N} \sum_{i=1}^{N} DO\ R_i \quad (6)$$

Unlike SORI, DOR explicitly captures redundancy depth *at* each bus, making it suitable for identifying weak buses.

- *Mathematical Relationship between SORI and DOR*

SORI is calculated by adding the DOR values of all buses in the system and can be expressed as:

$$\text{SORI} = \sum_{i=1}^{N} \text{DO R}_i = \text{N} \times \text{DOR}_{\text{avg}} \quad (7)$$

where:

DOR_irepresents the redundancy level of the i^{th}bus,

$\text{DOR}_{\text{avg}}$is the average Depth of Redundancy of the system,

N is the total number of buses.

This relation shows that SORI and average DOR are directly connected mathematically. However, both metrics provide different types of information. SORI gives only the overall redundancy level of the complete network, whereas DOR provides bus-wise redundancy information. Therefore, two systems may have the same SORI value but completely different redundancy distributions across buses.

For example, a system with SORI = 79 may have some buses observed through multiple PMUs while other buses are observed through only a single PMU. In such cases, the overall redundancy appears high, but certain buses remain vulnerable. DOR helps identify these weak points by showing the redundancy available at each bus. Buses with $DOR = 1$ are considered critical because the loss of a single PMU may make them unobservable. This is especially important for generator buses and weak buses, where reliable monitoring is essential for secure system operation. The DOR values considered in this work are based on the network topology, which remains unchanged under normal load variations. Therefore, load changes do not affect DOR unless a topological change, such as a transmission line outage, occurs.

3. OPP Algorithms

This section provides an overview of different approaches for solving the OPP problem. It includes both mathematical and heuristic approaches.

3.1. Integer Linear Programming (ILP)

ILP is a deterministic optimization technique used to find the OPP problem. The objective is to minimize the number of PMUs with maximum observability and hence minimize total cost.

Formulation problem [12] :

$$Min \sum_{i=1}^{N} C_i X_i \quad (8)$$

$$Subjected\ to: \quad AX \geq 1 \quad (9)$$

where:

N is the total number of buses

C_i : installation cost of PMU at bus i

X_i: binary decision variable, defined as

$$X_i = \begin{cases} 1 & if\ PMU\ is\ installed \\ 0 & otherwise \end{cases} \quad (10)$$

A is a binary connectivity matrix defined as

$$A_{ij} = \begin{cases} 1\ if\ i = j\ or\ i\ and\ j\ are\ connected \\ 0 \quad otherwise \end{cases} \quad (11)$$

The constraint $Ax \geq 1$ ensures that each bus is observable either directly (through a PMU installed at that bus) or indirectly (through adjacent buses). The problem of OPP is solved using the above procedure in MATLAB.

3.2. Quadratic Programming (QP)

QP serves as a fundamental tool in addressing non-linear programming, notably utilized in optimizing linearly constrained quadratic objective functions. Recent research, as exemplified by [17], employs linear QP for OPP, aiming firstly to attain a complete topologically observable system under normal operation or in cases of PMU or line outage, and secondly, to maximize measurement redundancy. Additionally, [18] utilizes 'Integer Quadratic Programming' for PMU placement, ensuring observability. Uncertainty evaluation within the system is addressed through various methods such as classical uncertainty propagation theory, Monte Carlo simulations, and Random Fuzzy Variable (RFVs), as explored in different studies. Furthermore, advancements include the extension of OPP to the 'Optimal PMU-Communication link Placement' (OPLP) problem, introduced by [19], where mixed-integer quadratically constrained programming is applied, considering PMU and communication link capacities.

3.3. Simulated Annealing (SA)

SA is a probabilistic optimization technique motivated by the annealing process in metallurgy, where an initial random configuration of PMUs is iteratively modified to find optimal placement by evaluating solutions based on reduced PMUs and improved observability. By gradually reducing the acceptance probability of worse solutions as the temperature decreases, SA escapes local optima and refines solutions until the stopping criterion is met, achieving near optimal or optimal PMU placements [1]. A practical solution for the communication-constrained PMU placement problem is the use of the SA method, which addresses the OPP problem by managing incomplete observability [20]. The case studies are simulated within the 'Power System Analysis Toolbox' (PSAT) using MATLAB software.

3.4. Depth First Search (DFS)

The DFS method is a stochastic optimization technique. In this algorithm, the initial step involves exploring the node connected to the highest number of branches within the network. Subsequently, a PMU is strategically placed on this bus. In cases where two or more nodes exhibit an equally substantial number of branches, the algorithm randomly selects nodes for the placement of PMUs. This iterative operation is systematically applied to all nodes throughout the network, ensuring comprehensive observability of the entire system [16], [21]. Initially, the case studies are simulated within the PSAT using MATLAB software. Subsequently, power flow analysis is executed to determine the power flows through the Newton-Raphson method. At last, the optimal placement is found utilizing the PMU placement toolbox within PSAT, employing the DFS.

3.5. Recursive Spanning Tree Search (RSTS)

RSTS is an algorithm that dynamically finds the best way from source to destination within a graph while avoiding bridge loops that could lead to inaccurate results. This technique involves the recursive exploration of multiple spanning trees, each starting from a different node (bus). In power systems, RSTS is particularly valuable for OPP problems. In [22] spanning trees use graphs to find the OPP based on the notion of depth of unobservability. This approach makes sure that the distance between unobserved and observed buses is minimized and that PMUs are distributed uniformly across the network, enhancing the efficiency and reliability of the monitoring system.

3.6. *Tabu Search*

Tabu search develops as a promising metaheuristic approach for optimizing PMU placement in power systems, aiming to minimize the number of PMUs while ensuring full observability and redundancy. Unlike conventional heuristic methods, this method utilizes a flexible memory mechanism to avoid revisiting previously explored solutions, thereby enhancing exploration efficiency.

Previous works have demonstrated the effectiveness of Tabu search in addressing the OPP problem, heuristic rules and prioritized lists to guide the search process towards optimal solutions [23]. Additionally, the recursive Tabu search method

(RTS) proposed in the literature introduces an iterative approach, where multiple executions of conventional Tabu search are performed with each initialization guided by the best solution obtained from prior executions, further enhancing the algorithm's convergence and solution quality [24].

3.7. Particle Swarm Optimization (PSO)

PSO harnesses swarm intelligence, mimicking the behaviour of natural phenomena like bird flocking or fish schooling, to iteratively optimize solutions. Specifically, Binary PSO (BPSO) has been introduced for discrete problem solving, proving effective for OPP. This approach, as demonstrated by recent research, aims to minimize the number of PMU installations while ensuring complete system observability and various constraints, such as a limited number of PMUs, critical bus measurements, the effect of ZIBs, PMU outages, line outages and contingency analysis have been included as an objective function in different PSO algorithms [25].

4. Review of OPP Techniques

The literature review for the OPP problem systematically describes diverse studies, methodologies, and outcomes related to the strategic locations and count of PMUs in power systems. Table 1 provides an overview of metaheuristic algorithms employed in studies, while Table 2 focuses on deterministic algorithms. The literature review indicates that most researchers aim to minimize the number of PMUs while optimizing their locations for maximum redundancy. The different solution has been obtained by researchers, considering different constraints such as zero injection buses, depth of unobservability, line and PMU outages, and fixed or variable channels for PMUs. Various techniques, ranging from single to multi-objective optimization, have been employed to address the OPP problem, emphasizing achieving complete system observability with a minimal number of PMUs and maximizing measurement redundancy simultaneously. Hence, obtaining the optimal position of PMUs involves addressing practical constraints such as zero injection effect, PMU outages, line outages, measurement redundancy, incorporating traditional measurements, accounting for existing PMUs,

Table1. Metaheuristic algorithms

Ref.	Method Adopted/ Bus system used	Description/Features	Focus
[23]	Tabu search algorithm (TSA)/ IEEE 14, 57, and NE 39 bus systems.	Utilizes an augmented incidence matrix to achieve full system observability with reduced computation time.	Minimize PMUs for full observability.
[24]	Recursive Tabu Search (RTS) / IEEE14, 30, 57, 118, New England (NE) 39, 2383 Bus system	Applies Tabu Search recursively, using the best solutions from previous runs; combines numerical observability with metaheuristics.	Minimize PMUs for full observability.
[25]	Binary search algorithm/ IEEE14,24,30, and NE 298 Bus system	An exhaustive search approach has been used to determine the minimum number and OPP for state estimation, considering single branch outages.	Global optimum with branch outage.
[26]	Modified Binary Particle Swarm Optimization (BPSO) / IEEE 14, 30, 39, 57, and 118 Bus system	Considers ZIBs, PMU loss, and line outages; introduces new topological rules for observability analysis.	Robust OPP considering ZIBs.
[27]	Multi-Stage Simulated Annealing (MSSA) / IEEE14, 30 Bus System.	Integrates PMUs with conventional meters using a spanning measurement subgraph approach.	OPP with conventional measurements.
[28]	Exhaustive Binary Search Algorithm (EBSA)/ IEEE 9, 14, 24, 30 Bus system	Reduces search space via connectivity and symmetry constraints; validated via Monte Carlo simulations.	Reduces search space and computation time with full observability.
[29]	Graph-based method / IEEE14, 30, 57, and 118 Bus system.	Provides a GUI-based tool for interactive system modelling and PMU placement with SORI.	OPP with visualization and SORI maximization
[30]	Genetic Algorithm (GA)/ IEEE30, 57, 118, and 114 Algerian Bus system.	Uses evolutionary optimization to minimize cost while achieving observability.	Cost-efficient OPP
[31]	Immunity Genetic Algorithm (IGA)/ IEEE14, 30, 39, 57, and 118 Bus system.	Improves upon GA by avoiding premature convergence using immune operators.	Faster convergence for full observability
[32]	Non-dominated Sorting Genetic Algorithm /IEEE 39 and 118 Bus system.	Multi-objective optimization balancing minimal PMUs and high redundancy; uses the Pareto front.	OPP with maximum measurement redundancy.

[33]	Non-dominated Sorting Differential Evolution (NSDE)/ IEEE 14 and 39 Bus system.	Multi-objective model to minimize PMUs and ensure N-1 measurement reliability.	OPP with reliability optimization
[34]	Genetic Algorithm-Based Approach/ IEEE 14 and NE 39 Bus System	Considers contingency scenarios using a Radial Basis Function network post GA-based placement.	Contingency-aware OPP.
[35]	Improved cuckoo and PSO Method / IEEE 14, 24, 30, and 57 Bus system	Evaluates radial systems and ZIBs; Cuckoo performs better in iterations.	OPP with radial topology and ZIBs.
[36]	BPSO with Integrated Mutation / IEEE 300 Bus system.	Addresses PMU channel limits, ZIBs, and PMU loss to reduce computational load.	Scalable OPP with constraints
[37]	Weighted Least Squares (WLS) - based OPP/ IEEE14, 30, 118, and China 211 Bus system.	Smart grid-oriented; considers ZIBs, channel limits, PMU failures, and state estimation errors.	OPP for smart grid & state estimation
[38]	Fuzzy Logic and Discrete Artificial Bee Colony (FL-DABC)/ IEEE14, 30, 57 Bus system	Suitable for microgrid systems; accounts for ZIBs, PMU loss, and line outages.	OPP under contingency in microgrids.

Table 2. Deterministic algorithms

Ref.	Method Adopted/ Bus system used	Description/Features	Focus
[39]	Integer Linear Programming (ILP) based algorithm / IEEE 14, NE 39, and Northern Region Power Grid (NRPG) 246 Bus systems	Incorporates voltage stability-based contingency scenarios into ILP to identify critical PMU placements in large networks.	Contingency-based OPP
[7]	ILP/ IEEE 14, 57, and 118 bus system.	Considers ZIBs and uses the Bus Observability Index (BOI) and SORI to rank optimal placements.	OPP with maximum observability.
[40]	Sequential Quadratic Programming (SQP) Algorithm/ IEEE 14, 24, 30, and NE-39 Bus system	Provides multiple OPP configurations while considering PMU channel limits. Compared with the BILP and MILP results.	OPP with channel constraints
[41]	Binary Integer Linear Programming (BILP)/ IEEE- 7, 14, 30, 39, 57, and 118 Bus systems.	Uses bus and line connectivity matrices, addressing PMU channel limits to minimize the total PMUs needed.	Minimized PMU count with connectivity matrix.
[42]	BILP/ IEEE 14, IEEE 30, and 57 Bus system	Generalized ILP approach supporting redundant and incomplete observability scenarios. Compares outcomes with and without ZIB consideration.	Comparison of ZIB impact
[16]	MILP/ Nairobi Region 30-bus system.	Combines MILP with Artificial Bee Colony (ABC) and Depth-First search to optimize PMU placement in Kenya's power grid. Evaluated using SORI.	OPP with maximum redundancy.
[43]	MILP/ NE 10-unit 39 and IEEE 118-Bus system	Formulates MILP with backbone-resilience planning. The observability index is applied to evaluate the impact on both the backbone and the full network.	OPP with ZIBs and resilience planning.

considering communication limits, and addressing weakly observable buses.

It is seen in literature that the OPP problem lacks a unique solution. Hence, to identify the best location, a key index, SORI, has been utilized [16]. Each optimal location has its own SORI, and the maximum SORI corresponding to the PMU location presents the maximum measurement redundancy. Therefore, a greater SORI value signifies a more reliable monitoring system.

5. Simulations and Results

The OPP techniques ILP, DFS, SA, and RSTS have been validated on IEEE 30-bus and IEEE 57-bus systems, and

implemented on the Indian Utility 62-bus system for a case study. The system models are developed using MATPOWER to ensure accurate network representation and observability analysis.

Since SA and RSTS are stochastic optimization methods employing probabilistic search mechanisms and random initialization, their obtained PMU placement configurations may vary across independent executions. Therefore, statistical analysis was performed using 30 independent runs with different random seeds to evaluate convergence consistency and solution robustness. The best 6 solutions based on PMU count and maximum SORI have been considered for analysis, as shown in Table 3.

In contrast, ILP and DFS are deterministic optimization approaches that produce identical solutions for repeated executions under the same network conditions. The DFS method implemented in this work follows a fixed in-degree–based search strategy without random exploration or probabilistic acceptance criteria.

The SORI is calculated for each optimal PMU configuration, and the newly introduced DOR metric is computed to quantify direct observability of voltage critical buses, particularly at generator buses and identified weak buses [15], [44].

5.1. Case study I: IEEE 30-Bus system

Figure 2 presents the PMU requirements and corresponding SORI values for the IEEE 30-bus system across different optimisation methods. The results show minimal variation in PMU requirements, with both ILP and RSTS achieving complete system observability using the minimum number of PMUs (7). ILP achieves superior performance, with a SORI of 35, compared with RSTS, which has a SORI of 31, indicating better redundancy distribution despite the identical PMU count. As shown in Table 3, SA requires one additional PMU (8) but achieves an improved SORI of 39. DFS deploys the maximum PMUs (10), achieving the highest SORI of 45, providing better redundancy than ILP but at a high cost.

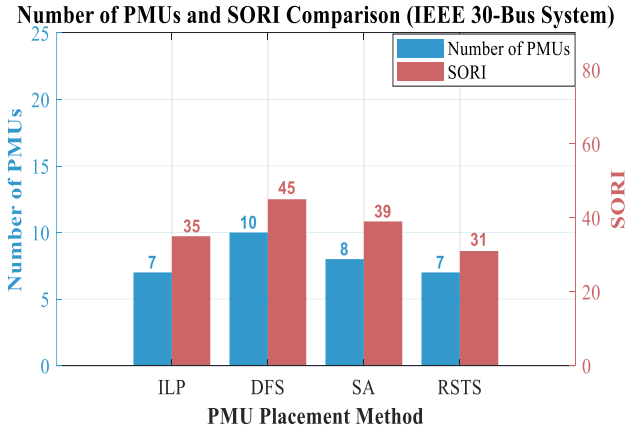


Fig. 2. Number of PMUs obtained for the IEEE 30 bus system using various techniques with their SORI.

5.2. Case Study II: IEEE 57-Bus system

Figure 3 presents the performance comparison of various methods of OPP for the IEEE 57-bus system, significantly showing variations in both PMU counts and SORI. It can be observed that ILP achieves observability with 11 PMUs and SORI of 54, representing the most economical solution. RSTS requires a marginally higher PMU count (12) but achieves lower SORI (51), indicating suboptimal redundancy distribution. SA provides a balanced approach with 14 PMUs and an SORI of 59. DFS demonstrates maximum redundancy capability with SORI of 85 using 22 PMUs, achieving higher redundancy requiring more PMUs, making it economically challenging for practical implementation.

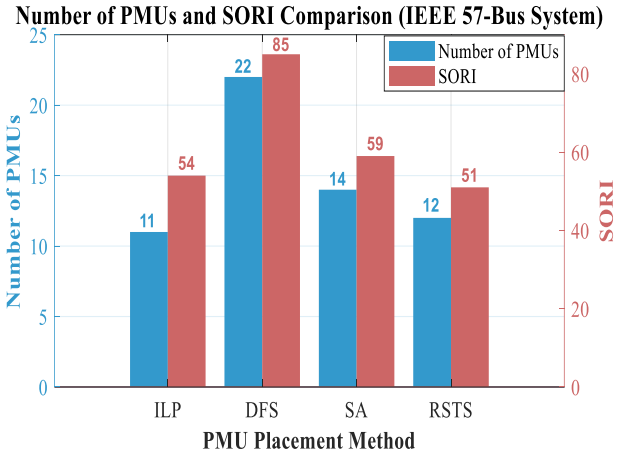


Fig. 3. Number of PMUs obtained for the IEEE 57 bus system using various techniques with their SORI.

5.3. Case Study III: Indian Utility 62 Bus System

The Indian Utility 62-bus system is the selected network to evaluate the effectiveness, robustness, and practical applicability of the proposed PMU placement methodologies under realistic operating conditions. Compared with standard IEEE test systems, the Indian Utility 62-bus system features a larger, more complex topology with multiple generation buses, interconnected corridors, and varying observability conditions. Therefore, this case study is used to assess not only optimal PMU placement performance but also resilience under contingency conditions.

5.3.1. System Description and PMU Placement Results

The Indian Utility 62-bus system is a real-world meshed transmission network representative of a regional grid in India. The system comprises 62 buses, 89 transmission branches, 11 transformers, 32 load buses, and 19 generating units distributed across 8 distinct generator buses (1, 14, 17, 25, 32, 37, 51, and 58). Multiple generating units are connected to common buses; therefore, PMU placement analysis focuses on generator buses rather than individual generating units. The network operates at three voltage levels: 220 kV, 132 kV, and 66 kV, interconnected through step-down transformers. Bus 1 serves as the slack (reference) bus. The system data, including bus admittance matrix, line parameters, weak buses (11, 20, 30 and 31) and

loading conditions, are available in [45]. Figure 5 presents the single-line diagram of the Indian Utility 62-bus system with optimal PMU placement locations. The diagram clearly illustrates the strategic positioning of PMUs to achieve comprehensive system observability while prioritizing critical generator and weak bus monitoring.

5.3.2. Multi-Objective Performance Comparison

The results obtained for the Indian Utility 62- Bus system in Table 3 have been presented in Figure 4, showing the variation in PMU count and corresponding SORI values obtained from multiple optimization runs of different OPP methods. The solution diversity is observed in SA and RSTS compared to ILP and DFS. The multiple configurations obtained from the SA and RSTS methods highlight the stochastic nature of these algorithms. SA generates solutions ranging from 15-17 PMUs with SORI values 69-79, while RSTS produces consistent solutions with 15 PMUs and SORI values ranging from 68-71. These findings emphasize that higher SORI alone does not guarantee fault tolerance. Average DOR and DOR of critical buses are better indicators of robustness in real-world grid monitoring. Furthermore, since multiple feasible solutions were obtained from the SA and RSTS methods, the PMU configurations corresponding to the highest SORI values were selected for further investigation. A comprehensive performance analysis was then carried out, considering not only the overall observability redundancy but also the direct observability of critical buses, including generator and weak buses.

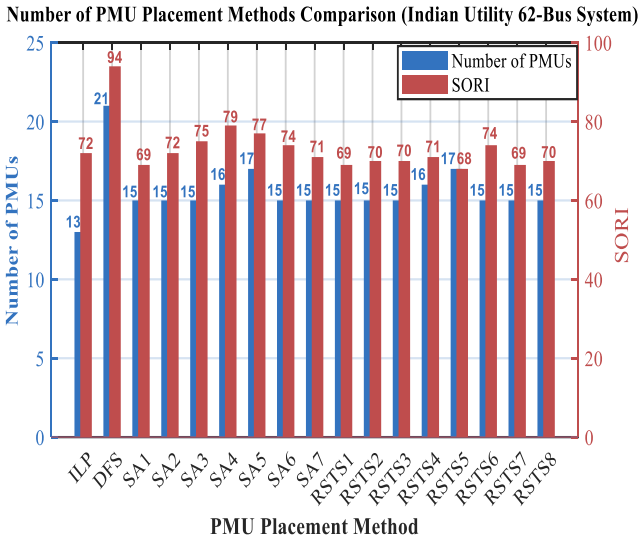


Fig. 4. Number of PMUs obtained for the Indian Utility 62-Bus System using various techniques with their SORI.

Table 4 presents the detailed comparison of all methods in terms of average DOR, SORI, generator bus observability, and weak bus coverage. The analysis identifies SA as the most balanced solution, achieving 100% generator bus observability with a 9.7% improvement in SORI compared to ILP, while requiring only 23.1% additional PMUs. This demonstrates that SA provides an effective trade-off between redundancy enhancement and implementation cost for reliable real-world grid monitoring.

The critical bus coverage analysis in Table 5 further highlights the effectiveness of the SA-based PMU placement strategy. Although all methods achieved complete system observability, major differences were observed in critical bus monitoring performance. SA achieved 100% generator bus coverage and complete weak bus observability, demonstrating superior monitoring capability for strategically important buses. In contrast, ILP and RSTS observed only 62.5% of generator buses, while DFS achieved high redundancy and 87.5% generator bus coverage at the expense of significantly higher PMU deployment. These results confirm that SA provides a better trade-off between redundancy, critical bus coverage, and installation cost for practical wide-area monitoring applications. To further evaluate the practical applicability of the obtained PMU placement solutions, the economic and reliability trade-off analysis presented in Table 6 compares the methods in terms of installation cost, redundancy, and fault tolerance capability. Although ILP provides the most economical solution, its redundancy and critical bus protection capability remain limited. DFS achieves the highest reliability and redundancy but requires substantially higher investment and maintenance complexity. In comparison, SA offers a balanced compromise by providing complete generator protection and high redundancy with only a moderate increase in installation cost, making it a more practical solution for reliable wide-area monitoring applications.

To further assess the robustness of the obtained PMU placement solutions under contingency conditions, PMU outage analyses were carried out. Table 7 presents the contingency analysis results under PMU outage condition for different OPP methods. The results show that ILP and RSTS are more vulnerable to PMU failures, where the worst-case PMU outage causes up to 7 buses to become unobservable. In contrast, SA reduces the maximum observability loss to 6 buses, while DFS demonstrates the highest resilience with only 4 buses becoming unobservable under the worst PMU outage scenario. Similarly, the average number of unobserved buses per PMU loss decreases from 3.77 in ILP to 1.24 in DFS, indicating improved redundancy and fault tolerance with increased PMU deployment. The SORI values after the worst PMU outage further highlight the superior robustness of DFS and SA compared to ILP and RSTS. These results confirm that redundancy-based PMU placement improves system robustness against PMU contingencies.

To provide a comprehensive quantitative comparison of the proposed OPP methods, Table 8 summarizes the key performance metrics, including observability redundancy, critical bus coverage, contingency resilience, PMU efficiency, and cost-effectiveness measures obtained for the Indian Utility 62-bus system. ILP provides the most economical solution with the minimum PMU count and highest PMU utilization efficiency; however, its redundancy and critical bus monitoring capability are comparatively lower. DFS achieves the highest SORI, Average DOR, and contingency robustness due to extensive PMU deployment, thereby offering superior fault

tolerance and observability performance. Meanwhile, SA demonstrates a balanced trade-off by achieving complete generator bus coverage and improved redundancy with only a moderate increase in PMU installation cost. The contingency analysis further shows that redundancy-based PMU placement significantly enhances system resilience against PMU outages. Overall, the results suggest that SA is a practical solution for reliable and cost-effective wide-area monitoring, whereas DFS is more suitable for applications requiring maximum redundancy and reliability.

Table 3. Performance comparison of ILP, DFS, SA, and RSTS algorithm

Methods / Bus System	Integer linear programming (ILP) [46]		Depth First Search (DFS) [6]		Simulated Annealing (SA) [21]		Recursive Spanning Tree Search (RSTS) [24]	
	Number of PMUs (Location)	SORI	Number of PMUs (Location)	SORI	Number of PMUs (Location)	SORI	Number of PMUs (Location)	SORI
IEEE30 BUS	7(2, 4, 10, 12, 19, 24, 27)	35	10 (1, 5, 6, 10, 11, 12, 18, 24, 26, 27)	45	8(1, 5, 6, 10, 12, 18, 24, 27)	39	7 (3, 2, 19, 12, 10, 30, 24)	31
IEEE 57BUS	11 (3, 9, 11, 12, 13, 14, 15, 19, 29, 35, 49)	54	22 (1, 4, 8, 10, 13, 19, 21, 24, 27, 29, 30, 32, 36, 38, 39, 43, 45, 47, 50, 53, 55, 56)	85	14 (1, 4, 9, 15, 19, 27, 30, 32, 38, 39, 41, 50, 52, 54)	59	12 (1, 4, 13, 19, 27, 29, 30, 32, 38, 51, 54, 56)	51
INDIAN UTILITY 62BUS	13 (1, 4, 12, 14, 21, 25, 30, 34, 41, 42, 48, 51, 58)	72	21 (1, 3, 6, 8, 11, 14, 17, 20, 22, 25, 30, 32, 38, 39, 41, 43, 44, 48, 51, 52, 58)	94	15 (1, 8, 14, 20, 22, 25, 30, 32, 38, 39, 41, 43, 49, 51, 58)	69	15 (8, 9, 11, 14, 20, 21, 24, 25, 30, 32, 38, 42, 48, 51, 58)	69
					15 (1, 6, 14, 20, 22, 25, 32, 38, 39, 40, 41, 43, 48, 51, 58)	72	15 (6, 9, 11, 14, 20, 22, 24, 25, 30, 32, 38, 42, 48, 53, 58)	70
					15 (1, 6, 14, 22, 23, 25, 30, 32, 38, 39, 41, 43, 48, 51, 58)	75	15 (6, 9, 11, 14, 20, 22, 24, 25, 30, 32, 38, 42, 48, 53, 58)	70
					16 (1, 3, 14, 17, 20, 25, 30, 32, 37, 38, 41, 43, 44, 48, 51, 58)	79	15 (6, 9, 11, 14, 17, 20, 24, 25, 30, 32, 38, 42, 48, 53, 58)	71
					17 (1, 3, 8, 14, 20, 22, 25, 30, 32, 38, 39, 41, 43, 48, 51, 52, 58)	77	15 (8, 9, 11, 14, 20, 22, 24, 25, 30, 32, 38, 42, 48, 53, 58)	68
					15 (1, 6, 14, 17, 20, 25, 30, 32, 38, 39, 41, 43, 48, 51, 58)	74	15 (6, 9, 10, 12, 14, 22, 24, 25, 30, 32, 38, 42, 48, 51, 57)	69

Table 4. Comprehensive performance analysis of PMU placement methods on the Indian 62-Bus system

Methods	Methods	No of PMUs	PMU locations	Avg. DOR	SORI	Generator Buses observed	DOR of Generator Buses	Total number of Weak Buses observed	DOR of weak buses
ILP	ILP	13	1, 4, 12, 14, 21, 25, 30, 34, 41, 42, 48, 51, 58	1.16	72	5 (1, 14, 25, 51, and 58)	2	1 (30)	1
RSTS	RSTS	15	6, 9, 11, 14, 17, 20, 24, 25, 30, 32, 38, 42, 48, 53, 58	1.15	71	5 (14, 17, 25, 32, and 58)	1	3 (11, 20, and 30)	1
SA	SA	16	1, 3, 14, 17, 20, 25, 30, 32, 37, 38, 41, 43, 48, 51, 58	1.21	79	8 (1, 14, 17, 25, 32, 37, 51, and 58)	1.63	3 (3, 20, and 30)	1
DFS	DFS	21	1, 3, 6, 8, 11, 14, 17, 20, 22, 25, 30, 32, 38, 39, 41, 43, 44, 48, 51, 52, 58	1.52	94	7 (1, 14, 17, 25, 32, 51, and 58)	1.43	3 (3, 20, and 30)	1

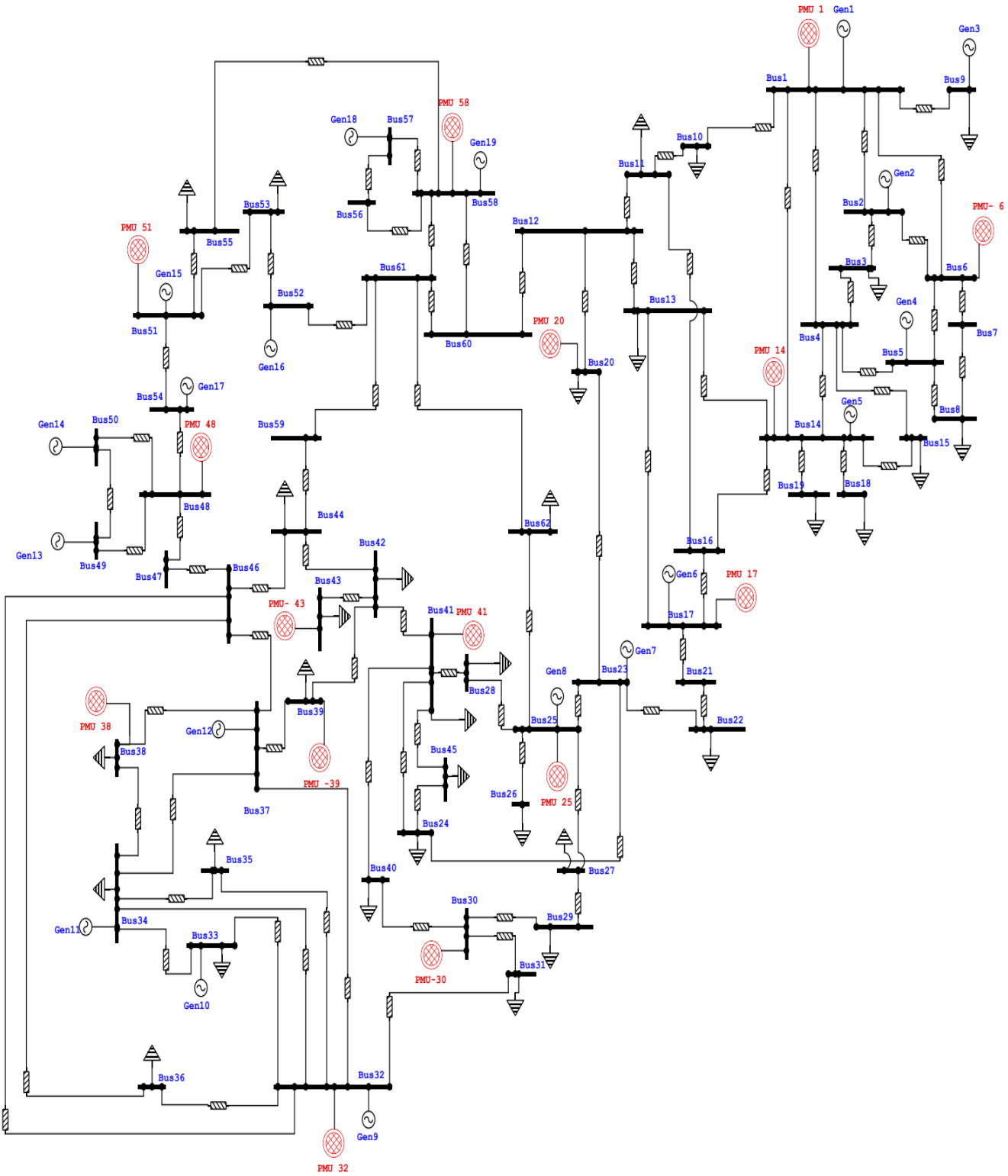


Fig. 5. Single-line diagram of the Indian utility 62 bus system with PMUs.

Table 5. Critical bus coverage analysis

Bus Type	Total Count	ILP Coverage	RSTS Coverage	SA Coverage	DFS Coverage	Best Coverage Method
Generator Buses	8	5 (62.5%)	5 (62.5%)	8 (100%)	7 (87.5%)	SA (Complete)
Weak Buses	3	1 (33.3%)	3 (100%)	3 (100%)	3 (100%)	RSTS /SA/DFS
Total Observable	62	62 (100%)	62 (100%)	62 (100%)	62 (100%)	All methods
Critical Redundancy	-	Low	Moderate	High	Highest	DFS

Table 6. Economic and reliability trade-off analysis

Economic Factor	ILP	RSTS	SA	DFS	Optimal Choice
*Installation Cost	$650K	$750K	$800K	$1,050K	ILP (Lowest)
Cost per SORI Point	$9.03K	$10.56K	$10.13K	$11.17K	ILP
Redundancy Investment	Baseline	+$100K	+$150K	+$400K	Context-dependent
Fault Tolerance	Low	Moderate	High	Highest	DFS
Generator Protection	Partial	Partial	Complete	Good	SA
Maintenance Complexity	Low	Moderate	Moderate	High	ILP

*Assuming $50K per PMU installation cost

Table 7. N-1 Contingency analysis (PMU)

Methods	No of PMUs	SORI baseline	Max unobserved buses (Worst PMU Loss)	Average unobserved buses (per PMU Loss)	SORI after the Worst PMU Loss	Avg. DOR after the Worst PMU Loss
ILP	13	72	7	3.77	69	1.14
RSTS	15	71	7	2.93	67	1.08
SA	16	79	6	2.69	74	1.19
DFS	21	94	4	1.24	86	1.87

*Critical PMU: The PMU whose outage causes at least one bus to lose observability
**Worst PMU: The PMU whose outage causes the maximum number of buses to go unobservable

Table 8. Comparative performance metrics and improvement

Performance Metric	ILP	RSTS	SA	DFS	Best Method	Improvement over ILP
PMU Count	13	15 (+15.4%)	16 (+23.1%)	21 (+61.5%)	ILP	Baseline (most economic)
Average DOR	1.16	1.15 (-0.9%)	1.21 (+4.3%)	1.52 (+31.0%)	DFS	+31.0%
SORI	72	71 (-1.4%)	79 (+9.7%)	94 (+30.6%)	DFS	+30.6%
Generator Coverage	5/8 (62.5%)	5/8 (62.5%)	8/8 (100%)	7/8 (87.5%)	SA	+60% coverage
Avg Generator DOR	2.0	1.0 (-50%)	1.63 (-18.5%)	1.43(-28.5%)	ILP	Baseline
Cost-Effectiveness Ratio*	6.42	5.61	5.98	7.25	DFS	+12.9%
PMU Efficiency**	4.77	4.13	3.88	2.95	ILP	Baseline
Contingency Resilience	Low	Moderate	High	Highest	DFS	Significant improvement

*Cost-Effectiveness Ratio = (SORI × DOR) / PMU Count
**PMU Efficiency = Total Observable Buses (62) / PMU

6. Key Performance Insights and Selection

The comparative analysis of the considered OPP techniques demonstrates that each method provides a different trade-off between installation cost, observability redundancy, and contingency resilience. Therefore, the selection of an appropriate PMU placement strategy should depend on the operational priorities and reliability requirements of the utility.

- *For Budget-Constrained Utilities:*

The ILP configuration is recommended for utilities prioritizing minimum installation cost and reduced PMU deployment. Although the redundancy level is comparatively lower, the method provides complete system observability with the minimum number of PMUs, making it suitable for initial PMU deployment planning and economically constrained applications.

- *For Reliability-Critical System Applications:*

The SA configuration offers a balanced compromise between deployment cost and monitoring reliability. The method achieves higher SORI and DOR values, improved generator bus observability, and stronger resilience under PMU outage contingencies. Therefore, SA is more suitable for real-time monitoring applications where redundancy and fault tolerance are critical operational requirements.

- *For Highly Secure or Defence-Oriented Applications:*

The DFS configuration provides the highest observability redundancy and the strongest contingency resilience due to the larger number of deployed PMUs. Although the installation cost is significantly higher, the improved fault tolerance and redundancy make this method appropriate for highly critical infrastructures requiring maximum monitoring reliability.

- *For Standard Utility Monitoring Applications:*

The RSTS configuration provides a practical balance between PMU count, redundancy enhancement, and implementation cost. The method maintains satisfactory observability performance while achieving improved robustness compared to minimum-PMU solutions, making it suitable for conventional utility monitoring environments.

- *For Long-Term Smart Grid Expansion Planning:*

Utilities may initially adopt economical ILP-based deployment and progressively transition toward redundancy-oriented configurations such as SA or RSTS as reliability requirements, contingency considerations, and monitoring infrastructure evolve.

Overall, the presented implementation guidelines indicate that optimal PMU placement should not be evaluated solely on minimum PMU count. Observability redundancy, critical bus coverage, and contingency resilience must also be considered to ensure reliable and secure real-time monitoring of modern power systems.

7. Conclusion

This paper presents a comprehensive review and analysis of various OPP techniques, introducing the novel DOR metric for direct observability of voltage-critical buses, alongside SORI, to provide an enhanced assessment of system observability Unlike existing approaches that primarily focus on achieving complete observability with minimum PMU deployment, the proposed DOR metric provides additional

insight into bus-level redundancy and critical bus protection under contingency conditions. The performance of different OPP methods is evaluated on IEEE 30-bus and IEEE 57-bus systems for validation, and subsequently implemented on the Indian Utility 62-bus system.

The obtained results show that ILP provides the most economical solution with minimum PMU deployment (13 PMUs), whereas stochastic approaches such as SA and RSTS achieve improved redundancy and contingency resilience. Among all methods, the SA-based configuration provided the best overall balance with 16 PMUs, SORI = 79, average DOR = 1.21, complete generator bus observability (100%), and full weak bus coverage. In contrast, DFS achieved the highest redundancy performance with SORI = 94, average DOR = 1.52, and the strongest resilience under PMU outage contingencies, although at significantly higher installation cost. The contingency analysis further confirmed that redundancy-oriented PMU placement significantly improves system robustness compared to minimum-PMU solutions. Overall, the results demonstrate that the selection of an appropriate OPP technique should depend on utility priorities such as cost, redundancy, and monitoring reliability. Future work could extend this redundancy-aware placement framework by integrating dynamic performance metrics such as voltage stability indices, contingency ranking, or communication latency impact, further refining PMU placement strategies with some advanced AI techniques in modern smart grids.

Acknowledgements

The authors would like to express their sincere gratitude to colleagues and others who supported and encouraged them and contributed directly or indirectly to this work.

Author Contributions

Ms. Nisha Sharma: Conceptualization, Methodology, Formal analysis, Investigation, Writing-Original Draft, Writing- review & editing, Visualization. **Dr Priti Prabhakar and Dr. Avneesh Verma**: Validation, review & editing.

Conflict of Interest

There is no conflict of interest regarding the publication of this paper.

References

[1] R. F. Nuqui and A. G. Phadke, "Phasor measurement unit placement techniques for complete and incomplete observability," *IEEE Trans. Power Deliv.*, vol. 20, no. 4, pp. 2381–2388, Oct. 2005, doi: 10.1109/TPWRD.2005.855457.

[2] H. Li, A. Bose, and V. M. Venkatasubramanian, "Wide-area voltage monitoring and optimization," *IEEE Trans. Smart Grid*, vol. 7, no. 2, pp. 785–793, 2016, doi: 10.1109/TSG.2015.2467215.

[3] T. Ilahi, T. Izhar, and M. Ali, "Comprehensive design analysis of economical e-bike charger with IoT-empowered system for real-time parameter monitoring," *J. Adv. Transp.*, vol. 2, pp. 1–14, 2024, doi: 10.1155/2024/2387983.

[4] R. Bhukya, S. M. Kuchibhatla, J. Venkata, N. Ramesh, and H. Rao, "Integrating AI and optimization algorithms to enhanced protection of smart power systems," *IAENG Int. J. Comput. Sci.*, vol. 52, no. 12, pp. 4738–4749, 2025.

[5] R. Babu, S. Raj, and B. Bhattacharyya, "Weak bus-constrained PMU placement for complete observability of a connected power network considering voltage stability indices," *Prot. Control Mod. Power Syst.*, vol. 5, no. 1, 2020, doi: 10.1186/s41601-020-00174-8.

[6] T. L. Baldwin, L. Mili, M. B. Boisen, and R. Adapa, "Power system observability with minimal phasor measurement placement," *IEEE Trans. Power Syst.*, vol. 8, no. 2, pp. 707–715, 1993, doi: 10.1109/59.260810.

[7] D. Dua, S. Dambhare, R. K. Gajbhiye, and S. A. Soman, "Optimal multistage scheduling of PMU placement: An ILP approach," *IEEE Trans. Power Deliv.*, vol. 23, no. 4, pp. 1812–1820, 2008, doi: 10.1109/TPWRD.2008.919046.

[8] H. A. Abd el-Ghany, I. A. Soliman, E. S. Ahmed, and A. E. ElGebaly, "Generalized wide-area fault detection and location algorithm for transmission system based on optimal PMUs allocation," *Int. J. Electr. Power Energy Syst.*, vol. 155, no. PB, p. 109634, 2024, doi: 10.1016/j.ijepes.2023.109634.

[9] I. Abdulrahman and G. Radman, "ILP-Based optimal PMU placement with the inclusion of the effect of a group of zero-injection buses," *J. Control. Autom. Electr. Syst.*, vol. 29, no. 4, pp. 512–524, Aug. 2018, doi: 10.1007/s40313-018-0389-4.

[10] N. Arora and S. T. Nagarajan, "Novel algorithm for optimal PMU placement for wide ranging power system observability," in *2021 6th International Conference for Convergence in Technology, I2CT 2021*, Institute of Electrical and Electronics Engineers Inc., Apr. 2021. doi: 10.1109/I2CT51068.2021.9417995.

[11] A. Gulraiz, S. S. H. Zaidi, H. Gulraiz, B. M. Khan, M. Ali, and B. Khan, "Leveraging dense layer hybrid graph neural networks for managing overvoltage in PV-dominated distribution systems," *Results Eng.*, vol. 27, p. 106169, doi: 10.1016/j.rineng.2025.106169.

[12] M. Ali, J. C. Vasquez, J. M. Guerrero, Y. Guan, and N. Bazmohammadi, "Microgrid an energy solution for remote islanded communities in Indonesia," *2024 IEEE 10th Int. Power Electron. Motion Control Conf. IPEMC 2024 ECCE Asia*, pp. 3799–3804, 2024, doi: 10.1109/IPEMC-ECCEAsia60879.2024.10567531.

[13] R. Maharjan and S. Kamalasadan, "Voltage Stability index for online voltage stability assessment," in *Proceedings of the 47th Annual North American Power Symposium*, Charlotte, NC, USA: IEEE, Oct. 2015. doi: 10.1109/NAPS.2015.7335245.

[14] B. Krishnakumar and R. Arthi, "Contingency assessment of Indian utility 62 bus system," *AICERA 2012 - Annu. Int. Conf. Emerg. Res. Areas Innov. Pract. Futur. Trends*, no. 2, 2012, doi: 10.1109/AICERA.2012.6306712.

[15] P. Prabhakar and A. Kumar, "Voltage stability assessment using phasor measurement technology," in *IEEE 6th India International Conference on Power Electronics (IICPE) 8-10 Dec. 2014, Kurukshetra, India*, 2014.

[16] E. O. Okendo, C. W. Wekesa, and M. J. Saulo, "Optimal placement of phasor measurement unit considering system observability redundancy index: case study of the Kenya power transmission network," *Heliyon*, vol. 7, no. 7, p. e07670, 2021, doi: 10.1016/j.heliyon.2021.e07670.

[17] C. Sailat, K. Elias, and G. E. Demetrios, "Placement of synchronized measurements for power system observability," *IEEE Trans. Power Deliv.*, vol. 24, p. 19, 2009, doi: 10.1109/TPWRD.2008.2008430.

[18] S. Chakrabarti, E. Kyriakides, and M. Albu, "Uncertainty in power system state variables obtained through synchronized measurements," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 8, pp. 2452–2458, 2009, doi: 10.1109/TIM.2009.2016387.

[19] X. Zhu, M. H. F. Wen, V. O. K. Li, and K. C. Leung, "Optimal PMU-communication link placement for smart grid wide-area measurement systems," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 4446–4456, 2019, doi: 10.1109/TSG.2018.2860622.

[20] M. Nazari-Heris and B. Mohammadi-Ivatloo, "Application of heuristic algorithms to optimal PMU placement in electric power systems: An updated review," *Renew. Sustain. Energy Rev.*, vol. 50, pp. 214–228, 2015, doi: 10.1016/j.rser.2015.04.152.

[21] A. I. Al-Odienat, B. Malahmeh, and A. R. Tarawneh, "The optimal PMU placement in the power systems for the enhancement of state estimation," in *2nd International Conference on Electrical, Communication and Computer Engineering, ICECCE 2020*, 2020, pp. 0–5. doi: 10.1109/ICECCE49384.2020.9179214.

[22] N. M. Manousakis, G. N. Korres, and P. S. Georgilakis, "Taxonomy of PMU placement methodologies," *IEEE Trans. Power Syst.*, vol. 27, no. 2, pp. 1070–1077, May 2012, doi: 10.1109/TPWRS.2011.2179816.

[23] J. Peng, Y. Sun, and H. F. Wang, "Optimal PMU placement for full network observability using Tabu search algorithm," *Int. J. Electr. Power Energy Syst.*, vol. 28, no. 4, pp. 223–231, May 2006, doi: 10.1016/j.ijepes.2005.05.005.

[24] N. C. Koutsoukis, N. M. Manousakis, P. S. Georgilakis, and G. N. Korres, "Numerical observability method for optimal phasor measurement units placement using recursive tabu search method," *IET Gener. Transm. Distrib.*, vol. 7, no. 4, pp. 347–356, 2013, doi: 10.1049/iet-gtd.2012.0377.

[25] J. Kennedy and R. C. Eberhart, "Discrete binary version of the particle swarm algorithm," *Proc. IEEE Int. Conf. Syst. Man Cybern.*, vol. 5, pp. 4104–4108, 1997, doi: 10.1109/icsmc.1997.637339.

[26] M. Hajian, A. M. Ranjbar, T. Amraee, and B. Mozafari, "Optimal placement of PMUs to maintain network observability using a modified BPSO algorithm," *Int. J. Electr. Power Energy Syst.*, vol. 33, no. 1, pp. 28–34, Jan. 2011, doi: 10.1016/j.ijepes.2010.08.007.

[27] P. Gopakumar, M. J. B. Reddy, and D. K. Mohanta, "Novel multi-stage simulated annealing for optimal placement of PMUs in conjunction with conventional measurements," in *12th International Conference on Environment and Electrical Engineering*, IEEE, May 2013, pp. 248–252. doi: 10.1109/EEEIC.2013.6549625.

[28] A. A. Ibrahim, K. Khalid, H. Shareef, and N. A. M. Kamari, "A bounded exhaustive search technique for optimal phasor measurement unit placement in power grids," *Symmetry, MDPI*, vol. 12, no. 12, pp. 1–13, Dec. 2020, doi: 10.3390/sym12122021.

[29] K. Schmaltz, G. Gilje, A. S. Nair, and P. Ranganathan, "An object oriented graphical user interface (GUI) for optimal placement of phasor measurement unit," in *IEEE International Conference on Electro/Information Technology(EIT)*, 2018, pp. 0741–0746.

[30] L. Mosbah, R.D. Mohammedi, and A. Kouzou, "Optimal placement of PMUs in Algerian network using genetic algorithm," in *15th International Multi-Conference on Systems, Signals & Devices , Hammamet, Tunisia.*, pp. 947–951.

[31] F. Aminifar, C. Lucas, A. Khodaei, and M. Fotuhi-Firuzabad, "Optimal placement of phasor measurement units using immunity genetic algorithm," *IEEE Trans. Power Deliv.*, vol. 24, no. 3, pp. 1014–1020, 2009, doi: 10.1109/TPWRD.2009.2014030.

[32] B. Milošević and M. Begović, "Nondominated sorting genetic algorithm for optimal phasor measurement placement," *IEEE Trans. Power Syst.*, vol. 18, no. 1, pp. 69–75, 2003, doi: 10.1109/TPWRS.2002.807064.

[33] C. Peng, H. Sun, and J. Guo, "Multi-objective optimal PMU placement using a non-dominated sorting differential evolution algorithm," *Int. J. Electr. Power Energy Syst.*, vol. 32, no. 8, pp. 886–892, 2010, doi: 10.1016/j.ijepes.2010.01.024.

[34] K. S. Sajan, V. Kumar, and B. Tyagi, "Genetic algorithm based support vector machine for on-line voltage stability monitoring," *Int. J. Electr. Power Energy Syst.*, vol. 73, pp. 200–208, Dec. 2015, doi: 10.1016/j.ijepes.2015.05.002.

[35] M. Patel and S. Talati, "Optimal PMU placement by improved Cuckoo & PSO method," in *Proceedings of the 2021 1st International Conference on Advances in Electrical, Computing, Communications and Sustainable Technologies, ICAECT 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi:

10.1109/ICAECT49130.2021.9392603.

[36] N. Hanis, A. Rahman, and A. F. Zobaa, "Integrated mutation strategy with modified binary PSO algorithm for optimal PMUs," *Ieee Trans. Ind. Informatics*, vol. 13, no. 6, pp. 3124–3133, 2017.

[37] T. Chen, H. Ren, Y. Sun, M. Kraft, and G. A. J. Amaratunga, "Optimal placement of phasor measurement unit in smart grids considering multiple constraints," *J. Mod. Power Syst. Clean Energy*, vol. 11, no. 2, pp. 479–488, 2023, doi: 10.35833/MPCE.2022.000003.

[38] V. Pattanaik, B. K. Malika, P. K. Rout, and B. K. Sahu, "Contingency-resilient PMU placement using fuzzy logic and discrete artificial bee colony algorithm for comprehensive network observability," *e-Prime - Adv. Electr. Eng. Electron. Energy*, vol. 5, no. June, p. 100275, 2023, doi: 10.1016/j.prime.2023.100275.

[39] Sodhi Ranjana, Srivastava S.C, and Singh S.N, "Optimal PMU placement to ensure system observability under contingencies," in *IEEE Power & Energy Society. General Meeting (2009 : Calgary, Alta.)*, IEEE, 2009.

[40] N. P. Theodorakatos, N. M. Manousakis, and G. N. Korres, "A sequential quadratic programming method for contingency-constrained phasor measurement unit placement," *Int. Trans. Electr. Energy Syst.*, vol. 25, no. 12, pp. 3185–3211, Dec. 2015, doi: 10.1002/etep.2027.

[41] V. Rokkamand and R. Bhimasingu, "A novel approach for optimal PMU placement considering channel limit," in *POWERCON 2014 - 2014 International Conference on Power System Technology: Towards Green, Efficient and Smart Power System, Proceedings*, Institute of Electrical and Electronics Engineers Inc., Dec. 2014, pp. 1164–1171. doi: 10.1109/POWERCON.2014.6993995.

[42] B. Gou, "Generalized integer linear programming formulation for optimal PMU placement," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 1099–1104, 2008, doi: 10.1109/TPWRS.2008.926475.

[43] Y. Bai, S. Li, and X. Gu, "Collaborative planning of resilient backbone grids and PMU placement for power systems," *Int. J. Electr. Power Energy Syst.*, vol. 131, p. 107106, Oct. 2021, doi: 10.1016/j.ijepes.2021.107106.

[44] P. Sharma and A. Kumar, "Thevenin's equivalent based P-Q-V voltage stability region visualization and enhancement with FACTS and HVDC," *Int. J. Electr. Power Energy Syst.*, vol. 80, pp. 119–127, 2016, doi: 10.1016/j.ijepes.2016.01.026.

[45] P. Harish Kumar and R. Mageshvaran, "Load flow analysis and optimal allocation of DG for Indian utility 62 bus power system," *Int. J. Emerg. Technol.*, vol. 11, no. 2, pp. 874–886, 2020.

[46] C. Lu, Z. Wang, M. Ma, R. Shen, and Y. Yu, "An optimal PMU placement with reliable zero injection observation," *IEEE Access*, vol. 6, pp. 54417–54426, 2018, doi: 10.1109/ACCESS.2018.2865513.

International Journal of Smart Grid (IJSmartGrid)
https://www.ijsmartgrid.org/ Online ISSN: 2602-439X
S. Sağıroğlu et al., Vol. 10, No. 3, September, 2026 Review Article

A Multi-Layer Conceptual Security Architecture Toward Robust and Resilient SDN-Enabled Smart Grids

Seref Sağıroğlu*‡, Mehmet Demirci*, Yılmaz Atay*, Murat Dener**, M. Sedef Demirci*, Ali Kara***, Samed Al**, Selcuk Yılmaz**

* Department of Computer Engineering, Faculty of Engineering, Gazi University, Ankara, Türkiye

** Department of Information Security Engineering, Graduate School of Natural and Applied Sciences, Gazi University, Ankara, Türkiye

*** Department of Electrical and Electronics Engineering, Faculty of Engineering, Gazi University, Ankara, Türkiye

({ss, mdemirci, yilmazatay, muratdener, sedefgunduz, akara, samedal, selcuk.yilmaz1}@gazi.edu.tr)

‡Corresponding Author; Seref Sagiroglu, Department of Computer Engineering, Faculty of Engineering, Gazi University, Ankara, Türkiye, Tel: +90 312 582 3124, Email: ss@gazi.edu.tr

Received:01.02.2026 Accepted:07.06.2026

Abstract- Smart grids integrate advanced communication and control technologies to improve the efficiency and reliability of power delivery, but their interconnected and heterogeneous nature expands the attack surface. Recent attacks ranging from malware to false data injection have demonstrated the need for robust, layered defenses spanning physical devices, network infrastructure, control systems, and applications. Software-defined networking (SDN) offers centralized, programmable control over grid communications to dynamically isolate threats and reroute traffic during attacks. Securing SDN-enabled smart grids requires a defense-in-depth approach that combines diverse emerging technologies. This paper proposes a multi-layer conceptual security architecture leveraging blockchain for data integrity, explainable AI for transparent decision-making, generative AI for adaptive threat detection, quantum-safe cryptography for future resilience, big data analytics for anomaly detection, and radio frequency fingerprinting for device authentication. We present our model and discuss how each component contributes to a cohesive security architecture. We also identify open research challenges for integrating these solutions seamlessly to move toward more secure grids in the future.

Keywords Software-defined networking (SDN), cybersecurity, smart grid, generative AI, radio frequency fingerprinting

1. Introduction

A smart grid (SG) is an advanced electrical network that enables the bidirectional flow of both electricity and data between power generation facilities and end users through modern communication technologies [1]. The SG infrastructure consists of two fundamental layers: the power system layer and the communication layer. The power system layer encompasses key operations such as electricity generation, transmission, and distribution. Meanwhile, the communication layer serves as the backbone that enables seamless information exchange between these power system components [2].

Cite this article as: S. Sağıroğlu M. Demirci, Y. Atay, M. Dener, M.S. Demirci, A. Kara, S. Al, and S. Yılmaz, "A Multi-Layer Conceptual Security Architecture Toward Robust and Resilient SDN-Enabled Smart Grids" *International Journal of Smart Grid (ijSmartGrid)*, Vol. 10, No. 3, pp. 225-236, September, 2026.

Traditional smart grid communication networks typically rely on legacy technologies like Multiprotocol Label Switching (MPLS). However, these conventional approaches are not fully aligned with the dynamic nature of SGs. One major limitation is their static architecture—whenever a new service needs to be integrated, the network components must be manually reconfigured. This lack of adaptability makes expansion and innovation within the grid challenging. Additionally, since different energy providers manage and operate their own networking equipment, the infrastructure becomes increasingly fragmented as more vendors enter the market. This heterogeneity leads to interoperability issues, network inefficiencies, and an increased risk of service disruptions. Moreover, the fragmented architecture creates security vulnerabilities, making the grid susceptible to cyber threats [3].

Software-Defined Networking (SDN) emerges as a promising solution to these challenges by transforming the communication layer into a programmable network. It decouples network control from data forwarding devices, centralizing management through a software-based controller. This approach allows networks to be dynamically programmed and adapted to changing requirements [4]. With SDN, the grid can dynamically adjust to changes, optimize data flow, and enhance security, ultimately ensuring a more resilient and efficient SG communication framework. Fig. 1 illustrates how SDN is integrated into the communication infrastructure of the SG.

The architecture consists of three layers: data plane, control plane, and application plane. The data plane comprises grid components such as smart meters, sensors, and power management units (PMUs), as well as programmable networking hardware that forwards grid traffic based on instructions from the control plane [5]. The control plane is responsible for network intelligence and decision-making. It consists of one or multiple SDN controllers that establish traffic paths and forwarding rules, which are then programmed into SDN switches located in the data plane. The communication between the control and data planes is facilitated through the Southbound API, with OpenFlow being a common protocol for this purpose. This ensures secure message exchange between the control and data layers. At the highest level, the application plane hosts both smart grid and SDN-specific applications that perform key functions such as device management, voltage and frequency control, grid status monitoring, load balancing, traffic filtering, intrusion detection/prevention, and deep packet inspection. Through the Northbound API, applications dictate their policies to the controller, which then translates them into enforceable rules for SDN switches within the data plane [1].

Although SDN can enhance the security of smart grid architectures, its open nature introduces new security challenges for software-defined smart grid systems. In this paper, we propose a multi-layer security concept model for SDN-enabled smart grids. The rest of the paper is organized as follows: Section II describes our proposed multi-layer security concept model for SDN-enabled smart grids. Section III highlights open issues and future research directions in the field. Section IV provides a discussion of all new concepts for smart grids. Finally, Section V summarizes the paper and gives concluding remarks.

2. Proposed Multi-Layer Security Concept Model

A security framework for SDN-enabled smart grids must incorporate solutions at multiple layers within the network to be effective and comprehensive. To this end, we propose a multi-layer security concept model that addresses threats at each layer of the system architecture, as illustrated in Fig. 2.

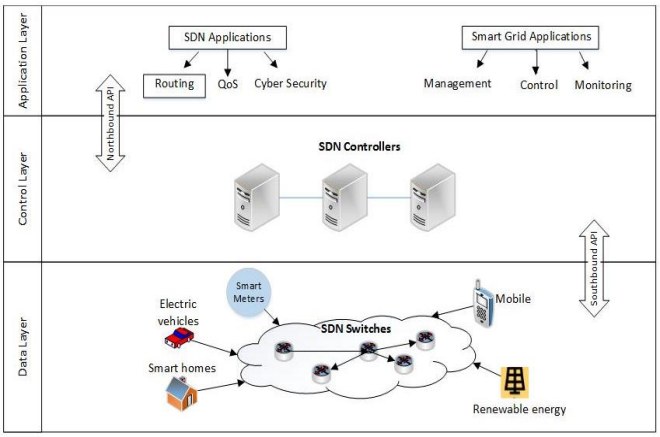


Fig. 1. Software-defined smart grid architecture (adapted from [1] and [6]).

The novelty of our framework can be explained in the following way:

- It incorporates security measures deployed at different layers to support defence in depth.
- It integrates SDN and NFV as central parts of the architecture.
- It focuses on emerging techniques such as radio frequency fingerprinting in the physical layer, which is often overlooked in existing proposals.

In this layered model, the smart grid's infrastructure is divided into the Physical/Perception layer, Network/Communication layer, Control layer, Application and Management layer for security governance. Each layer is secured using appropriate mechanisms, and cross-layer communication ensures that intelligence is shared between layers. This design follows the principle of defense in depth: even if an attacker penetrates one layer, additional layers of defense can contain the damage and protect grid functions.

Our proposed model should not be interpreted as a collection of independent security mechanisms, but as an integrated security architecture in which information and control flow continuously across layers. Security events detected at the Physical/Perception layer, such as the identification of a rogue device through radio frequency fingerprinting, are propagated to the Network/Communication layer, where the SDN controller can immediately update forwarding rules to isolate the affected device or redirect its traffic through additional security functions. Similarly, anomalies identified by

network monitoring or intrusion detection systems are communicated to the Control layer, allowing the controller to adapt routing policies, activate redundant communication paths, or deploy virtualized security services through NFV. At the Application and Management layer, information collected from all lower layers is aggregated by SIEM and analytics platforms, where AI-assisted correlation, blockchain-supported integrity verification, and risk assessment provide a comprehensive view of the system and generate coordinated response actions that are enforced throughout the architecture.

The proposed model therefore emphasizes orchestration rather than simply stacking security technologies. SDN serves as the central coordination mechanism that enables dynamic enforcement of security policies across all layers, while each emerging technology contributes a complementary capability. Radio frequency fingerprinting strengthens device identity assurance at the edge, blockchain provides integrity and traceability for critical operations, explainable and generative AI enhance intelligent detection and decision support, privacy-preserving analytics protect sensitive information, and quantum-safe cryptography prepares the communication infrastructure for future threats. By explicitly defining both the functional responsibilities of each layer and the interactions among them, the proposed model establishes a cohesive defense-in-depth architecture capable of adapting to evolving cyber threats while maintaining the operational requirements of smart grids. Each layer in our framework is further explained in the sections below.

2.1. Physical/Perception Layer

This layer includes field devices such as smart meters, sensors (e.g., PMUs), actuators, and intelligent electronic devices (IEDs) in substations. Securing the physical layer is crucial, as a compromised device can feed false data or disrupt grid operations at the source. Our model incorporates tamper-resistant hardware and physical security controls (like secure device enclosures and access restrictions) to reduce the risk of tampering. We also use device authentication mechanisms at this layer, including cryptographic authentication of device identities as well as emerging techniques like radio-frequency fingerprinting (RFF) as detailed in Sec. 3.8. Radio-based smart meters (sensors at the lowest level) transmitting data in real time is very typical. However, this may become challenging from a security perspective as rogue devices may be present. This could be adequately addressed in the physical layer with RFF, as illustrated in Fig. 3.

This configuration guarantees that all transmissions from the smart meters pass through the physical security barrier, i.e., RFF system, before reaching the upper layer systems. At the physical layer of smart grids, the star-of-stars network topology would be considered, which utilizes the fact that the data of all the sensors is concentrated at the relevant gateways. In this way, the RFF system hosted by the IoT gateways would serve as a barrier to not allowing readings from non-legitimate sources. Simply, it filters the rogue ones on a per-packet basis. By verifying that each sensor/actuator is legitimate and behaving as expected, the system can detect impostor or rogue devices injecting malicious signals. Moreover, lightweight anomaly detection agents can also run on local controllers, using machine learning to flag abnormal control actions or measurements at the source. These measures ensure that the data entering the grid from the physical layer is trustworthy, forming the first line of defense against attacks such as sensor spoofing, false data injection, and device impersonation.

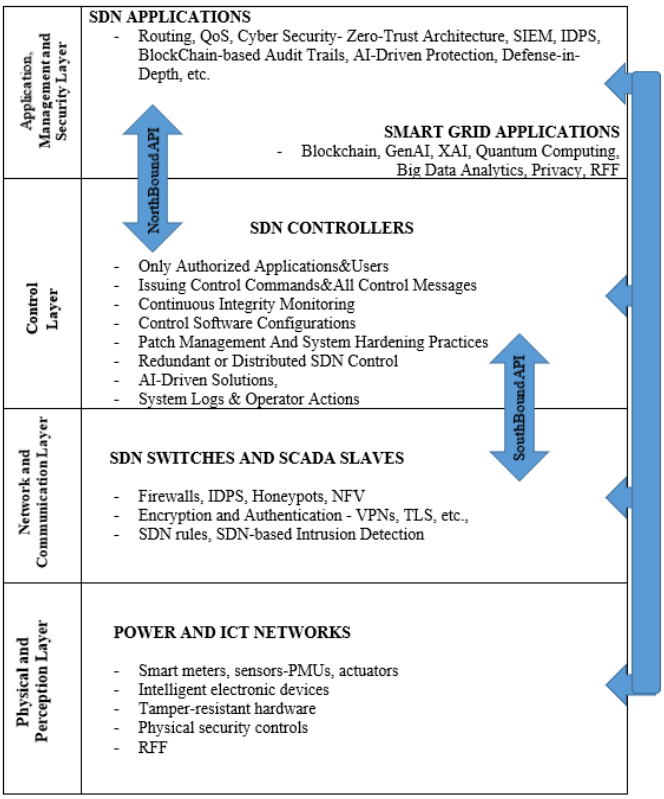


Fig. 2. A conceptual multi-layer security architecture for an SDN-enabled smart grid model (adapted from [1]-[5]).

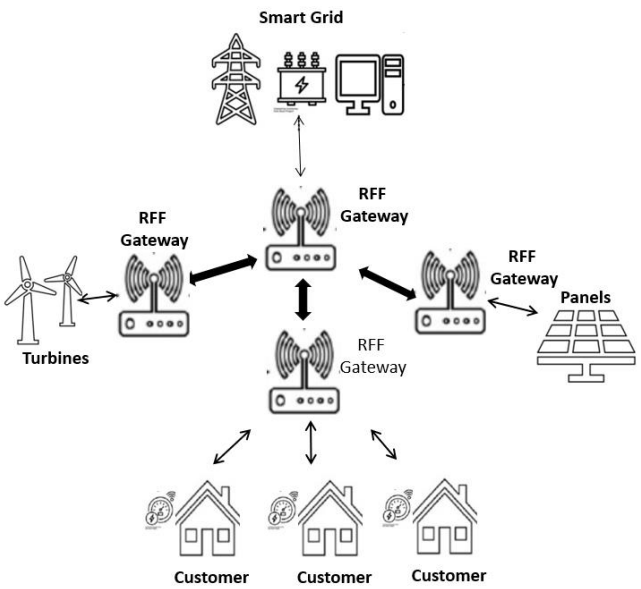


Fig. 3. Illustration of conceptual physical layer security (RFF) architecture for smart grid.

2.2. Network/Communication Layer

This layer comprises the communication links (wired or wireless) and networking hardware that connect grid devices and control centers. Here, the SDN controller plays a pivotal security role. All grid data traffic flows through SDN-enabled switches, which enforce network-wide security policies. Programmable switches are increasingly employed to improve flexibility, agility, and security [7, 8]. For instance, P4 switches can be leveraged to implement efficient defense against SYN flood attacks [9].

Our model uses the SDN controller to implement traffic whitelisting and segmentation, so that each device or subsystem only communicates with authorized endpoints. For example, a smart meter's data packets are only allowed to reach the metering database and not directly other meters, limiting lateral movement for an attacker. The controller also monitors traffic patterns for anomalies (e.g., sudden spikes that could indicate a DoS attack) and can dynamically reroute or block flows as needed. Communication links are secured with encryption and authentication (VPNs, TLS, etc.) to prevent eavesdropping and man-in-the-middle attacks. By segmenting and closely monitoring network traffic, the model contains malware propagation and mitigates threats like DDoS or routing attacks. If a substation network is compromised, SDN rules can isolate it from the rest of the grid while alternate paths maintain overall grid communication. This layered defense was demonstrated in the EU SDN-microSENSE project, which defined an SDN-based intrusion detection and self-healing framework to react to cyberattacks in real time. In essence, the network layer acts as the nerve center for detection and response, leveraging SDN's global view to enforce security across the grid [10].

2.3. Control Layer

The control layer refers to the central control systems and SDN controllers that manage grid operations. This includes SCADA master stations, distributed energy resource management systems, and the SDN controller itself (which links the communication network with grid applications). The SDN controller is also responsible for routing grid traffic through the security functions residing in the application plane. The decisions here can be optimized for minimizing link usage or energy consumption while preserving all security constraints. In order to do this, the controller software takes the topology of the SG communication network, the grid traffic matrix, along with their required security functions list that they must pass through, and a set of security function specifications as inputs. Then, it calculates paths for the flows and writes the calculated routing rules to the flow tables of the SDN switches. In this way, SG traffic can be steered through the desired points (e.g., trust nodes) if needed for in-path monitoring and security actions such as blocking, rate limiting, rerouting, etc. can be applied. Alternatively, if trust nodes are virtualized, they can be deployed dynamically at chosen locations. Placing these virtual trust nodes closer to the edge could be preferable for reduced latency, which is a common goal for most real-world SG deployments.

Securing the control plane is paramount because an attack here can potentially override grid commands or blind the operators to ongoing events. Our model ensures the integrity and availability of control layer components through several measures. First, strict access control and authentication are enforced for any interaction with the SDN controller or SCADA systems – only authorized applications and users can issue control commands, and all control messages are cryptographically signed to prevent forgery. Second, we incorporate continuous integrity monitoring for control software, using cryptographic checksums and anomaly detection on controller behavior (e.g., alert if the controller begins installing unexpected flow rules that deviate from its normal patterns). Third, robust patch management and system hardening practices are applied: keeping the controller and servers updated against known vulnerabilities, disabling unnecessary services, and running them in secure, isolated environments. Moreover, the model supports redundant or distributed SDN control: multiple synchronized controllers can take over if one fails or is compromised, significantly reducing the risk of a single point of failure. To detect sophisticated threats like stealthy malware or insider misuse, we utilize AI-driven solutions that analyse system logs and operator actions for anomalies (e.g., unusual sequence of commands at odd hours). In this way, our multi-layer model defends against attacks trying to hijack the grid's decision-making processes, such as false command injection, controller exploits, or timing manipulation attacks.

2.4. Application Management and Security Layer

The application layer encompasses the end-user services and grid management applications such as energy management systems, demand response programs, distributed energy resource (DER) trading platforms, and data analytics dashboards. It also includes databases storing smart meter data and other business systems. Furthermore, network security applications are integrated as virtualized functions in this layer, such as firewalls, intrusion detection/prevention systems (IDPS), and honeypots, which can be deployed on demand at strategic points via network function virtualization (NFV). The proposed framework is modular and extensible: New security functions can be added and updated independently as needed. In this way, shortcomings with regard to the security mechanisms can be addressed without the need to change the other layers or the general architecture.

Our model employs role-based access control and multi-factor authentication to ensure that only legitimate users and software services can access sensitive smart grid data or control functions. All transactions (e.g., energy trading records, meter reading updates) are logged on an immutable ledger to provide tamper-evident records of grid operations. This helps in post-incident forensics and in real-time detection of unauthorized changes. Big data analytics are heavily utilized in this layer to make sense of the massive volumes of grid data for both operational efficiency and

security insights. In our model, we integrate such analytic engines with an emphasis on explainability so that grid operators can validate and respond appropriately. We also embed privacy-preserving techniques in applications that handle customer data, ensuring compliance with regulations and protecting consumers. Our model also supports situational awareness dashboards that unify security information from all layers. By hardening applications and safeguarding data, the multi-layer model mitigates threats such as database breaches, phishing and social engineering, and data manipulation attempts.

Cutting across all the above layers is a dedicated management layer responsible for orchestrating security policies and responses – essentially the security brain of the architecture. This layer comprises components such as a Security Information and Event Management (SIEM) system, threat intelligence feeds, and the risk assessment and self-healing frameworks. It implements a holistic Zero-Trust Architecture (ZTA) throughout the grid. ZTA continuously verifies each user, device, and application action, rather than assuming any implicit trust based on network location. For example, even an internal control center user must re-authenticate and is only granted least-privilege access. The management layer also runs continuous risk assessments, evaluating the cybersecurity posture of grid assets in real time. If a high risk is identified for certain assets, the system can proactively adjust the controls. Correlation engines in this layer aggregate alerts from across the network using AI to recognize complex multi-stage attacks. Upon detecting a credible threat, the management layer can trigger the self-healing mechanisms: for instance, instructing the SDN controller to quarantine malicious traffic, re-configuring grid topology to maintain stability, or activating backup systems. By overseeing and unifying the defense strategy, the management layer ensures that the multi-layer defenses work in concert rather than in isolation and implements policies that span multiple layers.

Overall, the proposed multi-layer model creates a comprehensive security shield around the SDN-enabled smart grid. In the next section, we summarize how key security technologies are mapped onto this layered architecture, highlighting their benefits and challenges.

3. Open Research Challenges and Future Directions

This section examines the potential of technologies such as blockchain-based secure data sharing, big data analytics, quantum computing-enhanced encryption, XAI-driven transparent decision-making, and Generative Artificial Intelligence-enabled innovative data modeling in SDN-enabled smart grids.

3.1. Blockchain

In our multi-layer model, blockchain can serve as a distributed, tamper-proof ledger for recording events and transactions across the grid. For example, power generation data, grid control commands, and energy trading between prosumers can be logged in a blockchain to ensure transparency and integrity. Blockchain also introduces decentralization: critical security functions such as authentication or key management can be distributed among many nodes, reducing single-point-of-failure risks. Microgrid deployments have embraced blockchain for local energy trading to provide a secure and transparent platform for tracking energy transactions between homes or DER units. This decentralization aligns well with smart grid trends toward distributed generation and storage, enabling trust in transactions of power or data between entities that may not fully trust each other [11–13].

Recent studies have shown that blockchain can be integrated with SDN-based smart grids, cyber–physical systems (CPS), and IoT communication infrastructures to secure both control and data planes. In particular, blockchain has been used to maintain immutable records of SDN flow rules, ensuring consistency among distributed controllers and enabling version control of network policies [14,15]. By comparing hash values of flow tables stored on-chain with those deployed in switches, compromised or rogue network elements can be detected and isolated [16]. Furthermore, blockchain has been exploited to support collaborative intrusion detection and trustworthy data sharing in SDN-enabled CPS environments. Experiments show that such blockchain-assisted cooperative IDS frameworks achieve higher robustness and detection performance compared to conventional approaches [17].

Blockchain-supported SDN frameworks have also been applied to peer-to-peer energy trading and vehicle-to-grid (V2G) scenarios, where transactions are validated in a decentralized manner, thereby reducing latency and network overhead while preventing fraudulent trades and improving trust among participating entities [18, 19]. In addition, blockchain-based SDN communication architectures have been proposed to protect inter-controller communication and to guarantee secure dissemination of control messages across distributed smart grid domains [20].

Integrating blockchain into smart grids faces research challenges, primarily regarding scalability and performance: current public blockchain networks struggle with high throughput and low latency requirements. Processing all events generated by a national-scale smart grid in a blockchain could introduce delays that are unacceptable for real-time grid control. Solutions such as permissioned blockchains operating among grid stakeholders with more efficient consensus mechanisms (e.g., Proof-of-Authority or Practical Byzantine Fault Tolerance) are being explored to improve throughput [14, 16, 17, 19]. Another challenge is the energy and resource cost of blockchain mechanisms. Some consensus algorithms (e.g., Proof-of-Work) are too computationally intensive for the energy efficiency goals of smart grids. Moreover, data privacy on blockchains remains an open issue and techniques such as on-chain hashing with off-chain storage, blockchain-based access control, and privacy-preserving ledgers are proposed to reconcile transparency with confidentiality [17, 19, 20]. Therefore, further research is required to design lightweight consensus mechanisms and hierarchical or edge-assisted blockchain architectures that can meet the stringent performance requirements of future smart grids.

3.2. Explainable AI (XAI)

Artificial intelligence (AI) and machine learning (ML) are increasingly being used in smart grids for applications like load forecasting, fault detection, and cybersecurity. However, most AI models – especially powerful ones like deep learning (DL) – operate as "black boxes" that do not readily explain why they made a certain decision or prediction. In a critical infrastructure like the power grid, this lack of transparency is problematic. XAI addresses this challenge by providing techniques that make AI decisions more interpretable to humans. For smart grid security, XAI is crucial because it can bridge the gap between complex AI-driven defenses and the operators who oversee the grid by providing insights that allow a human operator to verify an alert and to understand the nature of the threat. It also helps in compliance and governance – regulators may require utilities to provide reasoning for automated decisions that affect customers. From a technical perspective, XAI can identify biases or blind spots in AI models by revealing their focus.

The challenge is that achieving explainability often comes at a cost. Simpler ML models are more interpretable but may not capture the complexity of grid data as well as DL models. Conversely, advanced models can achieve high accuracy but are intrinsically hard to interpret. In a smart grid, another difficulty is the scale and interdependency of the data – an AI's decision might be influenced by hundreds of input points, making comprehensive explanation a non-trivial task. XAI should produce concise real-time insights to help accelerate the adoption of sophisticated ML in smart grid operations. Moving forward, collaboration between AI researchers and power engineers is essential to develop XAI tools tailored for the grid's needs, potentially setting industry standards on how AI decisions in smart grids should be documented and audited.

3.3. Generative AI

Generative Artificial Intelligence (GenAI), particularly Large Language Models (LLMs) and Generative Adversarial Networks (GANs), has rapidly advanced in recent years and introduces both promising capabilities and unresolved challenges for smart grid cybersecurity. Generative AI refers to models that can create new data instances resembling their training data, enabling them to generalize beyond fixed patterns and adapt to evolving conditions. In the context of SDN-enabled smart grids, where communication flows, control logic, and attack surfaces are highly dynamic, such adaptability is especially attractive. One key security application of generative AI is the enhanced detection of novel or evolving cyber threats. LLMs can be fine-tuned to analyse heterogeneous data sources, including SDN controller logs, network traffic flows, SCADA events, and operational measurements, to identify suspicious behavior or infer coordinated attack patterns. Unlike traditional machine learning models that often require retraining when attackers change tactics, LLMs may generalize from broader contextual knowledge to flag unusual correlations or deviations from normal operational behavior [21, 22].

Beyond automated detection, generative AI enables natural language interaction between human operators and security systems. An LLM-based assistant could allow grid operators to pose high-level queries, such as whether there are indications of coordinated attacks on substation automation systems within a specific time window, and receive synthesized responses supported by relevant evidence extracted from logs and telemetry. Such capabilities may improve situational awareness and decision support, particularly during complex multi-stage attacks, while reducing the cognitive burden on human operators [23]. In addition, generative models can be leveraged for realistic synthetic data generation. GANs can produce simulated cyber-attack traffic, anomalous grid measurements, or false data injection scenarios that are difficult or risky to obtain from real-world smart grid deployments. These synthetic datasets can be used to train and evaluate intrusion detection and anomaly detection systems under a wider range of adversarial conditions, thereby improving robustness and generalization [24, 25].

Despite these potential benefits, the integration of generative AI into smart grid security architectures raises several open research challenges. A fundamental concern is the reliability and trustworthiness of LLM outputs. Generative models are often characterized as black boxes and may hallucinate plausible but incorrect information if insufficiently constrained or queried outside their effective knowledge domain. In a safety-critical environment such as a smart grid, fabricated alerts, inaccurate explanations, or misleading recommendations could lead to false alarms, unnecessary control actions, or delayed responses to genuine attacks [26]. Addressing this issue requires further research into constrained generation, uncertainty estimation, explainability mechanisms, and human-in-the-loop validation strategies.

Another unresolved issue is the computational and architectural feasibility of deploying generative models in smart grids. Large-scale LLMs typically require significant computational resources and may be impractical for continuous, real-time analysis at the grid edge or within substations. Running such models centrally may also introduce latency and single points of failure. Research into model distillation, quantization, hierarchical edge–fog–cloud deployment strategies, and event-driven or on-demand use of generative AI is necessary to balance performance, scalability, and resilience [27].

Finally, generative AI itself introduces new attack vectors. Adversaries may attempt to manipulate model behavior through data poisoning, adversarial prompts, or prompt injection attacks, potentially influencing security decisions or obscuring malicious activity. Understanding the robustness of generative AI under adversarial conditions and developing defenses against AI-specific attacks remain open research problems, particularly in the context of critical infrastructure protection [28].

3.4. Privacy

Preserving privacy in smart grid environments remains a significant open research challenge due to the inherently

heterogeneous nature of grid infrastructures and the widespread deployment of low-power, resource-constrained devices. Smart grids continuously collect and transmit large volumes of fine-grained data, including consumption patterns, device status, and operational telemetry, much of which may be considered sensitive or personally identifiable. Implementing comprehensive privacy-preserving frameworks, such as consent management mechanisms, encryption of personal data, access control policies, and data minimization strategies, inevitably increases system complexity and may introduce additional computational, communication, and latency overhead. These trade-offs are particularly problematic in time-critical grid operations, where delays or performance degradation can negatively impact stability and reliability.

Big data security analytics further complicates the privacy landscape. While large-scale data aggregation is essential for detecting anomalies, coordinated attacks, or long-term trends, indiscriminate collection and centralized analysis of raw data can raise serious privacy and regulatory concerns. Without appropriate safeguards, such practices may be perceived as overly intrusive by consumers or may conflict with data protection regulations, leading to public resistance or regulatory intervention. Consequently, security analytics must be designed to operate on aggregated and anonymized data whenever possible, or to ensure that explicit and transparent consent mechanisms are in place. Balancing the need for detailed security visibility with privacy preservation remains a non-trivial design challenge in smart grid architectures.

Federated learning has emerged as a promising and cost-effective approach to addressing these concerns. By enabling machine learning models to be trained locally on edge devices or regional nodes, federated learning minimizes the need to transmit raw data to centralized processing facilities. Instead, only model updates or gradients are shared, significantly reducing the risk of sensitive data exposure while also lowering network load and communication latency. This decentralized learning paradigm aligns well with the distributed nature of smart grids and supports scalability across large numbers of heterogeneous devices [1, 29]. However, deploying federated learning in smart grid security contexts introduces its own open issues. Challenges include handling non-independent and non-identically distributed (non-IID) data across devices, ensuring robustness against compromised or malicious participants, and managing the additional coordination and synchronization overhead required for model aggregation. Machine learning models that prioritize local learning and minimize data sharing can achieve a more favorable balance between security effectiveness, privacy preservation, and system performance. Therefore, prioritizing federated learning–based approaches in the development of security solutions is well aligned with the requirements of a secure, privacy-conscious smart grid platform [1, 29, 30].

3.5. Adversarial AI

Many common security applications rely on machine learning models, which can be directly targeted by adversarial attacks. The ability of a model to maintain good performance under adversarial attacks is called adversarial robustness [31]. While there are emerging solutions to improve the robustness of ML models, this is still a fairly nascent area and new attack methods are also being proposed to cripple naïve models. In recent years, researchers have proposed various methods to improve the robustness of ML-based IDPSs and other security systems [32, 33]. The proposed framework incorporates robust IDPS models from recent literature. Metrics and methodologies used in robustness evaluation must be determined and applied to any security solution at this layer before its deployment.

Poisoning attacks are generally considered highly dangerous because they compromise the fundamental decision-making process of ML models. A modern smart grid contains several points where careful injection of fake or falsely labeled data can cause crucial issues such as inaccurate forecasting, weakened anomaly detection, and wasteful or insufficient planning [34]. The adverse effects of poisoning attacks may linger for months after the contamination of the training data, and in the case of repeated poisoning, persistent unreliability and inefficiency become the norm for the grid. Best practices must be applied to combat poisoning, including the validation and sanitization of training data, and continuous monitoring of model inputs and outputs to detect concept drift before it leads to the normalization of subpar or dangerous operating conditions [35].

3.6. Quantum-Safe Security

Quantum computing poses a fundamental threat to existing cryptographic mechanisms used in smart grids, as quantum algorithms such as Shor's algorithm will be able to break widely deployed public-key systems including RSA, Diffie–Hellman, and ECC. This has motivated extensive research on post-quantum cryptography (PQC) and quantum-based security mechanisms for future grid infrastructures. In this context, a Self-Defensive Post-Quantum Blockchain Architecture (SD-PQBA) has been proposed to protect SCADA systems against both classical cyberattacks and quantum-enabled cryptographic threats. This architecture integrates lattice-based and code-based cryptographic schemes with a novel Proof of Derived Authority (PoDA) consensus mechanism and employs a three-key (primary–private–public) cryptographic structure along with Quantum Key Distribution (QKD) to ensure data integrity and immutability. The proposed framework demonstrates resilience against false data injection, man-in-the-middle, denial-of-service, and malware attacks while maintaining low computational overhead suitable for real-time SCADA environments [36].

Lattice-based cryptography has emerged as a core building block for quantum-resistant smart grid security. A comprehensive survey of lattice-based security and authentication schemes highlights that cryptosystems based on hard lattice problems provide strong resistance against both classical and quantum attacks. The study further analyzes lattice-based protocols for key exchange, authentication, secure data aggregation, and privacy-preserving communication in IoT-enabled smart grid networks, showing that these approaches offer a favorable

balance between security strength and computational/communication cost compared to traditional public-key cryptography [37].

In addition to software-based cryptographic solutions, quantum cryptography offers unconditional security through the physical properties of quantum mechanics. A quantum-secure data transmission model combines Quantum Key Distribution (QKD) with classical One-Time Pad (OTP) encryption to protect smart meter data exchanged with utility providers. By utilizing Quantum Random Number Generators (QRNG), the model generates truly random encryption keys, eliminating the predictability of pseudo-random number generators and enabling the detection of eavesdroppers through the quantum no-cloning theorem. Experimental evaluations conducted on both classical simulators and real quantum hardware using the IBM Qiskit platform demonstrate that the proposed approach achieves lower computational cost than DES, AES, and RSA while providing stronger security guarantees. These results confirm the practical feasibility of deploying QKD- and QRNG-based secure communication channels in smart grids [38].

Beyond key generation and encryption, the reliable distribution of quantum resources across grid communication networks is also critical. An SDN-based entanglement routing framework supported by reinforcement learning has been proposed to enable distributed quantum sensing and entanglement-based key distribution in smart power grids. In this architecture, the SDN control plane maintains a global view of network conditions, including link noise levels and available qubit resources, and dynamically selects optimal entanglement distribution paths. Simulation results on an IEEE-14 bus power grid topology show that the reinforcement learning-based routing strategy achieves a 100% success rate for entanglement delivery under degraded link conditions, significantly outperforming traditional shortest-path routing algorithms [39]. This demonstrates that intelligent SDN-assisted control is essential for sustaining quantum-secure communications in realistic grid scenarios.

Despite these advances, challenges remain for the large-scale deployment of quantum-safe security in smart grids. The integration of QKD infrastructure and quantum servers into existing communication networks introduces cost, deployment, and interoperability constraints. Moreover, hybrid security architectures that combine PQC with QKD will be required to support heterogeneous grid devices with limited computational capabilities. Therefore, future research must focus on lightweight post-quantum protocols, SDN-assisted quantum resource management, and gradual migration strategies toward fully quantum-safe smart grid communication infrastructures.

3.7. Big Data Analytics

The modern smart grid is characterized by an explosion of data, from high-resolution sensor measurements to millions of smart meters logging usage data, to network devices recording traffic flows. If harnessed effectively, big data analytics can significantly improve situational awareness and threat detection in the grid. By aggregating and correlating data across the electrical, IT, and physical domains, operators can identify anomalies or precursors to attacks that would be impossible to see in isolated data streams [40]. For example, a slight fluctuation pattern in voltage combined with unusual network traffic and some out-of-bounds control commands might together signal a coordinated cyber-physical attack.

Big data platforms can handle multi-source data streams in near real-time, applying ML or rule-based analytics to flag such complex events. In our multi-layer security model, a SIEM system enriched with big data capabilities can ingest logs and telemetry from all layers and produce a unified threat picture for analysts. However, realizing these benefits is challenging due to the classic “4 Vs” of big data (volume, velocity, variety, and veracity). The volume of data in a nationwide smart grid can be petabytes per day; storing and processing this cost-effectively requires scalable infrastructure and clever data reduction techniques. Velocity is also high – data streams need to be processed in seconds for effective intrusion detection, which calls for real-time stream processing engines rather than batch analysis. Variety is extreme: data formats range from time-series sensor data and IP network logs to unstructured data like maintenance reports or even social media. Ensuring data veracity (quality and trustworthiness) is non-trivial – sensors can be faulty or calibrated differently, and an attacker may try to feed false data to confound analytics. In short, big data analytics is a cornerstone for the future of grid security.

3.8. Enabling Technologies: Radio Frequency Fingerprinting

The nascent concept of radio frequency fingerprinting (RFF) for device identity has an important role to play in smart grid security. RFF is a signal intelligence method used to identify radio transmitting devices by extracting their unique characteristics. These unique features, or fingerprints, result from manufacturing imperfections in components like amplifiers, filters, and clock generators. These imperfections cause variations in signal characteristics such as phase offset, IQ imbalance, and clock skew. RFF enhances cybersecurity by providing a potential authentication technique for low-power, computation-constrained devices in next-generation networks, especially with the proliferation of IoT devices [41]. Unlike cryptographic methods with high computational overhead, RFF leverages a device’s unique physical characteristics for lightweight authentication. While earlier RFF techniques relied on human-extracted features, the current trend is to employ AI, particularly DL, for feature extraction and classification [42]. From a cybersecurity perspective, RFF offers a robust layer of physical-layer security [43]. It can be used as an access control and anti-fraud measure, effectively identifying and preventing unauthorized devices from accessing network resources. This is particularly important for critical infrastructures like smart grids, where the cybersecurity of IoT devices is paramount [44, 45]. As a result, RFF represents forward-looking elements of grid security and adds an innovative layer of device integrity checking that complements digital authentication. By addressing both who can see data (privacy) and what a device physically is (fingerprinting), we cover dimensions often overlooked in traditional solutions,

moving closer to a comprehensive security and trust framework for future SDN-enabled smart grids.

Current literature heavily focuses on DL-based RFF methods, demonstrating relatively high performances, and seems to be the standard approach. However, the practical implementation of RFF may still face significant challenges. For example, the long-term reliability of RFF is questionable due to hardware aging. Degradation in functionalities of electronic components, for instance, crystal oscillators, leads to slowly varying features resulting in alterations of device fingerprints. Questions regarding classification accuracy and the sustainability of the RFF system have been raised. In particular, the long-term impact of aging is often ignored or not modelled in existing literature, which results in a critical gap in understanding the long-term viability of the RFF systems for sustainable IoT security. Another challenge could be feature drift, introducing a more complex conceptual ambiguity for RFF-authentication-based security. If device fingerprints alter and evolve due to aging and environmental interference, then long-term device discrimination may be inherently dynamic rather than static. These issues challenge widespread assumptions about permanent hardware identifiers. It also suggests that both security guarantees and privacy risks need to be considered in a time-varying context. Consequently, RFF should be analysed as both a static biometric identifier and a dynamic signal feature subject to physical evolution.

4. Discussion

The explanations in this study highlight that resilience in SDN-enabled smart grids cannot be achieved through isolated security mechanisms alone; instead, it depends on a coordinated defense-in-depth architecture spanning physical, cyber, and operational domains. By structuring protections across the Physical/Perception, Network/Communication, Control, and Application/Management layers, the proposed framework positions SDN's centralized programmability as a cohesive security orchestration mechanism rather than merely a network optimization tool. This layered integration enhances resistance to false data injection, man-in-the-middle attacks, denial-of-service disruptions, and malware propagation while preserving the strict latency requirements inherent in real-time SCADA operations.

At both the network and control layers, SDN introduces a structural shift from passive connectivity toward active, policy-driven enforcement. Programmable forwarding elements enable detection and mitigation functions to operate directly within the data plane, reducing response latency and supporting early containment of volumetric attacks through dynamic isolation and rerouting. Simultaneously, the integrity of the control layer—whose compromise could destabilize the entire system—must be continuously verified rather than implicitly trusted. Multi-factor authentication, cryptographic validation of control messages, runtime integrity attestation, and controller redundancy collectively reduce systemic vulnerability, while AI-assisted behavioural monitoring helps identify subtle misuse or insider threats. Although these safeguards introduce operational overhead, SDN's automation capabilities embed security enforcement within routine control workflows, preserving responsiveness.

Security at the physical edge further underscores the need for distributed verification. Deploying radio frequency fingerprinting at IoT gateways provides an additional identity assurance mechanism for resource-constrained devices, complementing traditional cryptographic approaches. However, hardware signatures evolve due to environmental variation and component aging, necessitating adaptive enrolment strategies and multi-factor validation—particularly in open-set environments where previously unseen devices must be assessed without prior reference.

At higher operational layers, visibility and governance must be balanced with privacy considerations. Role-based access control, tamper-evident logging, and explainable analytics strengthen accountability while enabling timely response within a zero-trust management paradigm. In parallel, blockchain technologies offer targeted value in scenarios where provenance and distributed trust are essential, such as verifying policy consistency across controllers or enabling secure information exchange in decentralized energy ecosystems. Their deployment in time-sensitive environments, however, must be carefully engineered through permissioned architectures and hybrid on-chain/off-chain designs to avoid compromising operational responsiveness. Privacy-preserving analytics and federated learning further ensure that enhanced situational awareness does not come at the cost of excessive data centralization.

The integration of advanced analytics introduces both opportunities and constraints. Generative AI, when bounded by explainability, uncertainty estimation, and human oversight, can synthesize heterogeneous operational data and support scenario-based training; yet uncontrolled deployment risks latency, opacity, or erroneous outputs. Similarly, quantum-resilient cryptographic approaches represent a gradual transition rather than an immediate replacement of existing mechanisms. Lattice-based post-quantum schemes provide near-term protection, while selective use of quantum key distribution may enhance confidentiality in critical channels. SDN's ability to orchestrate both classical and emerging quantum resources reinforces its role as a unifying coordination layer.

Big-data analytics extend cross-domain awareness by correlating electrical, cyber, and operational signals, transforming raw telemetry into actionable intelligence despite challenges related to scale, velocity, and reliability. Within this framework, advanced SIEM platforms evolve into genuine decision-support systems rather than mere alert aggregators.

Implementation must therefore proceed incrementally. Hybrid validation environments combining simulation, hardware-in-the-loop testing, and field pilots are essential to quantify resilience gains using measurable indicators such as recovery time, detection accuracy, and service continuity. Early deployment efforts should prioritize segmentation, gateway-level identity assurance, and lightweight auditing before progressing toward AI-assisted decision support or quantum-secure communication. Variations in technological maturity, legacy integration constraints, and workforce readiness remain practical limitations, while the long-term

stability of physical-layer identification methods and the operational feasibility of blockchain and quantum solutions warrant further empirical investigation.

Overall, these findings support a convergence-oriented security strategy in which SDN provides the structural foundation for coordinated enforcement, blockchain enhances integrity and traceability, advanced analytics improve interpretability, and emerging cryptographic approaches prepare infrastructures for future threat landscapes. The effectiveness of this approach lies not in any single technology but in its deliberate alignment under measurable performance and privacy constraints, enabling smart grids to strengthen resilience while maintaining the reliability and responsiveness essential to critical infrastructure.

5. Conclusion

The evolution of power grids into smart infrastructures requires equally smart and adaptive security strategies. In this paper, we presented a multi-layer security concept model for SDN-enabled smart grids that incorporates a spectrum of emerging technologies – from blockchain and XAI to quantum cryptography and device fingerprinting – each reinforcing different layers of defense. The central theme is defense-in-depth: a coordinated architecture of complementary measures to significantly strengthen grid security. At the network level, SDN provides agility and oversight for threat response, while at the data and control levels, AI-driven analytics and blockchain-enabled integrity create a trustworthy environment for decision-making. Physical layer and human factor security are strengthened through innovations like RFF and privacy-preserving data handling. We have also discussed open research challenges in integrating these technologies, such as scalability in blockchain, explainability in AI, quantum resistance, managing big data, and maintaining privacy.

To achieve effective integration, emerging technologies should not be seen in isolation but as parts of a cohesive model. This convergence of technologies, orchestrated in an SDN framework, offers a promising route to proactively defend against both current and future threats. The main challenges include technical maturity, interoperability, and the need for upskilling the workforce to manage these advanced tools. However, each challenge is an opportunity for innovation, from pilot deployments of XAI in grid operation centers to field tests of RF fingerprinting for grid IoT devices. We hope that this work contributes to the development of comprehensive security solutions for the grids of tomorrow.

Acknowledgements

This work has been supported by YÖK ADEP through the Gazi University Scientific Research Projects Coordination Unit under grant number FGA-2022-8043.

Author Contributions

S. Sagiroglu: Supervision, project administration, conceptualization, methodology, validation.
M. Dener: Project administration, writing- review & editing.
A. Kara: Formal analysis, visualization, writing-review.
M. S. Demirci: Formal analysis, visualization, writing-review.
M. Demirci: Formal analysis, visualization, writing- review & editing, validation.
S. Yilmaz: Writing-original draft,
Y. Atay: Writing- review & editing.
S. Al: Writing-original draft.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] S. Demirci and S. Sağıroğlu, "Software-defined networking for improving security in smart grid systems," in *Proc. 2018 7th International Conference on Renewable Energy Research and Applications (ICRERA)*, Paris, France, 2018, pp. 1021–1026, doi: 10.1109/ICRERA.2018.8567005.

[2] Y. Su, P. Jiang, H. Chen, and X. Deng, "A QoS-guaranteed and congestion-controlled SDN routing strategy for smart grid," Applied Sciences, vol. 12, no. 15, p. 7629, Aug. 2022, doi: 10.3390/app12157629.

[3] M. H. Rehmani, A. Davy, B. Jennings, and C. Assi, "Software defined networks-based smart grid communication: a comprehensive survey," IEEE Communications Surveys & Tutorials, vol. 21, no. 3, pp. 2637–2670, thirdquarter 2019, doi: 10.1109/COMST.2019.2908266.

[4] D. Kreutz, F. M. V. Ramos, P. E. Veríssimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: a comprehensive survey," Proceedings of the IEEE, vol. 103, no. 1, pp. 14–76, Jan. 2015, doi: 10.1109/JPROC.2014.2371999.

[5] P. Ding, J. Li, L. Wang, M. Wen, and Y. Guan, "HYBRID-CNN: an efficient scheme for abnormal flow detection in the SDN-based smart grid," Security and Communication Networks, vol. 2020, p. 8850550, Aug. 2020, doi: 10.1155/2020/8850550.

[6] S. Demirci, "QoS-Aware Routing Approaches in Software-Defined Smart Grids," in Proc. 9th Int. Conf. Comput. Sci. Eng. (UBMK), 2024, pp. 982–987, doi: 10.1109/UBMK63289.2024.10773500.

[7] F. Hauser, M. Häberle, M. Schmidt, and M. Menth, "A survey on data plane programming with P4: fundamentals, advances, and applied research," Journal of Network and Computer Applications, vol. 212, p. 103561, Mar. 2023, doi: 10.1016/j.jnca.2022.103561.

[8] A. AlSabeh, J. Khoury, E. Kfoury, J. Crichigno, and E. Bou-Harb, "A survey on security applications of P4 programmable switches and a STRIDE-based vulnerability assessment," Computer Networks, vol. 207, p. 108800, Apr. 2022, doi: 10.1016/j.comnet.2022.108800.

[9] M. E. Sahin and M. Demirci, "ConPoolUBF: connection pooling and updatable Bloom filter based SYN flood

defense in programmable data planes," Computer Networks, vol. 231, p. 109802, Jul. 2023, doi: 10.1016/j.comnet.2023.109802.

[10] P. Radoglou-Grammatikis, P. Sarigiannidis, T. Lagkas, and I. Moscholios, "SDN-based resilient smart grid: the SDN-microSENSE architecture," Digital, vol. 1, no. 4, pp. 173–187, Nov. 2021, doi: 10.3390/digital1040013.

[11] Y. Y. Ghadi, T. Mazhar, T. Shahzad, M. A. Haq, and E. M. Tag-ElDin, "A hybrid AI-blockchain security framework for smart grids," Scientific Reports, vol. 15, no. 1, p. 20882, Feb. 2025, doi: 10.1038/s41598-025-05257-w.

[12] R. Alsaigh, R. Mehmood, and I. Katib, "AI explainability and governance in smart energy systems: A review," Frontiers in Energy Research, vol. 11, p. 1071291, Feb. 2023, doi: 10.3389/fenrg.2023.1071291.

[13] K. Singh and L. Sevukamoorthy, "Blockchain and AI-based threat detection for enhanced security in financial networks," in Proc. IEEE TEMSCON-ASPAC, Bengaluru, India, 2023, pp. 1–5, doi: 10.1109/TEMSCON-ASPAC59527.2023.10531316.

[14] P. Kumar, R. Kumar, A. Aljuhani, D. Javeed, A. Jolfaei, and A. N. Islam, "Digital twin-driven SDN for smart grid: A deep learning integrated blockchain for cybersecurity," Solar Energy, vol. 263, p. 111921, Oct. 2023, doi: 10.1016/j.solener.2023.111921.

[15] M. J. Islam, A. Rahman, S. Kabir, M. R. Karim, U. K. Acharjee, M. K. Nasir, and S. Wu, "Blockchain-SDN-based energy-aware and distributed secure architecture for IoT in smart cities," IEEE Internet of Things Journal, vol. 9, no. 5, pp. 3850–3864, Mar. 2022, doi: 10.1109/JIOT.2021.3107834.

[16] A. Rahman, M. J. Islam, A. Montieri, M. K. Nasir, M. M. Reza, S. S. Band, and A. Mosavi, "Smartblock-sdn: An optimized blockchain-sdn framework for resource management in IoT," IEEE Access, vol. 9, pp. 28361–28376, Feb. 2021, doi: 10.1109/ACCESS.2021.3058225.

[17] W. Li, Y. Wang, and J. Li, "A blockchain-enabled collaborative intrusion detection framework for SDN-assisted cyber-physical systems," International Journal of Information Security, vol. 22, no. 5, pp. 1219–1230, Oct. 2023, doi: 10.1007/s10207-023-00688-6.

[18] R. Chaudhary, A. Jindal, G. S. Aujla, S. Aggarwal, N. Kumar, and K. K. R. Choo, "BEST: Blockchain-based secure energy trading in SDN-enabled intelligent transportation system," Computers & Security, vol. 85, pp. 288–299, Aug. 2019, doi: 10.1016/j.cose.2019.05.006.

[19] A. Jindal, G. S. Aujla, and N. Kumar, "SURVIVOR: A blockchain based edge-as-a-service framework for secure energy trading in SDN-enabled vehicle-to-grid environment," Computer Networks, vol. 153, pp. 36–48, Apr. 2019, doi: 10.1016/j.comnet.2019.02.012.

[20] U. Ghosh, L. Njilla, S. Shetty, and C. A. Kamhoua, "A decentralized smart grid communication framework using SDN-enabled blockchain," in Proc. 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 2024, pp. 982–985, doi: 10.1109/CCNC57856.2024.10454790.

[21] H. Xu, S. Wang, N. Li, K. Wang, Y. Zhao, K. Chen, and H. Wang, "Large language models for cyber security: A systematic literature review," ACM Transactions on Software Engineering and Methodology, vol. 33, no. 6, pp. 1–38, Jul. 2024, doi: 10.1145/3648438.

[22] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," Transactions on Emerging Telecommunications Technologies, vol. 32, no. 1, p. e4150, Jan. 2021, doi: 10.1002/ett.4150.

[23] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, Sep. 2018, doi: 10.1109/ACCESS.2018.2870052.

[24] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 4, pp. 3313–3332, Apr. 2023, doi: 10.1109/TKDE.2021.3130191.

[25] M. Ozay, I. Esnaola, F. T. Yarman Vural, S. R. Kulkarni, and H. V. Poor, "Machine learning methods for attack detection in the smart grid," IEEE Transactions on Neural Networks and Learning Systems, vol. 27, no. 8, pp. 1773–1786, Aug. 2016, doi: 10.1109/TNNLS.2015.2474845.

[26] R. Zhang, H. W. Li, X. Y. Qian, W. B. Jiang, and H. X. Chen, "On large language models safety, security, and privacy: A survey," Journal of Electronic Science and Technology, vol. 23, no. 1, p. 100301, Mar. 2025, doi: 10.1016/j.jnlest.2024.100301.

[27] X. Yan and Y. Ding, "Are We There Yet? A Measurement Study of Efficiency for LLM Applications on Mobile Devices," in Proc. 2nd International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMCPS-IoT '25), May 2025, pp. 19–24, doi: 10.1145/3715015.3725062.

[28] I. Naik, D. Naik, and N. Naik, "Threat Landscape of Adversarial Attacks on Generative AI and Large Language Models (LLMs): Exploring Different Types of Adversarial Attacks, Associated Risks, and Mitigation Strategies," Authorea Preprints, Jan. 2025, doi: 10.22541/au.173711927.42438865/v1.

[29] K. Chahine, M. A. Sayed, A. Al-Ghadhban, and A. Al-Mogren, "Toward secure smart grid systems: risks, threats, challenges, and future directions," Future Internet, vol. 17, no. 7, p. 318, Jul. 2025, doi: 10.3390/fi17070318.

[30] H. Sun, R. Q. Hu, and Y. Qian, "Towards enhancing security for upcoming 6G-ready smart grids: a

hierarchical architecture," Cybersecurity, vol. 5, no. 1, p. 14, Dec. 2022, doi: 10.1186/s42400-022-00124-4.

[31] N. Abdulla, S. Demirci, M. Demirci, and S. Özdemir, "Adaptive learning in IoT-based smart city applications," in Principles and Applications of Adaptive Artificial Intelligence, Hershey, PA, USA: IGI Global, 2024, pp. 73–101, doi: 10.4018/979-8-3693-1950-5.ch004.

[32] A. Sarikaya, B. G. Kılıç, and M. Demirci, "RAIDS: robust autoencoder-based intrusion detection system model against adversarial attacks," Computers & Security, vol. 135, p. 103483, Dec. 2023, doi: 10.1016/j.cose.2023.103483.

[33] M. K. Roshan and A. Zafar, "Boosting robustness of network intrusion detection systems: a novel two phase defense strategy against untargeted white-box optimization adversarial attack," Expert Systems with Applications, vol. 249, p. 123567, Sep. 2024, doi: 10.1016/j.eswa.2024.123567.

[34] Y. Zhu, H. Wen, R. Zhao, Y. Jiang, Q. Liu, and P. Zhang, "Research on data poisoning attack against smart grid cyber–physical system based on edge computing," Sensors, vol. 23, no. 9, p. 4509, May 2023, doi: 10.3390/s23094509.

[35] A. Alajaji, "FortiNIDS: Defending smart city IoT infrastructures against transferable adversarial poisoning in machine learning-based intrusion detection systems," Sensors, vol. 25, no. 19, p. 6056, Oct. 2025, doi: 10.3390/s25196056.

[36] M. T. Naz, W. Elmedany, and M. Ali, "Securing scada systems in smart grids with iot integration: A self-defensive post-quantum blockchain architecture," Internet of Things, vol. 28, p. 101381, Dec. 2024, doi: 10.1016/j.iot.2024.101381.

[37] H. Shekhawat and D. S. Gupta, "A survey on lattice-based security and authentication schemes for smart-grid networks in the post-quantum era," Concurrency and Computation: Practice and Experience, vol. 36, no. 14, p. e8080, Jun. 2024, doi: 10.1002/cpe.8080.

[38] R. Rani, A. K. Singh, A. Parashar, and D. Saxena, "Quantum-Secure Data Transmission in Smart Grid," in Proc. 2024 IEEE 17th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), Singapore, 2024, pp. 475–481, doi: 10.1109/MCSoC64094.2024.00078.

[39] N. Skjellum, M. Shaban, and M. Ismail, "SDN-Based Entanglement Routing Strategy in Smart Power Grids Using Reinforcement Learning," in Proc. 2025 IEEE International Conference on Quantum Computing and Engineering (QCE), Montreal, QC, Canada, 2025, vol. 2, pp. 133–138, doi: 10.1109/QCE61605.2025.10234.

[40] Y. Maleh, Y. Qasmaoui, K. El Gholami, Y. Sadqi, and S. Mounir, "A comprehensive survey on SDN security: threats, mitigations, and future directions," Journal of Reliable Intelligent Environments, vol. 9, no. 2, pp. 201–239, Jun. 2023, doi: 10.1007/s40860-022-00171-8.

[41] A. Jagannath, J. Jagannath, and P. S. P. V. Kumar, "A comprehensive survey on radio frequency (RF) fingerprinting: Traditional approaches, deep learning, and open challenges," Computer Networks, vol. 219, p. 109455, Dec. 2022, doi: 10.1016/j.comnet.2022.109455.

[42] J. Zhang, G. Shen, W. Saad, and K. Chowdhury, "Radio frequency fingerprint identification for device authentication in the internet of things," IEEE Communications Magazine, vol. 61, no. 10, pp. 110–115, Oct. 2023, doi: 10.1109/MCOM.003.2200974.

[43] M. A. Awan, Y. Dalveren, F. O. Catak, and A. Kara, "Deployment and implementation aspects of radio frequency fingerprinting in cybersecurity of smart grids," Electronics, vol. 12, no. 24, p. 4914, Dec. 2023, doi: 10.3390/electronics12244914.

[44] M. Safi, S. Dadkhah, F. Shoeleh, H. Mahdikhani, H. Molyneaux, and A. A. Ghorbani, "A survey on IoT profiling, fingerprinting, and identification," ACM Transactions on Internet of Things, vol. 3, no. 4, pp. 1–39, Nov. 2022, doi: 10.1145/3539736.

[45] A. Kara, "Practical implications of the use of radio frequency fingerprinting (RFF) in cybersecurity of smart grids," in Proc. 4th IFSA Winter Conference on Automation, Robotics & Communications for Industry 4.0/5.0, Innsbruck, Austria, Feb. 2024, pp. 112–116.