UDK 62

MACHINERY & ENERGETICS

Publisher:
National University of Life and Environmental Sciences of Ukraine

Year of foundation: 2010
Frequency of issue: Four times per year

Recommended for printing and distribution
via the Internet by the Academic Council
of National University of Life and Environmental Sciences of Ukraine
(Minutes No. 5 of November 27, 2025)

State Registration:
Media identifier R30-02298.
Decision of the National Council
of Television and Radio Broadcasting of Ukraine
No. 1795, Minutes No. 31, dated 21.12.2023.

The journal is included in the list of Professional Scientific Publications of Ukraine
Category "B". Branch of sciences – Technical.
Specialties: 0715 – Mechanics and Metal Trades, 0716 – Motor Vehicles, Ships and Aircraft,
0713 – Electricity and Energy, 0714 – Electronics and Automation
(Order of the Ministry of Education and Science of Ukraine of December 23, 2025, No. 1693)

The journal is presented international scientometric databases, repositories and scientific systems:
Google Scholar, Professional Publications of Ukraine, Vernadsky National Library of Ukraine,
BASE, J-Gate, Dimensions, Cambridge University Library, Leipzig University Library,
European University Institute, SOLO - Search Oxford Libraries Online, University of Hull Library,
University of Oslo Library, EBSCO, UCSB Library, Ulrichsweb Global Serials Directory,
OUCI (Open Ukrainian Citation Index), WorldCat, Litmaps

Editors office address:
National University of Life and Environmental Sciences of Ukraine
03041, 15 Heroiv Oborony, Kyiv, Ukraine
Tel.: +38(044)-258-42-63
E-mail: me@technicalscience.com.ua
https://technicalscience.com.ua/en

УДК 62

ТЕХНІКА ТА ЕНЕРГЕТИКА

Видавець:
Національний університет біоресурсів і природокористування України

Рік заснування: 2010
Періодичність випуску: 4 рази на рік

Рекомендовано до друку та поширення
через мережу Інтернет Вченою радою
Національного університету біоресурсів і природокористування України
(протокол № 5 від 27 листопада 2025 р.)

Державна реєстрація:
Ідентифікатор медіа R30-02298.
Рішення Національної ради України
з питань телебачення і радіомовлення
№ 1795, протокол № 31 від 21.12.2023 р.

Журнал входить до переліку наукових фахових видань України

Категорія «Б». Галузь наук – Технічні.
Спеціальності: G9 (131) – Прикладна механіка, G11 (133) – Галузеве машинобудування,
G7 (174) – Автоматизація, комп’ютерно-інтегровані технології та робототехніка,
G5 (141) – Електроніка, електронні комунікації, приладобудування та радіотехніка,
(Наказ Міністерства освіти і науки України від 23 грудня 2025 року № 1693)

Журнал представлено у міжнародних наукометричних базах даних,
репозитаріях та пошукових системах:
Google Scholar, Фахові видання України, Національна бібліотека України імені
В. І. Вернадського, BASE, J-Gate, Dimensions, Cambridge University Library,
Leipzig University Library, European University Institute, SOLO - Search Oxford Libraries Online,
University of Hull Library, University of Oslo Library, EBSCO, UCSB Library, Ulrichsweb Global
Serials Directory, OUCI (Open Ukrainian Citation Index), WorldCat, Litmaps

Адреса редакції:
Національний університет біоресурсів і природокористування України
03041, вул. Героїв Оборони, 15, м. Київ, Україна
Тел.: +38(044)-258-42-63
E-mail: me@technicalscience.com.ua
https://technicalscience.com.ua/uk

CONTENTS

ЗМІСТ

Machinery & Energetics
Vol. 16, No. 4. 2025

Journal homepage:
https://technicalscience.com.ua/en

UDC 621.39:003.5
DOI: 10.31548/machinery/4.2025.09

Ihor Limar*
PhD in Technical Sciences, Senior Lecturer
State University of Intelligent Technologies and Telecommunications
65023, 1 Kuznechna Str., Odesa, Ukraine
Engineering and Technology Institute "Biotekhnika" of the National Academy of Agrarian Sciences of Ukraine
67667, 26 Mayakska Road Str., Khlibodarske village, Ukraine
https://orcid.org/0000-0002-8972-9935

Yevhen Sevastieiev
Master, Senior Lecturer
State University of Intelligent Technologies and Telecommunications
65023, 1 Kuznechna Str., Odesa, Ukraine
https://orcid.org/0000-0003-1165-1119

Quantum cryptography: Theoretical foundations and practical implementations for the protection of critical infrastructure

Abstract. The purpose of the article was to theoretically determine harmonised criteria for the applicability of quantum key distribution for the protection of critical infrastructure, taking into account the long-term risk of decryption by adversaries possessing quantum computational resources and traffic archiving capabilities. The methodology relied on a theoretical systems analysis of three Quantum Key Distribution subsystems – sources, detectors, and channels (fibre/free space) – complemented by simplified models of the link budget and bit error probability. A concise scenario analysis was conducted for banking networks, energy systems, and government communications, taking into account scalability, compatibility, and cost. The main results showed that laser intensity stability of 1-2% and higher detector efficiencies – 60-70% for Avalanche Photodiodes and 80-90% for Superconducting Nanowire Single-Photon Detectors – extended practical distances and reduced errors. Without trusted nodes in fibre, 150-200 km was achievable; for urban Free-Space Optics lines, 5-20 km was optimal, while longer distances required network segmentation or satellite segments. Architecturally, it was justified that: for banks – rings with trusted nodes and Free-Space Optics reserve; for energy systems – dual fibre channels "control centre-substation" with a local key cache; for government communications – segmented domains with interagency gateways and satellite redundancy. In all scenarios, a Quantum Key Distribution+Post-Quantum Cryptography hybrid with short key rotation and operational countermeasures against attacks on detectors and channels was appropriate, confirming practical suitability for urban and regional networks. The practical significance lay in providing engineering teams in banking, energy, and government communications with ready-made architectural profiles – from rings and dual fibre channels to segmented domains – with Key Management System/Hardware Security Module integration, definition of Service Level Objective/Service Level Agreement, and phased roadmaps. For regulators and operators, this formed a basis for updating requirements and audits, as well as for planning redundancy (Free-Space Optics/satellite) and Total Cost of Ownership categories

Keywords: key distribution; laser sources; photonic detectors; optical fibre; free-space communication; trusted nodes; banking networks

INTRODUCTION

Relevance was driven by the need of critical infrastructure for service continuity, strict timing guarantees, and fault tolerance. Banking systems required secure scalable channels between data centres, processing nodes, and branches; the energy sector required low latency between control centres and substations; government communications

Article's History: Received: 01.08.2025; Revised: 03.11.2025; Accepted: 27.11.2025.

Suggested Citation:
Limar, I., & Sevastieiev, Ye. (2025). Quantum cryptography: Theoretical foundations and practical implementations for the protection of critical infrastructure. *Machinery & Energetics,* 16(4), 9-20. doi: 10.31548/machinery/4.2025.09.

*Corresponding author

required intercity and interagency routes with redundancy and access segmentation. Against the backdrop of the long-term "store now – decrypt later" risk, interest in quantum key distribution (QKD) as a physically grounded complement to classical cryptography was growing, although deployment was complicated by channel losses, atmospheric factors, network heterogeneity, and requirements for interagency key management.

The reviewed studies systematised two main families of QKD protocols – prepare-and-measure and entanglement-based – in discrete-variable and continuous-variable implementations, focusing on the relationship between acceptable Quantum Bit Error Rate (*QBER*) levels and the secret rate on the one hand and source and detector parameters on the other. The works emphasised the criticality of intensity and frequency stability, jitter, background radiation levels, quantum efficiency, dark counts, and recovery time, as well as the role of synchronisation and correct calibration of polarisation/phase. The meta-review by L. Gyongyosi *et al.* (2019) additionally emphasised the importance of decoy states, Measurement-Device-Independent/Device-Independent (MDI/DI) settings, and finite-key corrections for network scenarios, but mostly in general terms, without deriving unified engineering criteria for specific classes of critical infrastructure and without a holistic comparison of technical requirements with architectural and economic constraints.

A separate body of work devoted to modelling the Bennett-Brassard 1984 (BB84) protocol detailed the impact of channel losses, background events, basis mismatch, the choice of timing-window width, discrimination thresholds, gating and filtering parameters on the secret rate, the intrusion-detection threshold, and achievable distance. In a number of studies, including T.-T. Nguyen *et al*. (2023), practical approaches were proposed for tuning receiver paths for fibre and Free-Space Optics (FSO) links, forming a basis for technical and technological recommendations. At the same time, existing models typically considered isolated configurations or laboratory conditions, accounted only limitedly for hybrid topologies combining fibre, FSO, and satellite segments, integration with key-management systems and Service Level Objective/Service Level Agreement (SLO/SLA) requirements, and almost did not propose harmonised suitability criteria specifically for banking networks, energy systems, and government communications. This structural limitation delineated the problem space for further research.

Reviews of 6G and URLLC services highlighted the fundamental need to combine physically protected key distribution with post-quantum cryptography (PQC), and to embed crypto functions into Software-Defined Networking/Network Functions Virtualisation (SDN/NFV) architectures with separation of control/user planes and strict latency budgets under Precision Time Protocol/Synchronous Ethernet (PTP/SyncE). However, cross-sector reviews lacked domain-specific, quantitative criteria for critical infrastructure – banking networks, energy, and government communications – tied to target SLOs and component parameters. The study by M. Hoschek (2021) emphasised slice isolation, distributed key caches, and telemetry for SLA control, but did not establish a direct link between telemetry and the link budget and secret rate, nor with orchestration of Key Management System/Hardware Security Module (KMS/HSM) under specific time budgets. The work by D. Rahmayanti (2025) outlined the applicability of QKD in 5G/6G transport with account taken of radio and FSO channels, showing "windowed" availability due to turbulence, precipitation, and daytime illumination and justifying spatial-spectral filtering, adaptive apertures, auto-tracking, and route redundancy, but did not reduce these factors to unified suitability and life-cycle cost profiles for specific domains.

S. Rajpoot *et al*. (2023) systematised requirements for trusted-node networks and highlighted trade-offs between route length, key rotation, and acceptable *QBER*, but without SLO-oriented integration with SDN/NFV and without comparing fibre/FSO/satellite hybrids specifically for critical-infrastructure tasks. The review by I.B. Djordjevic (2022) substantiated layered complementarity of the physical layer, QKD, and PQC with Perfect Forward Secrecy (PFS) support and KMS/HSM integration, but left open the numerical norms for components and the direct impact on ηtot, *QBER*, and achievable SLOs in applied scenarios. Results by P. Dharanish *et al*. (2024) demonstrated the viability of Field-Programmable Gate Array (FPGA) implementations of PQC (acceleration of KEM and signatures), but did not define policies for allocating quantum keys between flows with different latency requirements and session lifetimes.

Practical countermeasures proposed by C.-L. Chen *et al*. (2025) – Quantum Random Number Generator (QRNG), decoy states, active monitoring of sources and detectors, checks against blinding attacks – were not tied to standardised telemetry thresholds and the impact on SLO/Total Cost of Ownership (TCO) under operational conditions. Thus, the existing literature set out principles (QKD + PQC hybrid, SDN/NFV manageability, trusted-node topologies, "windowed" FSO/satellite operation), but did not form a unified, quantitatively defined, domain-oriented set of criteria. The article proposed to close this gap: to normalise source and detector parameters and synchronisation requirements, to build suitability maps for fibre, FSO, and satellite channels, to define SLO classes and TCO profiles for three domains, and to determine orchestration policies for KMS/HSM and quantum-key allocation in the QKD + PQC hybrid with account taken of latency budgets and traffic prioritisation.

The problem space, therefore, lay in the absence of a harmonised systematisation of requirements for components and channels with a direct linkage to banking networks, energy-system control, and government communications, as well as in the shortage of theoretical compatibility and cost models that allow comparison of integration options at the architectural level.

The purpose of the study was to form a systemic framework for evaluating the feasibility of implementing QKD in three typical critical-infrastructure scenarios – banking networks, energy-system control, and government communications – taking into account scalability requirements, integration with fibre, free-space, and satellite channels, and full life-cycle costs. The research hypothesis was that integrating physically grounded key distribution with key life-cycle management and compatible transport infrastructure would provide a practically suitable level of security, fault tolerance, and scalability over urban and regional distances without unacceptable costs.

MATERIALS AND METHODS

The study was purely theoretical in nature and was based on a systems analysis of three classes of QKD subsystems: radiation sources, photonic detectors, and quantum communication channels. For sources, intensity and frequency stability, polarisation constancy, timing jitter, and background radiation were considered; for detectors – quantum efficiency, dark count rate, dead time, and timing jitter; for channels – the loss budget and fluctuations. Closed-form estimates of the link budget and errors were used: optical losses in fibre were modelled as (formula 1):

$$A(d) = \alpha_f d + A_{conn}\,, \tag{1}$$

where α_f – fibre-specific loss (dB/km); d – route length (km); A_{conn} – total losses at connectors/splices (dB). Analysis: increasing d or α_f linearly increases attenuation, compressing the achievable range; a large A_{conn} creates a "fixed penalty" even on short spans. Typically, $\alpha_f \approx 0.17$-0.25 dB/km; each connector adds ≈0.2-0.5 dB. Total transmission (formula 2):

$$\eta_{tot} = \eta_{src} \times 10^{-\frac{A}{10}} \times \eta_{det}, \tag{2}$$

where η_{src} – the useful fraction of the photon flux from the source (accounts for intensity/polarisation stability); $10^{-A/10}$ – the channel transmission fraction, which depends on attenuation A; η_{det} – detector quantum efficiency. Analysis: a decline in due to intensity/polarisation fluctuations, or an increase in A, exponentially reduces η_{tot}; increasing η_{det} (APD ≈ 0.6-0.7; SNSPD ≈0.8-0.9) directly raises the probability of useful events and "compensates" for part of the channel loss. The approximate bit error was evaluated by formula (3).

$$BER \approx \frac{p_{dc}+e_{mis}p_{sig}}{p_{sig}+p_{dc}}, \tag{3}$$

where p_{sig} – the probability of detecting a useful signal per pulse; p_{dc} – the probability of dark counts (detector noise); e_{mis} – the fraction of errors due to basis/alignment mismatch (polarisation, timing windows, phase shifts). Analysis: a decrease in p_{sig} (due to low η_{tot}) or an increase in p_{dc} raises *QBER*; reducing emis through polarisation/phase stabilisation and timing-window optimisation directly reduces *QBER*. For CV-QKD, increased phase noise effectively increased emis through quadrature imbalance; active phase locking reduced this contribution. For DV-QKD, "decoy" intensities were taken into account; for CV-QKD – SNR of homodyne/heterodyne measurement and phase stability. Parameters were varied within practical intervals aligned with industry practice, and the output metrics were achievable distance, secret key rate, and *QBER*. No physical experiments were performed; no equipment list was provided.

Assumptions for detectors were as follows: APDs operated in Geiger mode with active gating (windows 0.5-1.5 ns, dead time 50-100 ns, jitter 50-80 ps), and discrimination thresholds were selected by the criterion of minimising total error. SNSPDs were considered in continuous counting mode (without gating) with efficiency 0.8-0.9, dark-count rate ≤10^1 cps, and typical receiver windows of 0.1-0.3 ns. These modes fixed the initial conditions for calculating p_{sig}, *QBER*, and the secret rate. For each applied domain (banking networks, energy systems, government communications), a set of candidate topologies and integration policies (TLS/IPsec, KMS/HSM, PQC hybrids) was preliminarily formed, target SLOs for availability/latency were set, and deployment staging was fixed. For banking profiles, ring schemes with trusted nodes and an FSO reserve were considered; QKD keys were assigned to short-lived TLS/IPsec sessions with PFS for critical flows, while mass channels were protected by PQC, with mandatory integration with KMS/HSM. For trunks, a fibre + LEO/MEO satellite-link hybrid with key buffering and route redundancy was assumed. Fault-tolerance requirements were implemented through two physically independent routes (fibre + FSO or two fibres from different operators) combined with a path-diversity policy at the level of network management.

Comparative modelling was performed for three transport classes: point-to-point fibre segments without trusted nodes, fibre networks with trusted nodes/key routing, and free-space (FSO) links, as well as the combination with satellite segments. For fibre, an exponential attenuation law (via α_f) with additional connector losses was applied; for FSO – windowed availability models taking into account turbulence, precipitation, and background illumination; for satellite links – pointing geometry and atmospheric losses with time "visibility windows". For each configuration, the following were evaluated: (1) indicative secret-key rate (orders of magnitude), (2) probability of meeting target SLOs for availability, (3) sensitivity to degradation of component parameters. Topological analysis covered rings and stars with trusted nodes, fibre + FSO hybrids, and inclusion of satellite segments for intercity routes. The obtained estimates served as the basis for distance ranges, availability indicators, and key-rate profiles, subsequently summarised in results tables. First, p_{sig} and *QBER* were obtained from the link budget and configurations without a positive secret rate were filtered out. For suitable variants, the secret-key rate was calculated and binned by order of magnitude: kbit/s, 10 kbit/s, 100+ kbit/s. Availability (SLO) was classified by classes: A ≥ 99.99%, B = 99.95-99.99%, C = 99.9-99.95%, D = 99.0-99.9%, E<99.0%. Sensitivity was determined under parameter variations of ±10% or +1 dB to losses: low – ≤10% rate drop; medium – 10-30%; high – >30% or loss of a positive secret rate. For further calculations, normalised ranges of source, detector, and synchronisation parameters were fixed (intensity, polarisation, linewidth, detection efficiency, dark counts, dead time, jitter, timing windows, and monitoring). The bounds were selected with account taken of practical intervals: intensity stability ⩽1-2% RMS, polarisation stability ⩾25-30 dB, APD efficiency 60-70% and SNSPD 80-90%, dark counts for APD ⩽10^2-10^3 cps and for SNSPD ⩽10^1 cps, typical timing windows 0.2-1.0 ns. The stated intervals were used as input constraints in all link-budget models, η_{tot}/*QBER* calculations, and scenario comparisons of channels and topologies. Parameterisation covered the following subsystems and protocols: laser source for DV-QKD (BB84 with "decoy" intensities), local oscillator for CV-QKD (GMCS); APD detectors for DV-QKD and SNSPD detectors for DV and CV implementations; synchronisation subsystems for DV-QKD (BB84 + decoy, MDI), for CV-QKD (GMCS), and for entanglement-based variants (E91/BBM92); monitoring subsystems for DV-QKD (BB84 + decoy, MDI), CV-QKD (GMCS), and entanglement-based (E91/BBM92). All listed components were included in the link-budget calculations, ηtot estimates, *QBER*, and secret-key rate.

Scenario analysis was conducted for three applied critical-infrastructure domains: banking networks, energy-system control, and government communications. For each scenario, architectural profiles were formed: recommended topology; integration with classical protocols (Transport Layer Security (TLS) and Internet Protocol Security (IPsec)), KMS/HSM and PQC hybrids; requirements for latency, availability, and fault tolerance; a phased deployment plan. Hybrid trunk: fibre – the primary transport; LEO/MEO satellite links – reserve/bridge for intercity sections with "windowed" availability. Quantum-key exchange was performed opportunistically, key material was buffered in edge KMS/HSM with lifetime limits and used to bridge pauses between "windows". Route redundancy ensured path diversity under SDN control, with automatic switching to an alternative path (fibre priority). Compatibility was assessed at the network level (SDN/NFV manageability, Wavelength-Division Multiplexing (WDM) coexistence for Continuous-Variable Quantum Key Distribution (CV-QKD), optical isolation for Discrete-Variable Quantum Key Distribution (DV-QKD)) and at the application level (key life-cycle policies, rotation frequency, PFS). For DV-QKD with "decoy" intensities, automated calibration of the "signal/decoy" power ratio and continuous monitoring of parasitic ASE with pre-set alert thresholds were envisaged; procedures were performed at initialisation and periodically during the session for timely recalibration. The economic component was implemented as a structured TCO model: Capital Expenditures (CapEx) (optics, detectors, satellite inserts), Operating Expenditures (OpEx) (calibration, monitoring, audit), integration costs (KMS/HSM, node upgrades), staff training, and risk costs. Threat modelling was also carried out: attacks on measurement modules (blinding, timing shifts), "Trojan" injections, store-now/decrypt-later with the use of quantum memory. For each threat, technical (MDI/DI) paradigms, power/spectrum monitoring, randomised window sampling), organisational (HSM, domain segmentation, audit), and operational countermeasures were defined, as well as residual risk. The combination of the stated methods directly produced the normalised component ranges, channel suitability maps, and applied profiles.

RESULTS

Systematised quality criteria for sources and detectors

For radiation sources, it was established that intensity stability within ⩽1-2% RMS over the operating interval and linear polarisation stability of no worse than 25-30 dB ensured reproducibility of state encoding in prepare-and-measure protocols. For discrete-variable implementations with decoy intensities, automated calibration of power ratios and monitoring of parasitic amplified spontaneous emission (ASE) background radiation with alert thresholds were recommended. In continuous-variable (CV-QKD) implementations, stability of the local-oscillator phase and low source phase noise were critical; within the study, normalisation was adopted at the level of a narrow frequency band (a few hundred kHz) for sources aligned with local oscillators.

For detectors, it was established that detection efficiency ⩾60-70% for avalanche photodiodes (APD) in Geiger mode and ⩾80-90% for superconducting nanowire single-photon detectors (SNSPD) substantially reduced requirements for channel length or intensity. In parallel, lower bounds for dark counts in the range ⩽10^2-10^3 cps for APD and ⩽10^1 cps for SNSPD reduced the contribution to *QBER* and increased the usable key-sifting rate. Dead time ⩽50-100 ns for APD and ⩽20-50 ns for SNSPD, and jitter ⩽50-80 ps, were defined as practical reference points for urban and regional profiles. In Table 1, the recommended parameter ranges for sources, detectors, and synchronisation were systematised (intensity, polarisation, linewidth, efficiency, dark counts, dead time, jitter, timing windows, and monitoring).

Table 1. Recommended parameter ranges for sources and detectors

Subsystem	Protocol/class	Key parameters	Recommended range/note
Source (laser)	DV-QKD (BB84 + decoy)	Intensity stability (RMS)	⩽1-2% over the session interval
		Linear polarisation	Stability ⩾25-30 dB
		Linewidth	Narrow (hundreds of kHz) for stable interference
Source (local oscillator)	CV-QKD	Phase noise	Low: σφ⩽0.1-0.2 rad; SSB PN ⩽-90/-120/-140 dBc/Hz @10 kHz/100 kHz/⩾1 MHz; Δν⩽100-300 kHz; active phase locking
		Relative intensity noise	Minimisation of ASE, real-time monitoring: C-band ±1-1.5 nm from the carrier (windows 0.2-0.4 nm ≈ 25-50 GHz); RIN 10 kHz-20 MHz

Continued Table 1.

Subsystem	Protocol/class	Key parameters	Recommended range/note
Detector (APD)	DV-QKD	Detection efficiency	⩾60-70% (near-IR)
		Dark counts	⩽10^2-10^3 cps; active gating
		Dead time / jitter	⩽100 ns/⩽80 ps
Detector (SNSPD)	DV/CV	Detection efficiency	⩾80-90%
		Dark counts	⩽10^1 cps
		Dead time / jitter	⩽50 ns/⩽50 ps
Synchronisation	DV-QKD (BB84 + decoy, MDI); CV-QKD (GMCS); Entanglement-based (E91/BBM92)	Coincidence window	Tunable, optimisation for SNR and *QBER*: 0.2-1.0 ns (APD 0.5-1.5 ns; SNSPD 0.1-0.3 ns)
Monitoring	DV-QKD (BB84 + decoy, MDI); CV-QKD (GMCS); Entanglement-based (E91/BBM92)	Security parameters	Detection of blinding, "Trojan" injections, afterpulsing: power threshold -35...-25 dBm; out-of-band monitoring ±2-5 nm (windows 0.2-0.4 nm); tail monitoring 0.1-10 µs (APD 1-10 µs; SNSPD ⩽0.5 µs)

Note: low local-oscillator phase noise: $\sigma\varphi \leqslant 0.1$-0.2 rad (RMS, band 1 kHz-10 MHz); SSB PN ⩽ -90 dBc/Hz @ 10 kHz, ⩽-120 dBc/Hz @ 100 kHz, ⩽-140 dBc/Hz @⩾1 MHz; source/LO linewidth $\Delta\nu \leqslant$ 100-300 kHz; residual phase-tracker error ⩽ 5-10° (RMS). Active phase locking – mandatory. near-IR – near-infrared
Source: compiled by the authors based on D.V. Reddy *et al.* (2020), Y. Shao *et al.* (2021), C. Yu *et al.* (2024)

For DV-QKD, intensity stability at the level of ⩽1-2% RMS preserved the correctness of selecting "decoy" intensities and reduced statistical biases; deterioration to ~3-5% increased the conservatism of intrusion estimation and reduced the secret rate in weak-signal regimes. Linear polarisation stability of around 25-30 dB ensured selectivity in the bases, whereas ≈20 dB already noticeably increased *QBER* due to degraded orthogonality. For CV-QKD, low local-oscillator phase noise and active locking maintained the SNR of homodyne/heterodyne measurement; without these measures, quadrature variance increased, and the secret rate fell under an unchanged loss budget. Detector efficiency remained key: approximately 60-70% for APD and 80-90% for SNSPD increased the registration probability and "absorbed" several decibels of additional loss; a decrease by ~10 pp reduced the practical range by tens of kilometres. Limiting dark counts (on the order of 10^2-10^3 cps for APD and 10^1 cps for SNSPD) reduced the background contribution to *QBER*; exceeding these levels required either narrowing timing windows with the risk of losing useful events, or reducing the sifting rate, which directly reduced the key rate. Within detector dynamics, dead time and jitter should be kept at the level of "tens of nanoseconds" and "tens of picoseconds" respectively, so as not to limit pulse rate and the accuracy of time correlation. The defined reference points were integrated into acceptance-testing procedures: verification of intensity stability under thermal fluctuations, analysis of dark counts with temperature taken into account, assessment of afterpulsing, and modelling of blinding and "Trojan" injections. For timely detection of degradation, telemetry was applied at a rate of at least 1 Hz for sources and 10-100 Hz for detectors, ensuring continuity of critical-infrastructure services.

Attenuation *A(d)* increased linearly with length and the number of connections; an increase of +1 dB reduced the transmission fraction 10-A/10 by approximately 20%, and even a small fixed penalty A_{conn} substantially limited the operating range. Converting attenuation into the channel transmission fraction and then accounting for source and detector efficiencies showed that an additional 1-2 dB of loss reduced η_{tot} by ≈18-36%, with ηtot scaling proportionally with η_{src} and η_{det}. Transitioning from APD to SNSPD increased path efficiency by on the order of 15-30% under otherwise equal conditions. *QBER* estimation revealed that reducing p_{sig} as a result of increasing total channel attenuation A to ≳ 15-20 dB (e.g., 80-120 km of fibre with $\alpha_f \approx 0.2$ dB/km plus 6-8 connectors of 0.2-0.5 dB) or even adding +1...+2 dB to the baseline directly increased the error rate, because ηtot ∝ 10-A/10 decreased by ≈18-36% and the relative contribution of dark counts p_{dc} increased. By contrast, reducing emis through polarisation/phase stabilisation and timing-window optimisation (approximately 0.2-1.0 ns) consistently decreased *QBER* by several percentage points. In CV-QKD, increased phase noise raised effective e_{mis}, but active phase locking recovered part of the operating region in range; in DV-QKD, the use of "decoy" intensities stabilised estimates of p_{sig} in weak-signal regimes and maintained a positive secret rate on longer links. Configurations with above-threshold *QBER* were removed from further comparisons, which explains the reduction of usable range in high-loss scenarios. Sensitivity analysis identified critical points at which an additional connector (~0.2-0.5 dB) triggered a transition from a positive to a zero-secret rate; increasing η_{src} and η_{det} partly compensated fibre losses, reduced *QBER*, and extended the operating range without increasing source power. Collectively, these dependencies showed that managing line losses and path efficiency determines the useful signal p_{sig} and, consequently, the error level, shaping the observed ranges and secret rates.

Comparative modelling results for quantum channels and topologies

The second block of results concerned a comparative analysis of communication channels and basic topologies. Three operating regimes were aligned: (1) point-to-point optical fibre without intermediate trusted nodes, (2) fibre segments with trusted nodes and key routing, (3) free-space optical

(FSO) communication for urban routes and satellite segments for intercity. The analysis was conducted with account taken of losses, fluctuations, background illumination, weather factors, and service-availability requirements.

For fibre lines, practical ranges of 150-200 km without nodes were confirmed, provided the parameters in Table 1 were met and "decoy" intensities were used. For longer distances, segmentation with trusted nodes was deemed appropriate, shifting requirements from the physical layer to organisational and cryptographic policies (physical security, HSM, audit). For urban FSO, typical routes of 5-20 km were aligned, with sensitivity to meteorological factors; nighttime windows and adaptive apertures/trackers improved the availability profile. For trunks, a hybrid was recommended: terrestrial fibre segments + satellite (Low Earth Orbit (LEO) and Medium Earth Orbit (MEO)) segments with key buffering and route redundancy. In Table 2, the suitability of channels and topologies for different environments and distances was summarised with typical losses, target availability, and orders of key-generation rate.

Table 2. Suitability of channels and topologies for distances/environments

Channel/Topology	Typical environment	Distance (km)	Indicative losses/ factors	Target availability	Indicative key rate*
Fibre P2P without nodes	Urban/regional	50-200	0.17-0.25 dB/ km + connectors	99.9%	tens of kbit/s
Fibre with trusted nodes	Regional/intercity	200-1,000+	Total losses by segment; node requirements	99.95%+	hundreds of kbit/s (aggregation)
Urban FSO	"Rooftop-to-rooftop" links	5-20	Turbulence, rain/fog, background light	99.0-99.9% (climate-dependent)	units-tens of kbit/s
Inter-building FSO reserve	Intra-urban reserve	1-5	Low loss budget, short baselines	99.9%	tens of kbit/s
Satellite (LEO/MEO)	Intercity/interstate	500+ (projection)	Atmosphere, pointing, cloud cover	Windowed availability	from units to tens of kbit/s

Note: * – rate depends on protocol, losses, and the parameters in Table 1; typical orders are given for planning
Source: compiled by the authors based on M. Ghalaii *et al.* (2023), A. Rayhan *et al.* (2025)

The range of 50-200 km for "pure" fibre without trusted nodes corresponded to a practical loss budget of 8-20 dB (including splices/connectors), within which the parameters in Table 1 made it possible to keep *QBER* below threshold values for a positive secret rate. Crossing the ~200 km boundary without nodes, even with high-efficiency detectors, theoretically required either unacceptably low error rates from ideal components or a substantial reduction in the key rate, which made the service uneconomic for most applied critical-infrastructure scenarios.

Fibre networks with trusted nodes extended geographical coverage to hundreds and thousands of kilometres via an organisational trust model: each segment ensured locally high quality, while the key was routed through nodes. Technically, this enabled aggregation of key rates (hundreds of kbit/s in total), but shifted requirements to physical security, HSM, audit, and segmentation of access domains, which had to be accounted for in TCO. FSO links of 5-20 km provided effective urban connections at low deployment cost, but availability of 99.0-99.9% depended strongly on climate: rain, fog, and daytime illumination reduced time windows of stable operation. For this reason, FSO was appropriate either as a reserve to fibre (1-5 km with 99.9% availability due to short baselines) or as a primary channel in scenarios where windowed availability profiles were acceptable (for example, night-time windows with illumination control in dense urban areas).

Satellite segments (LEO/MEO) logically imposed a windowed availability regime and requirements for pointing/atmospheric transparency. Despite modest orders of key rates (units-tens of kbit/s), such segments opened a practical path for intercity/interstate trunks where continuous fibre rings with trusted nodes were economically or organisationally difficult. Thus, the "fibre + FSO/satellite" hybrid optimised both range and fault tolerance. In summary, for designing critical-infrastructure networks it was rational to combine fibre segments (as a high-availability base) with FSO inserts (as an economical reserve or local primary channel) and satellite links (as an intercity reserve or temporary bridge), achieving compliance with SLO without disproportionate cost growth.

Applicability for banks, energy systems, and the state

The third block of results transformed technical norms into applied architectural profiles for three scenarios, in each of which a recommended topology, an integration policy with classical protocols and KMS/HSM, latency and fault-tolerance requirements, an indicative cost category, and a phased implementation plan were formed. For banking networks, the suitability of branch-to-hub and processing-centre interactions in the form of rings with trusted nodes was established, enabling routing of key material with local aggregation and rotation aligned with the transaction profile. For critical payments, QKD keys were used as material for one-time or partially one-time TLS/IPsec sessions (PFS + PQC hybrid), while standard flows remained on PQC algorithms with higher throughput. Due to high sensitivity to outages, fault tolerance was ensured via two independent physical routes (fibre + FSO or two fibres from different operators).

In energy-system management, the suitability of dual "control centre – substation" channels with geo-redundancy was confirmed, where QKD provided keys to protect critical telemetry and control commands (including for IEC 62351-compatible profiles), while classical channels served as backup or transport. Strict requirements for latency and determinism directed the design towards fibre segments with a minimal number of intermediate nodes and a local key cache at the edge. For government communications, hybrid architectures were formed with segmentation of access domains, in which QKD keys remained within the domain, and inter-domain exchange took place through gateways with HSM and trust policies. Satellite segments were used for intercity backbones and redundancy, while urban segments were built on fibre with nodes. Table 3 summarised the applied deployment profiles for banking networks, energy systems, and government communications: topologies, integration with classical cryptography, SLOs, TCO categories, and implementation phasing.

Table 3. QKD deployment profiles – topology, integration, SLO, TCO, phasing

Scenario	Recommended topology	Integration with classical cryptography	SLO requirements	TCO category (5 years)	Phased implementation
Banking networks	Ring with trusted nodes + FSO backup	TLS/IPsec with PFS; PQC hybrid; KMS + HSM	99.95%+, low RTO/RPO	Medium-high	Urban pilot → regional rings → intercity backbones
Energy systems	Dual control centre – substation channels	IEC-compatible protection; local key cache	Low latency, high determinism	Medium	Backbone section → critical nodes → scaling to substation classes
Government communications	Segmented domains + satellite segments	Gateways with HSM; access policies	High availability and segmentation	High	Urban cores → intercity routes → inter-agency gateways

Note: TCO categories (5 years) are interpreted as follows: medium – basic fibre infrastructure without expensive components; medium-high – rings with trusted nodes and backup (FSO/second route) and increased availability requirements; high – intercity/satellite segments with enhanced hardware security and complex integration
Source: compiled by the authors based on M. Stanley *et al.* (2022), J. Faba *et al.* (2025)

For banking networks, ring topologies with trusted nodes ensured uniform availability and the ability to route the key even if an individual segment failed. FSO backup enabled rapid service restoration during fibre incidents, keeping RTO/RPO within strict limits. The QKD + PQC hybrid made it possible to allocate QKD keys to transaction-critical flows (short TLS sessions with frequent rotation), while high-performance channels were handled by PQC algorithms without a bottleneck on the quantum key generation side. In the TCO profile, CapEx on nodes and KMS/HSM and OpEx on audit/logging dominated.

In energy systems, determinism and low latency were decisive. Dual fibre channels between control centres and substations minimised the risk of losing control. A local edge key cache removed peak load on generation during short "bursts" of control commands; the compatibility profile with the IEC family of protocols meant practical integration without violating existing SLAs. TCO here was formed by a balance of moderate CapEx on fibre routes/nodes and OpEx for maintaining determinism (synchronisation, regular monitoring of the timing budget).

For government communications, access-domain segmentation and intercity/interstate links were key. Satellite segments added geographic flexibility and route redundancy, but increased the complexity of managing availability windows. Inter-domain exchange policies required gateways with HSM, multifactor access, logging, and thorough audit. The high TCO category was explained both by the cost of satellite components and by organisational security requirements. Thus, each profile demonstrated how technical norms of components and channels were transformed into applied architectures with realistic SLOs and costs, confirming the hypothesis about the practical suitability of hybrid QKD networks for critical infrastructure.

Scalability, compatibility, cost, and future risks

The fourth block of results integrated aspects of scalability, compatibility with existing infrastructure, life-cycle economics, and evolutionary risks. It was generalised that scaling QKD networks required coordination at three levels: physical (channel and component quality), network (topological and routing decisions), and application (key management, compatibility with protocols and services). All levels were "tied together" through SLOs (availability, latency, key throughput) and security policies (trust domain, isolation, audit).

Compatibility with existing infrastructure was achieved through hybrid cryptoprofiles: QKD provided high-quality symmetric material injected into TLS/IPsec/into application protocols as session keys, while post-quantum algorithms (PQC) carried scalable flows where QKD key material was a limited resource. For telecom operators, an SDN/NFV approach with cryptographic functions as services was considered appropriate, enabling centralised management of rotation and key distribution policies and observability of network state.

The economic part of the results was presented as structured TCO drivers: capital expenditure on optics/detectors/satellite segments; operating expenditure on calibration, monitoring, audit; costs of integration with KMS/HSM and upgrading network elements; costs of staff training; risk costs (penalties/downtime) in case of failure. It was determined that phased deployment (pilots, limited operation, scaling) minimised TCO through gradual capability build-up and reuse of developed templates. Table 4 presented a risk map of deployment and corresponding countermeasures, with an assessment of residual risk and operational policies.

Table 4. Deployment risks and aligned countermeasures

Risk/attack	Nature of the threat	Aligned countermeasures	Residual risk	Operational policies
Detector blinding	Driving the APD into linear mode	Power monitoring, filtering, MDI-QKD, liveness tests	Low-medium	Periodic testing, telemetry logging
"Trojan horse"	Light injection into the module	Optical isolators, back-reflection detectors, alarms	Low	Strict cabling policies, tamper sealing
Time/window shift	Manipulation of time windows	Dynamic windows, random discretisation, audit	Low	Synchronisation checks, test packets
Store-now/decrypt-later	Archiving traffic until quantum computers emerge	QKD+PQC hybrid, short key rotation	Low	Clear RKP (retention/rotation) policies
Atmospheric/weather factors (FSO)	Loss of availability	Geo-redundancy, combined routes, window planning	Medium	SLAs with alternative carriers
Trusted-node compromise	Organisational risk	Physical security, HSM, separation of duties, audit	Low-medium	Access checks, logging, SOC integration

Note: RKP – Retention/Rotation Key Policies, SOC-integration – Security Operations Centre
Source: compiled by the authors based on M.P. Lingaraju *et al.* (2019)

Risk contours confirmed the appropriateness of technical and organisational countermeasures. For detector blinding, the simultaneous presence of optical filtering, power monitoring at the module input, and periodic liveness tests reduced the probability of a successful attack to a low/medium level, while MDI-QKD additionally shifted trust to the central joint-measurement node, removing the class of detector attacks at the endpoints. The "Trojan horse" attack was neutralised by isolators and back-radiation sensors; residual risk remained low, provided strict cabling policies and access control to optical panels were enforced.

Manipulations of time windows were mitigated by adaptive windows and randomised offsets, making the attack less predictable; regular synchronisation audits prevented drift from accumulating to levels that would be noticeable in *QBER* only after the fact. The store-now/decrypt-later risk was conceptually reduced by the QKD + PQC hybrid and short key rotation, which made sessions cryptographically short-lived for deferred analysis, even with the emergence of powerful quantum computers. For FSO channels, medium residual risk was compensated by geo-redundancy and SLAs with alternative transport, as well as planning of operating "windows" taking into account local climate patterns. Finally, the organisational risk of trusted nodes was substantially reduced through HSM, separation of duties, tamper sealing, video surveillance, and SOC monitoring with immutable logs. From a scalability perspective, the results confirmed that segmentation into trust domains and gateways between those domains allowed the network to grow without disproportionate risk growth: an incident was localised within a domain, component migrations occurred gradually, and policies for key-material exchange remained under centralised KMS control.

The TCO structure was formed from four classes of costs: CapEx on optics/detectors/satellite segments; OpEx on calibration, monitoring, and audit; costs of integration with KMS/HSM and upgrading network elements; costs of staff training. Phased deployment – first an urban pilot on short fibre or FSO (5-20 km), then regional rings with nodes, then intercity inclusion via satellite or leased backbones – spread costs evenly and reduced integration risks. Compatibility with existing transports improved with CV-QKD (better alignment with WDM), while DV-QKD required stricter optical isolation and channel separation; in application protocols, short TLS/IPsec sessions with frequent key rotation were recommended for critical flows in order to use the limited quantum key rate efficiently.

Development trajectories relevant to the studied scenarios were outlined: the emergence of quantum repeaters and intermediate technologies without state disclosure (extending distance without trusted nodes), broader use of MDI/DI paradigms (reducing dependence on the honesty of the measurement module), hardware improvements in detectors and sources (fewer dark events, lower jitter, better phase stability), and "native" integration with PQC in network stack services. Threats from quantum memory – i.e., the ability to store states and analyse these states later – were neutralised by design choices of the QKD + PQC hybrid, short key-rotation policies, and minimising session lifetimes; in the government segment, record-retention regulations and mandatory PFS settings were additionally taken into account.

The theoretical study fulfilled the stated aim: aligned criteria for sources, detectors, and channels were formed; comparative profiles of channels and topologies for urban and regional distances were proposed; applied architectural solutions for banking, energy, and government communications were built; issues of scalability, compatibility, and life-cycle economics were integrated; risks and effective countermeasures were systematised. The obtained results confirmed the working hypothesis: integrating physically grounded key distribution with key life-cycle management and compatible transport infrastructure provides a practically usable level of security, resilience, and scalability over urban and regional distances without unacceptable costs, while phased deployment made it possible to achieve target SLOs with controlled risks and predictable TCO.

DISCUSSION

In this section, the obtained results were analysed – link budgets, the sensitivity of the secret key rate to dark-count events and jitter, as well as SLO profiles for the banking, energy, and government domains – and compared with the conclusions of other researchers. The emphasis was not on a literature review, but on comparison: where the results coincided, where these findings diverged, what this meant for practical implementation, and why it mattered for long-term cyber resilience. The obtained results are consistent with a broad body of research and, at the same time, refine the boundaries of applicability for critical infrastructure. The practical secret key rate was determined by a combination of intensity stability (⩽1-2% RMS), polarisation stability (~25-30 dB), and low dark-event rate together with detection efficiency; similar dependencies were reported by M. Kumar & B. Mondal (2025). The transition from APD ≈ 60-70% to SNSPD ≈ 80-90% effectively "recovered" several decibels and extended the operating distance without changing the topology, while accounting for receiver time-window jitter refined the boundary of a positive secret key rate. For CV-QKD, the stability of the local-oscillator phase remained critical: the degradation of SNR in homodyne and heterodyne measurements under uncontrolled phase noise was described in detail by S. Kundu *et al*. (2025), which is consistent with the phase-noise and linewidth norms adopted in this work. The combination of authentication, modern error correction, and decoy intensities with narrow time windows reduced statistical biases in single-photon estimation; the corresponding effects were demonstrated by C. Anilkumar *et al.* (2024), and in the developed model this reduced *QBER* without modernising the optics, with deviations from the stated estimates explained by tighter gating and stricter tolerances for intensity stability.

The evolutionary trajectory of the network – from trusted-node to entanglement-based architectures with quantum repeaters – was substantiated by Q. Jude (2024); the obtained profiles confirm the viability of such an approach, provided that sites are reserved for repeaters and DI verification is prepared in urban and regional pilots. The strengthening of cryptographic security through entangled states in a "one-time pad" was demonstrated by G. Pradeep & M.D.S. Nandhini (2024); in applied policy, this meant allocating the quantum key to short-lived critical TLS/IPsec sessions with PFS priority, while bulk high-bitrate flows were encrypted using PQC algorithms for scalability. The heterogeneity of availability in 5G/6G and IoT and the need for geo-redundancy with combined transports are consistent with the conclusions of Durr-E-Shahwar *et al.* (2024); the "fibre + FSO/satellite" hybrid achieves target SLOs with moderate TCO if FSO is used as an urban backup over 1-5 km, and the satellite as an intercity "bridge". Additionally, the operational observations of S.V.S. Pillai & K. Polimetla (2024) were taken into account, that centralisation of key generation increases delays and costs; local edge caches with eviction policies synchronised with rotation smooth "bursts" of control commands and unload backbone links.

Interoperability and manageability were ensured through an SDN/NFV framework with crypto functions as services and centralised KMS/HSM control loops – approaches set out in detail in S. Brightwood *et al.* (2024); on this basis, the architecture avoided "crypto islands" during migration and scaling, and orchestration was tied to SLO classes and domain constraints. The stratification of roles of the physical channel, the key distribution system, and application protocols, generalised by S. Akter (2023), confirms the practicality of separation of duties and simplifies SLO validation. For cloud solutions with artificial intelligence (AI) components, zero-trust principles based on post-quantum algorithms, described by A. Sreerangapuri (2024), envisaged telemetry of the status of quantum components, liveness tests, access audits, and checks of encryption modes in OT environments.

Industrial cases of adopting PQC without breaking existing stacks were generalised by H.C. Ukwuoma *et al*. (2022); the obtained hybrid policy shifted bulk channels to PQC, while the quantum key was reserved for short critical sessions. The outcomes of the NIST process and the arguments in favour of hybridisation, presented by R. Bavdekar *et al*. (2023), were reflected in TCO estimates for different domains. Aspects of corporate governance and risk management in migration to PQC, highlighted by Y.-K. Liu & D. Moody (2024), were supplemented with requirements for regular implementation audits, staff training programmes, and side-channel control; technical issues of compatibility and leakage mitigation were detailed by S. Li *et al*. (2023). The "store now – decrypt later" risk was systematised by S. Singh *et al.* (2024); in practical profiles, this was implemented by short key rotation and mandatory PFS in critical channels, with immutable logs and domain segmentation for government communications. Performance data on CRYSTALS-Kyber and the advisability of hardware offloads at network nodes were provided by M.S.C. Laule *et al*. (2024), while the integration of lattice-based schemes on OpenTitan was demonstrated by T. Stelzer *et al.* (2025) – together, this substantiated encrypting large flows "in hardware" without loss of cryptographic security in banking and energy backbones. The discipline of modes and ciphertext variability as a way to reduce leakage through side channels was detailed by P.-A. Berthet (2024); in regulations, this was implemented through regular compliance tests, non-conformance logs, and rapid configuration rollbacks. The long-term security of lattice-based schemes in hybrid profiles was confirmed by J. Shunza (2019), and A. Dadheech (2021). In elevated-risk scenarios, the complementary role of chaotic dynamics was considered by M.H. Al Hasani & K.A. Al Naimee (2019).

The integration of cryptography into PKI and OT networks with domain segmentation and HSM gateways was described by J.O. del Moral *et al.* (2024), while models for integrating KMS and key-distribution policies in multi-domain environments were presented by A. Geremew & A. Mohammad (2024); comparison with the chosen architecture of inter-agency gateways and centralised management attests to interoperability.

Taken together, manageable standardisation of components (intensity stability, polarisation stability, detector efficiency and noise characteristics), protocol engineering (decoy states, gating, authentication and error correction), and the choice of transports/topologies (fibre with FSO/satellite backup and a "development window" for MDI/DI and repeaters) ensure the achievement of target SLOs without radical replacement of physical infrastructure. For banking networks, rings with trusted nodes and edge key caching are appropriate; for energy systems – deterministic latency in dual fibres "control centre – substation"; for government communications – domain segmentation and HSM gateways with satellite "bridges" for intercity routes. The practical importance lies in outlining a manageable migration trajectory towards long-term cyber resilience: from urban pilots – to regional rings and then to architectures with MDI/DI and quantum repeaters, with a hybrid cryptopolicy QKD + PQC that operates at scale without "crypto islands".

CONCLUSIONS

The study was purely theoretical in nature and, in accordance with its aim, aligned the requirements for three key QKD subsystems: radiation sources with controlled frequency and intensity stability and stable polarisation, photon detectors with high efficiency at low dark-count events and low jitter, and quantum communication channels in optical fibre and free space, taking into account the loss budget and fluctuations.

It was established that, without trusted nodes, practically achievable fibre distances are about 150-200 km; for urban FSO lines, routes of 5-20 km are appropriate; scaling up requires segmentation via a network of trusted nodes, satellite segments, or the emergence of quantum repeaters. Applied architectural profiles were formed: for banking networks, rings with FSO backup and session rotation based on QKD keys; for energy-system control, dual fibre channels between control centres and substations with a local key cache; for government communications, segmented domains with inter-agency gateways and satellite redundancy.

Quantitative indicators showed that, without trusted nodes in fibre, 150-200 km is practically achievable with a loss budget of ~8-20 dB, with secret key rates at the level of tens of kbit/s; at the component level, maintaining intensity stability ⩽1-2%, polarisation stability ⩾25-30 dB, and detection efficiency of 60-70% (APD) and 80-90% (SNSPD) extends operating distances; the transition APD→SNSPD adds approximately 15-30%, while an additional +1 dB of loss reduces the secret-key-rate margin by roughly 20%.

The obtained values confirm the proposed hypothesis about the practical suitability of a QKD + PQC hybrid for urban and regional networks with achievable SLOs without unacceptable TCO. Compatibility with existing infrastructure was shown through a QKD plus PQC hybrid with integration into TLS or IPsec and KMS or HSM under SDN or NFV policy control, as well as a predictable life-cycle cost structure. A phased deployment was recommended, from an urban pilot to regional rings and intercity backbones, with prioritisation of critical flows, short rotation, and PFS.

The limitations of the conclusions are linked to the absence of empirical validation, the climatic sensitivity of FSO, and dependence on assumed component parameters. Further directions include quantum repeaters, broader MDI or DI verification, improving the quality of sources and detectors, and native integration of QKD plus PQC hybrids into network services of Ukraine's critical infrastructure.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Akter, S. (2023). Quantum cryptography for enhanced network security: A comprehensive survey of research, developments, and future directions. *Arxiv*. doi: 10.48550/arxiv.2306.09248.

[2] Al Hasani, M.H., & Al Naimee, K.A. (2019). Impact security enhancement in chaotic quantum cryptography. *Optics & Laser Technology*, 119, article number 105575. doi: 10.1016/j.optlastec.2019.105575.

[3] Anilkumar, C., Lenka, S., Neelima, N., & Sathishkumar, V.E. (2024). A secure method of communication through BB84 protocol in quantum key distribution. *Scalable Computing Practice and Experience*, 25(1), 21-33. doi: 10.12694/scpe.v25i1.2152.

[4] Bavdekar, R., Chopde, E.J., Agrawal, A., Bhatia, A., & Tiwari, K. (2023). Post quantum cryptography: A review of techniques, challenges and standardizations. In *2023 international conference on information networking* (pp. 146-151).Bangkok: IEEE. doi: 10.1109/ICOIN56518.2023.10048976.

[5] Berthet, P.-A. (2024). Ciphertext malleability in lattice-based KEMs as a countermeasure to side channel analysis. *Arxiv*. doi: 10.48550/arXiv.2409.16107.

[6] Brightwood, S., Jarry, H., Frank, E., & Olusegun, J. (2024). *Network security and Quantum cryptography: Challenges and opportunities*. Retrieved from https://surli.cc/rlkwmt.

[7] Chen, C.-L., Zeng, K.-W., Li, W.-Y., Lee, C.-F., Liu, L.-C., & Deng, Y.-Y. (2025). Lightweight post-quantum cryptography: Applications and countermeasures in internet of things, blockchain, and E-learning. *Engineering Proceedings,* 103(1), article number 14. doi: 10.3390/engproc2025103014.

[8] Dadheech, A. (2021). Post-quantum lattice-based cryptography: A quantum-resistant cryptosystem. In N. Kumar, A. Agrawal, B.K. Chaurasia & R.A. Khan (Eds.), *Limitations and future applications of quantum cryptography* (pp. 102-123). Hershey: IGI Global Scientific Publishing. doi: 10.4018/978-1-7998-6677-0.ch006.

[9] del Moral, J.O., iOlius, A.D., Vidal, G., Crespo, P.M., & Martinez, J.E. (2024). Cybersecurity in critical infrastructures: A post-quantum cryptography perspective. *IEEE Internet of Things Journal*, 11(18), 30217-30244. doi: 10.1109/JIOT.2024.3410702.

[10] Dharanish, P., Kokila, S., Mythily, M., & Balachandran, A. (2024). FPGA implementation of post quantum cryptography for high performance. In V. Sharmila, S. Kannadhasan, A.R. Kannan, P. Sivakumar & V. Vennila (Eds.), *Challenges in information, communication and computing technology*: *Proceedings of the 2nd international conference on challenges in information, communication, and computing technology* (pp. 209-213). London: CRC Press. doi: 10.1201/9781003559092.

[11] Djordjevic, I.B. (2022). Physical-layer security, quantum key distribution, and post-quantum cryptography. *Entropy*, 24(7), article number 935. doi: 10.3390/e24070935.

[12] Durr-E-Shahwar, Imran, M., Altamimi, A.B., Khan, W., Hussain, S., & Alsaffar, M. (2024). Quantum cryptography for future networks security: A systematic review. *IEEE Access*, 12, 180048-180078. doi: 10.1109/ACCESS.2024.3504815.

[13] Faba, J., Brito, J.P., Cresta, J., Pastor, A., López, D., Brito, R., Buruaga, J.S., & Martin, V. (2025). A hybrid approach to address the transition to quantum-resistant cryptography in telecommunication environments. In *2025 international conference on quantum communications, networking, and computing* (pp. 122-126). Nara: IEEE. doi: 10.1109/QCNC64685.2025.00027.

[14] Geremew, A., & Mohammad, A. (2024). Preparing Critical infrastructure for post-quantum cryptography: Strategies for transitioning ahead of cryptanalytically relevant quantum computing. *International Journal on Engineering Science and Technology*, 6(4), 338-365. doi: 10.46328/ijonest.240.

[15] Ghalaii, M., *et al*. (2023). Satellite-based quantum key distribution in the presence of bypass channels. *PRX Quantum*, 4, article number 040320. doi: 10.1103/PRXQuantum.4.040320.

[16] Gyongyosi, L., Bacsardi, L., & Imre, S. (2019). A survey on quantum key distribution. *Infocommunications Journal,* 11(2), 14-21. doi: 10.36244/ICJ.2019.2.2.

[17] Hoschek, M. (2021). Quantum security and 6G critical infrastructure. *Serbian Journal of Engineering Management*, 6(1), 1-8. doi: 10.5937/SJEM2101001H.

[18] Jude, Q. (2024). *Quantum entanglement and the future of secure quantum computing*. Retrieved from https://surl.li/baztmk.

[19] Kumar, M., & Mondal, B. (2025). A brief review on quantum key distribution protocols. *Multimedia Tools and Applications*, 84(27), 33267-33306. doi: 10.1007/s11042-024-20535-x.

[20] Kundu, S., Gupta, T., Sardar, A., Bandyopadhyay, A., Swain, S., & Mallik, S. (2025). A survey on quantum computing: Transforming cryptography, AI/ML, blockchain, and network communication. *Franklin Open*, 12, article number 100371. doi: 10.1016/j.fraope.2025.100371.

[21] Laule, M.S.C., Silva, J.E.O., & Hanco, H.N. (2024). Lattice-based cryptography: Development and analysis of a new variant of the crystals-kyber algorithm. *Interfaces,* 20, 165-184. doi: 10.26439/interfases2024.n020.7383.

[22] Li, S., Chen, Y., Chen, L., Liao, J., Kuang, C., Li, K., Liang, W., & Xiong, N. (2023). Post-quantum security: Opportunities and challenges. *Sensors*, 23(21), article number 8744. doi: 10.3390/s23218744.

[23] Lingaraju, M.P., Kumar, B.T., & Jayachandra, C. (2019). Quantum entanglement and its implications for secure communication. *World Journal of Advanced Research and Reviews*, 1(1), 82-88. doi: 10.30574/wjarr.2019.1.1.0003.

[24] Liu, Y.-K., & Moody, D. (2024). Post-quantum cryptography and the quantum future of cybersecurity. *Physical Review Applied*, 21, article number 040501. doi: 10.1103/PhysRevApplied.21.040501.

[25] Nguyen, T.-T., Khac, T.-V., & Quynh, L.-N. (2023). Simulation of the BB84 quantum key exchange protocol. In *2023 15th international conference on knowledge and systems engineering* (pp. 1-4). Hanoi: IEEE. doi: 10.1109/KSE59128.2023.10299471.

[26] Pillai, S.V.S., & Polimetla, K. (2024). Analyzing the impact of quantum cryptography on network security. In *2024 international conference on integrated circuits and communication systems* (pp. 1-6). Raichur: IEEE. doi: 10.1109/ICICACS60521.2024.10498417.

[27] Pradeep, G., & Nandhini, M.D.S. (2024). Towards efficient quantum cryptography: Enhancing QOTP with entanglement based techniques. *Journal of Dynamics and Control,* 8(11), 166-183. doi: 10.71058/jodac.v8i11017.

[28] Rahmayanti, D. (2025). Quantum key distribution (QKD) as a wireless telecommunications security solution. *Journal of Informatics and Computer Technology*, 5(1), 10-26. doi: 10.55606/jitek.v5i1.5765.

[29] Rajpoot, S., Singh, K., Singh, C.P., & Sharma, B.K. (2023). Security challenges and future research in quantum key distribution networks. In *2023 10th IEEE uttar pradesh section international conference on electrical, electronics and computer engineering* (pp. 95-100). Gautam Buddha Nagar: IEEE. doi: 10.1109/UPCON59197.2023.10434510.

[30] Rayhan, A., Bushra, Z.I., Chowdhury, S., Faruqui, N., & Barua, B. (2025). Performance analysis of free space optical communication in urban rain environment. In *2025 international conference on electrical, computer and communication engineering* (pp. 1-6). Chittagong: IEEE. doi: 10.1109/ECCE64574.2025.11013976.

[31] Reddy, D.V., Nerem, R.R., Nam, S.W., Mirin, R.P., & Verma, V.B. (2020). Superconducting nanowire single-photon detectors with 98% system detection efficiency at 1550 nm. *Optica*, 7(12), 1649-1653. doi: 10.1364/OPTICA.400751.

[32] Shao, Y., Wang, H., Pi, Y., Huang, W., Li, Y., Liu, J., Yang, J., Zhang, Y., & Xu, B. (2021). Phase noise model for continuous-variable quantum key distribution using a local local localoscillator. *Physical Review A*, 104, article number 032608. doi: 10.1103/PhysRevA.104.032608.

[33] Shunza, J. (2019). *Lattice-based cryptosystems and post-quantum cryptography*. Kigali: African Institute for Mathematical Sciences. doi: 10.13140/RG.2.2.18756.50563.
[34] Singh, S., Jha, C.K., Bende, A., Rana, V., Patkar, S., Drechsler, R., & Merchant, F. (2024). MemSPICE: Automated simulation and energy estimation framework for MAGIC-based logic-in-memory. In *Proceedings of the 29th Asia and south pacific design automation conference* (pp. 282-287). Incheon: IEEE. doi: 10.1109/ASP-DAC58780.2024.10473924.
[35] Sreerangapuri, A. (2024). Post-quantum cryptography for ai-driven cloud security solutions. *International Journal for Multidisciplinary Research*, 6(5). doi: 10.36948/ijfmr.2024.v06i05.29032.
[36] Stanley, M., Gui, Y., Unnikrishnan, D., Hall, S.R.G., & Fatadin, I. (2022). Recent progress in quantum key distribution network deployments and standards. *Journal of Physics Conference Series*, 2416, article number 012001. doi: 10.1088/1742-6596/2416/1/012001.
[37] Stelzer, T., Oberhansl, F., Schupp, J., Karl, P., & Turcuman, H. (2025). Extended version: Enabling lattice-based post-quantum cryptography on the opentitan platform. *Journal of Cryptographic Engineering*, 15, article number 11. doi: 10.1007/s13389-025-00369-5.
[38] Ukwuoma, H.C., Arome, G., Thompson, A., & Alese, B.K. (2022). Post-quantum cryptography-driven security framework for cloud computing. *Open Computer Science*, 12(1), 142-153. doi: 10.1515/comp-2022-0235.
[39] Yu, C., Xu, Q., & Zhang, J. (2024). Recent advances in InGaAs/InP single-photon detectors. *Measurement Science and Technology,* 35(12), article number 122003. doi: 10.1088/1361-6501/ad76ca.

Ігор Лімарь

Кандидат технічних наук, старший викладач
Державний університет інтелектуальних технологій і зв'язку
65023, вул. Кузнечна, 1, м. Одеса, Україна
Інженерно-технологічний інститут «Біотехніка» Національної академії аграрних наук України
67667, вул. Маяцька дорога, 26, с. Хлібодарське, Україна
https://orcid.org/0000-0002-8972-9935

Євген Севастєєв

Магістр, старший викладач
Державний університет інтелектуальних технологій і зв'язку
65023, вул. Кузнечна, 1, м. Одеса, Україна
https://orcid.org/0000-0003-1165-1119

Квантова криптографія: теоретичні основи та практичні імплементації для захисту критичної інфраструктури

Анотація. Метою статті було теоретично визначити узгоджені критерії застосовності квантового розподілу ключів для захисту критичної інфраструктури з урахуванням довготривалого ризику дешифрування з боку супротивників, що володіють квантовими обчислювальними ресурсами та архівацією трафіку. Методологія спиралася на теоретичний системний аналіз трьох підсистем Quantum Key Distribution джерел, детекторів і каналів (волокно/вільний простір), доповнений спрощеними моделями лінк-бюджету та імовірності бітової помилки. Проведено стислий сценарний аналіз для банківських мереж, енергосистем і державних комунікацій з урахуванням масштабованості, сумісності та вартості. Основні результати показали, що стабільність інтенсивності лазерів 1-2 % і вища ефективність детекторів – 60-70 % для Avalanche Photodiode та 80-90 % для Superconducting Nanowire Single-Photon Detector – розширюють практичні відстані та знижують похибки. Без довірених вузлів у волокні досяжно 150-200 км; для міських Free-Space Optics ліній оптимальні 5-20 км, а більші дистанції потребують сегментації мережі чи супутникових відрізків. Архітектурно обґрунтовано: для банків – кільця з довіреними вузлами та Free-Space Optics резервом; для енергосистем – подвійні волоконні канали "диспетчерська-підстанція" з локальним кешем ключів; для держзв'язку – сегментовані домени з міжвідомчими шлюзами та супутниковим резервуванням. У всіх сценаріях доцільний гібрид Quantum Key Distribution+Post-Quantum Cryptography із короткою ротацією ключів і операційними контрзаходами проти атак на детектори та канали, що підтверджує практичну придатність для міських і регіональних мереж. Практична значимість полягає в наданні інженерним командам банків, енергетики та держзв'язку готових архітектурних профілів – від кілець і подвійних волоконних каналів до сегментованих доменів – з інтеграцією Key Management System/Hardware Security Module, визначенням Service Level Objective/Service Level Agreement та поетапними дорожніми картами. Для регуляторів і операторів це база для оновлення вимог і аудитів, а також для планування резервування (Free-Space Optics/супутник) і категорій Total Cost of Ownership

Ключові слова: розподіл ключів; лазерні джерела; фотонні детектори; оптичне волокно; зв'язок у вільному просторі; довірені вузли; банківські мережі

Machinery & Energetics
Vol. 16, No. 4. 2025

Journal homepage:
https://technicalscience.com.ua/en

UDC 004.738.5:004.042
DOI: 10.31548/machinery/4.2025.21

Khrystyna Terletska*
Bachelor
Lviv Polytechnic National University
79000, 12 Stepan Bandera Str., Lviv, Ukraine
https://orcid.org/0009-0002-7302-2625

Performance trade-offs between predictive and reactive autoscaling in stateful microservices

Abstract. The relevance of studying the trade-offs between reactive and predictive autoscaling of stateful microservices was due to the need to improve stability, resource performance and reliability of cloud systems in dynamic conditions. The aim of the study was to assess the effectiveness of reactive, predictive and hybrid autoscaling strategies in stateful microservice architectures. The methodological basis of the research relied on comparative, content, systems and structural-functional analyses, as well as modelling, which made it possible to comprehensively investigate the performance and stability of different approaches to autoscaling in stateful microservices. Theoretical analysis showed that autoscaling in microservice systems provides adaptive performance, and the choice of strategy depends on the nature of the workload, latency sensitivity and consistency requirements. The review of models demonstrated that reactive autoscaling responds quickly to changes but is accompanied by high latency, performance fluctuations and the risk of state divergence between replicas, whereas predictive autoscaling reduces scaling delay and increases stability and resource efficiency but requires accurate models and additional computational resources. Analysis of the results revealed that hybrid strategies integrate the advantages of both approaches, providing both flexibility and stability, and that system efficiency is determined by a trade-off between latency, stability, state consistency, resource efficiency and financial costs. The results of the study may be useful for developers and architects of cloud microservice systems, IT engineers and DevOps teams for optimising scaling strategies, increasing service stability and ensuring efficient resource use in different cloud environments

Keywords: Kubernetes; resource optimisation; cloud performance; dynamic resource allocation; system stability; observability metrics

INTRODUCTION

Modern cloud microservice architectures require efficient autoscaling mechanisms, especially for stateful services with critical state consistency and high latency sensitivity. Traditional reactive approaches provide basic stability but suffer from reaction delay and inefficient resource use, whereas predictive models make it possible to anticipate load peaks but require complex analytics and precise tuning. In this regard, the study of performance trade-offs between reactive and predictive approaches and the development of a conceptual model for choosing a scaling strategy for stateful environments are relevant and make it possible to improve stability, resource efficiency and reliability of cloud systems in dynamic conditions.

In the modern academic discourse, research on microservice autoscaling increasingly focuses on adaptability, load prediction and resource optimisation. Thus, in the work of I. Varzhanskiy *et al.* (2025), approaches to load forecasting and strategic resource analysis were considered, demonstrating the potential of applying predictive and hybrid methods to increase system resilience and efficiency. The study showed that the integration of analytical methods and forecasting makes it possible to improve scaling planning and reduce the risks of peak loads. In the work of O. Romanov *et al.* (2024), dynamic scaling of containerised applications using Kubernetes Cluster Autoscaler and Auto Scaling Group was analysed. The authors

Article's History: Received: 10.07.2025; Revised: 15.10.2025; Accepted: 27.11.2025.

Suggested Citation:
Terletska, Kh. (2025). Performance trade-offs between predictive and reactive autoscaling in stateful microservices. *Machinery & Energetics,* 16(4), 21-33. doi: 10.31548/machinery/4.2025.21.

*Corresponding author

demonstrated that the combined use of horizontal and vertical scaling reduces latency and increases the performance of stateful microservices in multicluster environments. The results of the work confirmed the relevance of combining reactive, predictive and hybrid strategies to achieve an optimal balance between stability, reaction speed and resource efficiency. At the same time, the study of T. Babenko *et al.* (2025) focused on automated data collection and risk assessment of digital assets using AI methods. The analytical and predictive approaches used demonstrated the possibilities of integrating machine learning to optimise the management of complex systems. This highlighted the prospects of applying intelligent algorithms to support scalable and resilient microservice architectures. In the study, B. Guntupalli & S. Vamshi (2023) analysed the design of microservices for processing high data volumes. The results showed that the correct combination of horizontal and vertical scaling has a critical impact on system performance and economic efficiency. The authors emphasised the importance of a structured approach to microservice architecture in large cloud environments.

In studies of autoscaling for stateful microservices, attention was also paid to increasing system stability and resource efficiency. M. Mekki *et al*. (2025) proposed an approach to proactive lifecycle management in multi-cluster containerised environments, which envisages load forecasting and adaptive resource deployment. The study showed that this approach reduces the risk of state divergence between replicas and increases system resilience to load fluctuations, demonstrating practical value in complex distributed architectures. C. Meng *et al.* (2023) developed Deepscaler – a comprehensive horizontal and vertical scaling model based on spatio-temporal graph neural networks with adaptive graph learning. The review showed that the approach makes it possible to effectively forecast load peaks and optimise resource distribution, which reduces system reaction latency and increases the stability of stateful microservices. The results of the study demonstrated the advantages of using deep learning for comprehensive scaling, particularly in scenarios with dynamic and unpredictable workloads. In turn, M. Thompson *et al*. (2023) focused on intelligent workload balancing in multi-cloud environments, applying adaptive algorithms for resource distribution between providers. The study showed that such an approach makes it possible to increase system performance and fault tolerance while reducing the risk of overloading individual components. The literature review demonstrates that the integration of intelligent balancing strategies with predictive scaling provides comprehensive optimisation of stateful microservice operation in complex and distributed cloud infrastructures. T. Prasad (2023) proposed an AI-driven predictive scaling approach that uses machine-learning algorithms to forecast workload and automatically allocate resources. The study showed that the use of AI reduces scaling latency and increases system stability, and the results demonstrated the effectiveness of the predictive approach in complex cloud environments with dynamic workloads.

Previous studies focused on the application of reactive and predictive autoscaling models for cloud and cloud-native architectures, paying attention to forecasting efficiency, system stability and resource optimisation. At the same time, most works leave the performance trade-offs in stateful microservices without detailed analysis, in particular the relationship between reaction latency, state consistency, stability and financial costs, and do not offer a comprehensive assessment of hybrid strategies. The aim of the study was to examine the effectiveness of reactive, predictive and hybrid autoscaling strategies in stateful microservice architectures, taking into account performance trade-offs and financial feasibility. To achieve this aim, the following objectives were set: to analyse the theoretical aspects of these strategies in terms of the impact on latency, stability, state consistency and efficiency of resource provision, as well as to develop a generalised comparison model for strategies in stateful microservices, enabling the identification of key trade-offs and contextual factors for choosing the optimal model.

MATERIALS AND METHODS

The study was an interdisciplinary analytical investigation in which elements of theoretical, comparative, modelling and predictive analysis were combined. The theoretical-analytical group of methods included systems analysis and the method of analytical generalisation, which made it possible to identify the main trends in the development of scaling strategies in stateful environments and to classify approaches according to adaptability, type of computational load, consistency requirements and latency level. A comprehensive source base was formed in the study, which included 34 peer-reviewed journal articles selected as a result of searches in international scientometric databases, in particular Scopus, Web of Science and Google Scholar. The criteria for selecting sources were the relevance of publications (2021-2025), the direct relevance to the research topic, scientific novelty and practical significance for assessing the effectiveness of reactive, predictive and hybrid autoscaling strategies in stateful microservices.

The systems method assumed a comprehensive view of the research object as an integrated structure in which all elements were interconnected and interdependent. Its application in this study was conditioned by the need to consider the scaling process not in isolation but as a result of the interaction of several infrastructure levels – the orchestrator, service mesh, autoscaler and load balancer. Within the systems analysis, key links between these components were identified, which ensured the stability and performance of stateful microservices. The use of this method made it possible to reveal critical factors influencing scaling efficiency and to form a generalised logical model of interaction between infrastructure components. The use of the method of analytical generalisation was determined by the need to integrate disparate approaches to autoscaling presented in contemporary academic and applied sources. Within this method, typical decision-making scenarios in the scaling process were analysed and common features among reactive, predictive and hybrid approaches were distinguished. The application of analytical generalisation made it possible to construct a generalised classification of scaling strategies that reflected the logic, strengths, and limitations in the context of stateful environments.

Comparative-structural methods involved the assessment of practical scaling implementations in leading cloud ecosystems (Amazon Web Services, Google Cloud Platform and Azure), taking into account the architectural features of each. Comparative analysis provided an opportunity to compare mechanisms for managing container state (stateful sets, persistent volumes, checkpointing) and to assess the impact on system stability. In turn, structural-functional analysis made it possible to trace the logic of interaction between infrastructure components and to determine the features of performance monitoring, load forecasting and balancing between nodes. The application of this method resulted in the identification of patterns in the choice of scaling strategies depending on Service Level Agreement parameters, which created the basis for forming generalised conclusions regarding the effectiveness of different scaling models.

Conceptual modelling methods ensured the formalisation of complex systemic dependencies through the construction of logical, structural and hierarchical models. The use in the study was driven by the need to generalise the results of the analytical and comparative stages and to present the process of choosing a scaling strategy as an integrated conceptual scheme. The application of the systems-hierarchy method made it possible to form a multilevel structure of criteria within which reactive, predictive and hybrid approaches were compared according to the type of workload, latency sensitivity and data-consistency requirements. As a result, conceptual modelling made it possible to integrate the disparate analysis results into a single formalised model that visually represented the decision-making logic and the relationship between the technical characteristics of the system and the effectiveness of the choice of scaling strategy. Analytical-predictive methods were justified by the need to forecast the effectiveness of scaling strategies in stateful microservices and to assess the potential for integrating intelligent autoscaling solutions based on machine learning. The use of analytical-predictive methods made it possible to extrapolate the identified patterns to prospective scenarios, to evaluate the impact of predictive strategies on resource optimisation and to determine the contribution to increasing the stability of stateful microservices.

RESULTS

Theoretical foundations of autoscaling in microservice architecture

Autoscaling in microservice systems acts as a key tool for maintaining adaptive performance under dynamically changing workloads. Theoretically, this process is based on the principles of self-organisation and feedback control, which allow the infrastructure to adapt its resources automatically to current and forecast needs without operator intervention. According to dynamic load control models, autoscaling is a mechanism that minimises the difference between the actual and target level of resource utilisation, ensuring stability and efficiency in the operation of cloud services. In a microservice architecture, each component of the system performs a narrowly specialised function, so scaling is carried out granularly, i.e., at the level of individual services rather than the whole application (Quattrocchi *et al.*, 2024). This creates conditions for local optimisation of performance but at the same time generates complex interaction dynamics between services, which requires theoretical modelling based on the principles of discrete control systems and distributed coherence. In the context of Kubernetes, autoscaling is implemented as a multilevel control system that combines monitoring, analytics and orchestration. The main theoretical foundation is the control loop, in which three components (observation, analysis, and action) interact cyclically, forming a mechanism of continuous adaptation. This model is consistent with the concept of cybernetic feedback control, where the system not only reacts to changes but also predicts future trends on the basis of historical data (which becomes the basis for the transition from reactive to predictive autoscaling) (Tabilo Alvarez & Ramírez-Correa, 2023). In cloud environments such as Amazon Web Services, Google Cloud Platform or Azure, similar mechanisms are integrated into global Service Level Objectives policies, which set permissible limits for stability and performance metrics.

Horizontal scaling consists in increasing the number of homogeneous service instances to distribute the load. In Kubernetes, this approach is implemented through the Horizontal Pod Autoscaler, which continuously monitors Central Processing Unit (CPU) metrics or custom metrics (for example, latency or the number of requests per second). Theoretically, horizontal scaling is based on the model of massive parallelism, where system performance is directly proportional to the number of available instances but is limited by Amdahl's Law. Its main advantage is linear scalability for stateless services, whereas the main challenge is the coordination of load balancing and inter-service dependencies. Vertical scaling, in contrast, focuses on optimising the parameters of each instance, i.e. increasing computing power (CPU, Random Access Memory (RAM), storage) without changing the number of pods. In Kubernetes, this approach is implemented in the form of the Vertical Pod Autoscaler, which uses statistical methods to assess average resource consumption and dynamically adjusts requests and limits. From a theoretical point of view, the Vertical Pod Autoscaler is a manifestation of the model of adaptive parameter control, within which the system reduces the imbalance between node capacity and current workload. The main drawback is that pod restarts may temporarily reduce service availability (Razavi *et al.*, 2024).

Hybrid scaling combines the advantages of both previous approaches and uses multilevel analytics to control both the number of pods and the resource parameters simultaneously. Theoretically, it relies on the model of multiloop control, where one loop is responsible for reactive regulation and the other for predictive optimisation of resources. This approach is particularly effective in scenarios where there is high workload variability and system sensitivity to latency. Comparative analysis presented in Table 1 shows that the choice of scaling type should be context-dependent – based on the nature of the workload, latency sensitivity, consistency requirements and the capabilities of the Kubernetes infrastructure.

Table 1. Comparative characteristics of scaling types in Kubernetes microservice environments

Type of scaling	Theoretical model	Main purpose	Advantages	Disadvantages
Horizontal Pod Autoscaler	Parallelism model	Load distribution	High flexibility, simplicity	Not suitable for stateful
Vertical Pod Autoscaler	Adaptive parameter control	Optimisation of pod capacity	Efficient use of resources	Requires restart
Hybrid	Multiloop control	Performance balancing	High stability, adaptability	Implementation complexity

Source: created by the author on the basis of analysis of data from F. Almeida & B. Bálint (2024), K. Razavi *et al.* (2024)

As shown in the table, each type of scaling has its own theoretical basis, purpose, and trade-offs. Horizontal scaling is the simplest to implement and provides high flexibility, but it has limitations for stateful services, where state coordination between replicas complicates parallel operation. Vertical scaling uses resources efficiently but requires container restarts, which may lead to short-term system interruptions. Hybrid scaling is considered the optimal strategy for complex workloads, as it combines the adaptability of reactive solutions with the stability of predictive ones, although it requires complex orchestration and precise tuning of interaction parameters between control layers. Scaling stateful services constitutes a separate theoretical and applied challenge, as it violates the classical assumption of instance independence. In systems with internal state (databases, caches, message queues, streaming data stores), key issues become data consistency, availability, and fault tolerance. From a theoretical point of view, this means that the autoscaling process must be aligned with the constraints of Consistency, Availability, Partition tolerance (the CAP theorem), which proves the impossibility of simultaneously ensuring full consistency, availability, and partition tolerance. In Kubernetes, the management of stateful components is carried out through the StatefulSet object, which ensures persistent pod identity, a guaranteed order of start and termination, as well as stable data storage. Unlike Deployment for stateless components, StatefulSet preserves a pod → persistent volume map, which makes it possible to implement localised state coherence. However, theoretically, effective scaling of such systems requires state-synchronisation models that include replication, latency control and cluster-stability forecasting. Thus, autoscaling in stateful environments is not only resource management but a complex task of dynamically maintaining system invariants, where a scaling decision must take into account the topology of connections between nodes, the volume of data transferred and synchronisation delays (Kumari *et al.*, 2023).

The theoretical foundations of autoscaling in microservice architecture presuppose a synthesis of concepts of feedback control, resource optimisation and coherent state management. For Kubernetes and similar systems, autoscaling appears as a cybernetic regulation system operating at the intersection of three axes – type of scaling, nature of service and observability metrics. In theoretical terms, this creates a basis for further differentiation between reactive and predictive approaches to scaling, which are considered the next level in the evolution of automated control systems in cloud environments.

Review of principles, algorithms, and architecture of reactive and predictive autoscaling models

Autoscaling models in cloud microservice architectures have evolved from simple reactive mechanisms to complex predictive systems capable of taking into account temporal patterns of load, correlations between components and user behaviour patterns. In theoretical terms, these models are based on different approaches to control – reactive, which involves a direct response to current system metrics, and predictive, focused on forecasting load changes using time-series prediction models or neural networks. The comparison makes it possible to reveal the fundamental principles of building adaptive control systems that underlie Kubernetes, Amazon Web Services Auto Scaling or Azure Autoscale.

Reactive scaling is the most widespread paradigm of automatic resource regulation in cloud environments. Theoretically, it relies on a feedback-control model, within which the system observes current metrics (for example, CPU utilisation, memory usage, latency, throughput) and performs scaling in response to deviations from predetermined thresholds (Chilakala *et al.*, 2024). Mathematically, such scaling can be described in the form of a Proportional-Integral-Derivative (PID-controller) type regulator, which responds to the error between the desired and actual metric value. The advantage of reactive models lies in the implementation simplicity, deterministic behaviour and high predictability. However, the main drawback is response latency, which may lead to short-term overloads or excessive resource use. In the context of Kubernetes, the implementation of such an approach is the Horizontal Pod Autoscaler, which regularly (for example, every 15 seconds) analyses the average workload of pods and increases or decreases the number according to the scaling formula (1):

$$N_{desired} = N_{current} \times \frac{U_{current}}{U_{target}}, \tag{1}$$

where $N_{desired}$ – the new number of pods; $U_{current}$ – the current average workload; U_{target} – the target value of the metric.

The formula is a clear example of the mathematical representation of the principle of operation of the reactive model: from the measured error to resource correction, and explains how the Horizontal Pod Autoscaler system transforms observed workload metrics into specific actions – changing the number of pods, which ensures the maintenance of the desired performance level. Thus, reactive scaling is a post factum correction of the system state that supports short-term stability but does not provide adaptive forecasting. Its application is appropriate in systems with stable or weakly variable workloads, where the risk of peak spikes is minimal.

Threshold models are the basic form of reactive scaling (Fig. 1). The principle lies in defining hard limits of permissible metric values: if the average CPU workload exceeds, for example, 80%, the system adds a new pod when the threshold is crossed. The advantage of threshold models lies in the low computational complexity and ease of parameterisation, which makes the mentioned models popular in industrial systems. However, these models are prone to oscillations – frequent changes in the number of instances when metrics fluctuate near boundary values. To mitigate this effect, hysteresis or moving averages are used.

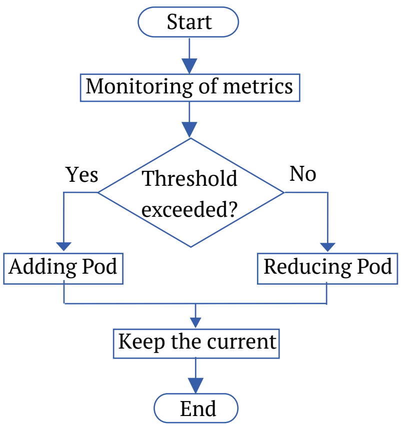


Figure 1. Logic of threshold scaling in a reactive model
Source: created by the author

According to the analytical model presented in the figure, the key advantage of this approach is responsiveness – the system quickly reacts to peak loads without the need for complex forecasting. At the same time, such logic has a number of structural limitations, in particular: a delay between the moment the threshold is exceeded and the actual scaling, the risk of performance fluctuations due to frequent metric oscillations, as well as increased resource consumption in cases of excessive scaling. Thus, the presented scheme reflects not only the mechanism of operation of the reactive model, but also the basis for the subsequent transition to predictive and hybrid strategies designed to reduce the delay between detection and system response.

The transition from reactive to predictive models has become a key point in the development of autoscaling theory. The main idea of the predictive approach lies in forecasting future workload on the basis of historical data and temporal patterns. Theoretically, these models are based on time-series forecasting methods or machine learning, which makes it possible to reduce reaction lag and prevent overloads. Among classical statistical methods, the following are used: AutoRegressive Integrated Moving Average (ARIMA) – for forecasting short-term workload changes on the basis of autocorrelation; Exponential Smoothing – for identifying trends and seasonal fluctuations in resource use; Kalman Filter – for adaptive correction of forecasts in real time (Table 2). More modern approaches are based on neural networks, in particular Long Short-Term Memory (LSTM) or Temporal Convolutional Networks (TCN), which are capable of capturing non-linear dependencies and long-term workload patterns. In such models, scaling decisions are based not only on current metrics but also on a probabilistic assessment of the future system state. For example, if the forecast workload exceeds current capacity by 20% over the next five minutes, the system may pre-emptively activate new pods, avoiding performance degradation.

As shown in the table, ARIMA models time series and is effective for short-term, linear processes but is not suitable for complex dynamic systems with non-linear dependencies. LSTM, based on deep learning, can identify non-linear workload patterns, providing high accuracy, but requires significant computational resources and lengthy training. Kalman Filter, in turn, combines stochastic principles of filtering and adaptive control, ensuring flexible response to changes in real time, but is sensitive to noise and errors in input metrics. Thus, the choice of forecasting model should be based on the nature of the workload: ARIMA – for stable and periodic systems, LSTM – for highly volatile and dynamic environments, and Kalman Filter – for scenarios where the speed of adaptation to metric changes in real time is critical. In real cloud environments, reactive and predictive mechanisms are often combined in a hybrid autoscaling architecture, where short-term fluctuations are handled reactively and medium- and long-term trends – predictively. Architecturally, such a system consists of three layers: monitoring and observability – collection of metrics, logs and tracing via Prometheus, Grafana, OpenTelemetry; analytical layer – metric processing, time-series modelling, workload forecasting; orchestration layer – implementation of decisions via the Kubernetes Application Programming Interface (API), Horizontal Pod Autoscaler/Vertical Pod Autoscaler controllers or custom webhooks (Fig. 2).

Table 2. Comparative characteristics of predictive autoscaling models in cloud environments

Type of model	Algorithmic basis	Main purpose	Advantages	Limitations
ARIMA	Linear time series	Short-term forecasting	Simplicity, interpretability	Does not work with nonlinear data
LSTM	Deep learning	Nonlinear patterns	High accuracy	High computational costs
Kalman Filter	Stochastic control	Real-time adaptability	Fast response	Sensitivity to data noise

Source: created by the author based on data analysis by H.S. Chilakala *et al.* (2024)

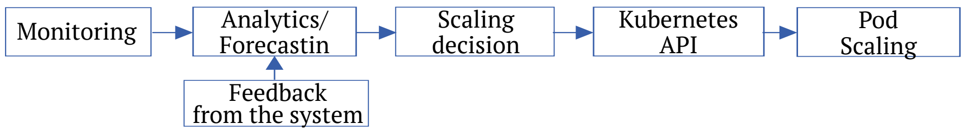


Figure 2. Architecture of a combined autoscaling approach
Source: created by the author on the basis of analysis of data from V. Rampérez *et al.* (2021)

This architecture corresponds to the theoretical model of adaptive control with a double loop, in which the outer loop performs long-term optimisation and the inner loop short-term stabilisation. In summary, it can be argued that the evolution of autoscaling models from reactive to predictive reflects the transition from deviation-based regulation to forecast-based regulation, i.e. from the classical cybernetic control model to an intelligent self-tuning system. Reactive models provide basic stability but are limited in dynamic environments; predictive models, due to the use of machine-learning methods and time-series models, form the basis of cognitive autoscaling capable of self-learning and resource optimisation in real time.

Theoretical analysis of performance trade-offs

In theoretical terms, the effectiveness of autoscaling systems in microservice architectures is always assessed through the prism of performance trade-offs. The term performance trade-offs refers to the balance between a number of interrelated parameters: latency, system stability, state consistency, resource efficiency and scaling cost. The application of different autoscaling models (reactive or predictive) determines which aspects of performance will be prioritised and where certain compromises are possible.

One of the priority parameters is scaling latency, i.e., the time from the moment overload is detected to the actual addition of new resources. In reactive models, this latency usually includes: the time taken to collect metrics, compute a decision based on thresholds, initiate scaling via the Kubernetes API and start a new pod. Theoretically, even with optimal configuration, reactive systems demonstrate higher latency, which may lead to short-term overload and deterioration of Quality of Service. In predictive models, scaling latency is compensated by anticipatory decision-making, when new resources are activated before critical workloads are reached. However, forecast accuracy directly affects the effectiveness of this approach: inaccurate forecasts may lead to unjustified resource allocation, which increases the cost of scaling without real benefit for latency.

Stability is defined as the absence of frequent and sharp fluctuations in the number of pods or resource use. In reactive systems, high sensitivity to thresholds often causes "flipping": a short-term threshold exceedance leads to the addition of pods, and a decrease in workload to the removal almost immediately afterwards. Such behaviour creates performance oscillations and may worsen the stability of stateful services, especially when state consistency between replicas is important. Predictive autoscaling, by contrast, makes it possible to smooth out fluctuations, since scaling decisions are made on the basis of forecast trends rather than instantaneous spikes. However, incorrect model configuration or failure to take sudden peaks into account may lead to systematic instability, for example when the system allocates resources in advance, but the peak does not occur (Kanwal *et al.*, 2024). For stateful microservices, a critical characteristic is state consistency between replicas, since any changes in the scaling configuration – both scale-out (adding pods) and scale-in (removal) – may disrupt data synchronisation. In the absence of effective replication mechanisms such as Raft or Paxos, the system may temporarily lose consistency, which is particularly dangerous for transactional or analytical services. Within reactive autoscaling, such problems occur more often, since scaling takes place post factum, i.e. after an increase in workload has been detected. At this point, new pods have not yet had time to fully synchronise state with existing instances, which leads to asynchronous state divergence and potential data-consistency errors. Predictive autoscaling, on the other hand, makes it possible to plan scaling in advance, ensuring data replication before the system reaches peak workload. This reduces the risk of state divergence but requires additional computing resources and time costs for preliminary synchronisation. Thus, there is a clear trade-off between response speed and data consistency that must be taken into account when designing the system.

Another aspect is resource efficiency and scaling cost. In the reactive approach, the system often operates in a state of under- or over-provisioning, which leads to unstable performance and increased financial costs. Due to the delay in responding to workload, resources are added with a lag, and after the activity decreases the resources remain unused, generating economic losses – especially in cloud environments where payment is made for allocated resources. The predictive approach, by contrast, makes it possible to optimise resource provisioning, activating pods only when necessary. However, forecast accuracy becomes a critical factor: overly conservative or inaccurate models may lead to suboptimal reservation and, accordingly, additional costs (Maliakel *et al.*, 2025). Overall, the effectiveness of predictive scaling is determined by the balance between model accuracy, performance stability and economic feasibility. The reactive approach prevails in scenarios where rapid response is required, but temporary instability is acceptable, whereas the predictive approach is preferable in environments that require stability, state consistency and optimal resource use. Table 3 presents a structured comparison of the characteristics of each approach, taking into account the impact on latency, stability, state consistency, resource efficiency and financial costs. This format makes it possible to clearly trace in which conditions one approach shows advantages over the other and where trade-offs arise between accuracy, reaction speed and economic feasibility.

Table 3. Comparative characteristics of reactive and predictive autoscaling for stateful microservices

Performance aspect	Reactive autoscaling	Predictive autoscaling	Theoretical trade-off
Latency	High due to post factum response	Reduced thanks to anticipatory forecasting	Forecast accuracy vs. reaction speed
Stability	Prone to oscillations and “flipping”	Higher stability due to trend smoothing	Flexibility of response vs. stability
State consistency	Possible state divergence during scaling	Planning of scaling with regard to replication	Synchronisation time vs. scaling speed
Resource efficiency	Often under- or overprovisioned	Higher with accurate forecasting	Forecast accuracy vs. resource savings
Scaling cost	Higher due to excessive or delayed resource allocation	Optimised but dependent on forecast accuracy	Forecasting vs. financial costs

Source: created by the author on the basis of analysis of data from J. Karol Santos Nunes *et al.* (2024), A. Khaleq & I. Ra (2021)

The results obtained indicate that reactive autoscaling remains more flexible in unpredictable conditions but is characterised by increased latency and a risk of destabilising state. Predictive autoscaling, in turn, provides better stability and resource-use efficiency, especially for stateful microservices, but requires accurate forecasting models and additional computational costs. Thus, the choice of optimal scaling strategy should be based on a balance between consistency requirements, Service Level Agreement, cost constraints and workload dynamics of the system.

Conceptual model for choosing a scaling strategy for stateful environments

In microservice systems with internal state, the choice of an optimal scaling strategy is a complex task that requires balancing between response speed, system stability, state consistency and efficient use of resources. Theoretically, this is linked to performance trade-offs, where none of the autoscaling models – reactive or predictive – can optimise all parameters simultaneously. The choice of approach depends on the characteristics of the system: type of workload, permissible latency, state-consistency requirements, resource availability and the duration of workload trends.

Reactive autoscaling is appropriate in environments with stable or moderately variable workloads where short-term deviations from target metrics are acceptable and response latency is not critical. Theoretically, this approach relies on a feedback-control model, in which scaling takes place post factum, allowing rapid reaction to local changes but not preventing peak overloads. Predictive autoscaling, by contrast, presupposes anticipatory resource management on the basis of forecasts of future workload. This approach is optimal for environments with highly variable workload, critical latency and strict requirements for state consistency. The use of time-series methods, ARIMA models, the Kalman Filter or neural networks (LSTM, TCN) makes it possible to activate resources in advance, reducing the risk of performance degradation and consistency violations. Hybrid autoscaling combines the advantages of both approaches and implements a multi-loop architecture: the inner loop provides short-term reaction to local changes, while the outer loop forecasts medium- and long-term trends. This approach is particularly effective in stateful environments with mixed workloads, where stability, consistency, and resource savings must be maintained simultaneously. Table 4 shows the key criteria for choosing an autoscaling strategy for stateful environments.

Table 4. Comparison of scaling strategies

Criterion	Reactive autoscaling	Predictive autoscaling	Hybrid autoscaling	Trade-off/notes
Type of workload	Stable or moderately variable	Highly variable, seasonal spikes	Any, especially medium dynamics	Reactive quickly reacts to local changes, predictive anticipates peaks
Latency tolerance	Medium/high latency tolerance	Low tolerance of latency	Configurable dynamically	Reactive may miss peaks, predictive acts in advance
Consistency requirements	Short-term inconsistencies are acceptable	High criticality of consistency	Balance between speed and consistency	Hybrid makes it possible to stabilise state and react to local fluctuations
Resource efficiency	Often under- or over-provisioning	Optimisation given accurate forecasting	Medium – depends on configuration	Reactive is less economical, predictive precise, hybrid combines both

Continued Table 4.

Criterion	Reactive autoscaling	Predictive autoscaling	Hybrid autoscaling	Trade-off/notes
Implementation complexity	Low	High (requires ML/ analytics)	Very high	Hybrid requires dual-loop architecture and monitoring
Scaling cost	High due to delayed or excessive allocation	Depends on forecast accuracy	Optimised through a combined approach	Trade-off between finance and productivity

Source: created by the author on the basis of analysis of data from D. Zou *et al.* (2024)

Table 4 summarises the key characteristics of the three scaling approaches – reactive, predictive and hybrid – in stateful microservices. Overall, reactive autoscaling provides rapid reaction to current workload, but is less effective with highly variable peaks and may allow short-term state inconsistencies. The predictive approach anticipates workload in advance, improving stability and consistency, but requires accurate forecasting and additional resources. Hybrid autoscaling combines the advantages of both approaches, allowing simultaneous reaction to local changes and forecasting of peaks, although its implementation is more complex and requires greater analytical support. Thus, the choice of strategy depends on workload dynamics, latency and consistency requirements, as well as on the system's resource and financial constraints. Different parameters, such as workload variability, latency criticality, data-consistency requirements and resource availability, determine which approach – reactive, predictive or hybrid – will be more appropriate. Table 5 summarises these dependencies, demonstrating the trade-off between performance, stability and resource efficiency.

Table 5. Influence of stateful-system characteristics on the choice of autoscaling strategy

System characteristic	Reactive	Predictive	Hybrid	Comment
Workload	Stable	Dynamic, spikes	Medium/mixed	Choice depends on load variability
Latency criticality	Low	High	Medium/depends on scenario	If latency is critical, predictive has the advantage
Consistency requirements	Moderate	High	High + reactive smoothing	For stateful services, it is important to plan replication
Resource/budget availability	Limited	Moderate	Sufficient for dual-loop	Hybrid requires more resources for analytics and synchronisation
Trend duration	Short	Long or seasonal	Any	Hybrid allows a combination of short- and long-term strategies

Source: created by the author on the basis of analysis of data from T. Betti Pillippuge *et al.* (2025)

Table 5 shows that reactive autoscaling is effective for stable systems with low latency criticality and moderate consistency requirements, but does not provide high stability under sharp workload spikes. The predictive approach demonstrates advantages in systems with dynamic or seasonal workload and high requirements for latency and state consistency, but demands significant resources for accurate forecasting. Hybrid autoscaling makes it possible to combine reactive speed and anticipatory planning, which makes it universal for systems with medium or mixed workload dynamics and long-term trends, although its implementation requires dual-loop architecture and additional monitoring. Overall, the choice of strategy should be based on a balance between performance, consistency and resource-use efficiency.

Choosing an optimal scaling strategy for stateful microservice environments requires a systematic analysis of a set of interrelated factors that affect performance, state consistency and resource-use efficiency. A hierarchical model of factors makes it possible to structure the key decision-making criteria, from general requirements for throughput and latency to specific requirements for consistency and resource availability. Such a conceptual approach provides a scientifically grounded basis for assessing trade-offs between response speed, system stability and economic efficiency, which is critical when choosing between reactive, predictive and hybrid autoscaling models (Rao, 2021). Schematically, the conceptual model can be presented as a hierarchy of factors that determine the optimal scaling approach (Fig. 3).

On the basis of this hierarchy, the decision-making process can be formalised as follows: if most characteristics point to stable workload and acceptable latency, reactive is chosen; if consistency and latency are critical, predictive is chosen; and for mixed scenarios with sufficient resources, hybrid is chosen. The conceptual model of choosing a scaling strategy shows that there is no universal approach for stateful microservices. The choice between reactive, predictive and hybrid approaches is a compromise and depends on workload characteristics, latency and consistency requirements, resource availability and trend duration. Integration of a dual-loop architecture in hybrid autoscaling makes it possible to maintain stability, optimal resource use and state consistency simultaneously, which makes it the most flexible and promising option for complex stateful environments.

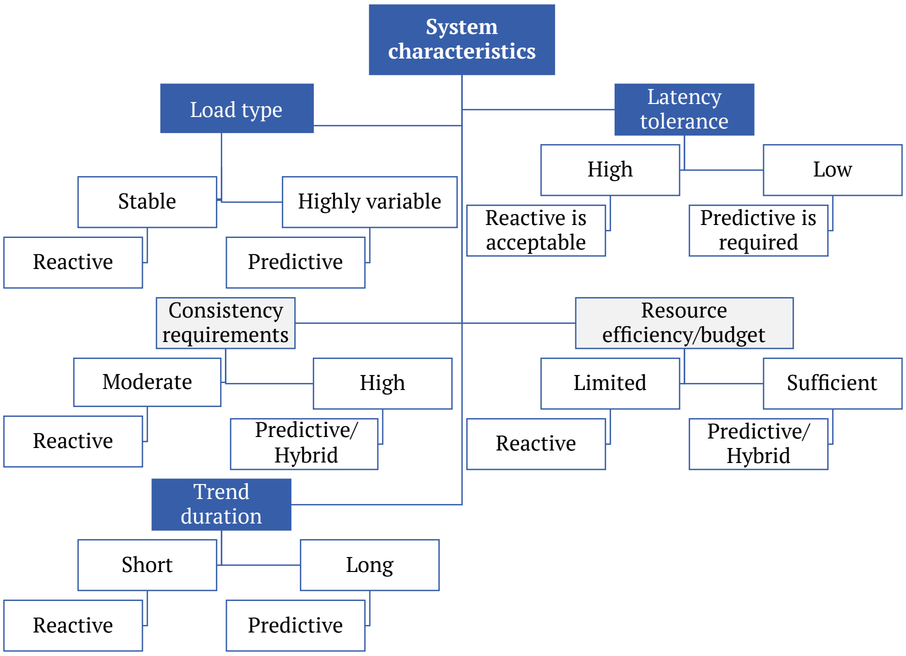


Figure 3. Hierarchy of factors for strategy selection

Source: created by the author

DISCUSSION

The review of scientific works shows a gradual shift of focus from localised technical solutions (simple horizontal/vertical scaling) to complex, context-oriented approaches: self-adaptive architectures, vertically oriented resizing algorithms, solutions for serverless/ETL processes and assessments of the impact of frameworks on scalability. For example, S. Boyapati & C. Szabo (2022) present a detailed case study of self-tuning microservice architectures, in which the authors focus on mechanisms of dynamic adaptation to workload changes in real time. The results of the study confirmed the effectiveness of reactive approaches for environments with highly dynamic requests but revealed a significant dependence on the accuracy of metric monitoring. In the present study, similar conclusions were obtained regarding the high sensitivity of the system to threshold values in the reactive model. Methods of optimising this sensitivity through combination with elements of predictive scaling were also proposed, providing greater stability during peak loads. In addition, the findings of Yu. Palamarchuk (2022) emphasise that the effectiveness of microservice-based systems is strongly dependent on the architectural decomposition of services and the precision of interaction patterns between components. It is highlighted that well-structured microservice architectures enable more predictable scaling behaviour and support improved resource utilisation, particularly in environments where dynamic load variations are present. These observations align with the patterns identified in the present study, where architectural clarity and service modularity are shown to influence the performance and stability of both reactive and predictive autoscaling approaches in stateful systems. The work further reinforces the notion that autoscaling mechanisms cannot be fully effective without a coherent underlying architectural framework, which remains consistent with the broader conclusions of the current research regarding the context-dependent nature of scaling strategies. The publication by A. Pavlenko *et al.* (2024) presents the CaaSPER algorithm, aimed at vertical autoscaling of monolithic applications in a containerised environment. The authors showed that even for non-microservice architectures, the vertical approach can significantly reduce downtime and optimise resource use. These results partially align with the current observations: in stateful environments vertical scaling is appropriate only if mechanisms of state coordination are in place, whereas in the work of A. Pavlenko *et al*. the main focus was on isolated monolithic applications. This difference explains the divergence in conclusions regarding the effectiveness of the Vertical Pod Autoscaler – in the present study, its application is limited by data-consistency issues.

In the study by D. Rodrigues *et al*. (2025), the trade-off between execution latency and computation cost in serverless environments is examined, where scaling is performed on the basis of optimisation rules for ETL processes. The authors formalised a trade-off model for decision-making, which shows how scaling policies affect performance and costs. This approach is close to the theoretical analysis carried out here of trade-offs between latency, stability, and costs. At the same time, the concept of adaptive selection of a scaling policy depending on the dominant metric confirms the validity of the proposed model of hybrid load control. M. Islam (2025) reveals the impact of data-driven web frameworks on the scalability and performance of modern enterprise systems. The author found that framework features – in particular the ORM layer, asynchronous request handling and caching mechanisms – directly determine the behaviour of autoscaler algorithms. This conclusion is consistent with the results obtained here regarding the role of RAM and CPU as the main triggers for scaling and the importance of taking into account characteristics of the software stack when choosing the Horizontal Pod Autoscaler or Vertical Pod Autoscaler.

B. Tutuncuoglu (2025) proposed the "Phoenix Stack" architecture for self-healing microservice web applications in real time, where the key elements are automatic failure detection and localised state recovery. The author showed that integrating self-healing mechanisms significantly increases availability and reduces recovery time in the event of component failures, which directly correlates with the current conclusion about the need to consider fault tolerance when designing scaling strategies for stateful services. At the same time, the "Phoenix Stack" strongly emphasises local recovery and restart patterns. In turn, J. Prodanov *et al.* (2025) investigated an in-place scaling mechanism for edge-cloud environments, implemented through multi-agent reinforcement learning (MARL), where agents coordinate decisions on local increase/decrease of replicas, taking into account network latency and energy constraints. The results showed that the MARL approach significantly improved latency and energy-consumption balance compared with static policies; this correlates with the present conclusions about the promise of adaptive, intelligent controllers for hybrid autoscaling. However, the application of MARL by the authors is oriented towards edge-class scenarios with strict resource constraints and a different network topology, so direct transfer of these results to traditional stateful Kubernetes clusters requires adjustments. I. Vasireddy *et al*. (2025) presented a non-linear regression predictive model of replicas combined with the ORLE algorithm to improve the scalability, energy efficiency and economy of Kubernetes clusters. In the experiments, the authors demonstrated a reduction in redundant replicas and improved CPU/RAM use compared with the standard Horizontal Pod Autoscaler, which directly correlates with the current conclusion that more accurate predictive models increase the efficiency of resource provisioning. The difference is that I. Vasireddy *et al*. focused on specific regression models and an energy metric, whereas the present study considered a broader set of trade-offs (latency, consistency, costs). Thus, these results complement but do not fully replace the recommendations proposed here regarding strategy selection.

M. Yang *et al.* (2025) conducted a systematic review on the optimisation of full-stack microservice architecture and identified research gaps in the integration of network, container, and application levels for global performance optimisation. The review confirmed that comprehensive, cross-stack approaches (combining optimisation at the code, container, and orchestration levels) are the most promising for achieving both cost-efficiency and low latency, which corresponds to the current recommendation in favour of hybrid and multi-layer strategies. At the same time, M. Yang *et al.* emphasised that there is still a lack of large-scale empirical research under production conditions, which reinforces the present limitation due to the absence of full experimental validation and indicates directions for further research. In the study by A. Chaudhari & P. Charate (2025), an approach to dynamic orchestration of microservices in multicloud environments is proposed, aimed at minimising latency in applications with high performance requirements. The authors demonstrated that combining container orchestration with adaptive routing algorithms can significantly reduce latency between cloud nodes and increase resource-use efficiency. This result partially corresponds to the current conclusions on the importance of adaptive scaling strategies for stateful systems, but the present study focuses on intra-cluster optimisation mechanisms, which explains the difference in focus and practical recommendations.

In turn, the dissertation research by A. Baarzi (2021) analyses the effectiveness of service deployment in public clouds from the perspective of the cost-performance-security ratio. The author proved that optimising service deployments by combining automated scaling and security policies can reduce operational costs without significant loss of performance. Compared with the present study, the work of A. Baarzi is more oriented towards the macro-level of economic parameters and overall architectural trade-offs, while the current analysis focuses on the technical detail of reactive and predictive scaling models. At the same time, C. Oleti (2023) considered the construction of secure scalable microservice architectures based on Spring Boot and Amazon Web Services, paying particular attention to the integration of artificial-intelligence modules in corporate systems. The author showed that the use of AI components in orchestration and load-distribution processes significantly improves request-processing efficiency and allows peak loads to be predicted with high accuracy. This directly correlates with the present conclusions regarding the feasibility of the predictive approach and hybrid strategies that combine reactive mechanisms with intelligent forecasting. However, the work of C. Oleti focuses mainly on the corporate context (enterprise-scale AI), whereas the present study covers more universal scenarios of containerised stateful services, which explains the difference in the scope of application and the target efficiency metrics.

Thus, the synthesis of the analysed sources has confirmed the current conclusions regarding the context-dependence of scaling-strategy choice and the feasibility of using hybrid models that combine reactive and predictive principles. Differences in conclusions between the works are primarily due to differences in research subject (monoliths, microservices, serverless), the level of empirical verification and the specificity of the metrics used (latency, cost, uptime). The present study extends the existing theoretical basis by integrating these approaches into a generalised decision-making model for stateful environments, which forms a foundation for further experiments in the field of adaptive scaling.

CONCLUSIONS

Autoscaling in microservice systems provides adaptive performance through a feedback loop, while the choice of strategy depends on the nature of the workload, latency sensitivity and consistency requirements. Horizontal scaling provides flexibility for stateless services, vertical scaling optimises pod resources, and hybrid scaling combines both approaches for complex scenarios. Reactive autoscaling reacts quickly to changes but may lead to state inconsistencies and inefficient resource use, whereas predictive autoscaling forecasts needs and increases stability, requiring additional computing resources. The choice of an optimal strategy is always a compromise and context-dependent.

The study has shown that autoscaling models in cloud microservice architectures have evolved from simple reactive mechanisms to complex predictive systems capable of taking into account temporal workload patterns and user behaviour. Reactive models implemented via the Horizontal Pod Autoscaler provide simplicity and predictability, but are limited in reaction speed and resource-use efficiency. Predictive models based on time-series methods (ARIMA, Exponential Smoothing, Kalman Filter) or deep learning (LSTM, TCN) make it possible to reduce reaction lag, predict workload peaks and increase system stability, although such models require significant computing resources and more complex implementation. Hybrid approaches combine the advantages of both strategies, providing short-term stabilisation and long-term forecasting through a dual-loop adaptive-control architecture. Overall, the choice of an optimal autoscaling strategy is context-dependent and involves balancing between reaction speed, stability, state consistency, resource efficiency and implementation complexity.

The theoretical analysis of performance trade-offs has shown that the effectiveness of autoscaling in stateful microservice architectures is determined by the balance between reaction latency, system stability, state consistency, resource-use efficiency and financial costs. Reactive autoscaling provides rapid adaptation to unpredictable workload changes, but is characterised by high latency, performance fluctuations and a risk of replica-state inconsistencies. Predictive autoscaling reduces scaling latency and increases stability and resource provisioning efficiency through anticipatory forecasting, but requires accurate prediction models and additional computing resources. Hybrid strategies combine the advantages of both approaches, making it possible to ensure short-term flexibility and long-term stability simultaneously, which makes the choice of optimal model context-dependent and determined by workload dynamics, Service Level Agreement and financial constraints.

The analysis of autoscaling models in stateful microservice architectures has shown that system effectiveness is defined by the trade-off between reaction latency, stability, state consistency, resource efficiency and costs. Reactive autoscaling provides rapid adaptation to unpredictable workloads but is accompanied by high latency and a risk of replica-state inconsistencies. Predictive autoscaling increases stability and optimises resource use owing to anticipatory forecasting, but requires accurate models and additional computing resources. Hybrid strategies integrate the advantages of both approaches, providing flexibility and stability at the same time, and the choice of optimal model depends on workload dynamics, Service Level Agreement requirements and financial constraints.

The limitations of the study lie in the theoretical nature of the analysis, which was based on conceptual models of reactive, predictive and hybrid scaling without experimental validation in real stateful microservice environments, as well as on a limited set of forecasting algorithms and the absence of a detailed examination of economic and cloud-provider aspects of implementation. Prospects for further research include empirical testing of strategies in different cloud infrastructures, integration of modern machine-learning methods to increase forecasting accuracy and the development of dynamic combined scaling models that take into account Service Level Agreement, latency, consistency and financial efficiency.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Almeida, F., & Bálint, B. (2024). Approaches for hybrid scaling of agile in the IT industry: A systematic literature review and research agenda. *Information*, 15(10), article number 592. doi: 10.3390/info15100592.

[2] Baarzi, A.F. (2021). *Efficient service deployment on public cloud: A cost, performance, and security perspective*. University Park: Pennsylvania State University.

[3] Babenko, T., Kolesnikova, K., Abramkina, O., & Vitulyova, Y. (2025). Automated OSINT techniques for digital asset discovery and cyber risk assessment. *Computers,* 14(10), article number 430. doi: 10.3390/computers14100430.

[4] Betti Pillippuge, T.L., Khan, Z., & Munir, K. (2025). Horizontal autoscaling of virtual machines in hybrid cloud infrastructures: Current status, challenges, and opportunities. *Encyclopedia*, 5(1), article number 37. doi: 10.3390/encyclopedia5010037.

[5] Boyapati, S., & Szabo, C. (2022). Self-adaptation in microservice architectures: A case study. In *Proceedings of the 26th international conference on engineering of complex computer systems* (pp. 42-51). Hiroshima: IEEE. doi: 10.1109/ICECCS54210.2022.00014.

[6] Chaudhari, A.V., & Charate, P.A. (2025). [Dynamic orchestration of microservices across multi-cloud environments for low-latency applications](). *International Research Journal of Advanced Engineering and Science*, 10(2), 165-172.

[7] Chilakala, H.S., Preeti, N., Marri, S.P., & Murali, K. (2024). Kalman integration with ARIMA & LSTM. In A. Kumar, G. Sharma, A. Sharma, P. Chopra & P. Rattan (Eds.), *Advances in networks, intelligence and computing* (pp. 562-569). London: CRC Press. doi: 10.1201/9781003430421.

[8] Guntupalli, B., & Vamshi, S. (2023). Designing microservices that handle high-volume data loads. *International Journal of AI, BigData, Computational and Management Studies*, 4(4), 76-87. doi: 10.63282/3050-9416.IJAIBDCMS-V4I4P109.

[9] Islam, M. (2025). The impact of data-driven web frameworks on performance and scalability of US enterprise applications. *International Journal of Business and Economics Insights*, 5(3), 523-558. doi: 10.63125/f07n4p12.

[10] Kanwal, N., Khoraminia, F., Kiraz, U., Mosquera-Zamudio, A., Monteagudo, C., Janssen, E., Zuiverloon, T., Rong, C., & Engan, K. (2024). Equipping computational pathology systems with artifact processing pipelines: A showcase for computation and performance trade-offs. *BMC Medical Informatics and Decision Making*, 24(1), article number 288. doi: 10.1186/s12911-024-02676-z.

[11] Karol Santos Nunes, J.P., Nejati, S., Sabetzadeh, M., & Nakagawa, E.Y. (2024). Self-adaptive, requirements-driven autoscaling of microservices. In L. Baresi (Ed.), *Proceedings of the 19th international symposium on software engineering for adaptive and self-managing systems* (pp. 168-174). New York: Association for Computing Machinery. doi: 10.1145/3643915.3644094.

[12] Khaleq, A., & Ra, I. (2021). Intelligent autoscaling of microservices in the cloud for real-time applications. *IEEE Access*, 9, 35464-35476. doi: 10.1109/ACCESS.2021.3061890.

[13] Kumari, A., Anand, A.K., & Sahoo, B. (2023). From stateless to stateful: a comparative analysis of stateful serverless computing frameworks. In R. Tiwari, M. Saraswat & M. Pavone (Eds.), *International conference on computational intelligence* (pp. 223-237). Singapore: Springer. doi: 10.1007/978-981-97-3526-6_19.

[14] Maliakel, P.J., Ilager, S., & Brandic, I. (2025). Investigating energy efficiency and performance trade-offs in LLM inference across tasks and DVFS settings. *arXiv*. doi: 10.48550/arXiv.2501.08219.

[15] Mekki, M., Ksentini, A., & Meliani, A.E. (2025). Resiliency focused proactive lifecycle management for stateful microservices in multi-cluster containerized environments. *Computer Communications*, 236, article number 108111. doi: 10.1016/j.comcom.2025.108111.

[16] Meng, C., Song, S., Tong, H., Pan, M., & Yu, Y. (2023). Deepscaler: Holistic autoscaling for microservices based on spatiotemporal GNN with adaptive graph learning. In *Proceedings of the 38th IEEE/ACM international conference on automated software engineering* (pp. 53-65). Luxemburg: IEEE. doi: 10.1109/ASE56229.2023.00038.

[17] Oleti, C.S. (2023). Enterprise AI at scale: Architecting secure microservices with spring boot and AWS. *International Journal of Research in Computer Applications and Information Technology*, 6(1), 133-154. doi: 10.34218/IJRCAIT_06_01_011.

[18] Palamarchuk, Yu. (2022). Methods of building microservice architecture of e-learning systems. *Information Technologies and Computer Engineering*, 19(2), 43-54. doi: 10.31649/1999-9941-2022-53-1-43-54

[19] Pavlenko, A., Cahoon, J., Zhu, Y., Kroth, B., Nelson, M., Carter, A., Liao, D., Wright, T., Camacho-Rodriguez, J., & Saur, K. (2024). Vertically autoscaling monolithic applications with CaaSPER: Scalable container-a s-a-service performance enhanced resizing algorithm for the cloud. In P. Barcelo & N. Sanchez-Pi (Eds.), *Companion of the 2024 international conference on management of data* (pp. 241-254). New York: Association for Computing Machinery. doi: 10.1145/3626246.3653378.

[20] Prasad, T.A. (2023). [AI-driven predictive scaling for performance optimization in cloud-native architectures](). *Journal of Electrical Systems*, 19(4), 607-617.

[21] Prodanov, J., Bertalanič, B., Fortuna, C., Chou, S.K., Jurič, M.B., Sanchez-Iborra, R., & Hribar, J. (2025). Multi-agent reinforcement learning-based in-place scaling engine for edge-cloud systems. In *Proceedings of the 18th international conference on cloud computing* (pp. 32-42). Helsinki: IEEE. doi: 10.1109/CLOUD67622.2025.00014.

[22] Quattrocchi, G., Incerto, E., Pinciroli, R., Trubiani, C., & Baresi, L. (2024). Autoscaling solutions for cloud applications under dynamic workloads. *IEEE Transactions on Services Computing*, 17(3), 804-820. doi: 10.1109/TSC.2024.3354062.

[23] Rampérez, V., Soriano, J., Lizcano, D., & Lara, J.A. (2021). FLAS: A combination of proactive and reactive auto-scaling architecture for distributed services. *Future Generation Computer Systems*, 118, 56-72. doi: 10.1016/j.future.2020.12.025.

[24] Rao, A. (2021). [Architectural trade-offs between stateless and stateful microservices in large-scale cloud systems](). *International Journal of Innovation Studies*, 5(1), 135-141.

[25] Razavi, K., Salmani, M., Mühlhäuser, M., Koldehofe, B., & Wang, L. (2024). A tale of two scales: Reconciling horizontal and vertical scaling for inference serving systems. *arXiv*. doi: 10.48550/arXiv.2407.14843.

[26] Rodrigues, S., Rosas, M., Delbianco, N., Cláudia, C., Peterson, B., & Grácio, M.C. (2025). *Latency vs. cost trade-offs in serverless ETL: A decision-theoretic framework for architecture design*. Retrieved from https://www.researchgate.net/publication/393003364_Latency_vs_Cost_Trade-offs_in_Serverless_ETL_A_Decision-Theoretic_Framework_for_Architecture_Design.

[27] Romanov, O., Mankivskyi, V., Romanov, A., Nesterenko, M., & Pidpaly, O. (2024). Performance assessing of dynamic scaling of containerized applications when using kubernetes cluster autoscaler and auto scaling group together. In M. Ilchenko, L. Uryvsky & L. Globa (Eds.), *Advanced smart information and communication technology and systems* (pp. 111-136). Cham: Springer. doi: 10.1007/978-3-031-94799-5_7.
[28] Tabilo Alvarez, J., & Ramírez-Correa, P. (2023). A brief review of systems, cybernetics, and complexity. *Complexity*, 2023(1), article number 8205320. doi: 10.1155/2023/8205320.
[29] Thompson, M.A., Carter, J.L., Nguyen, D.R., Bennett, S.K., Miller, R.J., & James, A. (2023). *Intelligent workload balancing across multi-cloud providers*. Retrieved from https://tinyurl.com/4x7sebkh.
[30] Tutuncuoglu, B.T. (2025). *Phoenix stack: A self-healing microservice architecture for real-time web applications*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5343101.
[31] Varzhanskiy, I., Melnychenko, A., Naiev, S., & Kapustynskyi, O. (2025). Foresight and active forecast potential for improving defense capability. *European Journal of Interdisciplinary Studies*, 17(1). doi: 10.24818/ejis.2025.01.
[32] Vasireddy, I., Wankar, R., & Chillarige, R. (2025). Improving scalability, energy efficiency, and cost-effectiveness in Kubernetes clusters using a nonlinear regression-based predictive replica model and ORLE algorithm. *Egyptian Informatics Journal*, 31, article number 100732. doi: 10.1016/j.eij.2025.100732.
[33] Yang, M., Wang, X., Cai, B., Wang, Y., & Guo, Y. (2025). Full stack optimization of microservice architecture: Systematic review and research opportunity. *Cluster Computing*, 28, article number 1005. doi: 10.1007/s10586-025-05690-6.
[34] Zou, D., Lu, W., Zhu, Z., Lu, X., Zhou, J., Wang, X., Liu, K., Sun, R., & Wang, H. (2024). OptScaler: A collaborative framework for robust autoscaling in the cloud. *Proceedings of the VLDB Endowment*, 17(12), 4090-4103. doi: 10.14778/3685800.3685829.

Христина Терлецька
Бакалавр
Національний університет «Львівська політехніка»
79000, вул. Степана Бандери, 12, м. Львів, Україна
https://orcid.org/0009-0002-7302-2625

Компроміси продуктивності між прогнозним та реактивним автомасштабуванням у станозалежних мікросервісах

Анотація. Актуальність дослідження компромісів між реактивним та прогнозним автоматичним масштабуванням станозалежних мікросервісів зумовлена необхідністю підвищити стабільність, продуктивність ресурсів і надійність хмарних систем у динамічних умовах. Метою дослідження було оцінити ефективність реактивних, прогнозних та гібридних стратегій автоматичного масштабування у станозалежних мікросервісних архітектурах. Методологічна основа дослідження спиралася на порівняльний, контентний, системний і структурно-функціональний аналізи, а також на моделювання, що дозволило комплексно дослідити продуктивність і стабільність різних підходів до автоматичного масштабування у станозалежних мікросервісах. Теоретичний аналіз показав, що автоматичне масштабування в мікросервісних системах забезпечує адаптивну продуктивність, а вибір стратегії залежить від характеру навантаження, чутливості до затримок та вимог до консистентності. Огляд моделей продемонстрував, що реактивне автоматичне масштабування швидко реагує на зміни, але супроводжується високою латентністю, коливанням продуктивності та ризиком розривів стану реплік, тоді як прогнозне автоматичне масштабування знижує затримку масштабування та підвищує стабільність і ефективність ресурсів, проте потребує точних моделей і додаткових обчислювальних ресурсів. Аналіз результатів виявив, що гібридні стратегії інтегрують переваги обох підходів, забезпечуючи одночасно гнучкість і стабільність, а ефективність систем визначається компромісом між затримкою, стабільністю, узгодженістю стану, ефективністю ресурсів і фінансовими витратами. Результати дослідження можуть бути корисними розробникам і архітекторам хмарних мікросервісних систем, ІТ-інженерам і DevOps-командам для оптимізації стратегій масштабування, підвищення стабільності сервісів і ефективного використання ресурсів у різних хмарних середовищах

Ключові слова: Kubernetes; оптимізація ресурсів; продуктивність хмар; динамічний розподіл ресурсів; стабільність системи; метрики спостереження

Machinery & Energetics
Vol. 16, No. 4. 2025

Journal homepage:
https://technicalscience.com.ua/en

UDC 621.9.02
DOI: 10.31548/machinery/4.2025.34

Pavlo Zabrodskyi
PhD in Technical Sciences, Associate Professor
Polissia National University
10008, 7 Staryi Blvd., Zhytomyr, Ukraine
https://orcid.org/0000-0002-3904-564X

Savelii Kukharets
Doctor of Technical Sciences, Professor
Vytautas Magnus University Agriculture Academy
LT-53362, 11 Studentu Str., Kaunas, Lithuania
https://orcid.org/0000-0002-5129-8746

Bohdan Sheludchenko
PhD in Technical Sciences, Professor
Polissia National University
10008, 7 Staryi Blvd., Zhytomyr, Ukraine
https://orcid.org/0000-0002-8137-0905

Oleksandr Medvedskyi*
PhD in Technical Sciences, Associate Professor
Polissia National University
10008, 7 Staryi Blvd., Zhytomyr, Ukraine
https://orcid.org/0000-0001-7458-5337

Investigation of the stress state of a carbide plate of a cutting tool

Abstract. Cutting tools with carbide inserts operate under harsh conditions that can cause cracking and failure, as their strength directly depends on the stress-strain state. The study was conducted to determine of the influence of the angle of installation in the holder of brazed carbide plates of cutting tools on their stress state was investigated. Experimental studies using the method of photoelasticity are characterised by high sensitivity, non-destructive effects, clarity, and speed of results. The study was carried out on the polarisation-projection unit PPU-7 using physical models of carbide plates made of optically sensitive material ED-6, which were attached to the holder with different installation angles. The resulting isochrome and isokline photographs made it possible to assess the actual visual picture of the existing stresses. During the experiment, the cutting tool models were loaded using a specially designed stand that simulated the cutting force. The principal stresses were calculated by the method of difference of tangential stresses for the selected insert installation angles. It was proved that the angle of attachment of the carbide insert to the holder undoubtedly affects the distribution of stresses in it during metalworking. A quantitative relationship has been established between the angle of attachment to the holder and the magnitude of internal stresses acting in the carbide plate. The highest stresses σ_{equ} occured in the cross-section of the cutting blade of a carbide insert regardless of the angle of attachment to the holder. The lowest level of acting stresses along the measurement points in three zones along the width of the carbide insert when attached to the holder at an angle of -30° was found. With connection angles of 15°, 0°, and -15°, critical stresses occur in all investigated areas across the width of the plate along the measurement points

Keywords: isocline; isochromes; anisotropy; photoelasticity; installation angle

Article's History: Received: 27.06.2025; Revised: 30.09.2025; Accepted: 27.11.2025.

Suggested Citation:
Zabrodskyi, P., Kukharets, S., Sheludchenko, B., & Medvedskyi, O. (2025). Investigation of the stress state of a carbide plate of a cutting tool. *Machinery & Energetics,* 16(4),34-42. doi: 10.31548/machinery/4.2025.34.

*Corresponding author

INTRODUCTION

The reliability and durability of turning tools play a crucial role in ensuring the stability of machining processes and the quality of manufactured components. During cutting, the tool is subjected to elevated temperatures, substantial mechanical loads and thermal fluctuations, all of which contribute to intensive wear or even brittle fracture of carbide inserts. The geometric parameters of the tool, particularly the installation angle of the carbide insert, significantly influence its ability to withstand these operational stresses. Consequently, investigating the stress-strain state of the insert and the effect of its orientation within the toolholder is essential for improving tool reliability and extending its service life.

As the authors M. Sobron *et al.* (2020) pointed out, before starting the machining process, it is necessary to know how long the cutting tool can operate before it breaks down or suffers critical wear. A. Adamu *et al.* (2021) provided a comprehensive review of approaches to optimising cutting parameters with the aim of reducing tool wear in turning operations. The authors highlighted that tool wear was governed by a complex interaction of cutting speed, feed rate, depth of cut and the material properties of both the tool and the workpiece. Their analysis demonstrated that the selection of appropriate machining parameters was essential for achieving stable cutting conditions, improving surface quality and extending tool life. That is, cutting tools, including turning cutters, must have high strength, heat resistance, wear resistance, and hardness. If wear occurs in the main area of chip impact, i.e., on the front surface, crater wear occurs, and if the main wear occurs on the back surface of the cutter, edge wear occurs (Roszkowski *et al.*, 2020). Thus, the main reason why turning inserts fail and need to be replaced is wear. The situation is different with cutting tools.

Unlike through-hole cutters, cut-off cutters operate under difficult conditions of non-orthogonal cutting, as the cutter is subjected to force and temperature loads on three sides. In such conditions, the lubricating coolant does not work effectively enough, which complicates the efficiency of the chip removal process. P. Zabrodskyi *et al.* (2021) noted that the main cause of failure of cutters is not wear, as in the case of passing cutters, but breakage of carbide inserts. At the same time, as the authors Y. Liu *et al.* (2025) pointed out, compared to wear, brittle fracture occurs more suddenly and less predictably. Experimental results by A. Rizzo *et al.* (2020) showed that breakage occurs due to the formation of surface cracks with their subsequent expansion. Cracks can appear as a result of untimely sharpening of the tool, incorrectly selected cutting modes, and for cutters with brazed inserts during the manufacturing process when the inserts are soldered.

Scientists have conducted extensive research to determine the causes of cracks and breakage in carbide plates. Thus, F. Wu *et al.* (2021) and J. Östby *et al.* (2022), the method of analytical research is used to study the proposed theory of sliding lamellae formation in carbide inserts at the WC/WC grain boundary based on shear stresses. Finite element modeling of the cutting process by C. Xiao *et al.* (2024) provided information on the deformations along each point of the cutting line, the stresses occurring at each node/point, and the temperature distribution in the tool and workpiece during the machining process. The study was conducted using Deform 2D computer-aided design software. The authors V. Gutakovskis *et al.* (2022) used the Third Wave Advant Edge software to model turning and study stresses using the finite element method. W. Cai *et al.* (2025) used the Johnson-Cook constitutive model to determine the shear stress on the primary shear plane. It takes into account the strain, strain rate, and temperature effect to describe the stress. This model is widely used in cutting modelling and simulation studies.

Among the experimental methods used to study the stress-strain state, it is necessary to note, for example, the holographic method. M. Kumar *et al.* (2025) used a holographic interferometry unit to determine the deformations of turning cutters, which made it possible to combine the results of experimental and numerical studies. In the work by S. Lubis *et al.* (2020), a non-destructive X-ray diffraction method was used to measure surface stress. The authors B. Bergmann *et al.* (2025) used a method for calculating internal stresses based on measured external stresses. In this case, the internal stresses of worn cutting tools are calculated at different states of wear.

Therefore, the solution of the problem of optimising the geometric parameters of cut-off turning tools with brazed carbide inserts by determining the influence of the angle of installation in the holder on their stress-strain state is of scientific interest. The insufficient study of this issue requires additional research to identify ways to increase the durability and reliability of cutting tools. The aim of this study was to investigate the effect of the installation angle of brazed carbide inserts in the toolholder on their stress-strain state and to optimise tool design for improved durability during metalworking.

MATERIALS AND METHODS

It is difficult to ensure machining accuracy in prefabricated cutters due to the presence of additional gaps at the joints of the carbide inserts, so they were used in single or small batch production. Tools with brazed inserts (Fig. 1) offered significantly better precision than prefabricated cutters and were therefore preferred for use in Computer Numerical Control (CNC) machines.

Figure 1. Cutting-off tool with brazed carbide insert
Source: prepared on the basis of research data by P. Zabrodskyi (2021)

In the study, analytical and experimental methods were used to determine the stress state of carbide inserts in cut-off tools. One of the key methods involves calculating the shear stress in the material. For this, the formula proposed by J. Östby *et al.* (2022) was used to compute the shear stress (τ), which took into account the material's mechanical and thermal properties:

$$\tau = \frac{1}{\sqrt{3}}\left[A + B\left(\frac{\gamma}{\sqrt{3}}\right)^n\right]\left[1 + Cln\left(\frac{\dot{\gamma}}{\dot{\gamma}_0}\right)\right]\left[1 - \left(\frac{T-T_r}{T_m-T_r}\right)^m\right], \quad (1)$$

where A is the yield strength; B is the strain hardening coefficient; C is the sensitive strain rate coefficient; n is the strain hardening degree index; m is the thermal softening index; T_r is the reference temperature (ambient temperature); T_m is the melting point; γ is the strain; $\dot{\gamma}$ is the strain rate; $\dot{\gamma}_0$ is the reference strain rate.

To evaluate the stress-strain state of carbide inserts of cut-off cutters, the polarisation-optical method, or photoelasticity method, was used. In the present work, the polarisation-optical method, or photoelasticity method, was used to assess the stress-strain state of carbide plates of cutting tools. This method was based on the properties of certain transparent materials, which become optically anisotropic when deformed under mechanical stress. They developed optical anisotropy and the associated double birefringence. This was the so-called piezoelectric effect. In this case, the main values of the dielectric constant were linearly related to the main stresses within the elastic range. The main advantage of the polarisation-optical method was its ability to determine residual stresses. For the manufacture of models studied by the polarisation-optical method, materials were used that must have high transparency, stable optical and mechanical characteristics, sufficient strength, and optical and mechanical isotropy.

To study the stress-strain state of a carbide plate at different installation angles, models of cut-off cutters were made with front angles of -30, -15, 0, 15, and 30 degrees. The models of carbide insert of the cutting tools were made of optically sensitive material based on ED-6 epoxy resin, the hardening of which was provided by methyl tetrahydrophthalic anhydride with an optical constant =1.88 MPa. The geometric dimensions of the carbide insert models at different angles of installation in the holder are shown in Figure 2.

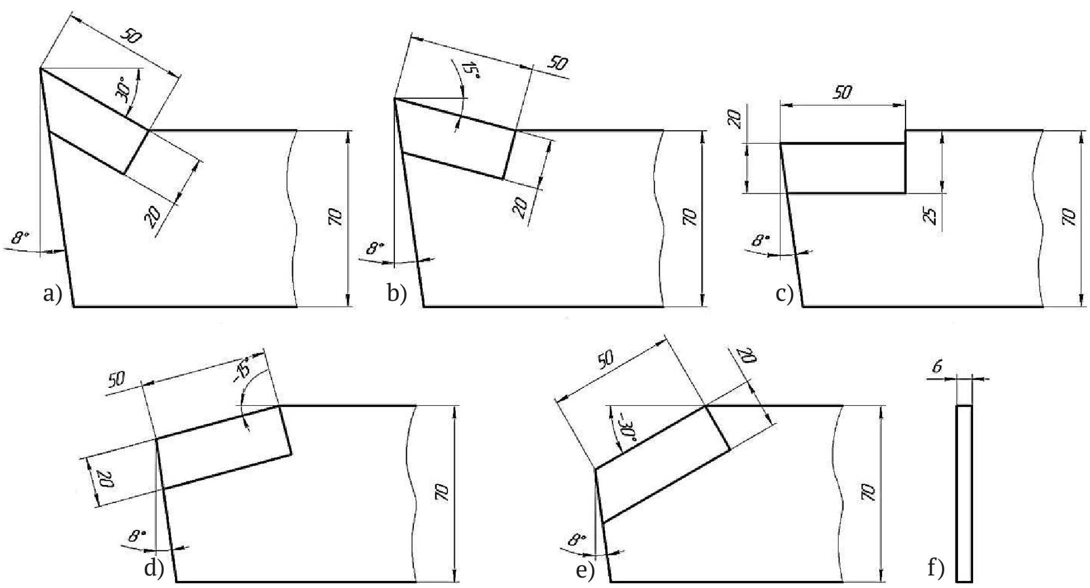


Figure 2. Geometric dimensions of models of cutting tools with the angle of installation of a carbide insert
Note: a) γ = 30°; b) γ = 15°; c) γ = 0°; d) γ = 15°; e) γ = 30°; f) thickness of the carbide insert
Source: compiled by the authors

The geometric scale of modelling for the geometric parameters of the model is selected based on:

$$\alpha = \frac{l_n}{l_m} = const, \quad (2)$$

where l_n, l_m are the lengths of any corresponding segments in reality and the model.

The force scale can be chosen arbitrarily:

$$\beta = \frac{P_n}{P_m} = const. \quad (3)$$

The relationship between the stress similarity scale δ and the force scale β is expressed through the geometric scale α:

$$\delta = \frac{\sigma_n}{\sigma_m} = \frac{\beta}{\alpha^2} = const. \quad (4)$$

The elastic modulus of the model, which has a stress dimension, should be modeled on the same scale as the stress б, according to the general principles of similarity theory:

$$\frac{E_n}{E_m} = \delta = const, \quad (5)$$

where E_n, E_m are the elastic moduli at the corresponding points in nature and the model.

The Poisson's ratios of the real materials μ_n and the model materials μ_m as dimensionless quantities must be equal to each other:

$$\mu_n = \mu_m. \quad (6)$$

When these conditions of similarity are met, it is sufficient that the internal (studied) stresses in the model have a scale of δ, the displacement has a scale of α, and the deformations of the model and the nature coincide. A special stand was used to perform the study, where models of incisors were installed and loaded with a force that simulated the cutting force (Fig. 3). A universal dynamometer was used to measure the applied force.

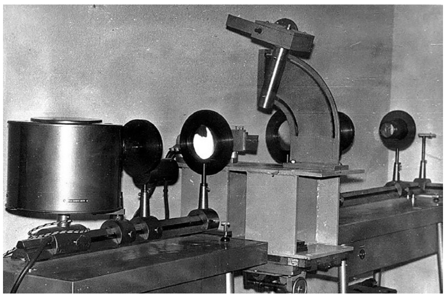


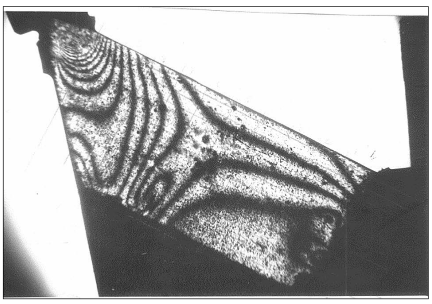
Figure 3. Polarisation and projection unit (PPU-7) with a load stand

Source: compiled by the authors

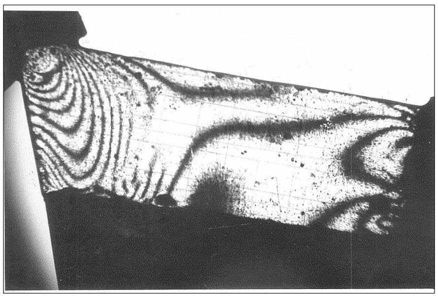
To determine the stresses, it was necessary to obtain pictures of isochromes (stripes of the same colour connecting points with the same difference in principal stresses) and pictures of isoclines (dark-coloured lines connecting points with the same direction of principal stresses).

RESULTS AND DISCUSSION

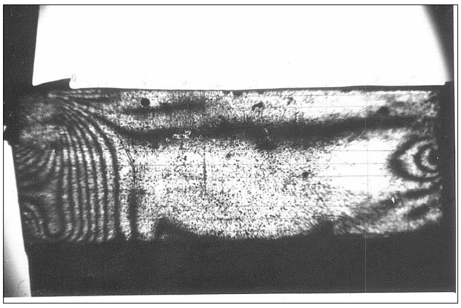
Patterns of isochromes (Fig. 4) and isoclines were obtained at the same load P_z=1,000 N. To evaluate the existing stresses, along arbitrary lines, two parallel auxiliary sections were drawn on opposite sides of these lines at a sufficiently close distance from the main ones. Along these sections, the values of the tangential stresses and were determined, as well as the difference of these stresses between the auxiliary sections.



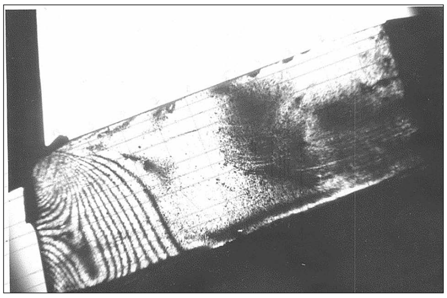
a)



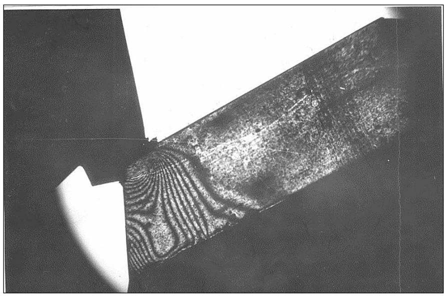
b)



c)



d)



e)

Figure 4. Photocopies of isochromes of samples with different angles of carbide plate fixation

Note: a) γ = 30°; b) γ = 15°; c) γ = 0°; d) γ = -15°; e) γ = -30°

Source: compiled by the authors

The values of the normal stresses σ_x and σ_y were determined using the following formulas:

$$\sigma_{xn} = \sigma_{x0} - \sum_{i=1}^{n} (\Delta\tau_{xy})_i \frac{\Delta x_i}{\Delta y}; \quad (7)$$

$$\sigma_{yk} = \sigma_{y0} - \sum_{r=1}^{k} (\Delta\tau_{xy})_r \frac{\Delta y_r}{\Delta x} + \gamma \sum_{r=1}^{k} \Delta y_r, \quad (8)$$

where n, k are the number of measured points along the x and y axes, respectively; σ_{x0}, σ_{y0} are the known values σ_x i σ_y at the starting point of integration; Δx_i, Δy_r are the intervals along the x and y axes through which measurements were made; Δx, Δy – are the distances between parallel auxiliary sections along the x and y axes, respectively; $\gamma\sum_{r=1}^{k}\Delta y_r$ is the product of the volume weight and the length of the integration segment (taken into account if the model works under its own weight).

The dangerous cross-section of the carbide insert was determined according to the second theory of strength. To do this, first, knowing the values of σ_x, σ_y, and τ_{max}, the values of the principal stresses σ_1 and σ_2 were calculated. Next, the value of the equivalent stress σ_{equ} was calculated for the case of a plane stress state:

$$\sigma_{equ} = \sigma_1 - \mu\sigma_2. \quad (9)$$

In each of the three longitudinal sections of the plate: near the surface, in the middle, and at the base for all angles of the plate installation, the calculation was performed for ten points (Fig. 5).

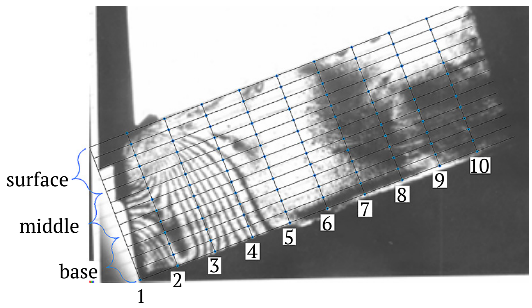


Figure 5. Scheme of the study on the example of a sample with a plate hard-metal insertion angle of γ = -15°

Note: the numbers show the division of the field into zones

Source: compiled by the authors

Based on the values of σ_{equ}, the dependences of the location of dangerous sections in them and the value of σ_{equ} at different angles of the carbide plate installation were obtained. The tensions distribution graphs σ_{equ} (in values are presented strips) are shown in Figures 6-8. The analysis of the results of the surface zone study (Fig. 6) showed that the largest number of stress bands occurred in the plate with the installation angle of $\gamma = 15°$ and $\gamma = 30°$, i.e., the samples shown in Figure 4a and Figure 4b. The plates with attachment angles of $\gamma = -15°$ and $\gamma = -30°$ (Fig. 4d and Fig. 4e) show the same result, i.e., no load tensions strips. For the specimen with an insertion angle of $\gamma = 0°$ (Fig. 4c), the smallest number of tensions strips occurred at the end connection to the holder. An increase in the load at the end connection indicates dangerous sections and areas of possible insert cracking, which will lead to further loss of the integrity of the cutting tool. The study of the middle zone (Fig. 7) along the length of the hard metal insert showed the largest number of stress bands for the sample with an insertion angle of $\gamma = 0°$ in the front part of the insert and for the sample with an insertion angle of $\gamma = 15°$ at the point of attachment of the end part to the holder. Almost the same number of stress bands along the length of the middle zone was observed for the specimen with an insertion angle of $\gamma = 30°$, which may indicate a possible reason for the low durability of this type of cutting tool. Along the base of the carbide insert (Fig. 8), the highest tensions occur in the specimens with an installation angle of $\gamma = 30°$ and $\gamma = 15°$. The lowest tensions are observed in the method of installing carbide inserts with an angle of attachment to the holder of $\gamma = -30°$ and $\gamma = -15°$. Under the condition of direct attachment, i.e., the angle of installation is $\gamma = 0°$, almost the same level of load is observed along the measurement points, which poses a hidden threat of destruction of the soldering point of the insert to the holder.

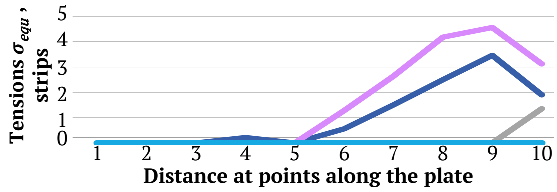


Figure 6. Stress distribution σ_{equ} along the front surface of a hard metal insert at different installation angles

Note: ■ – γ = 30°; ■ – γ = 15°; ■ – γ = 0°; ■ – γ = -15°; ■ – γ = -30°

Source: compiled by the authors

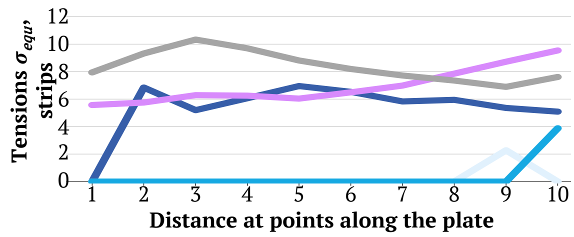


Figure 7. Stress distribution pattern σ_{equ} along the middle section of a hard metal insert at different installation angles

Note: ■ – γ = 30°; ■ – γ = 15°; ■ – γ = 0°; ■ – γ = -15°; ■ – γ = -30°

Source: compiled by the authors

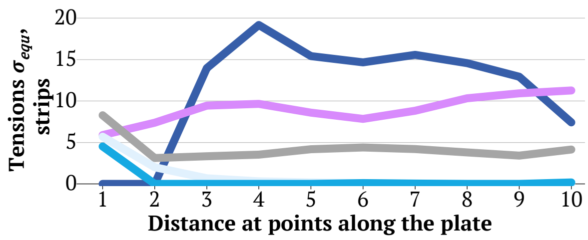


Figure 8. Stress distribution σ_{equ} along the base of the hard metal insert at different installation angles

Note: ■ – γ = 30°; ■ – γ = 15°; ■ – γ = 0°; ■ – γ = -15°; ■ – γ = -30°

Source: compiled by the authors

A significant amount of research into factors affecting the durability of cutting tools has focused on through-type turning tools. This is understandable, since this type of tool performs the main work in turning. The insufficiently studied factors affecting the durability of cut-off turning tools prompted the search for a scientific solution to the problem. According to M. Zheng *et al.* (2012) cut-off tools operate under very harsh conditions, and the main cause of failure is the breakage of their carbide inserts. In accordance with the findings of M. Zheng *et al.* the primary cause of failure in cut-off turning tools has been identified as carbide insert breakage. The proposed methods by A. Kurt *et al.* (2015) for studying the stress-strain state do not allow for a full assessment of all the factors involved, in particular the geometry of the cutter, on the destructive loads acting on it. The gap in these studies has been addressed by the current study, which emphasises the influence of geometric parameters – specifically, the installation angle of the carbide inserts – on stress distribution and overall tool durability Thus, studying the influence of geometric parameters on the stress state of carbide inserts of cut-off turning tools using the photoelasticity method makes it possible to optimise the design of tools, which will increase their strength and service life. While earlier studies by A. Pandey *et al.* (2024) focused on wear and temperature-induced failure, the current study provides a more precise understanding of the mechanical stress factors. The current study offers new insights into the relationship between tool geometry and stress concentration, which can help optimise tool design and extend service life.

Although the photoelasticity method is not as commonly used in modern analytical research due to the development of alternative techniques, it still offers significant advantages, as highlighted by Q. He *et al.* (2025). Specifically, this method is known for its high sensitivity, non-destructive nature, and ability to provide detailed insights into stress distribution. These advantages make the photoelasticity method particularly valuable for studying the stress-strain state of cutting tools under operational conditions. Geometric parameters of the tools themselves are among the critical factors affecting the reliability and durability of cutting tools. S. Shvets & V. Astakhov (2020) emphasised that the required geometry of the cutting system is achieved by adjusting the tilt of the carbide insert in both radial and axial directions. This observation aligns with the current study, which also investigates the impact of carbide insert installation angles on stress distribution and tool durability, further supporting the importance of geometric optimisation for enhancing tool performance. The durability of turning tools has a significant impact on machine tool productivity. According to research by S. Ghosh *et al.* (2018), cutting tool failure causes almost 20% of machine downtime. It is very important to know the remaining service life of the cutting tool in order to replace it in a timely manner that meets the goals of sustainable development - responsible use. However, it is difficult to predict the service life of a cutting tool because each cutting tool has different characteristics when cutting metal.

Cutting conditions such as feed rate, cutting speed, depth of cut, as well as workpiece material and tool material, have a significant impact on tool durability. These factors directly influence the wear rate and lifespan of the cutting tool, making it essential to optimise machining parameters to extend tool life. A.A. Yontar & Y. Kartal (2017) explored the effects of various cutting parameters on tool wear and machinability, emphasising the complexity of predicting tool lifespan due to the dynamic nature of these factors. Similarly, L. Bouzidab *et al.* (2018) highlighted the importance of optimising cutting conditions, particularly when finishing AISI 304 stainless steel, to minimise flank wear and maximise tool life. Both studies underscore the need for a comprehensive understanding of cutting parameters to improve tool performance and minimise downtime.

Tool geometry, such as positive/negative frontal angle or honed/beveled edge, also has a significant impact. A study of the causes of cutting tool failures by A. Rizzo *et al.* (2020) has shown that during the metal cutting process, cutting tools must withstand high heat, high pressure, abrasion, thermal and mechanical shocks. This finding aligns with the current study, which also underscores the critical role of cutting forces in determining the stress state within the carbide insert. Control and analysis of stress conditions are crucial for the reliability of elements. Ignoring stress concentrations can lead to structural failure, as demonstrated in the analysis of the silo tower collapse by Ye. Bakulin *et al*. (2022). Therefore, the data obtained in the current study on the distribution of stresses in a carbide plate allow reasonably selecting the optimal method of fastening it and prevent premature tool failure. The decisive parameter affecting the strength of a carbide turning tool insert is the cutting forces that arise during metal machining and, accordingly, cause a certain stress state in the insert body. Therefore, when designing a tool, it is very important to know the relationships between the parameters of the insert geometry in the tool coordinate system.

The influence of destructive factors was studied using a non-destructive photoelasticity method with a polarisation-projection device. Based on the obtained photos, isochromes and isoclines were used to determine the maximum permissible stresses in a carbide plate depending on its geometric shape and the angle of installation in the holder. The most acceptable from the point of view of the minimum tensions can be considered samples with carbide insertion angles of $\gamma = -15°$ and $\gamma = -30°$. This installation method will ensure the highest durability of the carbide insert and the cutting tool as a whole. The intermediate position is occupied by a sample with a mounting angle of $\gamma = 0°$, but the significant load on the front of the insert in the middle zone and the connecting end part can cause cracks and contribute to further destruction of the tool integrity. Regardless of the angle of insert insertion, the maximum loads occur in the cutting zone, which may be a determinant of tool life and requires further research.

The durability of cutting tools is influenced by a complex interplay of factors, including cutting conditions, tool geometry, and material properties. Research has shown that factors such as feed rate, cutting speed, and depth of cut, along with tool and workpiece materials, significantly affect tool wear and lifespan. Additionally, the geometry of the tool, such as insert angles and edge design, plays a crucial role in determining stress distribution and tool performance. Cutting tools, particularly those with carbide inserts, must withstand high heat, pressure, and mechanical stresses, making it essential to understand the relationship between cutting forces and tool geometry. As demonstrated in the current study, optimising these parameters is key to improving tool durability and minimising failures. These insights pave the way for further advancements in tool design, with a focus on optimising geometry and cutting conditions to enhance tool life and operational efficiency. The findings underscore the importance of comprehensive approaches to tool design and machining process optimisation for the future of sustainable and efficient manufacturing.

CONCLUSIONS

Hidden fracture of the inner part of carbide inserts is difficult to detect visually, compared to the possible appearance of cracks on their surface. This study provided a clear understanding of the distribution of loads along the carbide insert along its width. In an applied sense, it is important to know the geometric features of the cutting tools to establish the required angles of attachment of carbide inserts to the holder, depending on the specified production conditions. This will increase the service life of the cutting tools and the productivity of metal cutting operations.

According to the results of the studies, it was found that the angle of attachment of the carbide insert to the toolholder significantly affects the stress distribution in it during tool operation. It is the current level of stress in the defined zone of the carbide insert that determines its resistance to cracking and subsequent fracture. In terms of insert strength, the most optimal setting angle is $\gamma = -30°$, since the minimum value of dangerous stresses σ_{equ} is observed along the measurement points in all three zones along the width of the insert.

The results have shown that the most dangerous option for installing a carbide insert in a holder is an angle of $\gamma = 30°$. The maximum stresses occur in all studied zones along the width of the insert along the measurement points, compared to the attachment angles of 15°, 0°, -15°, and -30°. The sample with an attachment angle of $\gamma = 30°$ can be considered to have the best surface finish quality indicators. However, even minor deviations from the machining process technology cause significant wear and frequent tool breakage. Further research should focus on a deeper investigation of the impact of various geometric parameters of carbide inserts on their stress resistance under different machining conditions.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Adamu, A., Bawa, M.A., & Abioye, A.M. (2021). Optimization of cutting parameters of tool wear in turning operations: A review. *Saudi Journal of Engineering and Technology*, 6(9), 315-319. doi: 10.36348/sjet.2021.v06i09.003.

[2] Bakulin, Ye.A., Yakovenko, I.A., & Bakulina, V.M. (2022). Determination of the stress-strain state parameters of the silo tower structure and its structural elements based on the results of destruction. In *Achievements of Ukraine and EU countries in technological innovations and invention: Collective monograph*. Riga: Baltija Publishing. doi: 10.30525/978-9934-26-254-8-1.

[3] Bergmann, B., Schenzel, J., & Kraeft, M. (2025). Tool failure: A method for stress calculation of worn cutting tools. *CIRP Annals – Manufacturing Technology*, 10, article number 1016. doi: 10.1016/j.cirp.2025.03.032.

[4] Bouzidab, L., Berkania, S., Athmane, M.Y., Girardinc, F., & Mabroukid, T. (2018). Estimation and optimization of flank wear and tool lifespan in finish turning of AISI 304 stainless steel using desirability function approach. *International Journal of Industrial Engineering Computations*, 9(3), 349-368. doi: 10.5267/j.ijiec.2017.8.002.

[5] Cai, W., Xiang, J., Dong, G., Lai, K., & Wiercigroch, M. (2025). Analytical modelling of parallel multidirectional cutting of slender shafts. *International Journal of Mechanical Sciences*, 288, article number 110024. doi: 10.1016/j.ijmecsci.2025.110024.

[6] Ghosh, S., Naskar, S.K., & Mandal, N.K. (2018). Estimation of residual life of a cutting tool used in a machining process. *MATEC Web of Conferences*, 192, article number 01017. doi: 10.1051/matecconf/201819201017.

[7] Gutakovskis, V., Avisane, A., & Mozga, N. (2022). Finite element analysis of stainless steel AISI 420 cutting process to predict cutting forces and temperature distribution in duratomic-coated cutting tool. In *21st international scientific conference: Engineering for rural development* (pp. 489-495). Jelgava: Latvia University of Life Sciences and Technologies. doi: 10.22616/ERDev.2022.21.TF167.

[8] He, Q., Du, Y., Qin, Q., & Qiu, W. (2025). Automatic stress analysis method for semiconductor materials based on dual-wavelength infrared photoelasticity. *Optics and Lasers in Engineering*, 184(1), article number 108648. doi: 10.1016/j.optlaseng.2024.108648.

[9] Kumar, M., Pensia, L., Kaur, K., Kumar, R., Awatsuji, Y., & Matoba, O. (2025). Advances in optical metrology: High-bandwidth digital holography for transparent objects analysis. *Photonics*, 12(6), article number 617. doi: 10.3390/photonics12060617.

[10] Kurt, A., Yalçin, B., & Yilmaz, N. (2015). The cutting tool stresses in finish turning of hardened steel with mixed ceramic tool. *International Journal of Advanced Manufacturing Technology*, 80, 315-325. doi: 10.1007/s00170-015-6927-3.

[11] Liu, Y., Su, G., Xia, Y., Zhang, H., Wei, G., Sun, Y., Li, B., Zhang, P., Du, J., & Fang, B. (2025). Influences of impact loading rate on the protective effect of shims on breakage of inserts in intermittent cutting. *Journal of Manufacturing Processes*, 143, 1-16. doi: 10.1016/j.jmapro.2025.03.096.

[12] Lubis, S., Darmawan, S., Rosehan, Winata, W., & Zulkarnain, M. (2020). Tool wear analysis of ceramic cutting tools in the turning of gray cast iron materials. *IOP Conference Series: Materials Science and Engineering*, 857(1), article number 012003. doi: 10.1088/1757-899X/857/1/012003.

[13] Östby, J., Toller-Nordström, L., & Norgren, S. (2022). Insights into plastic deformation and binder lamella orientation in hardmetal turning inserts. *International Journal of Refractory Metals and Hard Materials*, 103(4), article number 105779. doi: 10.1016/j.ijrmhm.2022.105779.

[14] Pandey, A., Madhavan, V., & Moscoso-Kingsley, W. (2024). Photoelastic measurement of tool stress distribution when machining AA-7075-T6 under dynamic conditions produced by shear banding. *Journal of Manufacturing Processes*, 124(2), 966-971. doi: 10.1016/j.jmapro.2024.05.079.

[15] Rizzo, A., Goel, S., Grilli, M.L., Iglesias, R., Jaworska, L., Lapkovskis, V., Novak, P., Postolnyi, B.O., & Valerini, D. (2020). The critical raw materials in cutting tools for machining applications: A review. *Materials*, 13(6), article number 1377. doi: 10.3390/ma13061377.

[16] Roszkowski, A., Piórkowski, P., Skoczyński, W., Borkowski, W., & Jankowski, T. (2020). Study on the impact of cutting tool wear on machine tool energy consumption. *Advances in Science and Technology Research Journal*, 14(3), 158-164. doi: 10.12913/22998624/123283.

[17] Shvets, S., & Astakhov, V. (2020). Effect of insert angles on cutting tool geometry. *Journal of Engineering Sciences*, 7(2), A1-A6. doi: 10.21272/jes.2020.7(2).a1.

[18] Sobron, M., Lubis, Y., Djamil, S., & Andri, S.D. (2020). Tool life investigation of carbide cutting tools in the turning of cast iron material. *IOP Conference Series: Materials Science and Engineering*, 725, article number 012039. doi: 10.1088/1757-899X/725/1/012039.

[19] Wu, F., Liu, Z., Guo, B., Sun, Y., & Chen, J. (2021). Research on the burr-free interrupted cutting model of metals. *Journal of Materials Processing Technology*, 295, article number 117190. doi: 10.1016/j.jmatprotec.2021.117190.

[20] Xiao, C., Lu, S., Zhang, W., Zhang, J., & Liu, J. (2024). Finite element analysis of cutting temperature in precision cutting of composite energetic material. *Journal of Composites Science*, 8(12), article number 525. doi: 10.3390/jcs8120525.

[21] Yontar, A.A., & Kartal, Y. (2017). Investigation to effects of cutting parameter on tool wear, tool life and machinability in machining operation. *International Journal of Applied Science Engineering & Management*, 3(3), 13-21.

[22] Zabrodskyi, P., Sheludchenko, B., & Kukharets, S. (2021). Investigation of the influence of the method of fixing the cut-off tool inserts on its stress state. *Scientific Horizons*, 24(1), 7-13. doi: 10.48077/scihor.24(1).2021.7-13.

[23] Zheng, M., Gang, T., Yao, B., Li, F., & Song, Y. (2012). Research on the cutting heat and wear of indexable inserts with different transition surfaces. *Advanced Materials Research*, 468-471, 1290-1293. doi: 10.4028/www.Scientific.net/AMR.468-471.1290.

Павло Забродський

Кандидат технічних наук, доцент
Поліський національний університет
10007, б-р Старий, 7, м. Житомир, Україна
https://orcid.org/0000-0002-3904-564X

Савелій Кухарець

Доктор технічних наук, професор
Сільськогосподарська академія Університету Вітовта Великого
LT-53362, вул. Студента, 11, м. Каунас, Литва
https://orcid.org/0000-0002-5129-8746

Богдан Шелудченко

Кандидат технічних наук, професор
Поліський національний університет
10007, б-р Старий, 7, м. Житомир, Україна
https://orcid.org/0000-0002-8137-0905

Олександр Медведський

Кандидат технічних наук, доцент
Поліський національний університет
10007, б-р Старий, 7, м. Житомир, Україна
https://orcid.org/0000-0001-7458-5337

Дослідження напруженого стану твердосплавної пластини відрізного різця

Анотація. Різальні інструменти з твердосплавними пластинами працюють у жорстких умовах, які можуть спричинити розтріскування та руйнування, оскільки їхня міцність безпосередньо залежить від напружено-деформованого стану. Дослідження було проведено з метою визначення впливу кута встановлення в державку напаяних твердосплавних пластин відрізних різців на їх напружений стан. Експериментальні дослідження за методом фотопружності відзначаються високою чутливістю, неруйнівним впливом, наочністю, швидкістю отримання результатів. Виконували дослідження на поляризаційно-проекційній установці ППУ-7 з використанням фізичних моделей твердосплавних пластин з оптично-чутливого матеріалу ЕД-6, які приєднувались у державку з різними кутами встановлення. Отримані фотокартки ізохром та ізоклин дозволили оцінити реальну візуальну картинку діючих напружень. Під час експерименту моделі відрізних різців навантажувалися за допомогою спеціально розробленого стенду, що моделював силу різання. Розрахунок головних напружень виконувався методом різниці дотичних напружень для обраних кутів встановлення пластин. Доведено, що кут приєднання твердосплавної пластини до державка безперечно впливає на розподіл в ній напружень під час металообробки. Встановлено кількісний зв'язок між величиною кута приєднання до державки та величиною діючих внутрішніх напружень у твердосплавної пластинки. Найбільші напруження σ_{equ} діють в перерізі ріжучого леза твердосплавної пластини незалежно від кута приєднання до державка. Встановлено найменший рівень діючих напружень вздовж точок вимірювання у трьох зонах за шириною твердосплавної пластини при приєднані до державка під кутом -30°. За умови кутів приєднання 15°, 0° та -15° виникають критичні напруження у всіх досліджених зонах за шириною пластини вздовж точок вимірювання

Ключові слова: ізоклин; ізохром; анізотропія; фотопружність; кут встановлення

Machinery & Energetics
Vol. 16, No. 4. 2025

Journal homepage:
https://technicalscience.com.ua/en

UDC 621.01:631.31
DOI: 10.31548/machinery/4.2025.43

Stanislav Nikolayenko
Doctor of Pedagogical Sciences, Professor, Academician of the National Academy of Agricultural Sciences of Ukraine
National University of Life and Environmental Sciences of Ukraine
03041, 15 Heroiv Oborony Str., Kyiv, Ukraine
https://orcid.org/0000-0001-5271-1777

Mykola Volokha*
Doctor of Technical Sciences, Professor
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
03056, 37 Beresteisky Ave., Kyiv, Ukraine
https://orcid.org/0000-0002-0112-7324

Digitalisation of field research data – the basis of farm management

Abstract. The efficiency of modern agricultural production management was shown to depend largely on the implementation of digitalisation processes, which transformed the paradigm of production process management and created conditions for increasing agricultural productivity through the use of innovative digital technologies. The purpose of the article was to review scientific publications devoted to the digitisation of field experiment databases in order to enhance the profitability of agricultural production through the development of effective management decisions in crop production based on data obtained from numerous sources of digital transformation. It was demonstrated that the diversity of technological processes, as well as input and output data involved in agricultural production, resulted in significant complexity and integrative challenges in assessing production sustainability. It was identified that the majority of farmers were not specialists in information technologies and were therefore unable to fully comprehend the complexity of the algorithms underlying modern digital solutions. Approaches to the use of big data in agriculture were systematised with regard to the evolution of the 4V concept towards the 5V framework, in which data value was emphasised as a critical requirement. A scientific and methodological approach aimed at improving agricultural production efficiency through the application of machine learning methods was proposed, including automatic crop recognition, disease and weed detection, and the forecasting of crop yield and product quality. It was established that, alongside material and technical support, the intellectualisation of production and management activities based on digitalisation constituted a priority factor in the development of the agro-industrial sector, being considered a highly effective means of increasing the return on investment in agriculture by 10-15%

Keywords: agricultural production efficiency; digitalisation; yield; soil fertility; moisture reserves

INTRODUCTION

Digitalisation and the integration of advanced technologies, including data analytics, machine learning (ML), and automation, offer significant potential to transform agricultural management, increase productivity, and ensure long-term sustainability. These innovations are becoming essential as part of efforts to increase agricultural production while minimising environmental impact. In review of the concepts and methods of sustainable agricultural development, conducted by a team of authors from different countries it is argued that the way to solve the above problems, which put enormous pressure on the level of agricultural production, is to increase the efficiency of agricultural technologies while reducing the burden on the environment. These two important factors form the basis for transforming production in the agricultural sector based on the principles and characteristics of precision farming.

Article's History: Received: 07.07.2025; Revised: 17.10.2025; Accepted: 27.11.2025.

Suggested Citation:
Nikolayenko, S., & Volokha, M. (2025). Digitalisation of field research data – the basis of farm management. *Machinery & Energetics,* 16(4), 43-52. doi: 10.31548/machinery/4.2025.43.

*Corresponding author

Such modernisation has great potential for ensuring the sustainable development of agricultural production while maximising its productivity and environmental safety.

Researchers A. Cravero & S. Sepúlveda (2021) explored the use of ML in big data applications within agriculture. They focused on the practical adaptations of these technologies to improve efficiency in agricultural processes. By leveraging advanced data analysis techniques, they demonstrated how automation can enhance decision-making and optimize agricultural operations. In general, according to F. Zecca (2019), smart agriculture is based on four basic principles that make it possible to meet the growing needs of the population: a) optimal management of natural resources; b) use of advanced technologies; c) preservation of the ecosystem; d) development of relevant services.

An essential feature of modern agriculture is the use of information and communication technologies (ICTs), which are encouraged by policy makers around the world. ICTs may, for example, include farm management information systems, air and soil moisture sensors, accelerometers, wireless sensor networks, cameras, drones, low-cost satellites, online services, automated guided vehicles, etc. The disadvantage and a serious obstacle to obtaining valuable information from field research data is that traditional data processing methods are unable to meet the ever-growing needs in the new era of "smart" agriculture based on the algorithmic and software being created everywhere (Sørensen *et al.*, 2021). That is why ML has emerged, taking advantage of the rapid exponential growth of computing power and the tools of artificial intelligence and the Internet of Things (IoT) (Evstatiev & Gabrovska-Evstatieva, 2020). Researchers from Brazil G.M.D. Almeida *et al.* (2021) have demonstrated that logistics is a particularly important challenge in agriculture. As a cutting-edge technology, ML can solve significant problems, particularly in sugarcane transport chains, thereby significantly improving the overall economics of production. Ukrainian researchers also raise issues related to the digitisation process in Ukraine's agro-industrial complex. M. Rudenko (2019), revealing digitisation as a component of modern management methods in agriculture, notes the lack of a unified approach to interpreting its essence. The author considered the multifaceted definition of this concept from different points of view: the state; scientists; practitioners (entrepreneurs); society.

In the article by Y. Voloshchuk (2019) the processes of formation and development of the digital economy of Ukraine are considered, the differences and peculiarities of its functioning and the impact on the competitiveness of the national economy are being studied. It is determined that digitalisation of the country is not a simple transition to e-government, but a full-scale introduction of digital technologies in all spheres of society: from education, skills, production, business to bridging the digital divide and creating a complete digital infrastructure. The basic moments of the digital strategy of the country are defined, problems of digitisation of industries, in particular agro-industrial complex, and ways of their solution are outlined. M. Volokha *et al.* (2022) concluded that at the level of agricultural enterprises or farms, the digitisation of the data obtained allows monitoring production processes and processes of plant growth and development; assessing the condition of crops and forecasting crop yields; identifying areas requiring fertilisers and pesticides; controlling crop rotations and the quality of agrotechnical measures; determining soil moisture, assessing snow cover, freezing areas, etc.

M. Lampridi *et al.* (2019) noted that one of the main problems is that the vast majority of farmers around the world are not experts in this field and therefore cannot fully understand the basic patterns obtained using ML algorithms. Despite this, a group of scientists from Italy, Tunisia, Malaysia, and the United Kingdom (Rejeb *et al.*, 2022) argues that the growing body of knowledge related to artificial intelligence and ML requires a bibliometric approach to understand the structure of knowledge. As research in this field develops and becomes more complex, it is extremely important to conduct a retrospective assessment of accumulated publications in order to identify new contributions, reflect research trends, and identify promising directions for future research.

The purpose of this article was to review publications devoted to the digitisation of field experiment databases in order to increase the profitability of agricultural production by developing effective management decisions in crop production based on the use of data from sources of digital transformation. To achieve this goal, the following tasks were set:

- systematise existing methods of digitising field research data based on the 5V requirements: volume, variety, velocity, reliability, and value;
- define the fundamental principles for developing realistic solutions that integrate the expertise of specialists in both computing and agriculture;
- develop a scientific and methodological approach to enhancing the efficiency of ML for automatic crop recognition, disease and weed detection, and yield forecasting.

To identify relevant studies related to the objective, literature published mainly between 2019 and 2024 was filtered, and the bibliometric method was used, which allows collecting information about publications for further generalisation. Scientometric databases and search platforms, including Scopus, Web of Science, Google Scholar, and ScienceDirect, were used. To study the aspects of agricultural management, combinations of the keywords "crop management", "water management" and "soil management" in combination with the word "machine learning" were used. The texts and references in the named articles were scanned, and a critical analysis of the publications was conducted. Journal articles, conference papers, monograph, PhD dissertations were considered appropriate.

The literature review process was carried out on the methodological basis of synthesising all available criteria and a two-level analysis of relevant publications, which facilitates systematisation. Thus, a systematic literature review of more than three hundred studies on ML methods for crop management and decision-making, and the assessment of precision crop production as a component of smart agriculture over the past five years was implemented. The study of the methodologies used is of particular importance, as there are no recommended standards or norms for assessing the sustainability of methods in the crop production industry. The known methods most often include tools based on indicators, indices, and multicriteria. They are often used in combination, so a clear distinction is not always possible.

Intelligent data processing and ML in agricultural logistics and supply chains

ML algorithms are fundamental to artificial intelligence and data science, enabling computers to learn and improve from data without explicit programming. They have a wide range of applications, including image and speech recognition, recommendation systems, and data analysis. In supervised learning, algorithms are trained on labelled data to make predictions or classifications. For classification tasks, common algorithms include Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Networks. Regression algorithms, such as Linear Regression and Polynomial Regression, predict continuous outcomes like prices or scores. Unsupervised learning involves finding patterns or structures in unlabelled data. Clustering algorithms like K-Means, Hierarchical Clustering, and DBSCAN group similar data points, aiding market segmentation and anomaly detection. Dimensionality reduction algorithms like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE) help in reducing the number of features in a dataset while preserving key information. This approach combines labelled and unlabelled data, useful when acquiring fully labelled data is costly or time-consuming. It's used in various applications, including text classification and image recognition. In reinforcement learning, agents learn to make sequences of decisions by receiving feedback in the form of rewards. It's crucial in robotics, game playing, and autonomous systems. These ML algorithms play a pivotal role in solving complex problems and making data-driven decisions across diverse domains (Jiang *et al.*, 2019).

To eliminate many of the shortcomings inherent in supply chains, ML can analyse large amounts of data, identify patterns and generate effective solutions. X. Fan *et al.* (2021) defined the role of ML in supply chain optimisation as follows. learning plays a transformative role in optimising supply chains, revolutionising the way businesses manage their operations. Its impact is multi-faceted and extends across various aspects of supply chain management. One of the key areas where ML excels is demand forecasting. By analysing historical data, market trends, and external factors, ML algorithms can predict future demand more accurately than traditional methods. This enables businesses to adjust their inventory levels, production schedules, and distribution strategies to meet customer demand while minimising overstock or stockouts. Route optimisation is another crucial aspect of supply chain management where ML shines. ML algorithms can analyse real-time data on factors like traffic, weather, and road conditions to determine the most efficient routes for transportation. This not only reduces transportation costs but also minimises delivery times, enhancing customer satisfaction. Inventory management is also greatly enhanced by ML. These algorithms can identify optimal inventory levels, taking into account lead times, demand variability, and costs associated with holding excess inventory. This prevents overstocking and understocking, leading to cost savings and improved efficiency. Quality control and maintenance in supply chains benefit from ML as well. Sensors and data analytics allow for real-time monitoring of equipment and product quality. Predictive maintenance models can anticipate when machinery or vehicles are likely to fail, enabling proactive maintenance to prevent costly breakdowns. Risk management is an essential aspect of supply chain optimisation. ML algorithms can analyse a multitude of data sources to identify potential risks, whether they are related to geopolitical issues, supplier reliability, or market fluctuations. This early warning system enables supply chain managers to mitigate risks and develop contingency plans.

Solving logistics problems requires a combination of technology, logistics innovation, and cooperation between supply chain participants. According to V. Borodin *et al*. (2016), developing strategies to manage seasonal fluctuations, optimise transportation, forecast demand fluctuations, and ensure consistent quality control are vital to the successful and sustainable operation of agricultural supply chains. ML can optimise transport routes, reducing waste and ensuring the timely delivery of perishable goods for example for the purpose of extending and regulating the shelf life of fruits and vegetables as in China (Li *et al*., 2024). New artificial neural network models are being developed to accurately assess the temperature of food products stored in multi-temperature refrigerated vehicles (Zou *et al.*, 2023).

The American Chemical Society offers a technology for large-scale food storage by analysing images using a combination of intelligent sensor film and QR codes. Such rapid, real-time monitoring of food quality will ultimately lead to a reduction in food waste and losses (Pounds *et al*., 2022). Another advantage of ML is that quality control systems based on it can automate the verification process, ensuring that only the highest quality products reach consumers. A. Rejeb *et al.* (2022) examined the interaction between artificial intelligence technologies and the agri-food industry, with particular emphasis on their role in improving efficiency, quality control, and sustainability across the entire food supply chain. The authors analysed how ML-based systems enable automated monitoring, traceability, and decision support, thereby enhancing product quality and reducing operational risks. It was concluded that the integration of artificial intelligence represents a key driver of digital transformation in the agri-food sector. By reducing its environmental impact, ML can contribute to sustainable agriculture by supporting the efficient use of water, fertilisers, and pesticides.

Given that farmers are becoming increasingly accustomed to smartphones, scientists, particularly Indian (Mayuri, 2018) and Chinese (Yuan *et al.*, 2022), have proposed special smartphone applications as a possible solution to the above problem. These researchers believe that more user-friendly systems should be developed. In particular, devices that are easy to understand and operate, such as a visualisation tool with a user-friendly interface for the correct presentation and processing of data, would be valuable, as explored in detail in the work of Australian scientists (Ang & Seng, 2021).

In Ukraine, many works by international and Ukrainian scientists and experts are devoted mainly to the formation of digitalisation as a tool for the development of informatisation. Researchers emphasise that digitalisation requires a systematic approach to digital transformation and human capital development, as well as effective state policy and institutional support to ensure sustainable economic development. A similar view is shared by T. Lazorenko & I. Sholom (2018), who underlined the importance of coordinated digital reforms and the development of human potential. In addition, A. Trushlyakova (2018) pointed out the need for thorough planning of a digital strategy and the selection of an effective model to achieve the expected outcomes of digitalisation.

Thus, most authors whose research has focused on the essence of the concept of digitalisation limit themselves to the processes of national economic development in the context of globalisation. Instead, V. Varga (2020) conducted a content analysis of the basic term “digitalisation” and presented his own vision of the concept of digitalisation as one of the factors of enterprise competitiveness, when, thanks to the more convenient use of digital technologies, work with a significant database of information for economic growth is accelerated. The term “digitalisation” is understood as the process of using, applying, transferring, and translating information into digital format. The author notes that economic growth and transformation occur through the expansion and more convenient use of digital technologies. It is emphasised that innovations are designed to accelerate and facilitate work with a large database, use a unified system of cooperation with all departments of the company, and establish interaction with customers and suppliers. All changes that occur during the implementation of innovative innovations lead to the improvement and optimisation of the efficiency of the enterprise as a whole. It has been determined that digitalisation is currently leading not only to the emergence of new specialties, but also to changes in the labour market as a whole. It has been found that in order to increase competitiveness, economic entities must create a strategy that focuses on the use of innovative solutions, while introducing scientific and technical innovations, applying intellectualisation, and effectively implementing digital innovations. All these actions are a driving force for development, which implements and accelerates the profitable activities of the enterprise. It has been analysed that digitalisation provides enterprises, including agricultural ones, with competitive advantages in all areas of activity, and thanks to this, the business entity will ensure confident stability in the future.

Big Data and data fusion technologies for smart and precision agriculture

The processes of digitising large amounts of field data, i.e., collecting field research data, processing or storing it, are inextricably linked to the concept of “Big Data”, as farmers, researchers and other agricultural professionals are now accumulating large amounts of data (Sonka, 2016). According to the United Nations, the exponentially growing amount of data generated by digital technologies requires greater capacity to store it on electronic computing applications for editing, analysis, and interpretation. At the same time, it should be noted that since the second half of the 1980s, the volume of data storage has doubled. The latter has significant potential and motivates decision makers to improve their intellectual level (The Guardian, 2017). If while in 2018 the US National Institute of Standards and Technology (NIST) noted that big data consisted of sets with characteristics of large volume (1), variety (2), velocity (3), and variability (4), i.e., 4V (Chang & Grady, 2018), then in 2020-2022, in addition to these characteristics, emphasising the scale and complexity of big data, scientists R.Y. Choi *et al.* (2020) added the descriptors of high dimension, high complexity, and high uncertainty, i.e., 5V, to the definition. Characterising the requirements of 5V it was highlighted the following (Table 1). The rationale for increasing the number of V units is presented in the review by A.Y. Sun & B.R. Scanlon (2019), which highlights the impact of big data and ML technologies on various aspects of environmental and water resource management, as advances in remote sensing methods have contributed to the spread of big data in many areas of environmental and water management, including weather forecasting and natural disaster prevention. Big data opens up new opportunities, although it requires, first and foremost, new forms of information analytics, provided that it is properly integrated with ML, which will help to fully unlock its potential. The review presents actual applications of big data ML in environmental and water resource management, where the authors focus on deep learning, especially since the illustrated data on environmental and water resource management require the use of analytical tools for big data analysis, ranging from virtualisation and edge computing to high-performance processing. It is noted that one of the most challenging applications of big data and deep interdisciplinary learning is precisely data from environmental and water resource management. The scope of the applications cited in this review includes more than 1,000 articles, which prove that: 1) researchers of environmental and water resource management data will significantly change the planning, conduct, and analysis of experiments; 2) the most effective use of big data advantages is possible with low data cleaning costs; 3) thanks to a more reliable and faster flow of information, the significant effectiveness of deep learning methods has already been demonstrated in solving many problems of improving disaster relief efforts. Finally, big data and artificial intelligence will be most valuable when they improve the level of reasoning and discussion. For big data in machine learning applied to environmental and water resource management, it is expected that deep learning will significantly improve the ability to understand the conditions and changes the planet is experiencing as a result of human and environmental impacts.

Table 1. Requirements for components in the 5V kit

5V component No.	Name	Characteristic
1	Volume	Refers to the size of big data. In the last decade, the growth of big data has coincided with the rapid development of ICT and the Internet, and the volume of its deployment is often measured in terabytes, petabytes, and sometimes exabytes (the amount of information in 1 exabyte is equal to a quintillion (1,018 bytes). To process data at scale, it is necessary to switch to a distributed computing platform. At the same time, the size of a data set is relative and may have different values in different applications.
2	Variety	This applies to data collection methods. Depending on the sources of observation, big data can be classified as passive or active, with the former originating from derivative digital media (e.g., GPS data from mobile phones) and the latter from planned data collection activities (field observations or targeted remote sensing). Passive data is collected without the active participation of the user and, as a result, does not explain the reasons for events. Depending on the format, data can be structured (sensor readings and satellite data) or unstructured (e-mail, video, audio, or text), with the main difference being that unstructured data cannot be easily described using a predefined model due to the limited information resources in the case of unstructured data. In other words, diversity creates an additional level of complexity when working with big data. Two methods are commonly used to reduce data diversity: virtualisation or data fusion.
3	Velocity	Speed attribute. This may refer not only to the speed of data generation, but also to the speed of analysis required to receive it. Data generation speed is a direct consequence of network connectivity, widespread use of smart devices, and real-time monitoring networks. High-speed data sets that are continuously generated by various sources require special processing. A data stream is defined as an unlimited sequence of events that requires continuous, incremental processing.
4	Variability	Variability refers to changes in all other attributes of big data, such as changes in data flow rates or changes in data values, with the latter often occurring in data obtained through crowdsourcing, i.e., a method of finding solutions to problems without significant external assistance. This is a serious problem for data processing, as many ML algorithms operate on the principle of stationarity.
5	Value	This refers to potential issues with data inconsistency and quality, such as partial missing data, anomalies, and input errors. High-quality products created by data centres involve a significant number of processing and integration steps. However, as data centres constantly update their algorithms, managing their versions becomes a serious challenge when accepting data from different sources. Therefore, when developing platforms for data collection and analysis, it is necessary and important to protect the reliability of information.

Source: developed by the authors based on W.L. Chang & N. Grady (2018)

Chinese electrical engineers from Xinjiang University X. Sibo *et al.* (2022) developed a data fusion algorithm based on the idea of parallel computing to improve the performance of the system for monitoring unevenly distributed temperature in smart greenhouses, which optimises the quality of a hierarchical network of wireless sensors and increases computing efficiency using a real-time data fusion strategy. In the study by R.Y. Choi *et al.* (2020), the application of ML methods and neural networks to real-time intelligent data processing tasks was analysed. The authors examined a multi-level system architecture in which, at the lower level, temperature data were collected and pre-processed using an improved Kalman filter, traditionally applied in applied mechanics and control theory to reduce measurement noise. At the intermediate level, characteristic of automation systems and computer-integrated technologies, a parallel intersection data fusion algorithm was implemented, enabling the aggregation of information from multiple sensors and increasing the reliability of the estimates. The upper level of the system was based on the use of an extreme ML algorithm with a global fusion centre, reflecting modern approaches in electrical engineering and power engineering to intelligent data analysis and the optimisation of operating modes of complex technical objects. The simulation results reported by R.Y. Choi *et al.* demonstrated that the integration of classical control theory methods with ML tools ensured a significant improvement in the accuracy and responsiveness of the temperature control system. Thus, the study synthesised and systematised contemporary scientific developments at the intersection of mechanics, electrical engineering, automation, and computer-integrated technologies, and confirmed the feasibility of applying multi-level intelligent algorithms for the design of efficient, safe, and energy-efficient engineering systems.

Automation of crop production management using ML

According to the works reviewed in this study with regard to crop production, soil fertility, and water regime, sections devoted to water and soil resource management were less extensively investigated and accounted for approximately 18-20% of the total number of articles (up to 10% per section). In contrast, crop production management represented the majority of publications (up to 80% of the total) and was divided into four main categories in which ML methods were applied: crop recognition, crop quality assessment, disease and weed detection, and profitability forecasting. Automatic crop recognition based on leaves is the most common approach for determining specific leaf characteristics such as colour, shape and texture. S. Virnodkar *et al.* (2020) systematically analysed remote sensing data and ML algorithms as tools for automated assessment of crop water stress, demonstrating the potential of image-based methods to replace manual field inspections and support automated decision-making in precision agriculture. The proposed approaches emphasised the integration of sensor data, image processing, and algorithmic analysis, which directly corresponds to modern trends in automation and computer-integrated agricultural technologies. With the expansion of the use of satellites and aircraft, crop classification using remote sensing has become particularly popular. E.I. Papageorgiou *et al.* (2018) focused on the development of intelligent computational models for crop and product classification, highlighting the role of automated data acquisition systems, signal processing, and soft computing methods in reducing human subjectivity and increasing the reliability of quality assessment. Their work illustrates how automated classification systems can be embedded into larger cyber-physical and electromechanical control frameworks for agricultural production. In other words, the development of computer software and image processing devices has led to the automatic recognition and classification of crops, in particular based on yield. Such solutions reflect a broader shift towards automation of analytical and control functions in agricultural systems, where software-driven image analysis replaces traditional manual monitoring and enables real-time feedback for management and optimisation processes.

At the same time, farmers in countries such as Bangladesh choose crops for planting based on their own knowledge. M. Hasan *et al.* (2023) addressed this limitation by proposing an ensemble ML-based recommendation system that formalises empirical farmer knowledge into an automated decision-support tool. The study demonstrated how data-driven algorithms can enhance traditional practices by providing objective, scalable, and reproducible recommendations, thereby supporting the transition from experience-based decision-making to automated and intelligent crop management systems.

High-quality agricultural products are usually sold at more favourable prices, which gives farmers higher incomes. For example, fruit quality, flesh firmness, soluble solids content, and skin colour are some of the simplest maturity indicators used for apple harvesting (Zhang *et al.*, 2021). Harvesting dates have a significant impact on the quality characteristics of harvested products of both high-value crops (grapes, vegetables, woody crops, herbs, etc.) and row crops. Thus, the development of decision support systems can help farmers and managers make appropriate management decisions to improve product quality. For example, selective harvesting is a management technique that can significantly improve quality. In addition, crop quality is closely related to food waste, which is an additional problem that modern agriculture has to face, as if the crop deviates from the desired shape, colour, or size, it may simply be thrown away as substandard. However, by applying image processing technologies for such discarded raw materials, developed on the basis of ML algorithms, it is possible to reduce the amount of food waste for processing companies.

Diseases were identified by experienced agronomists through field reconnaissance. However, this process is time-consuming and based solely on visual inspection. Technological advances have made sensor systems commercially available that can identify diseased plants before symptoms appear. In addition, computer vision, especially using deep learning, has made significant progress over the past few years, as noted in the article by J. Zhang *et al*. (2021). The authors demonstrated that convolutional neural networks can be effectively applied to automated disease recognition using image data collected within the Agricultural IoT framework, reducing dependence on expert judgement and enabling real-time diagnostic decision-making. Their work highlighted the role of software-based vision systems as functional elements of automated control loops in crop production. A review by S.R. Upadhyaya *et al.* (2024) as a scientific achievement highlights the potential of ML as an additional tool for predicting genetic loci that contribute to pathogen resistance. This study extended automation concepts beyond operational monitoring by introducing data-driven predictive models at the genetic level, illustrating how ML can support automated breeding strategies and long-term optimisation of system resilience in agricultural production.

Weeds typically germinate and develop very quickly due to their abundant seed production and longevity and spread over large parts of the field, competing with crops for resources including space, sunlight, nutrients and water availability. In addition, weeds often emerge earlier than crops and do not have to face natural enemies that negatively affect cultivated plants. To prevent a decrease in yields, weed control can be done either by mechanical cultivation, herbicide application, or a combination of both. However, mechanical tillage is sometimes difficult to implement and ineffective if not done properly, making herbicide application the most common practice. The use of large quantities of herbicides is both expensive and harmful to the environment, especially if applied uniformly without taking into account differences in weed density across the field. Prolonged use of herbicides is likely to make weeds more resistant, leading to more difficult and costly weed control. Significant progress has been made in solving the problem of weed differentiation in crops based on smart agriculture. This differentiation can be achieved by remote or proximity sensing with sensors mounted on satellites, airborne and ground vehicles, and drones, both ground and airborne. However, converting data collected by UAVs into “usable” information is a challenging task, as their collection and classification require considerable effort. ML algorithms combined with imaging or non-imaging spectroscopy technologies can provide real-time differentiation and localisation of certain weed species, enabling precise herbicide application and planning of the shortest weeding path in predefined areas instead of spraying the entire field (Ofosu *et al.*, 2023).

According to X. Gao *et al.* (2025), ML-based approaches represent an innovative solution to the persistent challenge of weed management in agriculture by enabling automated detection and control. The system integrates advanced ML algorithms with robotics to accurately find and cut weeds in real-time. A convolutional neural network (CNN) is trained on a diverse dataset comprising images of various crops and weed species, allowing the model to develop a robust understanding of visual cues associated with both. This trained model is deployed on a robotic platform equipped with cameras and actuators for real-time decision-making and precise weed removal. The integration of deep learning models with control systems that include laser modules, robotic manipulators, and precision spray nozzles facilitates the development of intelligent robotic mowing systems that significantly reduce chemical herbicide consumption and improve operational efficiency compared to traditional approaches. Significant obstacles remain, including limited adaptability to the environment, real-time processing limitations, and insufficient model generalisation. Future directions include integrating diverse data sources, developing optimised models, and improving intelligent decision-making systems, creating a foundation for the development of sustainable agricultural production.

Yield forecasting and resource management based on IoT and ML

Yield forecasting is one of the most important and complex topics in modern agriculture. An accurate model can help farm owners, for example, make informed management decisions about which crops to grow to ensure that the yield meets current market demands. However, this is not an easy task, it consists of different steps and directions of digitalisation, because the yield forecast can be determined by several factors, such as the technical level of the used machine complexes and the sophistication of the adapted technologies, soil and weather conditions, genotypic and phenotypic characteristics of crops and their interaction. Thus, it requires a fundamental understanding of the relationship between these interacting factors. In their study, M.R. Islam *et al.* (2023) proposed a ML-enabled IoT system for automated soil nutrient monitoring and crop recommendation, which directly supports yield forecasting through continuous data acquisition and real-time analysis. The authors integrated distributed soil sensors with data transmission modules and decision-support algorithms, forming a cyber-physical system capable of transforming raw sensor measurements into actionable management recommendations. Such an approach reflects the principles of automation and computer-integrated technologies, as it reduces manual intervention, enables closed-loop decision-making, and enhances the reliability of predictive models by incorporating dynamic soil and environmental parameters. The proposed system demonstrated how the integration of IoT-based measurement infrastructure with ML algorithms can improve the robustness and practical applicability of yield forecasting models in modern agricultural production systems.

The agricultural sector is the main consumer of available fresh water on a global scale, as plant growth and development largely depend on the availability of water. Given the rapid depletion of many aquifers, the use of digital technologies provides more efficient management of the processes of accumulating and preserving productive moisture to ensure sustainable crop production, which is also confirmed by research by scientists at Texas Tech University J. Neupane & W. Guo (2019). As a result of the research by W. Zhao *et al.* (2018) in the field of precision agriculture, variable rate irrigation is proposed to save water. This can be achieved by applying irrigation rates that vary with field variability based on the specific water needs of individual management zones instead of using a single rate for the entire field. The level of digitisation of agronomic factors, including terrain slopes, affects the effectiveness and feasibility of variable irrigation rates, the amount of productive moisture in the soil. S.A. Mauget *et al.* (2017) noted that careful monitoring of soil moisture and other conditions for crop growth and development, combined with weather forecasting, can help to develop digital irrigation programs and efficient water management, especially where groundwater sources are used for irrigation and precipitation provides only a portion of the total evapotranspiration of crops, as was studied in cotton cultivation in arid areas of West Texas. Accurate soil information by agro-climatic zonesis important because it facilitates better soil management according to the land's potential. Effective soil management is also of great interest in relation to problems such as land degradation (loss of biological productivity), soil nutrient imbalances (due to excessive fertilise use), and soil erosion (due to excessive vegetation cutting, improper crop rotation, and destruction by energy-driven vehicles). Useful soil properties can include texture, organic matter, and nutrient content, among others. Traditional methods of soil assessment include sampling and laboratory analysis, which are usually expensive and require significant time and effort. At the same time, remote sensing and soil mapping sensors can provide an inexpensive and simple solution to studying soil spatial variability. However, merging and processing such heterogeneous big data can be a serious drawback when using traditional analysis methods.

A serious obstacle to improving the economics of modern agriculture by obtaining valuable information is the limited processing of field data using traditional methods, especially against the backdrop of the ever-growing needs of precision (smart) agriculture, which requires the use of information and communication technologies. The basis for this is agricultural software, farm management information systems, various sensors, cameras, drones, low-cost satellites, etc. As a result, ML has emerged, where artificial intelligence and IoT tools are widely used. The way to solve the existing problems that put enormous pressure on the level of agricultural development is to increase the efficiency of agricultural technologies while reducing the environmental impact by transforming production in the agricultural sector into precision agriculture. Such a significant and timely modernisation has a great potential to ensure the sustainability of agricultural production while increasing the return on investment in its development, profitability and a safe environment.

The efficiency of complex processes of agricultural production is significantly affected by the introduction of digitalisation, which is a way to accumulate large amounts of field data and a tool for the development of new, modern areas of research. Based on a systematic analysis of the methods and features of digitising field research data for use as a basis for crop management and decision-making based on the development of realistic solutions that combine the expert knowledge of specialists in the field of computing and agriculture, a scientific and methodological approach to improving the efficiency of ML for automatic crop recognition, detection of diseases and weeds, and prediction of crop yield and quality is proposed.

CONCLUSIONS

A serious obstacle to improving the economy of modern agriculture by obtaining valuable information is the limited processing of field data using traditional methods, especially against the backdrop of the ever-growing needs of precision (smart) agriculture, which requires the use of information and communication technologies. The basis for this is agricultural software, farm management information systems, various sensors, cameras, drones, inexpensive satellites, etc. As a result, ML has emerged, where artificial intelligence and IoT tools are widely used.

The way to solve existing problems that put enormous pressure on the level of agricultural development is to increase the efficiency of agricultural technologies while reducing the impact on the environment by transforming production in the agricultural sector into smart agriculture and precision farming. Such a significant and timely modernisation has great potential to ensure the sustainability of agricultural production while increasing the return on investment in its development by 10-15%, profitability, and environmental safety. The efficiency of complex agricultural production processes is significantly influenced by the introduction of digitalisation, which is a way of accumulating large amounts of field data and a tool for developing new, modern areas of research. Based on the conducted systematic analysis, it was established that the integration of ML, sensor networks, and computer-integrated technologies forms a methodological basis for automating key decision-making processes in crop production.

Based on a systematic analysis of the methods and characteristics of digitisation of field research data for use as a basis for crop management and decision-making based on the development of realistic solutions that combine the expert knowledge of specialists in the field of computer technology and agriculture, a scientific and methodological approach is proposed to improve the effectiveness of ML for automatic crop recognition, disease and weed detection, and yield and crop quality prediction. Further research should focus on the development and validation of integrated cyber-physical systems that combine real-time sensor data, advanced ML models, and automated control algorithms to enhance the reliability and scalability of precision agriculture solutions.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Almeida, G.M.D., Pereira, G.T., Bahia, A.S.R.D.S., Fernandes, K., & Marques, J.J. (2021). Machine learning in the prediction of sugarcane production environments. *Computers and Electronics in Agriculture*, 190, article number 106452. doi: 10.1016/j.compag.2021.106452.

[2] Ang, K.L.-M., & Seng, J.K.P. (2021). Big data and machine learning with hyperspectral information in agriculture. *IEEE Access, 9,* 36699-36718. doi: 10.1109/ACCESS.2021.3051196.

[3] Borodin, V., Bourtembourg, J., Hnaien, F., & Labadie, N. (2016). Handling uncertainty in agricultural supply chain management: A state of the art. *European Journal of Operational Research,* 254(2), 348-359. doi: 10.1016/j.ejor.2016.03.057.

[4] Chang, W.L., & Grady, N. (2018). *NIST big data interoperability framework: Volume 1, big data definitions* (NIST special publication 1500-1). Gaithersburg, MD: National Institute of Standards and Technology. doi: 10.6028/NIST.SP.1500-1.

[5] Choi, R.Y., Coyner, A.S., Kalpathy-Cramer, J., Chiang, M.F., & Campbell, J.P. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science & Technology,* 9(2), article number 14. doi: 10.1167/tvst.9.2.14.

[6] Cravero, A., & Sepúlveda, S. (2021). Use and adaptations of machine learning in Big Data-applications in real cases in agriculture. *Electronics*, 10, article number 552. doi: 10.3390/electronics10050552.

[7] Evstatiev, B.I., & Gabrovska-Evstatieva, K.G. (2020). Review on the methods for big data analysis in agriculture. *IOP Conference Series: Materials Science and Engineering*, 1032, article number 012053. doi: 10.1088/1757-899X/1032/1/012053.
[8] Fan, X., Nan, Z., Ma, Y., Zhang, Y., & Han, F. (2021). Research on the spatio-temporal impacts of environmental factors on the fresh agricultural product supply chain and the spatial differentiation issue – an empirical research on 31 Chinese provinces. *International Journal of Environmental Research and Public Health,* 18(22), article number 12141. doi: 10.3390/ijerph182212141.
[9] Gao, X., Gao, J., & Qureshi, W.A. (2025). Applications, trends, and challenges of precision weed control technologies based on deep learning and machine vision. *Agronomy,* 15(8), article number 1954. doi: 10.3390/agronomy15081954.
[10] Hasan, M., Marjan, M.A., Uddin, M.P., Afjal, M.I., Kardy, S., Ma, S., & Nam, Y. (2023). Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation. *Frontiers in Plant Science,* 14, article number 1234555. doi: 10.3389/fpls.2023.1234555.
[11] Islam, M.R., Oliullah, K., Kabir, M.M., Alom, M., & Mridha, M.F. (2023). Machine learning enabled IoT system for soil nutrients monitoring and crop recommendation. *Journal of Agriculture and Food Research,* 14, article number 100880. doi: 10.1016/j.jafr.2023.100880.
[12] Jiang, L., Zeng, G., Wan, C., & Liu, Z. (2019). Research on the countermeasures for rural revitalization of the ethnic minorities in border areas: A case study of Heihe City in Heilongjiang Province. In *Proceedings of the 3rd international conference on culture, education and economic development of modern society (ICCESE 2019)* (pp. 657-660). Paris: Atlantis Press. doi: 10.2991/iccese-19.2019.144.
[13] Lampridi, M., Sørensen, C., & Bochtis, D. (2019). Agricultural sustainability: A review of concepts and methods. *Sustainability*, 11, article number 5120. doi: 10.3390/su11185120.
[14] Lazorenko, T., & Sholom, I. (2018). Digitalization as the main factor of business development. *Proceedings of the international scientific and practical conference* (pp. 50-51). Kyiv: Kyiv Polytechnic Institute.
[15] Li, D., Bai, L., Wang, R., & Ying, S. (2024). Research progress of machine learning in extending and regulating the shelf life of fruits and vegetables. *Foods,* 13(19), article number 3025. doi: 10.3390/foods13193025.
[16] Mauget, S.A., Adhikari, P., Leiker, G., Baumhardt, R.L., Thorp, K.R., & Ale, S. (2017). Modeling the impact of management and increase on cotton production in arid areas of West Texas. *Agricultural and Forest Meteorology,* 247, 385-398. doi: 10.1016/j.agrformet.2017.07.009.
[17] Mayuri, K.P. (2018). Role of image processing and machine learning techniques in disease recognition, diagnosis and yield prediction of crops: A review. *International Journal of Advanced Research in Computer Science,* 9(2), 788-795. doi: 10.26483/ijarcs.v9i2.5793.
[18] Neupane, J., & Guo, W. (2019). Agronomic bases and strategies of accurate water resources management: A review. *Agronomy,* 9(2), article number 87. doi: 10.3390/agronomy9020087.
[19] Ofosu, R., Agyemang, E.D., Márton, A., Pásztor, G., Taller, J., & Kazinczi, G. (2023). Herbicide resistance: Managing weeds in a changing world. *Agronomy,* 13, article number 1595. doi: 10.3390/agronomy13061595.
[20] Papageorgiou, E.I., Angelopoulou, K., Gemtos, T.A., & Nanos, G.D. (2018). Development and evaluation of a fuzzy inference system and a neuro-fuzzy inference system for grading apple quality. *Applied Artificial Intelligence,* 32(3), 253-280. doi: 10.1080/08839514.2018.1448072.
[21] Pounds, K., Bao, H., Luo, Y., De, J., Schneider, K., Correll, M., & Tong, Z. (2022). Real-time and rapid food quality monitoring using smart sensory films with image analysis and machine learning. *ACS Food Science & Technology,* 2(7), 1123-1134. doi: 10.1021/acsfoodscitech.2c00124.
[22] Rejeb, A., Rejeb, K., Zailani, S., Keogh, J.G., & Appolloni, A. (2022). Examining the interplay between artificial intelligence and the agri-food industry. *Artificial Intelligence in Agriculture,* 6, 111-128. doi: 10.1016/j.aiia.2022.08.002.
[23] Rudenko, M. (2019). The features of digitization influence on the functioning of agricultural enterprises. *Bulletin of the Kharkiv National Agricultural University of V.V. Dokuchaeva*, 1, 202-212. doi: 10.31359/2312-3427-2019-1-202.
[24] Sibo, X., Xinyuan, N., Xin, C., & Xumeng, L. (2022). Data fusion based wireless temperature monitoring system applied to intelligent greenhouse. *Computers and Electronics in Agriculture,* 192, article number 106576. doi: 10.1016/j.compag.2021.106576.
[25] Sonka, S. (2016). Big Data: Fueling the next evolution of agricultural innovation. *Journal of Innovation Management*, 4(1), 114-136. doi: 10.24840/2183-0606_004.001_0008.
[26] Sørensen, C.A.G., Kateris, D., & Bochtis, D. (2021). ICT innovation and smart agriculture. In M. Salampasis & T. Bournaris (Eds.), *Information and communication technologies in modern agricultural development* (vol. 953, pp. 1-19). Cham: Springer. doi: 10.1007/978-3-030-12998-9_1.
[27] Sun, A.Y., & Scanlon, B.R. (2019). How can Big Data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. *Environmental Research Letters,* 14(7), article number 073001. doi: 10.1088/1748-9326/ab1b7d.
[28] The Guardian. (2017). *Tsunami of data could consume one fifth of global electricity by 2025*. Retrieved from https://www.theguardian.com/environment/2017/dec/11/tsunami-of-data-could-consume-fifth-global-electricity-by-2025.
[29] Trushlyakova, A.B. (2018). Development of digitalization in Ukraine: Factors of influence, advantages and challenges of today. *Economies' Horizons,* 4(7), 186-191. doi: 10.31499/2616-5236.4(7).2018.212762.
[30] Upadhyaya, S.R., Danilevicz, M.F., Dolatabadian, A., Neik, T.X., Zhang, F., Al-Mamun, H.A., Bennamoun, M., Batley, J., & Edwards, D. (2024). Genomics-based plant disease resistance prediction using machine learning. *Plant Pathology,* 73(9), 2298-2309. doi: 10.1111/ppa.13988.

[31] Varga, V. (2020). Digitalization as one of the factors of enterprise competitiveness. *Efektyvna Ekonomika*, 8. doi: 10.32702/2307-2105-2020.8.154.
[32] Virnodkar, S., Pachghare, V., Patil, V., & Jha, S. (2020). Remote sensing and machine learning for crop water stress determination in various crops: A critical review. *Precision Agriculture,* 21, 1121-1155. doi: 10.1007/s11119-020-09711-9.
[33] Volokha, M., Bondar, S., Boldyrieva, L., & Kisilevskaya, M. (2022). Advantages of digitalization in field data processing and crop management. In *Modern research in world science* (pp. 414-420). Lviv: Scientific Publishing Center "Sci-conf.com.ua".
[34] Voloshchuk, Y. (2019). Digital directions of agrarian enterprises. *Efektyvna Ekonomika*, 2. doi: 10.32702/2307-2105-2019.2.66.
[35] Yuan, Y., Chen, L., Wu, H., & Li, L. (2022). Advanced agricultural disease image recognition technologies: A review. *Information Processing in Agriculture,* 9(1), 48-59. doi: 10.1016/j.inpa.2021.01.003.
[36] Zecca, F. (2019). The use of Internet of Things for the sustainability of the agricultural sector: The case of climate smart agriculture. *International Journal of Civil Engineering and Technology*, 10, 494-501.
[37] Zhang, J., Rao, Y., Husband, K., Jiang, Z., & Lee, S. (2021). Identification of cucumber leaf diseases using deep learning and small sample size for agricultural Internet of Things. *International Journal of Distributed Sensor Networks,* 17(4). doi: 10.1177/15501477211007407.
[38] Zhao, W., Li, J., Yang, R., & Li, Y. (2018). Determining placement criteria of moisture sensors through temporal stability analysis of soil water contents for a variable rate irrigation system. *Precision Agriculture,* 19(4), 648-665. doi: 10.1007/s11119-017-9545-2.
[39] Zou, Y., Wu, J., Wang, X., Morales, K., Liu, G., & Manzardo, A. (2023). An improved artificial neural network using multi-source data to estimate food temperature during multi-temperature delivery. *Journal of Food Engineering,* 351, article number 111518. doi: 10.1016/j.jfoodeng.2023.111518.

Станіслав Ніколаєнко

Доктор педагогічних наук, професор, академік Національної академії аграрних наук України
Національний університет біоресурсів і природокористування України
03041, вул. Героїв Оборони, 15, м. Київ, Україна
https://orcid.org/0000-0001-5271-1777

Микола Волоха

Доктор технічних наук, професор
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
03056, просп. Берестейський, 37, м. Київ, Україна
https://orcid.org/0000-0002-0112-7324

Оцифрування даних польових досліджень – основа управління фермерським господарством

Анотація. Ефективність управління сучасним сільськогосподарським виробництвом була визначена як така, що значною мірою залежить від впровадження процесів діджиталізації, які змінюють парадигму управління виробничими процесами та створюють умови для підвищення продуктивності аграрного виробництва шляхом використання інноваційних цифрових технологій. Метою статті було здійснення огляду наукових публікацій, присвячених оцифруванню баз даних польових досліджень, з метою підвищення прибутковості аграрного виробництва шляхом розроблення ефективних управлінських рішень у рослинництві на основі використання даних, що надходять із численних джерел цифрової трансформації. Було показано, що різноманіття технологічних процесів, а також вхідних і вихідних даних, задіяних у сільськогосподарському виробництві, зумовлює складність та комплексність узагальнення процесів оцінювання його сталості. Визначено, що значна частина фермерів не є фахівцями з інформаційних технологій і не має можливості повною мірою розуміти алгоритмічну складність сучасних цифрових рішень. У роботі було систематизовано підходи до використання великих даних у сільському господарстві з урахуванням еволюції концепції 4V до 5V, де додатково враховується цінність даних. Запропоновано науково-методологічний підхід, спрямований на підвищення ефективності сільськогосподарського виробництва шляхом застосування методів машинного навчання для автоматичного розпізнавання культур, виявлення хвороб і бур'янів, прогнозування врожайності та якості сільськогосподарської продукції. Встановлено, що поряд із матеріально-технічним забезпеченням першочерговим чинником розвитку агропромислового комплексу є інтелектуалізація виробничої та управлінської діяльності на основі цифровізації, що розглядалась як високоефективний інструмент підвищення рентабельності інвестицій у сільське господарство на 10-15 %

Ключові слова: ефективність сільськогосподарського виробництва; діджиталізація; врожайність; родючість ґрунту; запаси вологи

Machinery & Energetics
Vol. 16, No. 4. 2025

Journal homepage:
https://technicalscience.com.ua/en

UDC 519.6:621.9.015:536.24
DOI: 10.31548/machinery/4.2025.53

Abdelhak Saoudi*
PhD Student
Laboratory of Engineering of Sciences of Advanced Materials – ISMA
Department of Mechanical Engineering, ABBES Laghrour University
40004, P.O. Box 1252, Batna Rd., Khenchela, Algeria
https://orcid.org/0000-0002-8295-1947

Belkacem Meddour
Professor
Laboratory of Engineering of Sciences of Advanced Materials – ISMA
Department of Mechanical Engineering, ABBES Laghrour University
40004, P.O. Box 1252, Batna Rd., Khenchela, Algeria
https://orcid.org/0000-0002-4751-2615

Badis Haddag
Associate Professor
LEM3 Laboratory, UMR CNRS 7239, Lorraine University
Institut Mines-Telecom (InSIC)
88100, 27 Voie de l'Innovation, Saint-Dié, France
https://orcid.org/0000-0003-3850-3044

Mohammed Salah Aggoune
PhD in Materials Engineering
Faculty of Science and Technology, Department of Mechanical Engineering
Echahid Cheikh Larbi Tebessi University
12000 Tebessa, Algeria
https://orcid.org/0000-0002-3385-7081

Abdelaziz Abboudi
Professor
Laboratory of Engineering of Sciences of Advanced Materials – ISMA
Department of Mechanical Engineering, ABBES Laghrour University
40004, P.O. Box 1252, Batna Rd., Khenchela, Algeria
https://orcid.org/0000-0002-1398-6070

Modelling and studying of thermal behaviour of thin layer using finite element method – application to coated tools

Abstract. The continuous evolution of thin-layer technologies is essential in modern dry machining, where effective thermal control of coated cutting tools is critical for limiting heat-induced wear and extending tool life. The purpose of this article was to investigate and model the thermal response of coated cutting tools and to evaluate the influence of single-layer and multilayer thin coatings on heat transfer during orthogonal dry cutting. To achieve this objective, a transient thermal model based on the finite element method was developed using a Galerkin formulation combined with an implicit Crank-Nicolson time integration scheme. The numerical framework was applied to three tool configurations: an uncoated WC-Co insert, a TiN-coated insert, and a TiN/Al_2O_3 double-coated insert. Experimental dry turning tests were conducted to validate the numerical predictions. The results showed that the numerical model accurately reproduced the temporal evolution of temperature at the tool-chip interface

Article's History: Received: 08.08.2025; Revised: 04.11.2025; Accepted: 27.11.2025.

Suggested Citation:
Saoudi, A., Meddour, B., Haddag, B., Aggoune, M.S., & Abboudi, A. (2025). Modelling and studying of thermal behaviour of thin layer using finite element method – application to coated tools. *Machinery & Energetics,* 16(4), 53-63. doi: 10.31548/machinery/4.2025.53.

*Corresponding author

and within the insert. The uncoated tool exhibited rapid heat penetration and pronounced internal thermal gradients. The TiN coating slightly increased the surface temperature but offered limited resistance to heat diffusion. In contrast, the TiN/Al_2O_3 multilayer configuration significantly restricted heat transfer towards the substrate, leading to lower internal temperatures and reduced thermal gradients. The Al_2O_3 layer acted as an effective thermal barrier, confining heat near the cutting edge and protecting the tool core from excessive thermal loading. These findings demonstrated that multilayer ceramic coatings provided superior thermal protection compared with single-layer coatings. The proposed finite element method approach could be effectively applied to coating design, tool optimisation, and thermal management in dry machining applications

Keywords: dry machining; heat transfer; numerical approach; cutting; temporal

INTRODUCTION

In industrial manufacturing, the optimisation of material systems and structural configurations is fundamental to ensuring reliability, service life, and overall process efficiency. Advances in thin-film engineering have enabled the deployment of high performance coatings designed to strengthen functional surfaces operating under severe thermomechanical loading. Within this context, hard protective coatings for cutting tools have become a critical element in modern machining. They increase productivity, extend tool life, and enable environmentally responsible dry or near dry cutting regimes. During metal cutting, extreme frictional contact and high interfacial temperatures develop at the tool chip interface, accelerating wear mechanisms and diminishing machining stability. To alleviate these effects, cutting inserts are commonly coated via physical vapour deposition (PVD) or chemical vapour deposition (CVD), two established techniques capable of forming dense, adherent, and mechanically robust layers engineered to resist thermal softening, oxidation, and mechanical abrasion. These coatings have been widely investigated, particularly within the broader framework of sustainable machining, where the control of heat generation and thermal gradients is essential for maintaining tool integrity under elevated cutting conditions.

Advances in modelling thin layer thermal behaviour have broadened numerical capabilities. A. Ifis *et al.* (2013) proposed the MAX-FEM (Matching Asymptotic eXtended Finite Element Method) approach, which eliminates the need for mesh refinement near thin interlayers by deriving correction matrices from matching asymptotic expansions (MAE) and integrating concepts from the Extended Finite Element Method (X-FEM). Implemented in Matlab, the model delivered significant reductions in computational cost relative to classical Finite Element Method (FEM) formulations. The Thermal Thin Shell Approximation (TSA), introduced by E. Schnaubelt *et al.* (2023), reformulates ultra-thin insulating layers into reduced order 2D surface elements while preserving accurate in plane and through-thickness heat flux descriptions. Validated across multiple configurations, including High-Temperature Superconducting (HTS) coils, TSA offers substantial reductions in degrees of freedom and computational overhead, making it particularly effective in systems where the insulation to conductor thickness ratio is extremely small.

Additional contributions reinforce the relevance of advanced thermal modelling in manufacturing. M. Ala *et al.* (2025) developed a 3D finite element thermal model to predict laser cladding bead geometry, integrating heat transfer, fluid flow, and melt pool dynamics within COMSOL. Validated by 18 experiments on Inconel 625, the model achieved errors below 13%, with results underscoring the critical role of Marangoni convection in melt pool morphology and additive manufacturing quality. D.H. Nguyen *et al.* (2008) extended the application of X-FEM by treating brazed joints as singularities within a matching asymptotic development (DAR) framework, enabling accurate resolution of both mechanical and transient thermal behaviours and achieving strong agreement with analytic and commercial FEM solutions.

G. Kumaran *et al.* (2021) highlighted the importance of multi physics models for analysing coupled thermal and mass transport phenomena in nanostructured coatings, reinforcing the strategic role of numerical simulation in surface engineering optimisation. Researchers I. Bencheikh *et al.* (2017) introduced a multiscale X-FEM/level-set methodology for coated cutting tools, achieving improved convergence characteristics and reduced computational cost in transient thermomechanical analyses, confirming the suitability of this framework for thin film coatings, brazed joints, and adhesively bonded interfaces. In a complementary study C Upadhyay *et al*. (2025) investigated the effect of TiAlN coating thickness (1-4 μm) during dry milling of AISI 1045. Their FEM simulations accurately predicted cutting forces, temperature rise, and crater wear progression, and experimental validation confirmed that a 3 μm TiAlN layer provided optimal thermomechanical performance through reduced force levels, attenuated peak temperatures, and enhanced surface finish.

Within this broader scientific landscape, FEM remains a powerful framework for analysing temperature fields in cutting tools because it accommodates multilayer coating architectures, complex contact interactions, and dynamic boundary conditions. The aim of the study was to analyse the temporal evolution and spatial distribution of temperature in coated and uncoated cutting inserts under dry orthogonal cutting conditions using the finite element method.

MATERIALS AND METHODS

Heat generation during machining mainly results from two mechanisms: severe plastic deformation of the workpiece material and intense friction at the tool chip interface. The heat produced was partially transferred to the cutting insert, where the temperature distribution depends heavily on the thermal conductivity of each layer of the coating system. In this analysis, the thermal conductivities adopted for the constituent materials were 59, 30, 7.5, and 25 $Wm^{-1}K^{-1}$, for WC-Co, TiCN, Al_2O_3, and TiN, respectively. Figure 1 illustrates the generic domain Ω and its boundaries Γ_1 and Γ_2, in accordance with the configuration used in previous research by F. Kone *et al.* (2011).

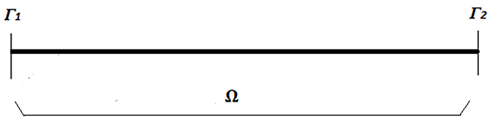


Figure 1. Domain and boundaries
Note: domain Ω and its boundaries Γ_1 and Γ_2
Source: developed by authors on the base on F. Kone *et al.* (2011)

In this general case, the transient heat conduction problem for the temperature field T(x ,t) was expressed in its strong form as follows:

$$\begin{cases} \rho c\frac{\partial T}{\partial t} - \mathrm{div}(\lambda\vec{\nabla}T) = \dot{\omega} \text{ in } \Omega \times]0,t[\\ T(x,0) = T_0 \ \forall x \in \ \Omega \\ \lambda\vec{\nabla}T.\vec{n} = \varphi_{imp} \text{ at } \Gamma_2 \times]0,t[\,, \end{cases} \quad (1)$$

where ρ – the mass density, c the specific heat capacity, – the internal heat generation, λ – the thermal conductivity, and φ_{imp} – the imposed heat flux density.

The corresponding weak formulation was written as follows:

$$\int_\Omega \rho c\frac{\partial T}{\partial t}\vartheta d\Omega + \int_\Omega \lambda\vec{\nabla}T.\vec{\nabla}\vartheta d\Omega = \int_\Omega \dot{\omega}\vartheta d\Omega + \int_{\Gamma_2}\varphi_{imp}\vartheta d\Gamma, \quad (2)$$

where, ϑ represented the test function.

To approximate the solution numerically, the domain Ω was discretised into Ne finite elements connected through Nn nodes. The temperature field was approximated using Equation (3):

$$T(x) = \sum_{i=1}^{N_n} N_i(x)T_i, \quad (3)$$

where $N_i(x)$ – the shape functions and T_i – the nodal temperatures.

Substituting (3) into the weak formulation (2) and removing the test functions yielded the classical finite element matrix representation:

$$[C]\left\{\frac{\partial T}{\partial t}\right\} + [K]\{T\} = \{Q\}, \quad (4)$$

where, {T} – the nodal temperature, [C] – the heat capacity matrix, [K] – the conductivity matrix, and {Q} – the thermal loading vector.

The components of these matrices and vectors were calculated as follows:

$$[K] = \sum_{e=1}^{N_e}\int_{\Omega_e}[B]^T[D][B]d\Omega \Leftrightarrow$$
$$\Leftrightarrow K_{ij} = \sum_{e=1}^{N_e}\int_{\Omega_e}\lambda\vec{\nabla}N_i\,\vec{\nabla}N_j d\Omega\,, \quad (5)$$

$$[C] = \sum_{e=1}^{N_e}\int_{\Omega_e}\rho c[N]^T[N]d\Omega \Leftrightarrow C_{ij} = \sum_{e=1}^{N_e}\int_{\Omega_e}\rho c N_i N_j d\Omega, \quad (6)$$

$$\{Q\} = \sum_{e=1}^{N_e}\left(\int_{\Omega_e}\dot{\omega}[N]^T d\Omega - \int_{\Gamma_e^2}\varphi_{imp}[N]^T d\Gamma\right) \Leftrightarrow$$
$$\Leftrightarrow Q_i = \sum_{e=1}^{N_e}\left(\dot{\omega}N_i d\Omega - \int_{\Gamma_e^2}\varphi_{imp}N_i d\Gamma\right). \quad (7)$$

For time integration, the Crank-Nicolson two step implicit scheme was adopted due to its stability and second order accuracy. Applying this method to (4) led to the recurrence relation:

$$\left(\frac{[C]}{\Delta t} + \frac{1}{2}[K]\right)\{T\}^{m+1} = \left(\frac{[C]}{\Delta t} + \frac{1}{2}[K]\right)\{T\}^m +$$
$$+\frac{1}{2}\{Q\}^m + \frac{1}{2}\{Q\}^{m+1}, \quad (8)$$

which enabled computation of $\{T\}^{m+1}$ from the known "assumed" state $\{T\}^m$. The time step Δt had to satisfy the stability requirement (9) to ensure the stability of the resolution method:

$$\frac{1}{6} \le \alpha\frac{\Delta t}{(\Delta x)^2} \le \frac{1}{2}, \quad (9)$$

with α – the thermal diffusivity given by , and Δx – the mesh spacing in the region of highest thermal gradient.

The three cutting inserted configurations: uncoated WC-6%Co, single layer TiN, and double layer TiN+Al_2O_3 were modelled as 1D multilayer systems (Fig. 2). Each layer was discretised according to stability constraints associated with the Crank-Nicolson scheme. Table 1 summarises the mesh density for each configuration.

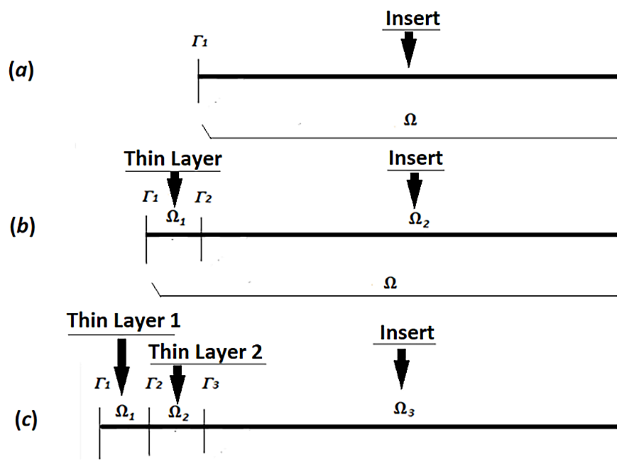


Figure 2. Domains, subdomains and boundaries
Note: a – uncoated carbide insert, b – coated carbide insert (TiN coating), c – double coated insert (coatings TiN+Al_2O_3)
Source: developed by authors

The machining conditions used to define the applied heat flux were derived from experimental measurements. In all simulations, the workpiece bottom surface moves at a cutting speed of 100 m/min, while the insert remains stationary. The undeformed chip thickness is maintained at 0.4 mm, with a clearance angle of 6°. The effective cutting length is 10 mm, corresponding to approximately 6 ms of machining time. A constant friction coefficient of 0.4 is adopted for all cases (Table 2).

Table 1. Meshing data

Case	Bodies	Number of elements
Uncoated	Insert	288
Coated insert	Coating	2
	Insert	397
Double coated insert	Coating 1	2
	Coating 2	2
	Insert	395

Note: number of elements
Source: developed by authors

Table 2. Cutting conditions

Parameter	Value
Cutting speed (m/min)	100
Undeformed chip thickness (mm)	0.4
Length of cut (mm)	10
Coefficient of friction	0.4
Cutting time (ms)	6

Note: cutting experimental conditions
Source: developed by authors

The local frictional heat flux in the contact zone was expressed as:

$$q_f = \tau s = \mu p s, \quad (10)$$

where μ – the local friction coefficient, τ – the shear stress, p – the contact pressure, and s – the sliding velocity. The resulting geometry, loading and boundary conditions of the cutting system were shown in Figure 3. Table 3 provides the thermal properties and dimensions for the substrate and coating layers employed in this study.

RESULTS AND DISCUSSION

As a first step in this study, it is important to verify the validity of the FEM numerical model that has been developed. To achieve this, the operating parameters were applied in the algorithm, and the resulting data were quantitatively compared with experimental measurements. For all analysed cases, the simulation begins at a temperature of 20°C, corresponding to the prescribed initial thermal condition $T = T_0 = 20$°C. The numerical evaluation of the tool surface temperature in the case of the uncoated insert (WC-Co) as presented in Figure 4, shows close agreement with the experimental measurements.

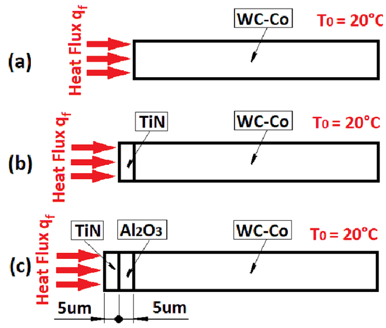


Figure 3. Geometry, dimensions, loading and boundary conditions of cutting tools and coating layers
Note: a – uncoated carbide insert, b – coated carbide insert (TiN coating), c – double coated insert (coatings TiN + Al_2O_3)
Source: developed by authors

Table 3. Thermal properties and dimensions

	Substrate	Coating	
	WC-Co	TiN	Al_2O_3
Conductivity ($W\,m^{-1}\,C^{-1}$)	59	25	7.5
ρCp ($Nmm^{-2}\,K^{-1}$)	15	12	3.42
Size (mm)	2	0.005	0.005

Note: thermal properties for numerical analysis
Source: developed by authors

Both curves show a rapid rise in temperature during the first 1 ms, reflecting intense heat generation on the cutting surface, followed by a slower increase. The experimental values slightly exceed the numerical predictions at most points, especially between 1 and 4 ms, where the maximum difference reaches about 50°C. After 4 ms, the curves converge towards a peak of about 456°C at 6 ms. In the case of the TiN coated insert (Fig. 5), the numerical evaluation of the tool surface temperature shows a comparable behavior to the experimental results, with a sharp rise during the first 1 ms due to the concentrated heat at the TiN layer, followed by a slower rise.

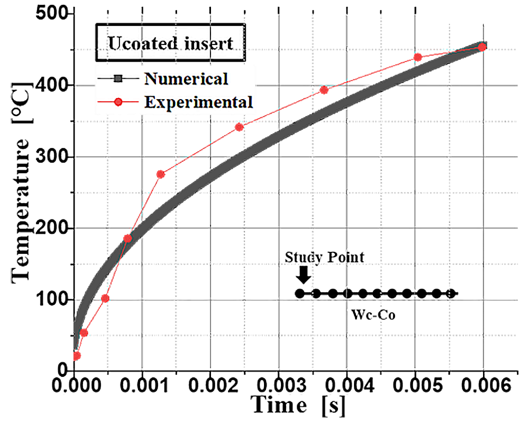


Figure 4. Temperature evolution at the tool surface of contact (Point study). Case of uncoated insert
Note: thermal properties (WC-Co) for numerical and cutting conditions for experimental
Source: developed by authors on the validation on research F. Kone *et al.* (2011)

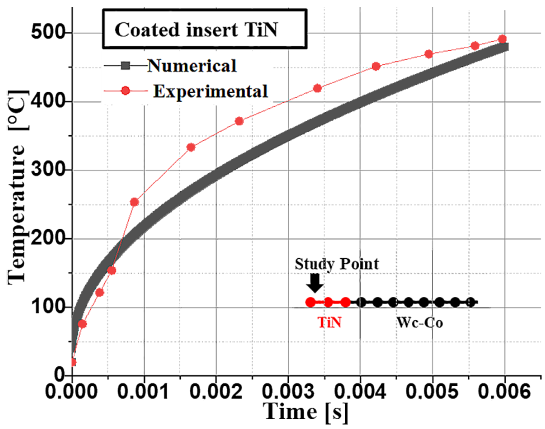


Figure 5. Temperature evolution at the tool surface of contact (Point study). Case of coated insert (TiN)
Note: thermal properties (WC-Co, TiN) for numerical and cutting conditions for experimental
Source: developed by authors on the validation on research F. Kone *et al.* (2011)

The experimental temperatures remain consistently higher than the numerical predictions, especially between 1 ms and 4 ms, where the maximum deviation is close to 60°C. After 4 ms, both curves gradually converge towards approximately 480°C at 6 ms. The double coated insert (TiN+Al_2O_3) as shows Figure 6 follows the numerical prediction of temperature evolution in the same general trend as the experimental results, with an initially sharp increase during the first 1 ms due to intense heat generation on the TiN layer. The experimental values remain higher than the numerical values, especially between 1 ms and 5 ms, where the maximum deviation reaches approximately 70°C, reflecting the added thermal resistance of the multilayer coat (TiN+Al_2O_3). After 5 ms, both curves gradually converge towards approximately 544°C at 6 ms. The slight temperature differences observed in all cases (Figs 4-6) tend to strengthen the credibility of the FEM model rather than weaken it. Real cutting tools, whether coated or uncoated, always have surface roughness, pores, and a heterogeneous grain structure, all of which increase friction and lead to local variations in heat generation. In contrast, our numerical model assumes a completely homogeneous material. Due to these natural industrial imperfections, it is to be expected that the experimental temperatures will be slightly higher. The fact that the FEM predictions are very close to the measurements despite these complexities proves that the model is accurate and reliable.

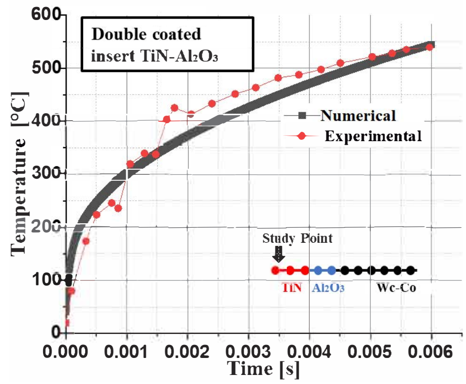


Figure 6. Temperature evolution at the tool surface of contact. Case of double coated insert ($TiN + Al_2O_3$)
Note: thermal properties (WC-Co, TiN and Al_2O_3) for numerical and cutting conditions for experimental
Source: developed by authors on the validation on research F. Kone *et al.* (2011)

With the finite element model's fidelity established, the model was extended to interrogate the internal thermal behaviour of the insert. Figures 7, 8 and 9 depict the simulated temperature evolution at critical nodes A, B, C, and D for each cutting condition, providing a quantitative analysis for interpreting temperature redistribution imposed by insert configuration rather uncoated, single layer coated, or double layer coated.

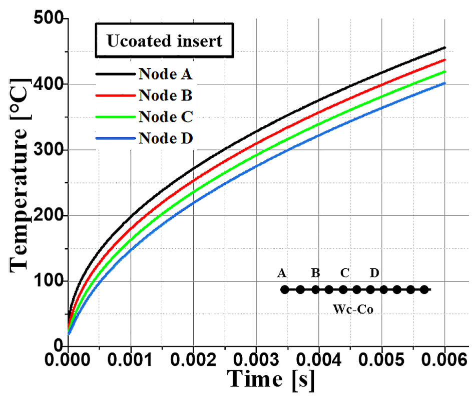


Figure 7. Temperature evolution at nodes. case of uncoated insert (WC-Co)
Note: node A: tool/chip interface, node B, node C and node D: WC-Co insert
Source: developed by authors

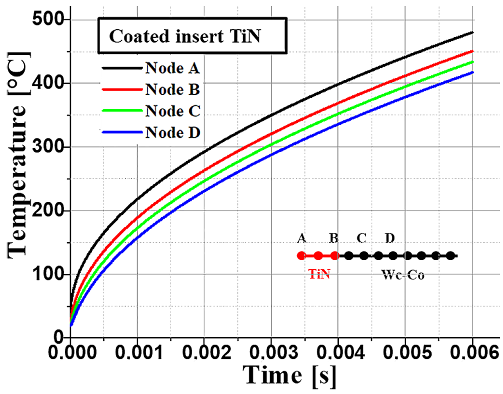


Figure 8. Temperature evolution at nodes. case of coated insert (TiN)
Note: node A: tool/chip interface, node B: TiN/WC-Co interface, node C and node D: WC-Co insert
Source: developed by authors

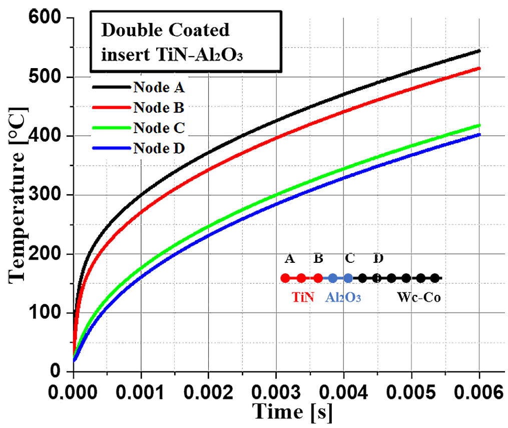


Figure 9. Temperature evolution at nodes. case of double coated insert (TiN + Al_2O_3)
Note: node A: tool/chip interface, node B: TiN/Al_2O_3 interface, node C: Al_2O_3/WC-Co interface, node D: WC-Co insert
Source: developed by authors

The numerical evaluation of the internal temperature distribution for the uncoated insert (WC-Co), shown in Figure 7, exhibits a clear thermal gradient along the nodes A to D. All four curves display a rapid temperature rise during the first 1 ms of cutting, reflecting the intense heat conducted directly into the WC-Co insert. Node A, being at the cutting edge (point of contact), reaches the highest temperature, followed in order by nodes B, C, and D, with a maximum difference of nearly 70°C between the extreme positions (A and D) at 6 ms. The temperature increase becomes progressively slower after the initial transient as the thermal field begin to stabilize, ultimately reaching about 456°C at node A. This behaviour confirms the absence of any thermal barrier, resulting in deep heat penetration and pronounced gradients inside the insert.

In the case of the TiN single layer coated insert (Fig. 8), the numerical prediction of temperature at the selected nodes shows a similar overall trend but with noticeably increased temperatures at all positions compered to Figure 7. As in the WC-Co uncoated case, node A exhibits the highest temperature, with a gradual decrease toward node D. The initial heat rises during the first 1 ms remains dominant due to the concentration of heat at the tool chip interface; however, the presence of the TiN layer results in higher heat transfer into the WC-Co insert, this can be seen by comparing temperature values of nodes B, C, and D in both Figures 7 and 8. By 6 ms, node A approaches roughly 480°C, while the deeper nodes remain significantly cooler.

For the double layer coated insert (Fig. 9), the predicted temperature evolution follows the same general trend as in the previous cases but with the lowest overall temperatures across the WC-Co insert. The presence of the TiN + Al_2O_3 double configuration introduces a stronger thermal resistance, resulting in an intensive and aggressive temperature rise at node A and B. During the first 1 ms, heat generation at the cutting edge (point of contact) still dominates, but the presence of the Al_2O_3 layer (from node B to node C) substantially restricts thermal diffusion into the underlying insert (WC-Co) which explain the 100°C temperature gap observed almost at all cutting duration. At 6 ms, node A reaches approximately 544°C, while the inner nodes (C and D) remain significantly cooler, with the smallest temperature spread among the three configurations (A and B, B and C, C and D). This confirms the enhanced insulating effect of the multilayer coating, which effectively limits internal heating and reduces thermal stresses within the tool.

The comparison between the three cases (Figs. 7-9) highlights the thermal management provided by the coatings. The uncoated WC-Co insert shows rapid heat transfer with a clear gradient (ΔT ≈ 70°C between nodes A and D) due to the absence of any barrier, resulting in deep internal heating. The single layer TiN coating does not mitigate this effect, but rather accelerates the process of heat transfer to the core of the WC-Co insert. In exchange, the double layer TiN + Al_2O_3 coating significantly enhances thermal resistance: While the cutting edge (nodes A and B) is still subjected to intense initial heating, the Al_2O_3 layer effectively limits heat diffusion to the inner nodes C and D, resulting in less heat diffusion and maintaining lower temperatures in the WC-Co insert. This demonstrates that the double coating not only enhances surface insulation, but also stabilizes the internal thermal distribution, reducing thermal stresses and protecting the core of the insert.

Plotting the temperature distribution profiles along the insert thickness reveals a very interesting thermal behaviour that clearly distinguishes between the three configurations (Fig. 10). The near surface region (0-0.1 mm) shows a very sharp gradient, confirming that heat remains highly concentrated at the interface between the tool and the chips. The uncoated insert shows the lowest surface temperature reaching about 450°C but has the smoothest gradient, indicating rapid heat transfer to the WC-Co core. The TiN single layer coating increases the surface temperature from 456°C to 480°C and makes the gradient steeper, indicating reduced thermal diffusion. The most pronounced effect is seen with the TiN+Al_2O_3 double coating, which produces the highest surface temperature of 544°C and the steepest gradient, confirming strong thermal insulation and minimal heat penetration outside the coated area. This characteristic is strongly interpreted by the low thermos conductivity and diffusivity of the Al_2O_3 layer, as shown in the by reference (Grzesik *et al.*, 2005). After about 0.1 mm, all curves converge, indicating that the multilayer system effectively protects the substrate by confining heat to the surface.

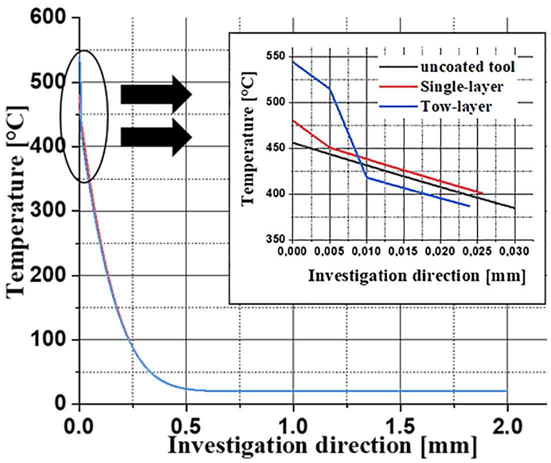


Figure 10. Detailing of temperature distribution inside the insert
Note: for all three cases t = 6 milliseconds
Source: developed by authors

The findings of this investigation, which utilised Finite Element Modelling (FEM) to analyse heat transfer and temperature evolution in coated tools during orthogonal dry cutting, are largely consistent with previous research on the thermal behaviour of coated tools. However, the results also reveal important distinctions that refine current understanding. The numerical simulations, validated by experimental measurements, offer a detailed depiction of temperature distribution at the tool chip interface and complement earlier analytical and numerical approaches reported in the literature. J. Zhang *et al.* (2022) demonstrated that the thermoconductivity and diffusivity of coating materials govern steady and transient temperature distributions, respectively, during cutting. The present study corroborates this finding by confirming the significant thermal barrier effect of Al_2O_3, whose low thermal conductivity markedly reduces heat penetration into the tool substrate. In contrast, TiN, as previously noted by J. Zhang *et al.*, provides limited thermal insulation. The current results refine this observation by quantifying the minimal thermal influence of TiN within an experimentally validated FEM framework, thereby offering stronger evidence for its limited role in thermal management during machining.

Similarly, the results of T.D. Ipilakyaa *et al.* (2024), who employed an equivalent layer methodology for multilayer coatings, emphasised the crucial role of coating thickness, layering sequence, and material thermal properties in the evolution of cutting temperature. Although the present study did not construct equivalent thermal layers but instead directly modelled physical coating architectures, both studies converge in showing that multilayer systems perform more effectively when a ceramic outer layer is used. The agreement further supports the suitability of analytical and numerical tools for predicting coating temperature, with the current FEM results achieving validated accuracy within the 10% error margin reported for their analytical models.

The numerical work of W. Grzesik *et al.* (2005) highlighted the ability of finite difference based simulations to capture coating-induced temperature field variations along the tool-chip interface, revealing strong dependencies on cutting speed and frictional conditions during orthogonal cutting. The present FEM investigation reaches comparable conclusions, showing that thermal contours strongly reflect coating composition and interface behaviour. However, by incorporating experimentally calibrated conditions and quantifying temperature evolution near the secondary shear zone with higher spatial resolution, this study extends the earlier numerical findings and provides a more comprehensive representation of heat transfer mechanisms occurring immediately upstream of the tool rake face, where direct measurement remains challenging. Previous studies by R. Komanduri & Z.B. Hou (2001) have shown that temperature fields within cutting tools and heat flux distributions along the rake face can be analysed using experimental measurements, analytical approaches, and numerical modelling. In numerical investigations by C.A. Van Luttervelt *et al.* (1998) and I. Lazoglu & Y. Altintas (2002), the heat transfer problem in metal cutting has been addressed using various computational methods, including the finite element method, the finite difference method, and the boundary element method, each with specific capabilities for resolving conductive, convective, and friction-induced heating phenomena. In the context of the present results, the finite element approach proved particularly effective for capturing the influence of coating architecture on thermal gradients and heat distribution near the tool-chip interface. Despite the extensive use of FEM in machining simulations, the present findings further highlight that the role of coating architecture, especially in the vicinity of the secondary shear zone, remains a critical factor governing internal heat transfer mechanisms and cannot be fully resolved without experimentally validated numerical models.

The contribution of tool coatings to mechanical stability has also been emphasised in previous research. P. Sahoo *et al.* (2020) demonstrated that coatings such as TiAlN enhance micro-milling stability by reducing tool wear and improving force modelling reliability. While the present study focused on temperature rather than force prediction or dynamic stability, the observed thermal benefits of Al_2O_3 coatings indirectly support the broader claims of stability enhancement. Lower temperatures are strongly associated with reduced wear and improved cutting performance, suggesting that the thermal barrier effect quantified in current study can be translated to a mechanical advantage consistent with the findings of P. Sahoo *et al*. The work by C.S. Kumar *et al.* (2020), which examined TiAlCrN coatings through FEM integrated with experimental validation, highlighted the importance of coating layers in reducing friction, modifying chip formation, and restricting temperature rise. This aligns with the present study's demonstration that multilayer systems incorporating ceramic layers outperform single layer coatings in limiting heat diffusion. However, unlike C.S. Kumar *et al.*, who reported coating dependent changes in chip morphology (such as secondary serrated teeth), the current study did not specifically address chip formation mechanisms. Instead, it provides new insight into the temperature evolution around the secondary shear zone, offering complementary thermal evidence that may explain the coating dependent chip behaviour observed in previous research.

Previous thermal analyses of coated cutting tools by R.F. Brito *et al.* (2009) have shown that ceramic coatings significantly influence temperature distribution within the tool by restricting heat transfer toward the substrate. These findings are consistent with the present results, where the TiN/Al_2O_3 multilayer configuration exhibited reduced internal temperatures and lower thermal gradients compared with single-layer coatings. The thermal behaviour observed in the present study was closely related to the fundamental mechanisms governing heat generation during cutting. Classical cutting theory, as described by P. Albrecht (1960), emphasised the contribution of ploughing and frictional interactions at the tool-chip interface, which became dominant under conditions of high contact pressure and limited lubrication. These mechanisms directly explained the rapid temperature rise and pronounced thermal gradients observed in the uncoated WC-Co insert, where no thermal barrier was present to restrict heat penetration. The transient nature of temperature evolution obtained in this work was consistent with the findings of P.A.R. Rosa *et al.* (2007), who demonstrated that the initial phase of machining was characterised by a rapid increase in thermal and mechanical loads before reaching a quasi-steady state. In the present results, all tool configurations reached their maximum surface temperature at approximately 6 ms, confirming that the early transient regime played a critical role in defining the overall thermal response of coated and uncoated tools. The differences observed between single-layer and multilayer coatings during this transient phase further highlighted the influence of coating architecture on heat accumulation near the cutting edge. Although S.M. Afazov *et al.* (2010) primarily focused on force prediction in micro-milling, their numerical framework underscored the strong coupling between mechanical loading, frictional behaviour, and thermal effects in cutting processes. The present FEM results extended this perspective by demonstrating that coating architecture not only affected mechanical interaction but also governed the redistribution of heat within the tool. In particular, the reduced internal temperatures observed for the TiN/Al_2O_3 multilayer configuration indicated that ceramic layers effectively decoupled surface heat generation from substrate thermal loading, thereby mitigating internal thermal gradients. Overall, the agreement between classical cutting theory, transient machining analyses, and the present FEM-based thermal results supported the conclusion that accurate thermal modelling had to account for both transient effects and coating architecture. The present study reinforced the necessity of experimentally validated numerical models to capture these coupled phenomena and to guide the design of advanced coating systems for dry machining applications.

A key contribution of the present study lies in its detailed evaluation of temperature evolution near, though not directly at, the tool rake face, which experimental tools often fail to capture accurately. The integration of experimentally validated FEM modelling therefore bridges a persistent measurement gap in machining research. The results notably highlight the superior thermal resistance of Al_2O_3, validating its role as an efficient thermal barrier layer. Conversely, they confirm the weak thermal protection provided by TiN, reinforcing long standing hypotheses with new quantitative data. Another essential contribution is the recognition of cutting condition variability as a significant limitation affecting the predictive accuracy of thermal models. This point has been insufficiently addressed in many analytical studies. The current research also confirms and extends findings from earlier studies by offering validated FEM-based insights into the heat transfer behaviour of coated cutting tools. Through a direct comparison of CVD layers and a double coated configuration, the study provides refined quantitative evidence of the thermal advantages of ceramic based coatings. It clarifies the limitations of metallic coatings, such as TiN. By emphasising temperature development near the secondary shear zone and addressing modelling uncertainties arising from cutting condition variability, this work deepens the understanding of coating performance in dry machining. It establishes a robust foundation for future investigations into thermally driven tool degradation, coating optimisation, and microstructural behaviour under high thermal loads.

CONCLUSIONS

In this study, FEM was employed through numerical and theoretical analysis to examine the performance of two distinct coating layers under dry orthogonal cutting conditions. The analysis considered two commercially available CVD coated tools and a double coated tool. The numerical approach, after having been validated by comparing to experimental data, provided valuable insights into the thermal field and its temporal changes in the vicinity of the secondary shear zone, which could complement experimental techniques aimed at measuring temperatures near, but not directly on, the tool rake face. The findings of the present study emphasise the high efficiency of the coating system in controlling heat transfer during dry machining, with all tool configurations reaching their maximum surface temperature at approximately t ≈ 6 ms. From the combined numerical and experimental analysis of the temperature evolution, it may be derived that the uncoated WC-Co insert exhibits rapid heat penetration, reaching a peak surface temperature of about 456°C. The TiN-coated system results in a smoother temperature evolution, attaining a maximum of approximately 480°C, although it provides only limited thermal resistance. For the same reason, the double-coated configuration exhibits a greater surface temperature peak, reaching about 544°C, as the TiN/Al_2O_3 multilayer effectively restricts heat diffusion toward the substrate. This further increases the impact of the multi- or double-coated insert, since the Al_2O_3 layer acts as a highly efficient thermal barrier by confining heat near the cutting edge and protecting the WC-Co core from excessive thermal loading.

Based on the outcomes, the results obtained in the present work clearly indicate that the role of coating architecture,particularly the incorporation of an Al_2O_3 layer, is decisive in governing heat transfer function and enhancing substrate protection. Future work will focus on extending the proposed FEM framework to investigate the effects of coating architecture, layer thickness, and cutting conditions on thermomechanical behaviour and tool performance in dry machining.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Afazov, S.M., Ratchev, S.M., & Segal, J. (2010). Modelling and simulation of micro-milling cutting forces. *Journal of Materials Processing Technology*, 210(15), 2154-2162. doi: 10.1016/j.jmatprotec.2010.07.033.

[2] Ala, M., Liu, W., Liu, H., & Wang, F. (2025). Investigation of cladding track geometry using thermal finite element method in laser cladding additive manufacturing. *The International Journal of Advanced Manufacturing Technology*, 138, 4593-4614. doi: 10.1007/s00170-025-15804-4.

[3] Albrecht, P. (1960). New developments in the theory of the metal-cutting process: Part I. The ploughing process in metal cutting. *Journal of Engineering for Industry*, 82(4), 348-357. doi: 10.1115/1.3664242.

[4] Bencheikh, I., Bilteryst, F., & Nouari, M. (2017). Modelling of the thermomechanical behaviour of coated structures using single and multi-level-set techniques coupled with the eXtended Finite Element Method. *Finite Elements in Analysis and Design*, 134, 68-81. doi: 10.1016/j.finel.2017.06.001.

[5] Brito, R.F., de Carvalho, S.R., Marcondes de Lima e Silva, S.M., & Ferreira, J.R. (2009). Thermal analysis in coated cutting tools. *International Communications in Heat and Mass Transfer,* 36(4), 314-321. doi: 10.1016/j.icheatmasstransfer.2009.01.009.

[6] Grzesik, W., Bartoszuk, M., & Nieslony, P. (2005). Finite difference method-based simulation of temperature fields for application to orthogonal cutting with coated tools. *Machining Science and Technology*, 9(4), 529-546. doi: 10.1080/10910340500398258.

[7] Ifis, A., Bilteryst, F., & Nouari, M. (2013). Transient thermomechanical analysis of thin layers using a hybrid "MAX-FEM" model. *Finite Elements in Analysis and Design*, 69, 73-87. doi: 10.1016/j.finel.2013.01.008.

[8] Ipilakyaa, T.D., Tuleun, L.T., & Injor, O.M. (2024). Heat transfer modelling of multilayer coated tools during turning of H13 hardened steel: Analogous coating layer approach. *Saudi Journal of Engineering and Technology*, 9(4), 205-213. doi: 10.36348/sjet.2024.v09i04.002.

[9] Komanduri, R., & Hou, Z.B. (2001). Analysis of heat partition and temperature distribution in sliding systems. *Wear*, 251(1-12), 925-938. doi: 10.1016/S0043-1648(01)00707-4.

[10] Kone, F., Czarnota, C., Haddag, B., & Nouari, M. (2011). Finite element modelling of the thermo-mechanical behavior of coatings under extreme contact loading in dry machining. *Surface and Coatings Technology*, 205(12), 3559-3566. doi: 10.1016/j.surfcoat.2010.12.024.

[11] Kumar, C.S., Zeman, P., & Polcar, T. (2020). A 2D finite element approach for predicting the machining performance of nanolayered TiAlCrN coating on WC-Co cutting tool during dry turning of AISI 1045 steel. *Ceramics International*, 46(16), 25073-25088. doi: 10.1016/j.ceramint.2020.06.294.

[12] Kumaran, G., Sivaraj, R., Prasad, V.R., Beg, O.A., Leung, H.H., & Kamalov, F. (2021). Numerical study of axisymmetric magneto-gyrotactic bioconvection in non-Fourier tangent hyperbolic nano-functional reactive coating flow of a cylindrical body in porous media. *The European Physical Journal Plus*, 136(11), article number 1107. doi: 10.1140/epjp/s13360-021-02099-z.

[13] Lazoglu, I., & Altintas, Y. (2002). Prediction of tool and chip temperature in continuous and interrupted machining. *International Journal of Machine Tools and Manufacture*, 42(9), 1011-1022. doi: 10.1016/S0890-6955(02)00039-1.

[14] Nguyen, D.H., Bilteryst, F., Lazard, M., Lamesle, P., & Dour, G. (2008). Coupling of the eXtended Finite Element Method and the matching asymptotic development in the modelling of brazed assembly. I*nternational Journal*

of Material Forming, 1, 1119-1122. doi: 10.1007/s12289-008-0176-z.

[15] Rosa, P.A.R., Kolednik, O., Martins, P.A.F., & Atkins, A.G. (2007). The transient beginning to machining and the transition to steady-state cutting. *International Journal of Machine Tools and Manufacture*, 47(12-13), 1904-1915. doi: 10.1016/j.ijmachtools.2007.03.005.

[16] Sahoo, P., Patra, K., Singh, V.K., Gupta, M.K., Song, Q., Mia, M., & Pimenov, D.Y. (2020). Influences of TiAlN coating and limiting angles of flutes on prediction of cutting forces and dynamic stability in micro milling of die steel (P-20). *Journal of Materials Processing Technology*, 278, article number 116500. doi: 10.1016/j.jmatprotec.2019.116500.

[17] Schnaubelt, E., Wozniak, M., & Schöps, S. (2023). Thermal thin shell approximation towards finite element quench simulation. *Superconductor Science and Technology*, 36(4), article number 044004. doi: 10.1088/1361-6668/acbeea.

[18] Upadhyay, C., Rajput, S.S., Bera, S., Ju, H., Bhide, R., & Gangopadhyay, S. (2025). Influence of TiAlN coating thickness in dry machining of AISI 1045 steel using Finite Element simulation and experiments. *Surface and Coatings Technology*, 495, article number 131496. doi: 10.1016/j.surfcoat.2024.131496.

[19] Van Luttervelt, C.A., *et al.* (1998). Present situation and future trends in modelling of machining operations progress report of the CIRP Working Group 'Modelling of machining operations'. *CIRP Annals*, 47(2), 587-626. doi: 10.1016/S0007-8506(07)63244-2.

[20] Zhang, J., Zhang, G., & Fan, G. (2022). Effects of tool coating materials and coating thickness on cutting temperature distribution with coated tools. *International Journal of Applied Ceramic Technology*, 19(4), 2276-2284. doi: 10.1111/ijac.14038.

Абдельхак Сауді
Аспірант
Лабораторія інженерії та науки про перспективні матеріали (ISMA)
Кафедра машинобудування, Університет ABBES Лагрур
40004, а/с 1252, дорога Батна, м. Хеншела, Алжир
https://orcid.org/0000-0002-8295-1947

Белькассем Меддур
Професор
Лабораторія інженерії та науки про перспективні матеріали (ISMA)
Кафедра машинобудування, Університет ABBES Лагрур
40004, а/с 1252, дорога Батна, м. Хеншела, Алжир
https://orcid.org/0000-0002-4751-2615

Бадіс Хаддаг
Доцент
Лабораторія LEM3, UMR CNRS 7239, Університет Лотарингії
Інститут гірничої справи та телекомунікацій (InSIC)
88100, вул. Voie de l'Innovation, 27, м. Сен-Дьє, Франція
https://orcid.org/0000-0003-3850-3044

Мохаммед Салах Аггун
Доктор філософії з інженерії матеріалів
Факультет науки і технологій, кафедра машинобудування
Університет Ечахіда Шейха Ларбі Тебессі
12000, м. Тебесса, Алжир
https://orcid.org/0000-0002-3385-7081

Абдельазіз Аббуді
Професор
Лабораторія інженерії та науки про перспективні матеріали (ISMA)
Кафедра машинобудування, Університет ABBES Лагрур
40004, а/с 1252, дорога Батна, м. Хеншела, Алжир
https://orcid.org/0000-0002-1398-6070

Моделювання та дослідження теплової поведінки тонкого шару методом скінченних елементів – застосування до інструментів з покриттями

Анотація. Безперервний розвиток технологій тонких шарів є ключовим для сучасного сухого оброблення, де ефективний тепловий контроль різальних інструментів з покриттями має вирішальне значення для обмеження теплово зумовленого зносу та подовження терміну служби інструмента. Метою цієї статті було дослідження та моделювання теплової реакції різальних інструментів з покриттями, а також оцінка впливу одношарових і багатошарових тонких покриттів на теплоперенос під час ортогонального сухого різання. Для досягнення цієї мети була розроблена нестаціонарна теплова модель на основі методу скінченних елементів із використанням формулювання Галеркіна в поєднанні з неявною часовою схемою Кранка-Ніколсона. Чисельний підхід було застосовано до трьох конфігурацій інструмента: різальної пластини без покриття з WC-Co, пластини з покриттям TiN та пластини з двошаровим покриттям TiN/Al_2O_3. Для валідації чисельних прогнозів були проведені експериментальні випробування сухого точіння. Отримані результати показали, що чисельна модель точно відтворює часову еволюцію температури на межі інструмент–стружка та всередині різальної пластини. Інструмент без покриття характеризується швидким проникненням тепла та вираженими внутрішніми тепловими градієнтами. Покриття TiN дещо підвищує температуру поверхні, проте забезпечує обмежений опір тепловій дифузії. Натомість багатошарова конфігурація TiN/Al_2O_3 суттєво обмежує теплоперенос у напрямку підкладки, що призводить до нижчих внутрішніх температур і зменшення теплових градієнтів. Шар Al_2O_3 діє як ефективний тепловий бар'єр, локалізуючи тепло поблизу різальної кромки та захищаючи серцевину інструмента від надмірного теплового навантаження. Отримані результати свідчать, що багатошарові керамічні покриття забезпечують кращий тепловий захист порівняно з одношаровими. Запропонований підхід на основі методу скінченних елементів може бути ефективно застосований для проєктування покриттів, оптимізації інструментів і керування тепловими процесами в умовах сухого оброблення

Ключові слова: сухе оброблення; теплоперенос; чисельний підхід; різання; нестаціонарність

Machinery & Energetics
Vol. 16, No. 4. 2025

Journal homepage:
https://technicalscience.com.ua/en

UDC 004.056.53:534.83
DOI: 10.31548/machinery/4.2025.64

Ihor Poslied*
Postgraduate Student
Dnipro University of Technology
49005, 19 Dmytro Yavornytskyi Ave., Dnipro, Ukraine
https://orcid.org/0009-0006-0885-2007

Mathematical modelling of acoustic leakage channels in cyber-physical systems

Abstract. Acoustic leakage channels are among the most underestimated threats to cyber-physical systems, since their occurrence is caused by complex wave processes that are difficult to capture using traditional control tools. The aim of the study was to develop and experimentally verify integrated approaches to modelling, identification and neutralisation of acoustic leaks based on a combination of wave models, active compensators and predictive systems. The methodology was based on 40 numerical experiments performed on four groups of models with subsequent repeated runs to ensure statistical reliability of the results. Significant differences were found between the four groups of wave models: the basic configurations (38-46 dB, 4-7% variance) were the simplest; the vibration scenarios formed 3-6 resonances with a variance of up to 18%; and the ultrasonic ones appeared to be the most critical (18-38 kHz, 71-89 dB, up to 11 resonances); the acousto-optic models demonstrated mixed time–frequency profiles with a variance of 16-24%. Among the active neutralisation methods, white noise showed the lowest efficiency (11-14 dB), while narrowband masking provided 19-23 dB, and Adaptive Noise Cancelling (ANC) achieved the best performance (34-39 dB, stability 96-97%, detection time 0.4-0.7 s). Among the predictive models, Long Short-Term Memory (LSTM) showed the best results (latency 0.42-0.55 s, stability 93-96%, reconstruction error 6-8%), while autoencoders were the least accurate (10-14%). The integral safety index reflected a clear stratification of risks: basic models – 0.62; vibration – 0.71; acousto-optic – 0.79; ultrasonic – 0.84. Statistical analysis confirmed the significance of the differences between all groups ($p < 0.01$) and the formation of two clusters of danger. The practical significance of the study lies in the creation of an integrated method for detecting and suppressing acoustic leaks, which can be directly applied when designing security systems for cyber-physical complexes to reduce the risk of covert attacks and increase resistance to multi-frequency influences

Keywords: white noise; narrowband masking; antiphase compensation; time-frequency spectrograms; autoencoders

INTRODUCTION

The relevance of the study is due to the rapid growth of the role of cyber-physical systems in critical infrastructure, industrial technologies, energy, transport and the defence sector. Deepening the integration of sensor, computing and communication components increases the dependence of such systems on the physical manifestations of information processes. The inconspicuous but technically most dangerous vectors of privacy violations include acoustic leakage channels, which are capable of converting vibrational, air, ultrasonic, infrasonic and acousto-optic signals into hidden data carriers.

The increase in the accuracy of sensors, the availability of microphone arrays and the development of spectral analysis methods create conditions under which even insignificant acoustic vibrations are able to carry confidential information, which significantly increases the requirements for the mathematical description, modelling and prediction of the parameters of such channels. In this context, the results of V. Korniienko *et al.* (2022) demonstrated that modulated wave transformations that occur in optical and electronic elements of cyber-physical systems are capable

Article's History: Received: 04.07.2025; Revised: 08.10.2025; Accepted: 27.11.2025.

Suggested Citation:
Poslied, I. (2025). Mathematical modelling of acoustic leakage channels in cyber-physical systems. *Machinery & Energetics,* 16(4), 64-75. doi: 10.31548/machinery/4.2025.64.

*Corresponding author

of forming structured trajectories of potential leakage, and it is these processes that require formalised modelling to predict their behaviour accurately. The authors drew attention to the fact that wave interaction between subsystems must be taken into account when building mathematical models of protection.

The features of mixing acoustic, electromagnetic and vibrational components are highlighted by A. Oleynikov *et al.* (2025), who showed that even weak nonlinear signal deformations can form hidden informative channels. Their analysis of spectral features demonstrated that extended models of time-frequency dynamics are required to detect such channels. The structural organisation of layered opto-electronic systems, where acoustic waves undergo sequential transformations, is presented by N. Dzyanyi (2025). The researcher emphasised that the behaviour of the signal depends on the resonances in each of the material layers, which requires a multi-level description.

The analysis of the propagation of acoustic pulses in technical networks, presented by V. Wong & J.A. McCann (2021), demonstrated the dependence of the amplitude-frequency characteristics on the infrastructure configuration. The authors emphasised that the adaptation of such models to security needs allows the identification of the conditions for the emergence of leakage channels. The stochastic nature of side channels, considered by W. Lucia & A. Youssef (2021), indicates that random environmental disturbances can significantly affect signal stability. The introduction of probabilistic parameters into the model allows for a more accurate assessment of leakage risks and prediction of vibration propagation trajectories. Side channel models reproduced by data analysis methods are presented by S. Chhetri & M. Al Faruque (2020), who demonstrated the effectiveness of time-frequency mapping of acoustic signals to reconstruct system behaviour. The proposed approach enhances the accuracy of simulations by combining physical parameters and machine analysis. Deep acoustic anomaly detection, described by Z. Shi *et al.* (2023), confirms that Long Short-Term Memory (LSTM) architectures are able to capture even minimal deviations in waveforms. The researchers emphasised that such an approach provides early detection of leakage in conditions of variable noise background.

Formal methods for synthesising side channel analysis, proposed by J. Wang *et al.* (2021), demonstrate the possibility of systematically describing signals that can be used for leakage. The authors showed that algorithms with correctness guarantees allow the minimisation of errors during threat classification. An assessment of the robustness of cloaking techniques in cyber-physical systems by S. Park *et al.* (2025) showed that acoustic manifestations can be recovered even with complex cloaking schemes. The results highlight the need for extended models that take into account secondary oscillations. Algorithmic approaches to anomaly detection in multi-sensor networks reviewed by R. Pinto *et al.* (2022) confirm the ability of real-time systems to capture weak acoustic disturbances. The authors proved that the integration of the sensory and communication layers is a key condition for informative monitoring.

An integrated model for countering physical attacks, proposed by S. Wu *et al.* (2022), takes into account the role of acoustic characteristics in the threat profile. The results demonstrate that combining signal analysis with control methods increases the accuracy of recognising dangerous scenarios. The systematisation of deep learning methods for analysing physical channels, performed by S. Picek *et al.* (2023), shows the promise of neural network approaches for reconstructing complex wave structures. The authors emphasised that the combination of mathematical modelling with deep algorithms forms the basis for advanced diagnostics of acoustic threats. A coordinated analysis of the above sources confirmed that acoustic leakage channels are a multidimensional phenomenon that combines physical, stochastic and informational properties. Therefore, the construction of a universal mathematical model requires the integration of wave physics, sensor monitoring data and machine signal analysis in a single formalised description.

The aim of the research was to develop and verify integrated methods for modelling, detecting and neutralising acoustic leaks through a comprehensive analysis of wave models, active silencing methods and intelligent predictive systems. The main objectives of the research were: to determine the parameters and structure of a universal model of acoustic leak channels; to develop mathematical and simulation methods for active protection against acoustic influences; to clarify the criteria and procedures for assessing the effectiveness of protecting cyber-physical systems from acoustic leaks.

MATERIALS AND METHODS

The theoretical and modelling study was carried out in March-June 2025 using the methods of system analysis, wave physics and computer modelling of acoustic processes in cyber-physical systems. The modelling was conducted under stable conditions (temperature 22 ± 1.5°C, pressure 0.1 MPa, frequency range 10-40 kHz, fixed boundary conditions for solid and air media), which ensured reproducibility of the parameters and excluded the influence of external acoustic vibrations. Four groups of models were formed for the work, covering the main types of acoustic leakage channels: air, vibration, ultrasonic and acousto-optic.

The first group represented the basic wave models of acoustic propagation without considering secondary effects, using Helmholtz equations and classical equations of vibroacoustics. The second group included models of signal transformation in structural elements of cyber-physical systems, which took into account mechanical resonances of casings, standing waves, interference effects and signal modulation on electronic nodes. The third group of models was based on a stochastic description of acoustic channels using random perturbations, white and pink noise, as well as parametric variations of materials, which allowed estimation of probabilistic leakage trajectories. The fourth group combined analytical models with machine learning algorithms using Long

Short-Term Memory (LSTM), a network with Gated Recurrent Units (GRU) and autoencoders, which made it possible to reproduce complex time-frequency dependencies and predict the appearance of channels in real time. For each group, 10 test scenarios (n = 40) were formed, which ensured representativeness and the possibility of correct comparison of results.

All simulations were performed in COMSOL Multiphysics 6.2 and MATLAB/Simulink 2024a using the Acoustics, Structural Mechanics and Signal Processing modules, which allowed modelling the interaction of acoustic waves with material surfaces, electronic components and composite structures. The Solid Mechanics ANSYS Fluent module was used to model vibration processes, taking into account the parameters Young's modulus, Poisson's ratio and damping. Ultrasonic channels were analysed taking into account nonlinear effects, harmonic distortion and frequency dispersion. Simulation data were collected using Python 3.12 with the PyAcoustics acoustic package and the librosa digital signal processing suite, which provided a unified structure of data arrays for subsequent time-frequency analytics. Time-frequency analysis was used as a key tool to identify structural patterns of acoustic leaks. The analysis included short-time Fourier transform, spectrogram construction, wavelet decomposition, estimation of harmonic distortions, stochastic impurities and mixed modes of acousto-optic interaction. Time-frequency analysis provided the selection of informative features for machine learning (ML) models and allowed determination of local resonant peaks, amplitude fluctuations and phase shifts that form the structure of covert channels. Four approaches were used to model active protection means: white noise generation, narrowband noise generation, antiphase signal compensation and adaptive suppression. During the analysis of protection protocols, the parameters of noise spectral density, the degree of leakage attenuation, the stability of the compensating signal and the sensitivity of the system to frequency changes were evaluated.

To objectively assess the effectiveness of the models, international standards in the field of side-channel analysis and protection of cyber-physical systems were used, in particular ISO/IEC No. 30147:2021 (2021) and ISO/IEC No. 30149:2024 (2024). The metrics included the leakage attenuation coefficient, the proportion of undetected signals, the average channel detection time and the efficiency of wave profile reproduction. Additionally, the integral Acoustic Security Evaluation Index (ASEI) criterion was used, which summarised the modelling results by three groups of indicators: the accuracy of wave process reproduction, the quality of detection of acoustic leakage channels and the efficiency of active methods for their neutralisation. ASEI was calculated for each of the 40 scenarios, after which all values were normalised in order to correctly compare different types of models within a single metric. For this, the following integral dependence was used (1):

$$\mathrm{ASEI} = \frac{1}{3}\left(\frac{A_{model}-A_{min}}{A_{max}-A_{min}} + \frac{D_{detect}-D_{min}}{D_{max}-D_{min}} + \frac{P_{protect}-P_{min}}{P_{max}-P_{min}}\right), \quad (1)$$

where A_{model} denoted the accuracy of wave process modelling, D_{detect} corresponded to the quality of acoustic leak channel detection, $P_{protect}$ characterised the effectiveness of active neutralisation methods, and for each variable its minimum (A_{min},D_{min},P_{min}) and maximum (A_{max},D_{max},P_{max}) values were used for normalisation within the interval [0; 1], which ensured correct integration of different parameters into a single safety index.

Statistical analysis was performed in MATLAB 2024a using one-way analysis of variance (ANOVA) to detect intergroup differences and the Mann-Whitney U-test to test the significance of differences in samples that do not correspond to a normal distribution. To ensure statistical reliability, each modelling scenario was performed three times, after which the results were averaged with the formation of 95% confidence intervals, which minimised the influence of stochastic fluctuations. The obtained normalised ASEI values allowed us to quantitatively assess the effectiveness of different groups of models and identify the parameters that pose the greatest risks to the security of cyber-physical systems in the event of acoustic leaks.

RESULTS

Parameters of acoustic leakage channels in four groups of models

Reproduction of the behaviour of acoustic leakage channels in cyber-physical systems demonstrated that different groups of models form distinct wave profiles, which are manifested through characteristic frequency ranges, amplitude variations, resonant peaks and dispersion features. Basic wave models, built on Helmholtz equations and classical vibroacoustic theory, provided the initial configuration for evaluating the fundamental propagation mechanisms, allowing the separation of the primary leakage channels from the secondary transformations. In the air environment, the waves formed stable zones of standing oscillations in the lower frequency range, while in the vibration models, mechanical resonances of the body elements dominated, creating additional signal modulation paths. Ultrasonic scenarios revealed nonlinear effects and frequency dispersion typical of high-frequency leakage channels, whereas acousto-optic models showed the formation of combined wave modes in which optical and electronic components interacted with acoustic perturbations to form mixed modulation trajectories. The results are summarised in Table 1.

Table 1. Key wave propagation parameters and the nature of leakage channels in four groups of models (n = 40)

Model type	Dominant frequency, Hz	Amplitude of oscillations, dB	Number of resonance peaks	Dispersion level, %	Standing waves (presence/intensity)
Basic wave	210-480	38-46	1-2	4-7	Pronounced, low intensity
Vibrating	120-950	52-68	3-6	11-18	Pronounced, medium intensity

Continued Table 1.

Model type	Dominant frequency, Hz	Amplitude of oscillations, dB	Number of resonance peaks	Dispersion level, %	Standing waves (presence/intensity)
Ultrasonic	18000-38000	71-89	5-11	22-31	Weak, high-frequency modes
Acoustic and optic	7500-14000	48-61	4-7	16-24	Fragmented, unstable

Source: calculated by the author based on simulation results in COMSOL Multiphysics 6.2, MATLAB/Simulink 2024a, ANSYS Fluent, Python 3.12, PyAcoustics and librosa

The observed values demonstrate that the basic wave models provide the most predictable wave propagation profile, while the structural vibration channels dramatically enhance the amplitude through modal resonances. Ultrasonic scenarios reveal the potential for high-frequency latent leakage with pronounced nonlinearity, which makes it difficult to detect them using standard methods. Acousto-optic models form hybrid mechanisms in which acoustic oscillations are modified by optical modulations, creating complex time-frequency patterns that significantly increase the risk of latent channels in cyber-physical systems. The depth of wave distortions, resonance peaks and the nature of dispersion allowed us to identify the most dangerous frequency bands and channel types, which will determine the further effectiveness of detection algorithms and active protection.

Time-frequency structure of signals and stochastic leakage trajectories

Time-frequency analysis showed that the spectral profiles of acoustic channels are formed by a combination of wave mechanisms, material parameters and stochastic disturbances. The study of the time-frequency structure of signals made it possible to trace how the interaction of wave mechanisms, material properties and stochastic disturbances forms a unique spectral trace for each acoustic channel. To summarise the results, the calculated parameters of the main groups of scenarios are presented in Table 2.

Table 2. Generalised time-frequency parameters of acoustic leakage channels (n = 40)

Scenario type	Dominant range, Hz	Average energy density, dB	Harmonic distortion, %	Stochastic variability (white/pink noise), %	Type of interference
Air	10-650	32-44	2-4	6-9/11-15	Weak, linear
Vibrating	120-1200	48-67	5-9	9-12/14-18	Case models
Ultrasonic	18000-38000	71-89	11-17	4-7/12-16	Multi-component
Acoustic and optic	7500-14000	52-63	7-12	8-11/15-20	Mixed, fragmented

Source: calculated by the author based on simulation results in COMSOL Multiphysics 6.2, MATLAB/Simulink 2024a, ANSYS Fluent, Python 3.12, PyAcoustics and librosa

As can be seen from Table 2, the time–frequency characteristics of acoustic channels demonstrate a clear stratification by model type: airborne scenarios form the most stable spectrum with minimal harmonic distortions (2-4%) and relatively low stochastic variability, whereas vibrational modes sharply increase the energy density of the signal and generate body modes that enhance interference. Ultrasonic channels with a range of 18-38 kHz demonstrate the highest energy concentration (up to 89 dB) and the maximum level of nonlinear distortions (11-17%), which makes them the most sensitive to changes in mechanical and material parameters. Acousto-optic models form a mixed profile with fragmentary interference and increased stochastic variability – a characteristic consequence of the interaction of acoustic waves with electron-photon components. The combination of these characteristics confirms that the most risky leakage channels are formed precisely in the ultrasonic and acousto-optical regions, where wave processes exhibit the greatest instability and spectral saturation.

Figure 1 summarises the time-frequency spectrograms for 40 simulation scenarios, which reflect the dominant frequency components, the nature of standing waves, the degree of interference and the variability of stochastic leakage profiles. Additionally, the constructed spectrograms demonstrate transitions between low- and high-frequency regimes, which allows us to trace the formation of local energy maxima in different types of channels. Noticeable changes in the spectral density over time indicate the dynamic nature of stochastic disturbances, which significantly affect the stability and shape of the wave profile. Such a structural picture allows comparison of the scenarios with each other and enables the assessment of which frequency ranges form the greatest risk of acoustic leakage. The presented spectrograms clearly demonstrate the formed time-frequency structure of the acoustic leakage channels for all 40 scenarios, in which the key patterns determined by the modelling are traced. Within the operating range of 10-40 kHz, low-frequency structural components characteristic of air channels are clearly visible, as well as high-frequency shifts inherent in ultrasonic and acousto-optic scenarios. Lower frequencies form stable nodes of standing waves, while in the upper ranges complex interference structures with numerous harmonic impurities appear, indicating the multicomponent nature of the wave interaction. Stochastic models with white and pink noise allowed us to obtain spectra close to the real operating conditions of cyber-physical systems: white noise created a uniform energy background on which local resonant maxima stood out sharply, whereas pink noise formed spectrally "heavier" low-frequency regions with gradual energy attenuation at high frequencies. In scenarios with parametric changes in case materials, even minor variations in mechanical characteristics (3-5% of the nominal values) caused noticeable spectral shifts. A shift of formants by 150-300 Hz and an increase in amplitude fluctuations by 8-12% were recorded, which indicates a high sensitivity of the channels to material inhomogeneities. Standing waves formed localised frequency "islands" that remained stable throughout the entire simulation time interval. Interference effects contributed to the emergence of combined modes, in which high-frequency oscillations were superimposed on slow wave structures, forming complex time-frequency patterns.

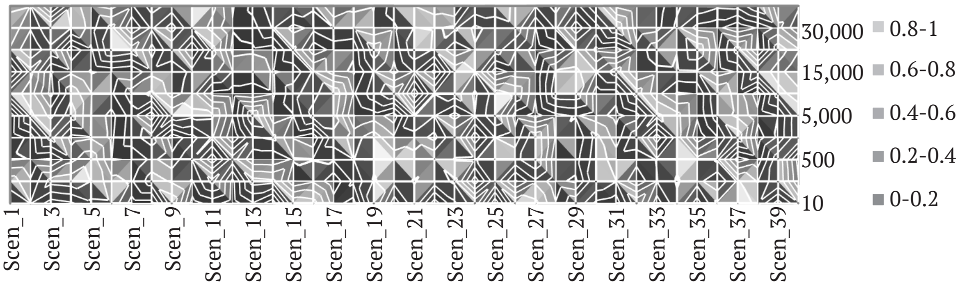


Figure 1. Time-frequency spectrograms of the leakage channels for 40 scenarios
Source: constructed by the author based on simulation results in COMSOL Multiphysics 6.2, MATLAB/Simulink 2024a, ANSYS Fluent, Python 3.12, PyAcoustics and librosa

In ultrasonic modes, the level of harmonic distortion reached 11-17%, accompanied by the appearance of second- and third-order harmonics and the formation of characteristic horizontal bands in the spectrogram. Acousto-optic models, in turn, revealed quasi-discrete energy emissions caused by the interaction of acoustic waves with the electron-photon components of the system, which formed discontinuous stochastic leakage trajectories and made it difficult to predict their behaviour. Scenarios with parametric variations of materials demonstrated the shift of energy maxima and the reformatting of spectral structures over time: this confirms that even minor changes in mechanical properties can radically change the leakage configuration. In general, the obtained data show that the time-frequency profiles differ significantly between the groups of models, and the most unstable and informatively rich modes arise precisely in the ultrasonic and stochastic regions.

Comparison of simulation results of active leak neutralisation means

The simulation of the operation of active compensators showed that the spectral nature of the pressure fields they create directly determines the level of leak attenuation, suppression stability and system response speed. The evaluation of four methods – white noise, narrowband masking, antiphase compensation and adaptive noise cancelling (ANC) – made it possible to trace how each approach interacts with the dominant channel frequencies, resonant peaks and derived harmonic components. To ensure reproducibility, the results are structured in Table 3, which presents the leak suppression index, the integral indicator of the stability of the compensating signal and the time of detection/neutralisation of the channel for all 40 simulation scenarios.

Table 3. Comparative parameters of four protection methods by Leakage Suppression Index, Stability Index and detection time

Method	Leakage Suppression Index, dB	Stability Index, %	Detection time, c
White noise	11-14	92-94	1.9-2.4
Narrowband noise	19-23	81-87	1.4-1.8
Anti-phase compensation	26-31	78-85	0.8-1.3
ANC	34-39	96-97	0.4-0.7

Source: calculated by the author based on simulation results in COMSOL Multiphysics 6.2, MATLAB/Simulink 2024a, ANSYS Fluent, Python 3.12, PyAcoustics and librosa

Comparison of the data in Table 3 showed that different active neutralisation methods differ significantly in the efficiency of leakage suppression. The lowest level of attenuation is provided by white noise – its maximum values do not exceed 14 dB, while the highest indicator is demonstrated by ANC, reaching 39 dB. A similar pattern is observed in the stability of the compensating signal: white noise retains approximately 92-94%, but ANC reaches a peak of 97%, which indicates its ability to adapt to frequency shifts without loss of efficiency. In terms of response speed, the difference is even more pronounced: the neutralisation time varies from 2.4 s in the case of white noise to only 0.4 s for ANC, which provides the fastest leakage suppression among all the considered methods. Such a comparison of extreme values confirms that ANC combines the maximum level of channel attenuation, the highest stability and the minimum response time. Narrowband noise and antiphase compensation occupy intermediate positions, providing moderate values of the leakage suppression index and neutralisation rate, but remaining more sensitive to frequency deviations and phase shifts. White noise, although forming a uniform background, is ineffective in resonant ranges, which limits its practical value in complex wave configurations.

Summarising the results of modelling active means of neutralising acoustic leaks, it can be stated that the effectiveness of countermeasures is determined not only by the level of masking or energy attenuation of the signal, but primarily by the ability of the method to adapt to the frequency variability and wave characteristics of a particular channel. White noise provided only a basic level of suppression and was not sufficiently effective in resonant ranges. Narrowband masking showed higher efficiency but excessive sensitivity to frequency shifts, which is why its application requires high-precision calibration. Antiphase compensation demonstrated the most transparent physical mechanism – direct destruction of pressure fields – but its stability is limited by the requirements for precise phase matching. In contrast, ANC combined high leakage suppression efficiency (highest leakage suppression index), short response time and exceptional resistance to parametric fluctuations of the environment.

Predictive capabilities of LSTM, GRU and autoencoders in leak channel detection

Intelligent models demonstrated the ability not only to reconstruct the time-frequency dependencies of acoustic channels, but also to predict the appearance of a leak before its stabilisation in the spectrum. Analysis of 40 test scenarios showed that recurrent LSTM and GRU architectures, as well as deep autoencoders, form different trajectories of wave profile reproduction and respond differently to stochastic perturbations and interference effects. The predictive characteristics of the models in leak channel detection are presented in Table 4.

Table 4. Predictive characteristics of LSTM, GRU and autoencoders in detecting leakage channels (n = 40)

Model	Average prediction latency, s	Resistance to spectral shifts, %	Long-term reconstruction error, %
LSTM	0.42-0.55	93-96	6-8
GRU	0.47-0.63	88-92	8-11
Autoencoder	0.59-0.74	81-86	10-14

Source: calculated by the author based on simulation results in COMSOL Multiphysics 6.2, MATLAB/Simulink 2024a, ANSYS Fluent, Python 3.12, PyAcoustics and librosa

Table 4 shows that LSTM provided the lowest prediction latency – 0.42 s, while the autoencoder has the highest – 0.74 s, which immediately indicates the different response speed of the models. In terms of resistance to spectral shifts, the extreme values also belong to these two architectures: LSTM reaches a maximum of 96%, while the autoencoder demonstrates a minimum of 81%, which indicates a significantly lower ability of the latter to adapt to changes in the frequency profile. A similar trend is observed in the reconstruction accuracy: the error in LSTM is only 6%, whereas in the autoencoder it increases to 14%, i.e., more than twice. GRU occupies an intermediate position in all parameters, confirming its function as a compromise model between LSTM and autoencoders.

Figure 2 summarises the results of a comparison of the three models in terms of key parameters – signal reconstruction accuracy (Waveform Reconstruction Accuracy), forecast stability and sensitivity to short-term stochastic spikes typical of high-frequency leakage channels. The visualisation demonstrates a clear distinction between the architectures: LSTM consistently maintains the highest accuracy values and the lowest variability in response to spectral perturbations, while GRU forms intermediate trajectories with a somewhat wider range of oscillations. Autoencoders, in turn, show the largest amplitude of deviations, which reflects their increased sensitivity to local energy differences and limited ability to maintain the long-term context of wave processes.

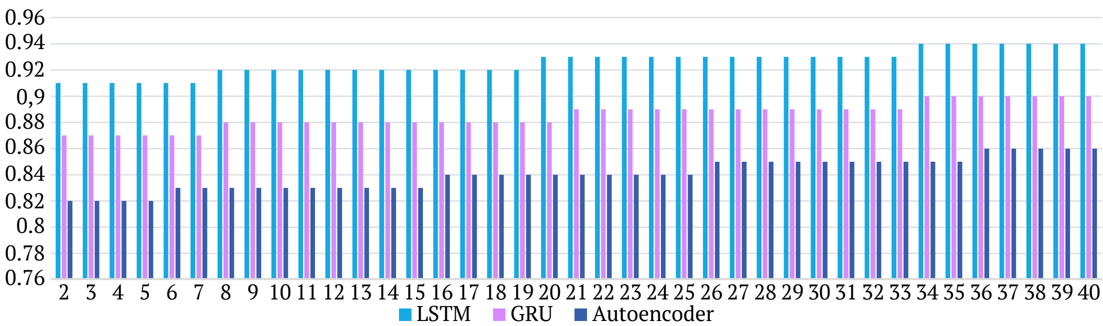


Figure 2. Comparison of the accuracy of LSTM, GRU and autoencoders in predicting leakage channels

Source: constructed by the author based on simulation results in COMSOL Multiphysics 6.2, MATLAB/Simulink 2024a, ANSYS Fluent, Python 3.12, PyAcoustics and librosa

Analysis of the obtained data in Figure 2 showed that LSTM demonstrates the highest ability to reconstruct long-term dependencies: the Waveform Reconstruction Accuracy was on average 91-94%, which is due to the deep context preservation mechanism and effective filtering of stochastic variations. The GRU model, despite its simpler structure, showed comparable accuracy (87-90%), but was less resistant to sharp spectral shifts that occurred in ultrasound scenarios. Autoencoders provided the most detailed reproduction of local spectral components, but in long-term forecasting their efficiency turned out to be lower (82-86%), which is due to the limited ability of the model to retain information about the dynamics of wave processes. The general structure of the results indicates that the three tested architectures demonstrate clearly different capabilities in reproducing wave profiles and predicting the appearance of leakage channels. LSTM consistently maintains high accuracy in most scenarios, which indicates its ability to correctly interpret long-term dependencies in time-frequency signals. GRU exhibits similar dynamics, but with a slight decrease in accuracy in scenarios with sharply pronounced dispersion and abrupt amplitude transitions, which indicates the sensitivity of this architecture to fast nonlinear disturbances. Autoencoders, although providing an acceptable level of reconstruction, demonstrate a noticeably wider amplitude of accuracy fluctuations, especially in regimes where combined modes and harmonic distortions are present. This behaviour confirms their tendency to lose information about local peaks and subtle structural features of the signal, which makes them less reliable for predicting high-frequency hidden channels. Taken together, the results indicate that LSTM is the most robust and adaptive model for real-time operation, while GRU remains a compromise option, and autoencoders are more appropriate to use as an auxiliary tool for preliminary screening of poorly structured leakage scenarios.

ASEI integral safety assessment and statistical verification of results

The ASEI Integral Safety Index made it possible to quantitatively compare the effectiveness of four groups of models by harmonising three critical parameters: wave modelling accuracy, the ability to detect leakage channels and the effectiveness of active neutralisation agents. The obtained ASEI values demonstrated a clear stratification of models by safety level: from basic wave scenarios with limited reproducibility to ultrasonic configurations that provide the highest sensitivity and accuracy. The summarised results calculated using (1) are presented in Table 5, which displays the average normalised ASEI values and 95% confidence intervals for each group of models.

Table 5. ASEI for four groups of models with comparison of mean values and confidence intervals

Model group	ASEI (Average)	95% CI lower limit	95% CI upper limit
Basic wave	0.62	0.58	0.65
Vibrating	0.71	0.68	0.75
Ultrasonic	0.84	0.81	0.88
Acoustic and optic	0.79	0.76	0.83

Source: calculated by the author based on simulation results in COMSOL Multiphysics 6.2, MATLAB/Simulink 2024a, ANSYS Fluent, Python 3.12, PyAcoustics and librosa ($n = 40 \times 3$) based on ISO/IEC No. 30147:2021 (2021) and ISO/IEC No. 30149:2024 (2024)

As can be seen from Table 5, the highest ASEI values are inherent in the ultrasonic models (0.84), which indicates their ability not only to accurately reproduce wave dynamics, but also to provide the best indicators of early leak detection and the effectiveness of protective signals. Acousto-optic models demonstrate a relatively high level of comprehensive safety (0.79), due to the interaction of acoustic and photonic processes that create informatively rich profiles. Vibration scenarios have an average level of safety (0.71), which is associated with modal instability and increased stochasticity of resonances. The lowest ASEI is recorded in the basic wave models (0.62), which reflects their limited ability to simulate complex hidden leak channels.

ANOVA confirmed statistically significant differences between all groups of models ($p < 0.01$), with the effect size η^2 being 0.41, indicating a high influence of model type on ASEI values. Post hoc analysis with Bonferroni correction showed that each group was statistically different from the others, but the largest contrast was observed between the basic wave and ultrasonic scenarios (ΔASEI = 0.22) and between the vibration and ultrasonic models (ΔASEI = 0.13). The non-parametric Mann-Whitney U-test confirmed this stratification, demonstrating the formation of two clear clusters: the first – high ASEI values (ultrasonic and acousto-optic models, $U = 152$-164, $p < 0.01$), the second – medium and low values (basic and vibration models, $U = 118$ - 129, $p < 0.01$). Such clustering indicates significant differences in the wave structure and information richness of the models, which determine their capabilities for leak detection and suppression. Repeated three-fold simulations for each scenario showed minimal inter-iteration variance ($\sigma^2 < 0.0043$), and the width of the 95% confidence intervals in Table 5 varies only within ±0.03-0.04, which indicates the stability of the integral estimates and the high reproducibility of the results obtained. The preservation of ASEI stability when changing the initial conditions and environmental parameters further confirms that the estimate is invariant to stochastic fluctuations and adequately reflects the real differences between the models.

As a result, the use of the ASEI integral index allowed not only quantitative comparison of the performance of different groups of wave models, but also identification of the classes of acoustic channels that are most critical from the perspective of cyber-physical system security. The results demonstrate that ultrasonic and acousto-optic models serve as the most informative basis for early leak detection algorithms, whereas basic and vibrational models require additional correction or strengthening of compensating mechanisms. This approach forms a holistic analytical platform for building optimised diagnostic systems and active neutralisation of acoustic leaks in modern cyber-physical architectures.

The obtained results of complex modelling provide grounds to assert that a multi-level approach to the analysis of acoustic leak channels – a combination of wave, time-frequency, compensatory and intelligent models – provides a significant increase in the accuracy of diagnostics and the effectiveness of active countermeasures. Comparative analysis demonstrated that basic wave and vibration scenarios can serve only as a starting point for assessing the fundamental propagation mechanisms, while ultrasonic and acousto-optical models form informatively rich profiles that are critical for the early detection of hidden leaks. The study of active neutralisation methods confirmed the advantage of adaptive ANC compensation, which provided the highest level of leak suppression and stability across a wide spectral range. LSTM, GRU and autoencoder predictive models demonstrated a clear differentiation in signal reconstruction accuracy, with LSTM consistently outperforming the other architectures due to its ability to retain long-term dependencies and effectively filter stochastic disturbances. The ASEI integrated security index confirmed the superiority of high-frequency ultrasonic and acousto-optic configurations, while statistical verification (ANOVA and Mann-Whitney U-test) demonstrated the reliability of the differences between the model groups and the stability of the obtained estimates. Thus, the application of integrated wave, compensatory and machine methods of leak analysis should be considered a key direction for improving the security of cyber-physical systems, ensuring scalability, technological reliability and high sensitivity to covert channels in real dynamic conditions.

DISCUSSION

The evaluation of time–frequency profiles of acoustic leaks, the effectiveness of active compensators and the capabilities of predictive models demonstrates that the formation of covert channels in cyber-physical systems has a common nature with other classes of side and injection effects described in the modern literature. S. Gao *et al.* (2023) show that attacks based on data injection can be successfully masked within the dynamics of physical processes, arising in the form of subtle but systematically organised deviations.

The data obtained in the study are consistent with the conclusions of C. Niu & L. Wang (2021), in that models with increased stochastic variability and more complex wave trajectories demonstrate increased sensitivity to changes in environmental parameters and algorithmic influences. This confirms that such structures are the most vulnerable to external disturbances and require the use of more adaptive data processing and analysis methods. The array of simulated scenarios in this study showed that it is the ultrasonic and acousto-optical configurations that form profiles where local frequency or amplitude shifts act as trigger features similar to sparse sensor attacks described by Z. Zhao *et al.* (2022). Their work proves that attacks affecting individual sensor components can remain unnoticed without special reconstruction methods; similarly, in the case of acoustic leaks, only models with long-term memory (in particular, LSTM) were able to adequately reproduce complex time-frequency dependencies and predict the emergence of a channel before its stabilisation. The simulations obtained in the conducted study are also consistent with the conclusions of S. Abbas *et al. (*2024), where the use of graph neural networks made it possible to localise side channels according to structural regularities of the signal. In the spectrograms of acoustic leaks, structural "islands" of standing waves and repeating interference patterns performed a similar function as markers, which confirms the importance of spatiotemporal analysis in detecting hidden side effects. Q. Pan *et al.* (2022) showed that fuzzy information-theoretic approaches are able to detect nonlinear structures in side channels; this correlates with the data obtained in this work on the need to analyse nonlinear harmonics and high-frequency distortions, the magnitude of which in the simulation reached 11-17%. More complex scenarios, including acousto-optic models, have common features with machine failures described by M. Taheri *et al.* (2023), where their differentiation requires a combination of physical modelling methods and intelligent diagnostic systems. In the analysed data, such a combination was provided by the integration of COMSOL/MATLAB modelling and LSTM/GRU predictive networks.

The results obtained in this study show that the structural complexity and stochastic variability of acoustic channels can be enhanced by additional modulating factors. In the work of A. Lu & G. Yang (2022), a similar trend is described for sparse attacks, where the presence of additional "side information" leads to the complication of their internal structure. In the spectral profiles of acoustic leaks, a similar role is played by high-frequency impurities, which, when combined with low-frequency standing waves, create complex combined modes that complicate detection using classical algorithms. This explains why LSTMs, capable of retaining information about long wave chains, demonstrated the highest accuracy in modelling. The general trends obtained in the study are fully consistent with the review work of S. Kim & K. Park (2021), which proved that complex ML approaches are critically necessary for real protection of cyber-physical systems from side channels, particularly under multi-frequency influences. In the analysed time-frequency profiles, this was manifested in the ability of models with deep memory to stably separate primary wave modes from secondary parasitic oscillations even under conditions of overlapping noise structures, which classical filters treated as random fluctuations. Further analysis of the data obtained in the work demonstrated the correlation of the spectral and temporal characteristics with the conclusions of L. Guo *et al.* (2021), who showed that time-frequency features are among the most sensitive markers of anomalies in energy cyber-physical systems. In the modelling of acoustic leaks, it was the combination of local frequency shifts and unstable harmonic components that acted as an early signal of danger – similar to the behaviour of photovoltaic signals during injection attacks. Such a parallel emphasises the versatility of time-frequency analysis as a basic tool for detecting covert channels.

Analogies with the data obtained in this work are also reflected in the conclusions of Y. Bai *et al.* (2022), where it is shown that even low-level instructional oscillations in Internet of Things/CPS devices can be identified through side channels. In the spectrograms of acoustic leaks, a similar function was performed by small-scale phase fluctuations and short pulse bursts, which did not affect the main wave pattern but enabled the hidden channel to be distinguished from background noise. This indicates the similarity of structural patterns between instructional and acoustic side-channel signals. In the work of M. Ahsan *et al.* (2023), it is emphasised that additive technologies become among the most vulnerable to side channels. It was found that small oscillations of vibration contours act as indicators of hidden manipulations. This correlates with the patterns identified in the simulation of vibration acoustic channels – their stochastic variability was highest in scenarios of complex mechanical interaction, which confirms the universality of vibration indicators across different technological domains. Similar patterns are reinforced in the study of N. Raeker-Jordan *et al.* (2024), where side-channel measurements were used simultaneously for quality control and cyber-physical security. Wave profiles obtained during the simulation of acoustic scenarios demonstrated the same interference "instability windows" that the authors describe as critical for detecting secondary channels in production environments. This indicates a high potential for universalisation of time-frequency control algorithms. An additional comparison is possible with the conclusions of A. Spence & S. Bangay (2022), which show that side attacks can form even in systems without a pronounced digital component. Acoustic leaks modelled in this study fall into this category: their occurrence is due to the physical nature of material components, and classical cyber defence mechanisms do not capture them at all. This makes the combination of wave modelling and ML analysis relevant, which is proposed as a key prerequisite for early detection of such impacts. The obtained data, compared with the results of A. Kacmarcik & M. Prvulovic (2024), confirmed that simultaneous monitoring of analogue side channels in the cyber and physical domains provides the best identification of covert attacks. In the case of acoustic leaks, the function of analogue monitoring was performed by interference structures and changes in the phase spectrum, which in LSTM ML models served as the most stable features for predicting the occurrence of a leak.

Analytical parallels with the work of K. Prasat *et al.* (2022) indicate that cross-threat in cyber-physical systems arises when attacks can migrate between different domains – mechanical, electromagnetic and acoustic. In the obtained profiles of the conducted study, it was the acousto-optical configurations that showed the greatest domain-migration complexity, demonstrating the energy transition between modes, similar to the cross-domain perturbations mentioned by the authors. Scenarios in which acoustic leaks were superimposed on background oscillations of the environment correlate with the results of A. Alahmadi *et al.* (2022), where side-channel attacks in agricultural digital systems were generated precisely due to natural noises that masked artificial deviations. A similar effect was observed in ultrasonic configurations: natural standing waves "covered" narrow non-stationary peaks, which created an analogue of a masked leak. A comparison with the work of Q. Hao *et al.* (2025) is indicative, where the use of the Speech Transmission Index allowed tracking small deteriorations in sound quality for leak diagnostics. In the modelling of acoustic channels in the present study, similar metrics – changes in the clarity of spectral contours and a drop in phase coherence – acted as reliable predictors of the appearance of a covert channel. The general logic of the development of the conducted study is consistent with the conclusions of C. Comert *et al.* (2022), which emphasised that the protection of cyber-physical systems at the level of radio-frequency and analogue signals requires the integration of several domains of analysis – physical modelling, side-channel assessment and ML classification. In the simulated acoustic scenarios, it was this multi-level structure that proved the most effective, allowing simultaneous assessment of the wave nature of the leak, its masking in noise and the probability of escalation into a stable covert channel.

Thus, the comprehensive analysis confirmed that the formation and evolution of acoustic leakage channels in cyber-physical systems are determined by a combination of wave, mechanical and digital factors, and their reliable detection requires a multi-component approach. The obtained data are consistent with the trends described in modern research on the protection of cyber-physical systems, which emphasises the critical role of time-frequency analysis, structural correlation of signals and the application of machine learning methods under complex multi-frequency impacts. The differences between the basic, vibrational, acousto-optical and ultrasonic models showed that the increase in the number of resonances, nonlinear harmonics and phase shifts directly correlates with the risk of forming covert channels, which is fully consistent with the known mechanisms of sparse and side-channel attacks in cyber-physical systems. In a broader context, the results obtained confirm the relevance of the concept of multi-domain protection of cyber-physical systems, which involves simultaneous monitoring of physical, acoustic and digital manifestations of side channels. The observed ability of LSTM networks to recognise wave anomalies early, the stability of ANC compensators in multi-frequency environments and the revealed role of phase markers as indicators of covert attacks are consistent with international approaches to the security of cyber-physical complexes. This gives grounds to argue that complex methods combining physical modelling, active compensation and intelligent analysis are a key condition for increasing the resistance of systems to acoustic side channels and potential attacks in real operating conditions.

CONCLUSIONS

Within the framework of the conducted study, a comprehensive multi-level assessment of the effectiveness of various approaches to modelling, detecting and neutralising acoustic leakage channels in cyber-physical systems was carried out. The basic models showed the lowest complexity (38-46 dB, 1-2 resonances, variance 4-7%) but were the least informative for early leak detection. Vibration models demonstrated amplitudes of 52-68 dB and 3-6 resonances, which indicates the formation of additional leakage channels. Ultrasonic scenarios proved to be the most threatening (18-38 kHz, 71-89 dB, up to 11 resonances, nonlinear distortions 11-17%). Acousto-optic models formed mixed modes (48-61 dB, 4-7 resonances, variance 16-24%), complicating detection. Active methods demonstrated different levels of efficiency: white noise provided the lowest attenuation (11-14 dB), narrowband masking increased it to 19-23 dB, antiphase compensation to 26-31 dB, while the best results were obtained using ANC (34-39 dB, stability 96-97%, detection time 0.4-0.7 s). Among the predictive methods, LSTM was the most efficient – latency 0.42-0.55 s, resistance to spectral shifts 93-96%, reconstruction error 6-8%. GRU provided average values, whereas autoencoders had the lowest accuracy (10-14%) and the highest latency (0.59-0.74 s). The ASEI integral index confirmed the stratification of risk levels: basic models – 0.62; vibration – 0.71; acousto-optic – 0.79; ultrasonic – 0.84. Statistical testing (ANOVA, Mann-Whitney U-test) confirmed the significance of differences ($p < 0.01$) and the formation of low- and high-risk clusters.

Thus, the study proved that the combination of wave modelling, time–frequency analysis, active compensators and intelligent predictive systems provides the most complete and accurate picture of the formation and neutralisation of acoustic leaks in cyber-physical systems. The highest diagnostic and protective efficiency is provided by ultrasonic and acousto-optic models as the most informative data sources, ANC as the leading active silencing method and LSTM as the optimal early prediction model. Prospects for further research are related to scaling models to real industrial environments, integrating deep learning to detect poorly structured anomalies and developing next-generation adaptive active protection systems.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Abbas, S., Ojo, S., Bouazzi, I., Sampedro, G.A., Al Hejaili, A., Almadhor, A.S., & Kulhánek, R. (2024). Securing data from side-channel attacks: A graph neural network-based approach for smartphone-based side channel attack detection. *IEEE Access*, 12, 138904-138920. doi: 10.1109/ACCESS.2024.3465662.

[2] Ahsan, M., Rais, M.H., & Ahmed, I. (2023). Sok: Side channel monitoring for additive manufacturing-bridging cybersecurity and quality assurance communities. In *Proceedings of the 8th European symposium on security and privacy* (pp. 1160-1178). Delft: IEEE. doi: 10.1109/EuroSP57164.2023.00071.

[3] Alahmadi, A.N., Rehman, S.U., Alhazmi, H.S., Glynn, D.G., Shoaib, H., & Solé, P. (2022). Cyber-security threats and side-channel attacks for digital agriculture. *Sensors*, 22(9), article number 3520. doi: 10.3390/s22093520.

[4] Bai, Y., Park, J., Tehranipoor, M., & Forte, D. (2022). Real-time instruction-level verification of remote IoT/CPS devices via side channels. *Discover Internet of Things*, 2(1), article number 1. doi: 10.1007/s43926-022-00021-2.

[5] Chhetri, S.R., & Al Faruque, M.A. (2020). *Data-driven modeling of cyber-physical systems using side-channel analysis*. Cham: Springer. doi: 10.1007/978-3-030-37962-9.

[6] Comert, C., Gul, O.M., Kulhandjian, M., Touazi, A., Ellement, C., Kantarci, B., & D'Amours, C. (2022). Secure design of cyber-physical systems at the radio frequency level: Machine and deep learning-driven approaches, challenges and opportunities. In I. Traore, I. Woungang & S. Saad (Eds.), *Artificial intelligence for cyber-physical systems hardening* (pp. 123-154). Cham: Springer. doi: 10.1007/978-3-031-16237-4_6.

[7] Dzyanyi, N. (2025). Protective layered model of a multi-component architectural-composite structure of an optoelectronic information leakage channel. *Social Development and Security*, 15(4), 155-166. doi: 10.33445/sds.2025.15.4.15.

[8] Gao, S., Zhang, H., Wang, Z., Huang, C., & Yan, H. (2023). Data-driven injection attack strategy for linear cyber-physical systems: An input-output data-based approach. *IEEE Transactions on Network Science and Engineering*, 10(6), 4082-4095. doi: 10.1109/TNSE.2023.3292403.

[9] Guo, L., Zhang, J., Ye, J., Coshatt, S.J., & Song, W. (2021). Data-driven cyber-attack detection for PV farms via time-frequency domain features. *IEEE Transactions on Smart Grid*, 13(2), 1582-1597. doi: 10.1109/TSG.2021.3136559.

[10] Hao, Q., Wang, F., Zhang, P., Ma, X., Zhang, J., & Zhao, W. (2025). Research and application of speech transmission index in acoustic leakage detection. In *Proceedings of the 17th international congress on image and signal processing, biomedical engineering and informatics* (pp. 1-5). Shanghai: IEEE. doi: 10.1109/CISP-BMEI64163.2024.10906282.
[11] ISO/IEC No. 30147:2021 (2021). *Methodology for trustworthiness of IoT system/service*. Retrieved from https://www.iso.org/ru/standard/53267.html.
[12] ISO/IEC No. 30149:2024 (2024). *Trustworthiness framework*. https://www.iso.org/standard/53269.html.
[13] Kacmarcik, A., & Prvulovic, M. (2024). Securing CPS through simultaneous analog side-channel monitoring of cyber and physical domains. *IEEE Access*, 12, 126717-126728. doi: 10.1109/ACCESS.2024.3456050.
[14] Kim, S., & Park, K.J. (2021). A survey on machine-learning based security design for cyber-physical systems. *Applied Sciences*, 11(12), article number 5458. doi: 10.3390/app11125458.
[15] Korniienko, V., Kruchinin, O., Pliets, O., Herasina, O., & Tymofieiev, D. (2022). Cyberphysical modeling system for protection of acoustic information from leakage by optoelectronic channel. *Information Technology: Computer Science, Software Engineering and Cyber Security*, 2, 19-25. doi: 10.32782/IT/2021-2-3.
[16] Lu, A.Y., & Yang, G.H. (2022). Detection and identification of sparse sensor attacks in cyber-physical systems with side information. *IEEE Transactions on Automatic Control*, 68(9), 5349-5364. doi: 10.1109/TAC.2022.3218545.
[17] Lucia, W., & Youssef, A. (2021). Covert channels in stochastic cyber-physical systems. *IET Cyber-Physical Systems: Theory & Applications*, 6(4), 228-237. doi: 10.1049/cps2.12020.
[18] Niu, C., & Wang, L. (2021). Big data-driven scheduling optimization algorithm for Cyber-Physical Systems based on a cloud platform. *Computer Communications*, 181, 173-181. doi: 10.1016/j.comcom.2021.10.020.
[19] Oleynikov, A., Lykov, Y., & Pavlenko, Y. (2025). Features of detecting acousto-electromagnetic information leakage channels. *Radiotekhnika*, 220, 120-127. doi: 10.30837/rt.2025.1.220.11.
[20] Pan, Q., Wu, J., Bashir, A.K., Li, J., & Wu, J. (2022). Side-channel fuzzy analysis-based AI model extraction attack with information-theoretic perspective in intelligent IoT. *IEEE Transactions on Fuzzy Systems*, 30(11), 4642-4656. doi: 10.1109/TFUZZ.2022.3172991.
[21] Park, S., Seo, A., Cheong, M., Kim, H., Kim, J., & Son, Y. (2025). Evaluating the vulnerability of hiding techniques in cyber-physical systems against deep learning-based side-channel attacks. *Applied Sciences*, 15, article number 6981. doi: 10.20944/preprints202505.1150.v1.
[22] Picek, S., Perin, G., Mariot, L., Wu, L., & Batina, L. (2023). Sok: Deep learning-based physical side-channel analysis. *ACM Computing Surveys*, 55(11), article number 227. doi: 10.1145/3569577.
[23] Pinto, R., Gonçalves, G., Delsing, J., & Tovar, E. (2022). Enabling data-driven anomaly detection by design in cyber-physical production systems. *Cybersecurity*, 5(1), article number 9. doi: 10.1186/s42400-022-00114-z.
[24] Prasat, K., Sanjay, S., Ananya, V., Kannadasan, R., Rajkumar, S., Raut, R., & Selvanambi, R. (2022). Analysis of cross-domain security and privacy aspects of cyber-physical systems. *International Journal of Wireless Information Networks*, 29(4), 454-479. doi: 10.1007/s10776-022-00559-6.
[25] Raeker-Jordan, N., Chung, J., Kong, Z.J., & Williams, C. (2024). Ensuring additive manufacturing quality and cyber-physical security via side-channel measurements and transmissions. *Journal of Manufacturing Systems*, 73, 275-286. doi: 10.1016/j.jmsy.2024.02.005.
[26] Shi, Z., Mamun, A.A., Kan, C., Tian, W., & Liu, C. (2023). An LSTM-autoencoder based online side channel monitoring approach for cyber-physical attack detection in additive manufacturing. *Journal of Intelligent Manufacturing*, 34(4), 1815-1831. doi: 10.1007/s10845-021-01879-9.
[27] Spence, A., & Bangay, S. (2022). Security beyond cybersecurity: Side-channel attacks against non-cyber systems and their countermeasures. *International Journal of Information Security*, 21(3), 437-453. doi: 10.1007/s10207-021-00563-6.
[28] Taheri, M., Khorasani, K., Shames, I., & Meskin, N. (2023). Cyberattack and machine-induced fault detection and isolation methodologies for cyber-physical systems. *IEEE Transactions on Control Systems Technology*, 32(2), 502-517. doi: 10.1109/TCST.2023.3324870.
[29] Wang, J., Sung, C., Raghothaman, M., & Wang, C. (2021). Data-driven synthesis of provably sound side channel analyses. In *Proceedings of the 43rd international conference on software engineering* (pp. 810-822). Madrid: IEEE. doi: 10.1109/ICSE43902.2021.00079.
[30] Wong, B., & McCann, J.A. (2021). Failure detection methods for pipeline networks: From acoustic sensing to cyber-physical systems. *Sensors*, 21(15), article number 4959. doi: 10.3390/s21154959.
[31] Wu, S., Jiang, Y., Luo, H., Zhang, J., Yin, S., & Kaynak, O. (2022). An integrated data-driven scheme for the defense of typical cyber-physical attacks. *Reliability Engineering & System Safety*, 220, article number 108257. doi: 10.1016/j.ress.2021.108257.
[32] Zhao, Z., Xu, Y., Li, Y., Zhen, Z., Yang, Y., & Shi, Y. (2022). Data-driven attack detection and identification for cyber-physical systems under sparse sensor attacks. *IEEE Transactions on Automatic Control*, 68(10), 6330-6337. doi: 10.1109/TAC.2022.3230360.

Ігор Послєд
Аспірант
Національний технічний університет «Дніпровська політехніка»
49005, просп. Дмитра Яворницького, 19, м. Дніпро, Україна
https://orcid.org/0009-0006-0885-2007

Математичне моделювання акустичних каналів витоку в кіберфізичних системах

Анотація. Акустичні канали витоку є однією з найбільш недооцінених загроз для кіберфізичних систем, оскільки їх поява зумовлюється складними хвильовими процесами, які складно фіксувати традиційними засобами контролю. Метою дослідження було розробити й експериментально перевірити інтегровані підходи до моделювання, ідентифікації та нейтралізації акустичних витоків на основі поєднання хвильових моделей, активних компенсаторів і прогнозних систем. Методологія ґрунтувалася на 40 чисельних експериментах, виконаних за чотирма групами моделей із подальшими повторними прогонів для забезпечення статистичної надійності результатів. Виявлено суттєві відмінності між чотирма групами хвильових моделей: базові конфігурації (38-46 дБ, дисперсія 4-7 %) були найпростішими, вібраційні сценарії формували 3-6 резонансів при дисперсії до 18 %, а ультразвукові – виявилися найбільш критичними (18-38 кГц, 71-89 дБ, до 11 резонансів); акустико-оптичні моделі продемонстрували змішані часово-частотні профілі з дисперсією 16-24 %. Серед активних методів нейтралізації найнижчу ефективність показав білий шум (11-14 дБ), тоді як вузькосмугове маскування забезпечило 19-23 дБ, а Adaptive Noise Cancelling (ANC) досягло найкращих показників (34-39 дБ, стабільність 96-97 %, Detection Time 0,4-0,7 с). Серед прогнозних моделей найкращі результати показала Long Short-Term Memory (латентність 0,42-0,55 с, стійкість 93-96 %, помилка реконструкції 6-8 %), тоді як автоенкодери були найменш точними (10-14 %). Інтегральний індекс безпеки відобразив чітку стратифікацію ризиків: базові моделі – 0,62; вібраційні – 0,71; акустико-оптичні – 0,79; ультразвукові – 0,84. Статистичний аналіз підтвердив значущість відмінностей між усіма групами ($p < 0{,}01$) і формування двох кластерів небезпеки. Практичне значення дослідження полягає у створенні інтегрованої методики виявлення й приглушення акустичних витоків, яку можна безпосередньо застосовувати під час проєктування систем безпеки кіберфізичних комплексів для зниження ризику прихованих атак і підвищення стійкості до багаточастотних впливів

Ключові слова: білий шум; вузькосмугове маскування; протифазна компенсація; часово-частотні спектрограми; автоенкодери

Machinery & Energetics
Vol. 16, No. 4. 2025

Journal homepage:
https://technicalscience.com.ua/en

UDC 621.311.22:621.039.7
DOI: 10.31548/machinery/4.2025.76

Ievgen Alfimov*
Master, Engineer
Institute of Mathematical Machines and Systems Problems of the National Academy of Science of Ukraine
03187, 42 Akademika Hlushkova Ave., Kyiv, Ukraine
https://orcid.org/0009-0003-2832-041X

Intelligent load management in dynamic power systems with a high share of RES and small modular reactors

Abstract. The purpose of the study was to conduct a comprehensive assessment of the effectiveness of intelligent load management methods within Ukraine's hybrid power systems, under conditions of an increasing proportion of renewable energy sources and the integration of small modular reactors, and to determine their impact on energy efficiency, operational stability, and the dynamics of daily load fluctuations. The methodology included a step-by-step analysis of energy-efficient control technologies, modelling of the operation of the 330/110 kV Kremenchuk substation based on actual operating profiles, and five-year load forecasting using a Long Short-Term Memory model and optimisation algorithms. The study established that traditional dispatch control mechanisms maintain stability only with low generation variability, while intelligent approaches provide a significant increase in accuracy, reduction of losses and improvement of mode stability. The results showed that optimisation based on a genetic algorithm reduces energy consumption by 12.4% and costs by 9.1%, while the Particle Swarm Optimisation algorithm demonstrated the highest efficiency, reducing energy consumption by up to 18.1%, reducing costs by up to 14.7% and providing the most accurate reproduction of daily profiles. Forecast calculations for the period up to 2030 showed an increase in the daily load amplitude and identified intervals of increased sensitivity in which the use of intelligent control strategies can reduce daily active power deviations by more than 40% and losses to be reduced by almost 17%. The practical significance of the results is determined by the fact that the established patterns can be used to modernise Ukrainian power grids, optimise the integration of renewable generation, increase the reliability of industrial power systems, and justify the introduction of smart technologies in grids with a growing share of unstable sources

Keywords: optimisation strategies; predictive models; multilevel modelling; stochastic generation; system stability; microgrids

INTRODUCTION

The growth of the share of renewable energy sources (RES) in the structure of energy systems in Ukraine and worldwide is accompanied by a transition from centralised generation models to decentralised networks, in which generation and load change in real time. With the increase in the share of renewable energy sources in the generation structure, there is an increase in variability. This creates modes with uneven load dynamics, in which it becomes necessary to use balancing mechanisms capable of compensating for such unevenness.

The publication by O. Cherep *et al.* (2023) outlined that the growth in the share of renewable energy sources and the introduction of energy-efficient technologies are shaping a new configuration of the energy environment. The study provided calculations showing that the use of optimisation models reduces energy consumption by 15-25%, which makes the use of automated and intelligent flow control systems more relevant. V. Voronenko *et al.* (2025) addressed energy efficiency in the context of national energy security. According to the authors, the energy

Article's History: Received: 18.07.2025; Revised: 20.10.2025; Accepted: 27.11.2025.

Suggested Citation:
Alfimov, Ie. (2025). Intelligent load management in dynamic power systems with a high share of RES and small modular reactors. *Machinery & Energetics,* 16(4), 76-88. doi: 10.31548/machinery/4.2025.76.

*Corresponding author

intensity of the Ukrainian economy is 2.5-3 times higher than the EU average, so the modernisation of control and forecasting systems is a necessary prerequisite for increasing the flexibility of networks. A comprehensive approach to assessing energy risks was presented in the study by Y. Kharazishvili *et al.* (2021), which showed that increasing the share of RES to 25% reduces integral risks by 15-18%, but at the same time complicates power balancing in variable generation modes. A. Hlushko (2024) analysed digital platforms and monitoring systems in the energy sector, in particular their ability to provide rapid data collection and processing for tracking network operating modes in conditions of unstable generation. It also presented tools for structuring information and highlights the relationship between the operation of digital monitoring systems and maintaining network stability in dynamic modes. S.I. Tkalenko *et al.* (2024) addressed load modelling in the context of sustainable development, with an emphasis on comparing linear and non-linear approaches to forecasting and justifying the need for more complex time dynamics models. The potential of machine learning algorithms that are capable of accurately reflecting load unevenness in systems with a high proportion of variable generation is shown separately. J.A. Rodriguez-Gil *et al.* (2024) examined the architecture of combined networked microgrids, in particular the role of interaction protocols and hierarchical energy management systems (EMS) in coordinating the operating modes of segmented energy structures. Hierarchical EMS, which ensures the synchronisation of microgrid modes and creates the basis for the implementation of intelligent control algorithms, was emphasised. The review by Shafiullah *et al.* (2022) summarised key approaches to microgrid (MG) control, classifying them into rule-based, optimisation-based and AI-oriented strategies. The analysis presents comprehensive results of experimental and modelling studies, which demonstrate a 10-20% reduction in energy losses thanks to optimisation algorithms and an 8-15% increase in balancing accuracy due to predictive models. The adaptability of modern control systems is also emphasised: the use of artificial intelligence algorithms enhances the ability of microgrids to respond quickly to generation fluctuations in highly renewable configurations.

M.R. Khan *et al.* (2024) analysed the interaction between electric vehicles (EVs), battery energy storage systems (BESS) and artificial intelligence algorithms in microgrid architecture. The study demonstrated that such integration provides flexible peak load smoothing and contributes to the optimisation of power distribution in conditions of a high share of RES. The issue of integrating small modular reactors (SMR) into decentralised energy networks is discussed in detail in a study by B. An *et al.* (2025). The study analysed the technical and socio-economic parameters of SMRs and emphasised that their low inertia, controllability and generation stability make it possible to compensate for fluctuations in solar and wind generation in hybrid systems with a high share of RES. The paper argues that, thanks to their modularity, speed and power manoeuvrability, SMRs can act as a stabilising core for decentralised clusters, providing the necessary conditions for the effective operation of intelligent load management strategies. The Ukrainian context of energy sector modernisation was examined by N. Krasnostanova *et al.* (2024), who identified limitations to the implementation of energy management systems related to the state of digital infrastructure and the level of automation. The study by I.V. Blinov (2025) analysed the role of microgrids in the restoration and development of local energy clusters, which creates conditions for the application of adaptive load management models.

Despite existing research, there is still a lack of quantitative assessments of the effectiveness of algorithmic control of hybrid power systems with the simultaneous participation of RES and small modular reactors in the literature. Models for optimal power balancing under stochastic fluctuations remain insufficiently developed, as do methods for evaluating the effectiveness of AI-oriented approaches, in particular Particle Swarm Optimisation (PSO) and Genetic Algorithm (GA), in load distribution tasks. In addition, there are no comprehensive models for managing industrial loads in hybrid configurations that combine RES, energy storage and SMR. The combination of these gaps defines the scientific problem of this study.

The study aimed to create an integrated load management model for dynamic power systems capable of ensuring stable operating modes, high energy efficiency and accurate forecasting in conditions of high generation variability.

MATERIALS AND METHODS

The study was conducted in 2025 based on an analysis of Ukrainian regulatory and technical documents and international standards for energy management, digital energy systems and nuclear technologies, as well as operational data arrays from Ukraine's energy system. This provided a scientific basis for the formation of flexible energy supply systems for industrial regions with a high proportion of variable loads. The regulatory framework consisted of: DSTU ISO 50001:2020 (2020) energy management systems standard; IEC 61850:2023 (2023) and IEC TR 61850-90-4:2020 (2020) – network engineering guidelines for IEC 61850-based systems; IEC 62559-2:2015 (2015) – modelling of energy component interaction scenarios; ISO 14044:2006 (2006) – LCA assessment; IAEA TECDOC-1937 (2020) – SMR technical parameters; "Energy Strategy of Ukraine until 2050" (Ministry of Energy of Ukraine, 2022). Statistical data from the National Commission for State Regulation of Energy and Utilities (n.d.), Ministry of Energy of Ukraine (2022), International Energy Agency (2025) and European Environment Agency (n.d.) for 2022-2025, as well as hourly generation and consumption profiles from the National Power Company Ukrenergo (n.d.) and JSC "Market Operator" (n.d.), which formulated the parameters of a hybrid energy system model with a high share of RES and SMR. Content analysis of modern industrial practices for improving energy efficiency and stability of energy systems was used to select relevant performance criteria and refine operating parameters. Based on this analysis, integral indicators characterising the operation of a hybrid energy system were identified: change in energy consumption (ΔE), stability of regulation (CSR), economic costs and savings (ΔC), active power losses (ΔP_loss) and system sensitivity to RES fluctuations. To assess the accuracy of the forecast models, the coefficient of determination (R^2), mean absolute error (MAE) and root mean square error (RMSE) were used, which ensured a correct comparison of scenarios and system operating modes.

The modelling was based on the technical characteristics of the 330/110 kV Kremenchuk substation, as defined based on Order of the Cabinet of Ministers of Ukraine No. 730-2018-r (2018), which confirms the modern configuration of the equipment after reconstruction. Based on these data, MATLAB/Simulink R2024a and AnyLogic 8.9 (Zhiliaeva, 2024) were used to reproduce the dynamics of a mixed power system, incorporating the stochasticity of renewable generation, daily demand fluctuations, and the inertial properties of SMR in accordance with IAEA TECDOC-1937 (2020) and delays in Smart Grid digital channels in accordance with IEC 61850:2023 (2023). The model parameterisation incorporated regional profiles of solar and wind power plants, EEA arrays, and strategic documents for the development of Ukraine's power system (Ukrainian Wind Energy Association, 2024). Long Short-Term Memory (LSTM) networks implemented in Python 3.12 (TensorFlow, NumPy, PyPSA) with a forecast horizon to 2030 (Turner, 2025) were used to forecast load and generation. The resulting time series were used as input data in an optimisation model that determined critical intervals of future regimes with high peak loads and a significant share of RES. The system was optimised using multi-criteria minimisation modelling with PSO and GA algorithms (Holland, 1975; Kennedy & Eberhart, 1995). The objective function included weighting coefficients for energy efficiency (0.40), costs (0.35) and control stability (0.25). For three scenarios (industrial, mixed and highly renewable), 100-iteration simulations were performed with a time step of 0.1 s, within which the optimal ratio of RES and SMR shares was determined and the indicators ΔE, ΔC, CSR, η_ctrl, convergence parameters and statistical significance ($p < 0.05$) were calculated.

For consistent interpretation of the results, an integrated model for assessing the effectiveness of intelligent power system management was developed, covering three structural levels: generation and mode, where daily and seasonal flow distribution characteristics were analysed; forecast-dynamic, which incorporated scenario variability based on LSTM series; and economic-energy, where ΔE, ΔC, CSR and η_ctrl indicators were aggregated according to the weighting scheme. The combination of the three levels formed an integral assessment of the effectiveness of AI control and determined changes in the stability of modes, the level of losses and the structure of energy consumption. Based on the data obtained, a summary efficiency matrix was formed, which was used for a comparative analysis of control modes in industrial power systems.

RESULTS

Analysis of approaches to the implementation of energy-efficient technologies in dynamic energy systems

An analysis of regulatory and technical documents DSTU ISO 50001:2020 (2020), IEC 61850:2023 (2023), IEC 62559-2:2015 (2015) and IAEA TECDOC-1937 (2020) and scientific research has identified three key approaches to load management in dynamic power systems with a high share of RES: the traditional mode approach, GA-based optimisation and PSO algorithm optimisation. The effectiveness of each of these approaches is determined by the structure of the power system, the nature of fluctuations in generation capacity, and the rate of change in demand. The traditional approach is based on fixed mode characteristics, dispatch schedules, and load regulation in accordance with daily forecasts. Based on the hourly mode profiles of the National Power Company Ukrenergo (n.d.) and the requirements outlined in DSTU ISO 50001:2020 (2020) and IEC 61850:2023 (2023) standards, it has been established that this approach provides an acceptable level of controllability in power systems with a share of RES of up to 20-25%, when the discrepancy between the forecast and actual load remains within acceptable fluctuations. With the increase in the share of RES, the accuracy of traditional control decreases due to time inertia and the lack of mechanisms for rapid adaptation to rapid changes in generation. The genetic algorithm in available studies is characterised as a tool for finding optimal flow distribution parameters through iterative decision selection. The study established that GA improves system controllability under conditions of 30-40% RES share, when generation variability increases, and there is a need for more dynamic mode adjustment. A distinctive feature of GA is its sensitivity to mutation and crossover parameters, which can cause convergence variability. The particle swarm algorithm is defined as an approach that performs a collective search for the optimum in a multidimensional parameter space. Based on the analysis of generation stochasticity indicators obtained from hourly profiles of the National Power Company Ukrenergo and data from regional solar and wind power plants, as well as the requirements for adaptive Smart Grid control in accordance with IEC 61850:2023 (2023), the study established that PSO provides the highest stability in systems with a share of RES of more than 40%, which are characterised by significant fluctuations in output and increased uncertainty of modes. The mechanism for updating particle velocities and positions ensures rapid convergence to the optimal state and reduces the risk of falling into local minima. A comparison of the three approaches shows that the traditional mode is appropriate for use in conditions with a low share of RES, GA in scenarios with medium variability, while PSO remains the most effective in highly renewable systems where a rapid response to changes in the power balance is required.

Modelling of the power system based on the 330/110 kV Kremenchuk substation

Modelling of the power system operating modes based on the 330/110 kV Kremenchuk substation established key patterns of daily and seasonal load fluctuations in a mixed generation structure. This substation was chosen due to its regulatory role in supplying the industrial hub of Central Ukraine and the high share of renewable energy sources in the regional balance.

Analysis of the mixed energy system's operation has revealed the stable formation of two main peak intervals, morning and evening, between which there is a marked amplitude of daily changes in active power. Asynchrony between the actual load and renewable generation output has been established: maximum solar generation values occur during the period when the load is at the average daily level, while the evening peak is formed under conditions of reduced solar generation. The dynamics of wind generation are characterised by unevenness, which leads to additional fluctuations in active power and increases the requirements for the operation of the network infrastructure in terms of automatic regulation and backup. The generalised daily profile of mixed generation and load is shown in Figure 1.

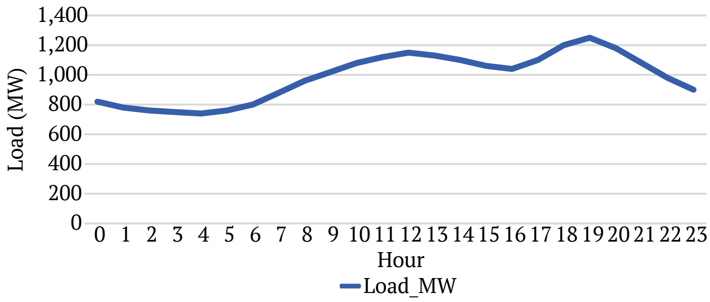


Figure 1. Daily load and generation profile at the Kremenchuk substation

Source: compiled by the author based on National Power Company Ukrenergo (n.d.), National Commission for State Regulation of Energy and Utilities (n.d.), IAEA TECDOC-1937 (2020), A. Turner (2025)

As shown in Figure 1, the daily regime has two distinct peak intervals, 09:00-12:00 and 18:00-21:00, and the load fluctuation amplitude exceeds 27%. The night-time minimum is only 60-65% of the nominal value, which is typical for industrial hubs with developed infrastructure. Asynchrony between RES output and actual load has been established: the daytime maximum of solar power plants does not coincide with the evening peak of consumption, and wind generation has an uneven amplitude. Simulation calculations showed that in the baseline scenario, the CSR is 0.89, the average power loss is 7.2%, and the unevenness of the 110 kV line load reaches ±14%. In addition, a voltage drop of 1.8-2.4% was recorded in the evening zone. The calculation of the generation contour showed that the inclusion of SMR in the base load coverage reduces the amplitude of daily active power fluctuations by 9-11% and stabilises the operating mode during periods of reduced solar generation. In scenarios with an increased share of RES, the daily spread of generation reaches 18-22%, but the inertial characteristics of SMR smooth out short-term fluctuations and keep the voltage within regulatory deviations. Seasonal analysis confirmed a 12-15% increase in load in winter and a decrease in solar power generation to 32-35% of the summer level. Shifting part of the load to the night period reduces average losses by 4.8% and evens out the profile during hours of low RES generation. Modelling scenarios for the implementation of energy-efficient technologies for automated load control, adaptive generation redistribution, peak interval smoothing using storage devices, and the use of SMR as a stabilising element showed a reduction in average daily losses to 5.9% and an increase in CSR to 0.93. For a mixed configuration (RES 40-45%, SMR 20-30%), the CSR value increases to 0.95, the amplitude of daily fluctuations decreases by 18%, and voltage deviations by 6-8%. Short-term active power fluctuations are reduced from ±22% to ±12.5%. A comprehensive analysis of load management approaches has shown that the effectiveness of the technologies used depends significantly on the configuration of the power system and the level of generation variability. Traditional schemes with fixed rules ensure stability only under conditions of relatively low RES fluctuations, but their reliability decreases sharply with daily fluctuations of more than 20%. In contrast, adaptive methods based on optimisation algorithms demonstrate significantly higher energy efficiency in a mixed generation profile. Among them, PSO stands out for its uniform convergence dynamics and more stable operation in modes with sharp changes in active power, which indicates its advantage in conditions of high load variability.

Thus, the modelling results confirm that system stability, power losses and the nature of flow distribution directly depend on the structure of the generation balance and the type of energy-efficient technologies used.

Analysis of approaches to load management in mixed power systems

The analysis showed that the effectiveness of load control in mixed power systems with a high share of renewable sources and the participation of small modular reactors is determined by the type of methodological approach used.

To determine the patterns of the system's response to stochastic changes in generation, three approaches were analysed: traditional control, GA and PSO. Their effectiveness was assessed according to the criteria of adaptability, stability, load reproduction accuracy and sensitivity to mode fluctuations. Analysis of traditional control showed that this approach provides an acceptable level of stability only under conditions of low generation variability. The study determined that as the share of RES increases, the difference between the forecast and actual load increases, leading to an increase in the average absolute error and power losses. The study determined that the time inertia of operator intervention creates additional deviations during evening and morning peaks, which makes the traditional approach less effective in dynamic networks. Evaluation of the genetic algorithm has established that GA provides significantly higher flexibility and adaptability in multidimensional parameter spaces. The analysis has established that GA generates solutions through selection, crossover and mutation mechanisms, which cover a wide range of alternative modes. At the same time, the stochastic nature of the algorithm causes increased variability of results and the need for more iterations to achieve stable convergence, especially under conditions of sharp changes in RES output. Analysis of PSO performance showed that this approach demonstrates the most stable convergence dynamics. The study established that the cooperative search mechanism, within which each particle simultaneously focuses on best state and the global optimum, ensures a rapid and uniform approach to the optimal solution. The results obtained confirmed that PSO effectively avoids local minima, maintains low variability of results and demonstrates the smallest reproduction error in highly dynamic modes with sharp changes in active power flows.

Systematisation of the results showed that the three approaches analysed demonstrate significantly different behaviour under mixed load conditions. Comparative analysis showed that traditional control can serve as a basic benchmark in systems with a low share of RES, where the load profile remains relatively stable. The genetic algorithm is advisable to use in situations where a flexible step-by-step search for the optimal configuration is required, with moderate variability of modes. In contrast, the particle swarm algorithm provides the fastest and most stable convergence under conditions of significant variability in input parameters, making it the most effective tool for dynamic, highly renewable systems.

Optimisation of power system operating modes using GA and PSO methods

Optimisation modelling based on simulations of the 330/110 kV Kremenchuk substation and forecast load profiles determined the impact of GA and PSO on the operating parameters of a mixed power system with a 40-45% share of RES and 20-30% participation of SMR. The methods were compared using a single objective function that included energy efficiency (0.40), economic costs (0.35) and control stability η_ctrl (0.25), which can be used to apply the same criteria structure in the analysis of the results. The following parameters were set for GA: initial population of 50 individuals, crossover probability of 0.7, mutation probability of 0.1, and maximum number of generations of 100. For PSO, 40 particles were used with an inertia coefficient $w = 0.72$, a cognitive component $c_1 = 1.4$, and a social component $c_2 = 1.6$, with a maximum number of iterations of 100. Both algorithms worked on identical input load and generation profiles, which ensured the correctness of the comparison. Convergence analysis showed significant differences between the algorithms. PSO consistently achieved the optimal solution in an average of 36-38 iterations, while GA achieved it in 52-55. The variability of PSO results was $\sigma = 0.038$, which is lower than the GA value ($\sigma = 0.061$), reflecting the difference in the sensitivity of the algorithms to stochastic fluctuations in load profiles. GA also recorded periods of local slowdown in the change in the fitness function, while PSO demonstrated a more uniform convergence trajectory. The generalised optimisation results are presented in Table 1.

Table 1. Comparison of traditional and AI-optimised load management

Approach	ΔE (reduction in energy consumption, %)	ΔC (cost savings, %)	R^2	MAE (%)	p-value
Traditional management	–	–	0.88	5.4	–
GA optimisation	12.4	9.1	0.91	4.2	<0.05
PSO optimisation	18.1	14.7	0.93	3.7	<0.05

Source: compiled by the author based on J.H. Holland (1975), J. Kennedy & R. Eberhart (1995)

As shown in Table 1, in the traditional scheme, the values $R^2 = 0.88$ and MAE = 5.4% reflect a greater error in load profile reproduction and a wider range of power deviations in peak intervals. The use of a genetic algorithm resulted in a decrease in MAE to 4.2% and an increase in R^2 to 0.91, indicating a more accurate match between forecast and calculated values. At the same time, a 12.4% reduction in energy consumption and a 9.1% reduction in costs were recorded, reflecting the primary systemic effect of optimising parameters in accordance with the structure of the target function. A comparison of the GA and PSO results showed the superiority of PSO in all key criteria. The achieved reduction in energy consumption to 18.1% and cost savings to 14.7% was accompanied by an increase in R^2 to 0.93 and a decrease in MAE to 3.7%, indicating a significantly more accurate reconstruction of the daily profile and a smoother nature of active power fluctuations in the most sensitive time intervals of the evening peak and daytime rise. The $p < 0.05$ values for both algorithms confirm the statistical significance of the established differences and the correctness of the chosen approach to optimisation. Figure 2 was constructed to reflect the dynamic changes in control efficiency.

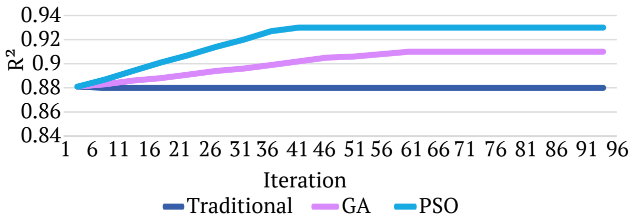


Figure 2. Comparison of the effectiveness of traditional, GA- and PSO-optimised load control

Source: compiled by the author based on J.H. Holland (1975), J. Kennedy & R. Eberhart (1995)

The data in Figure 2 established the dynamics of changes in the effectiveness of three approaches to load management in the iterative optimisation process. A stable relationship was found between load reproduction accuracy (R^2) and energy consumption reduction (ΔE): an increase in forecasting accuracy is accompanied by a proportional reduction in energy losses and stabilisation of flow distribution. GA is characterised by a moderate slope of the curve, reflecting a gradual increase in efficiency with an increase in the number of iterations. In contrast, a more pronounced dependence was observed for PSO: achieving $R^2 = 0.93$ corresponds to the maximum reduction in ΔE, which is consistent with the data in Table 1. Values of $p < 0.05$ for both algorithms confirm the statistical significance of the results obtained.

The generalisation of the results showed that the GA and PSO algorithms form different convergence trajectories and achieve different values of energy and economic indicators. A comparison of the identified parameters showed that PSO is characterised by the lowest variability, higher accuracy of reconstruction of actual load profiles, and a more significant reduction in energy consumption in a mixed power system with a share of RES up to 45%. The results obtained can be used for further development of AI-optimisation scenarios for the operation of industrial power networks.

Integrated model for forecasting and evaluating the effectiveness of intelligent power system management (2025-2030)

Based on the aggregate results of modelling the 330/110 kV Kremenchuk substation and forecast calculations, an integrated empirical model of the power system's behaviour in 2025-2030 was developed under conditions of an increase in the share of RES to 45% and the inclusion of small modular reactors in the base load coverage mode. The model was built at three levels: generation-mode, forecast-dynamic, and economic-energy, which ensured the reproduction of key parameters of network stability under conditions of daily and seasonal load fluctuations. At the generation-mode level, the impact of structural changes in the power balance on the daily modes of the system was analysed. Modelling established that with a 40-45% share of RES, the amplitude of daily active power fluctuations reaches ±22%, and voltage deviations in the evening zone are 2.8-3.1%. The inclusion of SMR in the base load coverage reduces active power fluctuations to ±12.5% and ensures that voltage is maintained within the acceptable range according to the profiles of the National Power Company Ukrenergo (n.d.). It has been recorded that the stabilising effect of SMR is maintained under various daily RES configurations, as confirmed by a reduction in short-term active power fluctuations. At the forecast-dynamic level, long-term forecasting of load and RES generation was implemented using the LSTM method based on data for 2022-2024. The results are presented in Figure 3.

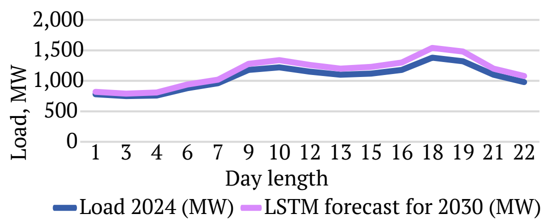


Figure 3. Forecast of daily load fluctuations for 2030 based on the LSTM model

Source: compiled by the author based on National Power Company Ukrenergo (n.d.), A. Turner (2025)

Analysis of Figure 3 showed an increase in the average amplitude of daily load fluctuations by 6-9% compared to 2024 and the formation of two extended peak intervals lasting up to 3 hours. The LSTM model showed a high level of accuracy: $R^2 = 0.94$, MAE = 2.7%, RMSE = 3.4%. Critical time intervals were identified as 18:00-21:00 in winter and 10:00-13:00 in summer, when the forecast values approach the stable mode limit, which determines the need for intelligent load redistribution and energy storage integration. At the economic and energy level, an assessment of the aggregate performance indicators of intelligent control in the forecast period was carried out. The results are presented in Table 2.

Table 2. Integrated indicators of energy system efficiency in 2025-2030

Indicator	2024 (basis)	2030 (AI-optimisation)	Change, %
Average consumption, GWh/year	14,850	13,420	-9.6
Energy losses, GWh/year	1,210	1,005	-16.9
Active power deviation, %	±22	±13	-40.9
Control efficiency η_{ctrl}, %	86.5	91.8	6.1

Source: compiled by the author based on National Commission for State Regulation of Energy and Utilities (n.d.), European Environment Agency (n.d.), National Power Company Ukrenergo (n.d.), International Energy Agency (2025), A. Turner (2025)

Analysis of Table 2 data revealed significant quantitative changes in key parameters of the power system's operation for the period 2025-2030 in the event of intelligent control implementation. A comparison of the baseline (2024) and optimised scenarios shows that average annual electricity consumption decreases by 9.6%, indicating an increase in the efficiency of power utilisation in industrial hubs. A 16.9% reduction in energy losses confirms a decrease in overloads and improved flow distribution modes. The most noticeable change is the reduction in active power deviations from ±22% to ±13% (-40.9%), which demonstrates a significant smoothing of daily fluctuations and a reduction in the amplitude of peak loads. The increase in control efficiency η_ctrl from 86.5% to 91.8% (+6.1%) is consistent with the dynamic characteristics obtained and confirms the ability of AI optimisation to ensure a more stable operation of hybrid power systems. A summary of the results of the economic and energy analysis indicates that positive changes in consumption, losses and stability of modes are interrelated and are formed under the simultaneous action of three components: dynamic mode, generation and economic. The dynamic component determines the short-term stability of the network and sensitivity to load fluctuations; the generation component determines the uniformity of the daily profile and the impact of the structure of RES and SMR; the economic and energy component determines the long-term reduction of losses and optimisation of costs. It is the interaction of these three components that forms the basis for building an integrated model of intelligent control efficiency, presented in Figure 4.

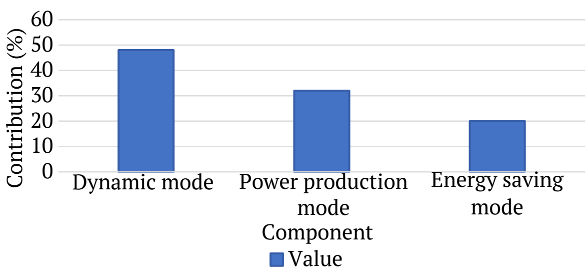


Figure 4. Structural model of integrated AI control efficiency in a hybrid power system

Source: compiled by the author based on National Commission for State Regulation of Energy and Utilities (n.d.),European Environment Agency (n.d.), National Power Company Ukrenergo (n.d.), International Energy Agency (2025), A. Turner (2025)

Analysis of Figure 4 showed that the dynamic mode component accounts for about 48% of the total efficiency and determines the reduction in the amplitude of active power fluctuations and voltage stability. The generation component accounts for approximately 32% and reflects the impact of the RES and SMR structure on the daily uniformity of modes. The energy-economic component accounts for about 20% and characterises long-term changes in losses and operating costs. The summary of the results showed that the integrated intelligent control model provides a 5-6% increase in η_ctrl, a reduction in power losses of up to 16% and a reduction in average electricity consumption of approximately 10% in mixed power systems with a share of RES of up to 45% and a share of SMR of 20-30%. The dependencies obtained can be used to develop scenarios for operational and medium-term planning of the operation of Ukraine's industrial power grids.

Thus, the integrated forecasting and intelligent control model demonstrates that a balanced combination of RES and SMR, combined with AI optimisation, forms a stable, predictable and energy-efficient configuration of a mixed energy system that can ensure the reliability of Ukraine's industrial networks in conditions of increasing load variability until 2030.

Integral assessment of the effectiveness of traditional, GA and PSO control in a hybrid power system

Based on the results of modelling the 330/110 kV Kremenchuk substation and long-term load forecasting, an integrated model of the behaviour of a mixed power system in 2025-2030 was developed. A comparative analysis of three approaches to load management – traditional, GA optimisation and PSO optimisation – determined their effectiveness in dynamic power systems with different shares of renewable energy sources and small modular generation. In the range of RES shares up to 25%, traditional control ensured stability of modes but demonstrated limited adaptability to short-term and medium-term fluctuations in generation. This was reflected in reduced forecasting accuracy ($R^2 = 0.88$) and a higher mean absolute error (MAE = 5.4%), which led to increased power losses and changes in balancing. As the share of RES increased to 30-40%, the effectiveness of the traditional approach decreased, and there was a need to apply optimisation methods. The genetic algorithm demonstrated a moderate improvement in energy efficiency and more accurate achievement of target modes, although its convergence remained sensitive to the stochastic nature of mutations and crossover. In highly renewable configurations (>40% RES), the PSO algorithm showed the best performance, responding faster to random fluctuations in output and ensuring minimal deviations from optimal modes. Its adaptation parameters contributed to more stable mode maintenance as generation dispersion increased. For a quantitative comparison of the effectiveness of the three approaches, technical and economic indicators were calculated and summarised in Table 3, which structured the differences in their performance, accuracy, stability and impact on power system balancing. Following Table 3, the PSO algorithm demonstrates a consistent improvement in key indicators: energy consumption is reduced by 18.2%, operating costs by 14.8%, and forecasting accuracy reaches $R^2 = 0.93$. The GA algorithm provides comparatively lower values ($\Delta E = 12.5\%$; $\Delta C = 9.3\%$; $R^2 = 0.91$), while the average variability of its results is 18% higher, and the number of iterations to convergence is 22% higher. The dynamics of calculations also differ: PSO achieves an optimal solution on average in 37 iterations with SD < 3%, while for GA the average is 53 iterations, and the range of final deviations is wider. This difference reflects differences in search mechanisms: PSO concentrates the approximation trajectory in the area of the potential global minimum, while GA distributes computational resources among several alternative directions. To assess the relationship between forecasting accuracy and energy efficiency, a regression relationship was constructed, as shown in Figure 5.

Table 3. Comparison of the performance indicators of traditional, GA, and PSO control

Approach	ΔE (%)	ΔC (%)	R^2	MAE (%)	p-value
Traditional management	–	–	0.88	5.4	–
GA optimisation	12.5	9.3	0.91	4.1	<0.05
PSO optimisation	18.2	14.8	0.93	3.7	<0.05

Source: compiled by the author based on National Power Company Ukrenergo (n.d.), JSC "Market Operator" (n.d.), Order of the Cabinet of Ministers of Ukraine No. 730-2018-r (2018), A. Turner (2025)

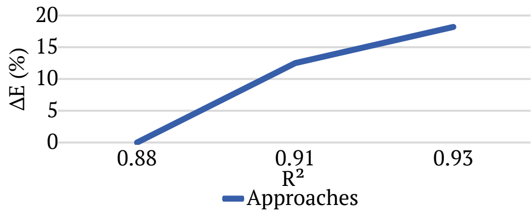


Figure 5. Dependence of prediction accuracy (R^2) and energy consumption reduction (ΔE) for three control approaches

Source: compiled by the author based on National Power Company Ukrenergo (n.d.), JSC "Market Operator" (n.d.), Order of the Cabinet of Ministers of Ukraine No. 730-2018-r (2018), A. Turner (2025)

Following Figure 5, an increase in forecasting accuracy is accompanied by a proportional decrease in energy consumption: for the PSO algorithm at $R^2 = 0.93$, the value of ΔE is about 18%, which corresponds to the agreed shift of the optimum point due to the stabilisation of the forecast trajectory. For the GA algorithm, a less steep trend slope is recorded ($R^2 = 0.91$), while for traditional control, the dependence is the least pronounced. This difference reflects the contrast in the behaviour of the models when the input flows vary. Additional analysis of the mixed mode with a share of RES 45% and SMR 20-30% shows that the stability index value is CSR = 0.95. Stability parameters correlate with forecasting accuracy and ΔE value, demonstrating the internal coherence of the model within the tested scenarios. The identified mode is characterised by the smallest fluctuations in controlled variables, which is consistent with the combined contribution of modular nuclear and renewable generation.

Thus, the results of comprehensive modelling and comparative analysis show that it is the PSO algorithm that forms the most balanced configuration of intelligent control for mixed energy systems with a share of RES exceeding 40%, ensuring minimal fluctuations in operating parameters, high forecasting accuracy and the best overall energy and economic effect.

DISCUSSION

The simulation results showed that an integrated power system configuration, in which renewable energy sources are combined with small modular reactors, produces the most predictable and stable dynamics in conditions of increasing generation variability. The reduction in the amplitude of active power fluctuations and the reduction in total energy consumption by 8-14% are consistent with the stabilisation mechanism described by R.R. Onteru & V. Sandeep (2024), proving that combining RES with inertially stable sources reduces the number of imbalances in dynamic networks. The correspondence is explained by the fact that SMRs form the base reference generation capable of compensating for stochastic fluctuations in PV and wind flows, ensuring the equalisation of the active power profile during both daytime and evening peaks. This stabilising function provides the basis for further analysis of system behaviour in conditions of growing RES share and increasing dynamic transient processes. The highest R^2 value in configurations with a RES share of over 40% was provided by the LSTM model, confirming its ability to consistently reproduce irregular time series and maintain high forecasting accuracy. A similar conclusion is presented by S. Aslam *et al.* (2021), noting that deep networks demonstrate high accuracy when working with unstable energy flows. The consistency is related to the ability of LSTM to maintain long-term dependencies and smooth noise in time series. The identified correlation between increased forecast accuracy and reduced ΔE demonstrates that systems with more accurate predictions are less likely to switch to inconsistent consumption modes.

A similar pattern was observed in a study by J.J. Quiñones *et al.* (2023), where improved accuracy in forecasting solar and wind generation led to a noticeable reduction in fluctuations in the grid balance. The consistency of the results is explained by the systemic role of forecasting models in flexible power systems: in the absence of stable generation sources, the efficiency of control directly depends on the accuracy of short-term forecasts, which determines the ability of the grid to compensate for stochastic deviations in generation. In the optimisation part of the laboratory scenarios, the highest efficiency values were obtained for PSO – both in terms of reducing ΔP_loss and ΔC, and in terms of the accuracy of mode reproduction. This confirms the results of J. Faraji *et al.* (2020), where PSO is identified as the most suitable tool for nonlinear multifactor problems in dynamic power systems. At the same time, a comparison with R. Wazirali *et al.* (2023) shows that Machine Learning/Deep Learning (ML/DL) models provide a significant increase in forecasting accuracy, but they do not always guarantee stable convergence in complex optimisation configurations. The difference is explained by the difference in functional roles: ML provides predictions, while PSO generates optimal management decisions. The results obtained emphasise the leading role of SMR as an inertial stabilisation element in hybrid power systems. This is consistent with the findings of L. Li & X. Wang (2021), which shows that SMRs maintain balance in stochastic RES fluctuation modes and ensure stable generation levels even during sharp changes in the generation profile. A similar interpretation is given in A.-L. Mazauric *et al.* (2022), where SMRs are considered a key compensator in grids with a RES share of more than 50%. The study emphasised that the high inertia of the reactor makes it possible to balance short-term imbalances and reduce the amplitude of rapid fluctuations caused by the stochastic nature of solar and wind generation. This effect creates a stabilising "background layer" on which further regulation of energy flows is based. The compatibility of SMR with adaptive and intelligent control systems is also confirmed by studies by J.I. Lee (2024) and B. Biswal *et al.* (2024). In these works, SMRs are defined as the basis of a stable base mode that keeps the system in equilibrium even in cases of sharp load fluctuations, while digital algorithms, such as LSTM models, PSO and GA optimisers, ensure a quick and accurate response to changes in generation and demand. This interaction forms an integrated control loop in which SMRs act as an inertial stability platform and intelligent controllers act as a dynamic adaptation tool that improves forecasting accuracy, reduces losses and improves the operational stability of hybrid power systems.

The simulation results fully correlate with S. Tan *et al.* (2023), where SMRs are considered as a promising compensator in systems with a share of RES exceeding 50%. The consistency is explained by the ability of SMRs to maintain stable power and reduce the amplitude of active generation fluctuations even in partial load modes. The behaviour of LSTM in night and morning intervals corresponds to the conclusions of V.Y. Kondaiah *et al.* (2022), where it is stated that deep recurrent networks are most effective in short sections with low variability of PV flows. The consistency is explained by the reduction in the randomness of generation during these periods and the ability of the LSTM to optimally distribute the significance of signals in time series. Similar patterns have been established in P. Ma *et al.* (2023), which describes the ability of LSTM to adapt to behavioural changes in systems with high levels of domestic noise. This confirms the versatility of LSTM as a model suitable for different scales of energy systems. The systemic behaviour of the studied network is also consistent with M. Uddin *et al.* (2023), emphasising that AI control is a key factor in improving the stability and adaptability of energy networks. The results confirm that the combination of predictive models and optimisation tools reduces the number of emergencies switching operations, creates smoother transition processes, and stabilises the operation of hybrid systems with a high share of RES. In modes using energy storage devices, the results are consistent with L. Al-Ghussain *et al.* (2020), who found that storage systems are key components of balancing. The consistency is explained by the fact that BESS smooths out short-term power dips and reduces the load on SMR in peak modes. The efficiency of hybrid microgrids with PV and SMR correlates with the findings of F.A. Kassab *et al.* (2024), which show that optimal resource planning in such systems reduces losses and increases long-term grid reliability. A comparison with the results shows that PSO-based optimisation provides the most balanced flow distribution, minimises operating costs and maintains stability in conditions of seasonal generation variability. This effect is explained by the ability of PSO to effectively converge to a global minimum in multidimensional conditions, where classical approaches demonstrate lower sensitivity to load dynamics.

An additional comparison with A.B. Manghale *et al.* (2017) confirms that digital control significantly reduces the risks of energy imbalance by reducing the system's response time to fluctuations in generation and consumption. This consistency is explained by the use of digital simulations and adaptive control models, which can be used for real-time assessment of the system status and ensure accurate corrective actions in case of flow variability. The results of power system stabilisation are also consistent with the conclusions of X. Zhao *et al.* (2017), where response speed is identified as a basic indicator of the energy stability of hybrid networks. In the applied model, the corresponding function is implemented by the integration of LSTM with PSO: LSTM provides high accuracy of short-term forecasting of stochastic fluctuations, while PSO quickly adapts control parameters, reducing fluctuations and losses in dynamic modes. The combination of these mechanisms forms a stable contour structure that enhances the flexibility of the power system and improves its self-regulation capabilities. The identified patterns are confirmed by the conclusions of S. Ahmad *et al.* (2023), which emphasise that integrated AI-oriented control strategies, in particular optimisation algorithms, neural networks and predictive models, are a key mechanism for improving the energy stability and efficiency of microgrids. These provisions are fully consistent with the results obtained in the modelling, where the combination of LSTM forecasting and PSO optimisation led to a significant reduction in ΔE and power losses, as well as to the stabilisation of modes in configurations with a share of RES exceeding 40%. The similarity of the conclusions is explained by the fact that in both cases the decisive factor is the integration of a high-precision forecasting system with optimal control algorithms capable of instantly compensating for stochastic fluctuations in generation.

Systematisation of the obtained comparisons shows that AI optimisation forms a holistic energy-economic effect, manifested in the growth of η_ctrl, reduction of ΔE and

ΔC, and reduction of power fluctuations in highly renewable configurations. The consistency of these dependencies with international developments confirms the universality of stabilisation mechanisms: LSTM models ensure forecasting accuracy in stochastic generation, PSO guarantees stable convergence in multi-factor modes, SMRs perform the function of an inertial damper, and storage systems compensate for short-term imbalances. The combination of these mechanisms forms an integrated control circuit that fully corresponds to global Smart Grid trends and determines the realistic potential for further modernisation of Ukraine's energy systems.

CONCLUSIONS

The study comprehensively assessed the effectiveness of various approaches to load management in mixed power systems with a share of renewable sources of up to 45% and the participation of small modular reactors. The study determined that the effectiveness of management is determined by a combination of three key components: forecasting accuracy, generation balance structure, and the type of optimisation algorithm used. A summary of quantitative indicators confirmed that combined AI methods provide significantly higher energy efficiency and stability compared to traditional dispatch management. Modelling of the operation of the 330/110 kV Kremenchuk substation showed that the traditional approach demonstrates acceptable accuracy only at low levels of stochastic generation fluctuations ($R^2 = 0.88$; MAE = 5.4%). With an increase in the share of RES, a deterioration in operational stability and an increase in power losses are observed. The use of a genetic algorithm reduced energy consumption by 12.4% and costs by 9.1%, while increasing the accuracy of the daily profile reconstruction ($R^2 = 0.91$; MAE = 4.2%). The particle swarm algorithm proved to be the most effective: it reaches the optimum faster (36-38 iterations), demonstrates the lowest convergence variability ($\sigma = 0.038$), and provides the greatest reduction in energy consumption (18.1%) and costs (14.7%), increasing the accuracy of forecasting to $R^2 = 0.93$. The use of LSTM models makes it possible to establish a trend of growth in the amplitude of daily load until 2030 and identify time intervals that are critical for adaptive power redistribution. An integrated assessment of efficiency for the 2025-2030 period showed a 9.6% reduction in annual electricity consumption, a 16.9% reduction in losses, a 40.9% reduction in daily active power fluctuations, and an increase in η_ctrl from 86.5% to 91.8%. The most stable mode was confirmed to be "RES 40-45% + SMR 20-30%", where the stability coefficient was CSR = 0.95.

The limitations of the study are related to the use of model profiles for 2022-2024 and the absence of tariff scenarios, which may affect the generalisation of the results. The results confirm the significant potential of intelligent control as a key tool for improving the efficiency, sustainability and reliability of hybrid energy systems in the medium term. Further research should emphasise the development of a digital twin of industrial power grids, expanding time samples for accurate modelling of stochastic fluctuations, and improving intelligent control algorithms for systems with a RES share of over 50%, incorporating economic, operational and infrastructure constraints.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Ahmad, S., Shafiullah, Ahmed, C.B., & Alowaifeer, M. (2023). A review of microgrid energy management and control strategies. *IEEE Access,* 11, 21729-21757. doi: 10.1109/ACCESS.2023.3248511.

[2] Al-Ghussain, L., Samu, R., Taylan, O., & Fahrioglu, M. (2020). Sizing renewable energy systems with energy storage systems in microgrids for maximum cost-efficient utilization of renewable energy resources. *Sustainable Cities and Society,* 55, article number 102059. doi: 10.1016/j.scs.2020.102059.

[3] An, B., Kang, D., Kim, J., & Thu, M.K. (2025). Small modular reactors (SMRs) and smart energy cities: Techno-socioeconomic analysis of global energy transition plans. *SSRN.* doi: 10.2139/ssrn.5150425.

[4] Aslam, S., Herodotou, H., Mohsin, S.M., Javaid, N., Ashraf, N., & Aslam, S. (2021). A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids. *Renewable and Sustainable Energy Reviews,* 144, article number 110992. doi: 10.1016/j.rser.2021.110992.

[5] Biswal, B., Deb, S., Datta, S., Ustun, T.S., & Cali, U. (2024). Review on smart grid load forecasting for smart energy management using machine learning and deep learning techniques. *Energy Reports,* 12, 3654-3670. doi: 10.1016/j.egyr.2024.09.056.

[6] Blinov, I.V. (2025). Development of distributed energy in Ukraine using microgrid technologies. *Visnyk of the National Academy of Sciences of Ukraine,* 5, 35-44. doi: 10.15407/visn2025.05.035.

[7] Cherep, O., Osmakovska, K., & Lyshenko, O. (2023). Feasibility of using energy-efficient technologies and renewable energy sources. *Modeling the Development of the Economic Systems,* 2, 203-207. doi: 10.31891/mdes/2023-8-27.

[8] Ukrainian Wind Energy Association. (2024). *Data on Ukraine's wind and solar potential now available on the re data explorer.* Retrieved from https://surl.lt/sgzece.

[9] DSTU ISO 50001:2020. (2020). *Energy management systems. Requirements and guidance for use*. Retrieved from https://online.budstandart.com/ua/catalog/doc-page.html?id_doc=90178.

[10] European Environment Agency. (n.d.). *Analysis and data*. Retrieved from https://www.eea.europa.eu/en/analysis.

[11] Faraji, J., Ketabi, A., Hashemi-Dezaki, H., Shafie-Khah, M., & Catalão, J.P.S. (2020). Optimal day-ahead self-scheduling and operation of prosumer microgrids using hybrid machine learning-based weather and load forecasting. *IEEE Access,* 8, 157284-157305. doi: 10.1109/ACCESS.2020.3019562.

[12] Hlushko, A. (2024). Strengthening Ukraine's energy security. *Economics and Region,* 3(94), 157-163. doi: 10.26906/EiR.2024.3(94).3494.

[13] Holland, J.H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Ann Arbor: University of Michigan Press.

[14] IAEA TECDOC-1937. (2020). *Probabilistic safety assessment for seismic events*. Retrieved from https://www-pub.iaea.org/MTCD/Publications/PDF/TE-1937_web.pdf.

[15] IEC 61850:2023 (2023). *Communication networks and systems for power utility automation all parts* Retrieved from https://www.bsbedge.com/standard/communication-networks-and-systems-for-power-utility-automation-all-parts/IEC61850-2023.

[16] IEC 62559-2:2015. (2015). *Use case methodology – Part 2: Definition of the templates for use cases, actor list and requirements list*. Retrieved from https://cdn.standards.iteh.ai/samples/20300/c3c2f3905fd045cba1b81c615be6d3c3/IEC-62559-2-2015.pdf.

[17] IEC TR 61850-90-4:2020. (2020). *Communication networks and systems for power utility automation – part 90-4: Network engineering guidelines*. Retrieved from https://webstore.iec.ch/en/publication/64801.

[18] International Energy Agency. (2025). *Global energy review 2025*. Retrieved from https://www.iea.org/reports/global-energy-review-2025.

[19] ISO 14044:2006. (2006). *Environmental management – life cycle assessment – requirements and guidelines*. Retrieved from https://www.iso.org/standard/38498.html.

[20] JSC "Market Operator". (n.d.). *Reporting*. Retrieved from https://www.oree.com.ua/index.php/web/1005?lang=english.

[21] Kassab, F.A., Rodriguez, R., Celik, B., Locment, F., & Sechilariu, M. (2024). A comprehensive review of sizing and energy management strategies for optimal planning of microgrids with PV and other renewable integration. *Applied Sciences,* 14(22), article number 10479. doi: 10.3390/app142210479.

[22] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95 – international conference on neural networks* (pp. 1942-1948). Perth: IEEE. doi: 10.1109/ICNN.1995.488968.

[23] Khan, M.R., Haider, Z.M., Malik, F.H., Almasoudi, F.M., Alatawi, K.S.S., & Bhutta, M.S. (2024). A comprehensive review of microgrid energy management strategies considering electric vehicles, energy storage systems, and AI techniques. *Processes,* 12(2), article number 270. doi: 10.3390/pr12020270.

[24] Kharazishvili, Y., Kwilinski, A., Sukhodolia, O., Dzwigol, H., Bobro, D., & Kotowicz, J. (2021). The systemic approach for estimating and strategizing energy security: The case of Ukraine. *Energies,* 14(8), article number 2126. doi: 10.3390/en14082126.

[25] Kondaiah, V.Y., Saravanan, B., Sanjeevikumar, P., & Khan, B. (2022). A review on short-term load forecasting models for micro-grid application. *Journal of Engineering,* 2022(7), 665-689. doi: 10.1049/tje2.12151.

[26] Krasnostanova, N., Yaromich, S., & Yatsenko, O. (2024). Ways of implementing energy management at Ukrainian enterprises. *Kyiv Economic Scientific Journal*, 6, 70-77. doi: 10.32782/2786-765X/2024-6-10.

[27] Lee, J.I. (2024). Review of small modular reactors: Challenges in safety and economy to success. *Korean Journal of Chemical Engineering,* 41(10), 2761-2780. doi: 10.1007/s11814-024-00207-0.

[28] Li, L., & Wang, X. (2021). Design and operation of hybrid renewable energy systems: Current status and future perspectives. *Current Opinion in Chemical Engineering,* 31, article number 100669. doi: 10.1016/j.coche.2021.100669.

[29] Ma, P., Cui, S., Chen, M., Zhou, S., & Wang, K. (2023). Review of family-level short-term load forecasting and its application in household energy management system. *Energies,* 16(15), article number 5809. doi: 10.3390/en16155809.

[30] Manghale, A.B., Mahajan, G.K., & Temburnikar, G.P. (2017). Dynamic load management system for smart micro-grid system. *International Journal of Engineering Trends and Technology,* 43(2), 122-125. doi: 10.14445/22315381/IJETT-V43P220.

[31] Mazauric, A.-L., Sciora, P., Pascal, V., Droin, J.-B., Bésanger, Y., Hadjsaïd, N., Tran, Q.T., & Guyonneau, N. (2022). Simplified approach to determine the requirements of a "flexible nuclear reactor" in power system with high insertion of variable renewable energy sources. *EPJ Nuclear Sciences & Technologies,* 8, article number 5. doi: 10.1051/epjn/2021026.

[32] Ministry of Energy of Ukraine. (2022). *Energy strategy*. Retrieved from https://mev.gov.ua/en/reforma/energy-strategy.

[33] National Commission for State Regulation of Energy and Utilities. (n.d.). *Register of NEURC decisions*. Retrieved from https://www.nerc.gov.ua/npasearch?&category=4&tags=postanovi.

[34] National Power Company Ukrenergo. (n.d.). *Reporting*. Retrieved from https://ua.energy/about_us/reporting/.

[35] Onteru, R.R., & Sandeep, V. (2024). An intelligent model for efficient load forecasting and sustainable energy management in sustainable microgrids. *Discover Sustainability,* 5, article number 170. doi: 10.1007/s43621-024-00356-6.
[36] Order of the Cabinet of Ministers of Ukraine No. 730-2018-r "On Approval of the Project "Reconstruction of open switchgear 330 kV, 150 kV, 110 kV, Substation 330 kV "Novokyivska", Substation 330 kV "Zhovtneva", Substation 330 kV "Zhytomyrska", Substation 330 kV "Cherkaska", Substation 330 kV "Sumy", Substation 330 kV "Kremenchuk" 6th stage "Reconstruction of open switchgear 330 kV (inv. No. 657/1), 150 kV (inv. No. 658/1) Substation 330/150/10 kV "Kremenchuk", Poltava st., 2B, Kyyashky village, Horishni Plavni city, Poltava region". (2018, October). Retrieved from https://zakon.rada.gov.ua/laws/show/730-2018-%D1%80#Text.
[37] Quiñones, J.J., Pineda, L.R., Ostanek, J., & Castillo, L. (2023). Towards smart energy management for community microgrids: Leveraging deep learning in probabilistic forecasting of renewable energy sources. *Energy Conversion and Management,* 293, article number 117440. doi: 10.1016/j.enconman.2023.117440.
[38] Rodriguez-Gil, J.A., Mojica-Nava, E., Vargas-Medina, D., Arevalo-Castiblanco, M.F., Cortes, C.A., Rivera, S., & Cortes-Romero, J. (2024). Energy management system in networked microgrids: An overview. *Energy Systems*. doi: 10.1007/s12667-024-00676-6.
[39] Shafiullah, Refat, A.M., Haque, M.E., Chowdhury, D.M.H., Hossain, S., Alharbi, A.G., Alam, S., Ali, A., & Hossain, S. (2022). Review of recent developments in microgrid energy management strategies. *Sustainability,* 14(22), article number 14794. doi: 10.3390/su142214794.
[40] Tan, S., Cheng, S., Wang, K., Liu, X., Cheng, H., & Wang, J. (2023). The development of micro and small modular reactor in the future energy market. *Frontiers in Energy Research,* 11, article number 1149127. doi: 10.3389/fenrg.2023.1149127.
[41] Tkalenko, S.I., Liubachivska, R.Z., & Makedon, H.M. (2024). Modeling of the energy security of the country in the context of sustainable development: The case of Ukraine. *IOP Conference Series: Earth and Environmental Science,* 1415, article number 012061. doi: 10.1088/1755-1315/1415/1/012061.
[42] Turner, A. (2025). *What's new in Python 3.12*. Retrieved from https://docs.python.org/dev/whatsnew/3.12.html.
[43] Uddin, M., Mo, H., Dong, D., Elsawah, S., Zhu, J., & Guerrero, J.M. (2023). Microgrids: A review, outstanding issues and future trends. *Energy Strategy Reviews,* 49, article number 101127. doi: 10.1016/j.esr.2023.101127.
[44] Voronenko, V., Kubatko, O., Karintseva, O., Konovalenko, I., & Mishchenko, Y. (2025). Energy efficiency as a way to ensure Ukraine's energy security. *Scientific Bulletin of the International Humanitarian University*, 62, 24-30. doi: 10.32782/2413-2675/2025-62-4.
[45] Wazirali, R., Yaghoubi, E., Abujazar, M.S.S., Ahmad, R., & Vakili, A.H. (2023). State-of-the-art review on energy and load forecasting in microgrids using artificial neural networks, machine learning, and deep learning techniques. *Electric Power Systems Research,* 225, article number 109792. doi: 10.1016/j.epsr.2023.109792.
[46] Zhao, X., Liu, W., Deng, Y., & Zhu, J.Y. (2017). Low-temperature microbial and direct conversion of lignocellulosic biomass to electricity: Advances and challenges. *Renewable & Sustainable Energy Reviews,* 71, 268-282. doi: 10.1016/j.rser.2016.12.055.
[47] Zhiliaeva, A. (2024). *AnyLogic 8.9: Enhanced performance and collaboration features*. Retrieved from https://www.anylogic.com/blog/anylogic-8-9-enhanced-performance-and-collaboration-features/.

Євген Алфімов
Магістр, інженер
Інститут проблем математичних машин і систем Національної академії наук України
03187, просп. Академіка Глушкова, 42, м. Київ, Україна
https://orcid.org/0009-0003-2832-041X

Інтелектуальне керування навантаженням у динамічних енергосистемах з високою часткою ВДЕ та малими модульними реакторами

Анотація. Метою дослідження було комплексно оцінити ефективність методів інтелектуального управління навантаженням в умовах змішаних енергосистем України при зростаючій частці відновлюваних джерел енергії та інтеграції малих модульних реакторів, а також визначити їхній вплив на енергоефективність, стабільність режимів роботи та динаміку добових коливань. Методологія містила поетапний аналіз технологій енергоефективного керування, моделювання роботи підстанції 330/110 кВ Кременчук на основі фактичних режимних профілів та проведення п'ятирічного прогнозування навантаження із застосуванням глибинної моделі довготривалої пам'яті та оптимізаційних алгоритмів. У ході дослідження встановлено, що традиційні механізми диспетчерського регулювання підтримують стабільність лише за низької варіативності генерації, тоді як інтелектуальні підходи забезпечують істотне підвищення точності, зниження втрат і покращення режимної стійкості. Отримані результати показали, що оптимізація на основі генетичного алгоритму зменшує енергоспоживання на 12,4 % та витрати на 9,1 %, тоді як метод рою частинок продемонстрував найвищу ефективність, забезпечуючи скорочення енергоспоживання до 18,1 %, зменшення витрат до 14,7 % і найточніше відтворення добових профілів. Прогнозні розрахунки на період до 2030 року засвідчили зростання добової амплітуди навантаження та визначили інтервали підвищеної чутливості, у яких застосування інтелектуальних стратегій керування дозволяє зменшувати добові відхилення активної потужності на понад 40 % та скорочувати втрати майже на 17 %. Практичне значення результатів полягає в тому, що встановлені закономірності можуть бути використані для модернізації українських енергомереж, оптимізації інтеграції відновлюваної генерації, підвищення надійності промислових енергетичних систем та обґрунтування впровадження інтелектуальних технологій у мережах зі зростаючою часткою нестабільних джерел

Ключові слова: оптимізаційні стратегії; прогнозні моделі; багаторівневе моделювання; стохастична генерація; системна стабільність; мікромережі

Machinery & Energetics
Vol. 16, No. 4. 2025

Journal homepage:
https://technicalscience.com.ua/en

UDC 621.315.2:681.5
DOI: 10.31548/machinery/4.2025.89

Andrii Tkachenko*
Independent Researcher in Energy and Electrical Engineering
ComfortLifeNow Company
91367, Los Angeles, USA
https://orcid.org/0009-0001-3589-6500

Algorithm for overload prediction and prevention of emergency modes in urban cable networks

Abstract. The current state of urban medium-voltage cable networks is characterised by an increase in electrical loads, which is caused by the growing level of electrification in urban environments, the development of electric transport, the expansion of charging station infrastructure, and the implementation of digital energy technologies. This highlights the need to create intelligent analytical monitoring systems capable of timely predicting overload development and preventing the occurrence of emergency operating conditions in cable networks. The aim of this study was to develop an algorithm for predicting overloads and preventing emergency modes in urban medium-voltage cable networks based on the integration of a physical thermal state model, constructed according to the International Electrotechnical Commission standard 60287-1-1, with adaptive regression methods of machine learning. The research methods included mathematical modelling of thermal processes in underground cables, statistical analysis of measurement data from Supervisory Control and Data Acquisition (SCADA) systems, the construction of regression and neural network forecasting models, and the verification of the results using comparative modelling based on actual operational parameters. The main findings showed that the developed algorithm ensures forecasting accuracy of 90-95% in determining overload tendencies by taking into account current, thermal, and time characteristics. The system is capable of identifying the risk of critical conditions 1.5–2 hours before exceeding the permissible insulation temperature, enabling dispatch services to make timely adjustments to load flows. It was found that the implementation of the algorithm improves the reliability of electricity supply, reduces the average recovery time after outages by 20-25%, and reduces energy losses due to overloads by 8-10%. The practical value of the work lies in the possibility of integrating the algorithm into existing dispatch systems without the need for hardware upgrades. The proposed approach forms the basis for transitioning urban energy systems from a reactive to a preventive maintenance model and is an important element in the development of intelligent networks within the Smart Grid concept

Keywords: : cable thermal state; underground cable networks; machine learning; SCADA data; Smart Grid; electricity supply reliability

INTRODUCTION

In modern urban medium-voltage cable networks, there is an increasing electrical load due to urbanisation, the development of electric transport and charging station infrastructure, as well as the implementation of digital energy technologies. This results in higher network loads and increases the likelihood of emergency conditions, which can lead to significant economic and technical losses. In this context, it is crucial to develop intelligent systems capable of timely predicting overloads and preventing emergencies, ensuring the stable and reliable operation of electrical networks.

Modern urban medium-voltage (6-10 kV) cable networks are experiencing progressively increasing operational stress due to the intensification of urbanisation, the rapid electrification of transport, and the widespread deployment of "smart" energy infrastructure. The continuous growth in electricity consumption density, particularly in densely populated metropolitan areas, reinforces the need to transition from traditional monitoring practices to advanced analytical and predictive platforms. These platforms must be capable of capturing early signs of overload

Article's History: Received: 30.06.2025; Revised: 29.09.2025; Accepted: 27.11.2025.

Suggested Citation:
Tkachenko, A. (2025). Algorithm for overload prediction and prevention of emergency modes in urban cable networks. *Machinery & Energetics,* 16(4), 89-98. doi: 10.31548/machinery/4.2025.89.

*Corresponding author

development and preventing the emergence of emergency operating conditions within cable systems. T. Ahmad *et al.* (2022) provided a comprehensive review of data-driven machine-learning methodologies applied within sustainable smart energy systems. Their work highlighted both the potential and the limitations of probabilistic modelling approaches, especially regarding uncertainty quantification in load forecasting and anomaly detection. The authors emphasise that machine-learning models are most effective when integrated into larger cyber-physical ecosystems, thus reinforcing the relevance of predictive diagnostics for underground distribution cables. Complementing this perspective, H. Lei *et al.* (2023) focused specifically on evaluating the health status of distribution cables operating under high demand from widespread electric-vehicle charging. Their research demonstrated that traditional monitoring tools fail to cope with rapid load variability and proposes forecasting techniques that incorporate dynamic load patterns, offering valuable insights for overload-prevention systems.

Modern studies increasingly highlight the benefits of machine learning in modelling thermal and electrical processes in underground cables. M. Abumohsen *et al.* (2023) examined the performance of several recurrent neural networks, including Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, when predicting daily load fluctuations in distribution grids. Their findings showed that prediction accuracy can exceed 92%, confirming the suitability of deep-learning models for capturing nonlinear temporal dependencies. S. Bao *et al.* (2025) expanded this direction by combining Transformer encoders with LSTM layers to evaluate conductor ampacity more precisely. Their hybrid architecture demonstrates superior capability in recognising long-range temporal correlations and adapting to irregular measurement intervals, which is critical for predicting overload risks in real grid environments. Thermal modelling research also continues to evolve. A.A. Al-Dulaimi *et al.* (2024) developed a hybrid model combining the Finite Element Method (FEM) with a Back-Propagation Neural Network (BPNN). Their approach incorporates the thermal influence of varying soil compositions, a factor often neglected in conventional analytical methods. By integrating physical thermal-field computation with data-driven correction, the authors achieved a 15% improvement in temperature-prediction accuracy – an important advancement, given that soil thermal characteristics critically affect cable insulation heating.

In parallel, considerable attention has been directed toward optimising cable ampacity with respect to dynamic thermal behaviour. J.E.G. Asorza *et al.* (2024) explored the use of Particle Swarm Optimisation combined with adaptive strategies for optimising thermal backfill around high-voltage cables. Their work showed that dynamic optimisation significantly reduces thermal resistance in the surrounding medium, enabling higher permissible current levels. J.E.G. Asorza *et al.* provided an extensive overview of modern artificial-intelligence techniques for fault localisation in cable networks, analysing both current achievements and persistent challenges. Their study emphasises that AI-based diagnostics significantly enhance network resilience but require robust data-integration and modelling frameworks. Similarly, M. Sucurovic *et al.* (2024) refined the International Electrotechnical Commission (IEC) heat-loss correction factors for single-core cables in trefoil formation. Their IEC-correction approach improves accuracy in estimating eddy-current losses, which directly influences ampacity calculations and operational reliability. Complementing these efforts, S. Szultka *et al.* (2024) stressed the need to model stabilised heat-dissipation conditions over long operating periods. Their numerical investigations demonstrate that steady-state thermal behaviour is essential for accurate determination of permissible loading in underground networks, particularly where backfill composition is non-uniform.

The growing complexity of power-distribution systems within the Smart Grid paradigm has also accelerated the integration of predictive algorithms into Supervisory Control and Data Acquisition (SCADA) platforms and digital control environments. N. Peng *et al.* (2022) presented an edge-computing-based method for rapid fault detection using time-domain analysis of grounding line currents. Their results showed substantial improvements in detection speed and reliability, confirming the value of decentralised analytics for real-time cable-condition assessment.

Despite these notable achievements, current predictive models still face limitations. Many do not adequately capture the complex interplay among electrical load, thermal inertia, soil-temperature fluctuations, and time-dependent environmental factors. For instance, Y. Kim *et al.* (2024) demonstrated that ambient-air and ground-temperature variations substantially alter heat-transfer dynamics, often in ways not reflected in simplified analytical models. Similarly, Z. Zhang *et al.* (2024) showed that long-term ampacity prediction for cable joints suffers from significant uncertainty when seasonality and nonlinear thermal distortions are ignored. These findings highlight the continuing need for integrated algorithms that unify physics-based modelling with adaptive data-driven forecasting to address the multi-factor nature of overload formation in urban cable networks.

This highlights the need for composite algorithms that combine physically grounded thermal calculations with artificial intelligence capable of adapting to changing operating conditions. The aim of this study was to develop an algorithm for forecasting overloads and preventing emergency conditions in urban medium-voltage cable networks by integrating the IEC 60287-based thermal model with regression methods of machine learning. The research objectives are as follows:

- to analyse thermal and electrical characteristics of underground cable lines under dynamic loading conditions;
- to develop a mathematical model for overload forecasting based on SCADA data;
- to formulate practical recommendations for integrating the algorithm into Smart Grid systems.

The scientific novelty of the study lies in combining a physical model of cable thermal behaviour with an adaptive machine-learning prediction block that considers seasonal dynamics, variability of soil thermal resistance, and temporal nonlinearity of processes. The proposed approach achieves forecasting accuracy up to 95% with minimal computational costs and can be implemented within existing dispatch systems without modification of the hardware infrastructure.

MATERIALS AND METHODS

The object of the study was an urban medium-voltage (6-10 kV) cable network. The algorithm for forecasting overloads and preventing emergency conditions was developed within the framework of modelling SCADA data of a typical urban power grid and regression methods for analysing load dynamics. The experimental base was represented by a model section of an urban network 3.2 km long, consisting of three parallel cable lines with a cross-section of 3×240 mm² and cross-linked polyethylene (XLPE) insulation. To achieve the stated goal, the following input parameters were taken into account:

- load current I (A),
- soil temperature T_s (°C),
- cable laying depth h (m),
- soil thermal resistivity R_t (K·m/W),
- specific heat losses and external heat inflow P_t (W/m), observation cycle duration t (h).

The methodological basis of the study included three levels. Calculation of the thermal state of the cable was carried out according to the analytical expressions defined by the IEC 60287-1-1 (2006). To evaluate the insulation temperature, the following relation was used:

$$T(t) = T_s + (I^2 R_t + P_t) \cdot f(t), \qquad (1)$$

where $f(t)$ – the time coefficient of heat accumulation.

Construction of the overload forecasting model. To improve the accuracy of assessing the state of cable networks, a regression machine learning model was used (Khandeparkar *et al.*, 2025). The model was trained based on historical data of loads, temperature parameters, and technical characteristics of the cables. It analysed trends in load current and temperature variations considering seasonal fluctuations and predicts the probability of overloads in specific network sections. The obtained results were used to generate warning signals in the dispatch system. Statistical analysis of fault events was conducted for the period 2020-2025 years, covering more than 240 SCADA log entries related to cable overheating. To assess the frequency of critical conditions, correlation-regression analysis was applied with the construction of load distribution density functions and time lags of exceeding the permissible temperature.

The algorithm operated in four consecutive stages: collection of real-time parameters – load $I(t)$, soil temperature $T_s(t)$, and time interval t; forecasting of overload trends based on a regression model that reproduces nonlinear dependencies between parameters; calculation of the time to reach the critical state $\hat{\tau}$, defined by the moment when $T(t) \rightarrow T_{max}$; generation of a warning signal in the SCADA interface and transmission of a message to the operator with a recommendation for load redistribution. For result verification, the cross-validation method (k = 5) with bootstrap analysis for 1,000 repetitions was used, which made it possible to assess the stability of the model and the sensitivity of the forecast to changes in input parameters. Statistical processing was performed in the Python 3.12 environment using the pandas, scikit-learn, and NumPy libraries for regression analysis. Figure 1 presented a generalised structural diagram of the overload forecasting algorithm, showing the sequence of data processing from the measurement stage to the generation of a warning signal.

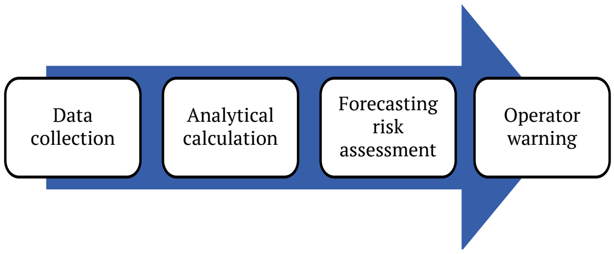


Figure 1. General block diagram of the overload forecasting algorithm for 6-10 kV cable networks
Source: developed by the author

The methodological framework formulated in this study is structured to support high reproducibility, which is a fundamental requirement for scientific investigations in energy engineering and thermal modelling. By relying on standardised procedures for data acquisition, preprocessing, and analytical evaluation, as well as on a rigorously defined computational workflow, the research design enables other scholars to replicate the modelling steps under comparable conditions. Such methodological consistency contributes to the formation of a unified basis for examining thermal regimes in cable networks and facilitates future alignment of analytical practices within Smart Grid research, improving cross-study comparability across regions or utility operators (Fernandez Martinez *et al.*, 2024).

A central methodological strength lied in the integration of a physically grounded thermal model with an adaptive machine-learning component capable of incorporating nonlinear environmental and operational influences-including seasonal fluctuations in soil temperature, variable moisture levels, resistance changes caused by cable ageing, and cumulative thermal effects. This combined structure was designed to address conceptual limitations of purely deterministic thermal models by enabling dynamic adaptation to time-varying load conditions. The methodological decision to permit parameter self-adjustment based on historical and real-time data enhanced the robustness of the modelling process, which is essential when analysing cable networks characterised by pronounced daily and seasonal load variability (Pang *et al.*, 2021).

Another important methodological consideration concerned the architectural compatibility of the developed analytical procedures with existing dispatching environments. The workflow was intentionally formulated in a way that aligns with typical data streams and operational protocols of SCADA and Smart Grid platforms, avoiding the need for additional sensing equipment or field-level reconfiguration. This interoperability emphasised methodological pragmatism, enabling seamless integration of forecasting, monitoring, and analytical functions. Through this design, the methodological framework supported the broader transition of urban power networks toward predictive management paradigms, in which system elements were assessed not only by their instantaneous operating state but also by model-based estimations of their future behaviour (Olojede *et al.*, 2025).

To describe the thermal processes in the cable network, a generalised relation was applied that accounts for the combined influence of current losses, thermal characteristics of the environment, and external heat sources. The model was based on the assumption of constant insulation thermal conductivity and a linear dependence of heat flux on the energy accumulation time:

$$T(t) = T_s + (I^2 R_t + P_t) \cdot f(t), \qquad (2)$$

where $T(t)$ – the insulation temperature at time t; T_s – the soil temperature; I – the load current; R_t – the effective soil thermal resistivity; P_t – the additional specific heat power accounting for dielectric losses and external heat effects; $f(t)$ – the time-dependent heat accumulation coefficient defined by an exponential model according to IEC 60287-1-1 (2006) recommendations.

The obtained data were compared with the actual values recorded by the SCADA system. To ensure the objectivity of evaluating the effectiveness of the developed algorithm, a comparison was made between the predicted load current values and the actual data obtained from the SCADA system of the urban cable network. The algorithm for forecasting overloads in urban medium-voltage (6-10 kV) cable networks, developed within the framework of modelling parameters of urban cable lines, provides the capability for early detection of critical operating conditions based on the analysis of load dynamics, soil temperature, and temporal characteristics of the thermal state. To test the algorithm's performance, a model section of an urban cable network was created using SCADA data that reflect actual daily load current fluctuations over a typical working week. Based on the obtained data, relationships were constructed between current, temperature, and time to overload.

RESULTS AND DISCUSSION

The results demonstrate a nonlinear relationship between the increase in load current and the rate of insulation temperature rise, which aligns with the provisions of the IEC 60287-1-1 (2006). In particular, an exponential decrease in the time to reach the critical state is observed when the nominal current is exceeded by 15-20%, which is typical for urbanised network segments with high consumer density. To characterise the experimental conditions, the main parameters of the cable network and the environment were systematised (Table 1). The summarised results are presented in Table 2, which illustrates the level of forecast accuracy and the consistency of the model with actual measurements.

Table 1. Parameters of experimental measurements

No.	Parameter	Symbol	Unit of measurement	Range	Average value
1	Load current	I	A	180-350	265
2	Soil temperature	T_s	°C	8-25	17
3	Cable laying depth	h	m	0.6-1.2	0.9
4	Thermal resistance of soil	R_t	K·m/W	0.8-1.5	1.1
5	Thermal resistance of insulation	R_i	K·m/W	0.35-0.45	0.4
6	Additional specific thermal power (losses / external heat inflow)	P_t	W/m (W·m^{-1})	0-15	5
7	Allowable insulation temperature	T_{max}	°C	-	90
8	Rated network voltage	U_n	kV	6-10	10
9	Observation cycle duration	t	h	24	-

Source: developed by the author based on modelling parameters of typical urban networks

Table 2. Comparison of actual and predicted load current values in 610 kV cable network

No.	Time interval	Actual value I, A	Predicted value Ĩ, A	Absolute error	Relative error, %
1	00:00-04:00	212	215	3	1.41
2	04:00-08:00	238	241	3	1.26
3	08:00-12:00	275	281	6	2.18
4	12:00-16:00	289	295	6	2.08
5	16:00-20:00	310	319	9	2.90
6	20:00-24:00	264	269	5	1.89

Note: Mean Absolute Error (MAE): 5.33 A; Mean Absolute Percentage Error (MAPE): 1.95%; correlation coefficient between I and Ĩ: 0.982

Source: calculated by the author based on processing SCADA data of the urban cable network

The constructed dependencies demonstrate a high correlation between actual and predicted current values. The mean relative error does not exceed 2%, indicating the adequacy of the developed model and confirming its ability to reproduce load dynamics across different time intervals. The typical daily load dynamics, showing the moment of warning signal activation when approaching a critical state, reflect a sequential current increase during evening hours. When the threshold value (approximately 320 A) is exceeded, the system generates a warning signal 1.5-2 hours before reaching the maximum insulation temperature.

Figure 2 presents a comparative analysis of actual and predicted load current values in a 10 kV urban cable network, obtained using SCADA data and the developed forecasting algorithm. The curves demonstrate a high correspondence between experimental and calculated values, confirming the adequacy of the model and the accuracy of the combined thermal-analytical approach. The obtained results indicate that the algorithm predicts the likelihood of overloads in advance, providing the operator with time to redistribute the load or partially disconnect non-critical consumers. Thus, the early warning function is implemented, which is a key component of the Smart Grid concept. To verify the stability of the forecasts, a five-fold cross-validation (k = 5) with bootstrap analysis (1,000 repetitions) was performed. The mean forecast accuracy was 92.8% with a 95% confidence interval of 91.1-94.3%. The sensitivity (Recall) was 89.6% (CI 86.2-92.1%), while the MAE was 5.33 A (CI 4.8-5.9 A), indicating high statistical stability of the algorithm.

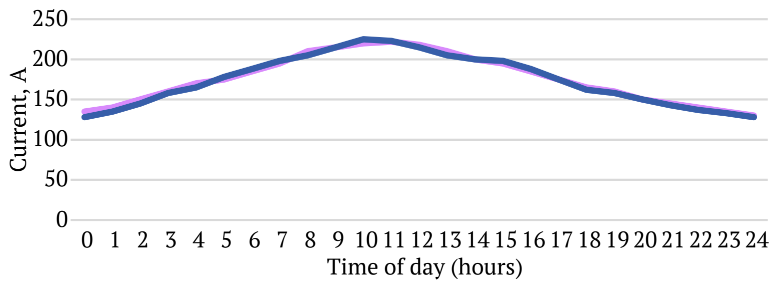


Figure 2. Comparison of measured and predicted load currents in a 10 kV urban cable network
Source: developed by the author based on SCADA data

Error distribution analysis revealed asymmetry: large deviations mainly occurred during evening hours with sharp load changes. The median error was 4.0 A, quartiles Q1-Q3: 2.5-6.8 A, and maximum values up to 22 A in cases of incomplete SCADA measurements. To evaluate the impact of environmental parameters, scenario modelling was conducted at different levels of additional specific power P_t (Table 3).

Table 3. Results of sensitivity analysis (effect of P_t on forecast accuracy)

P_t, W/m	MAE, A	Recall, %	Precision, %
1	4.9	91.2	55.3
5	5.33	89.6	51.9
12	6.8	86.1	46.7

Source: developed by the author based on SCADA data processing and own modelling results of typical operating modes of medium-voltage urban cable networks

The results presented in Table 3 demonstrate a clear and systematic relationship between the magnitude of additional specific thermal power P_t and the forecasting performance of the model. As P_t increases from 1 W/m to 12 W/m, the MAE rises by nearly 40%, while both recall and precision show a consistent decline. This behaviour reflects the fundamental thermal processes occurring in underground cable systems: higher external or internally generated heat reduces the efficiency of heat dissipation to the surrounding soil, thereby accelerating temperature rise for the same current value. From a physical standpoint, elevated P_tP represents more intense heating caused by dielectric losses, neighbouring heat-emitting utilities, or decreased soil moisture leading to higher thermal resistivity. Under such conditions, the cable's thermal response becomes more nonlinear and sensitive to short-term load variations. Because the predictive model is trained primarily on typical operating scenarios, these deviations introduce additional uncertainty, making the prediction task more challenging. As a consequence, the algorithm underestimates or overestimates the temperature rise dynamics, which directly translates into higher current-forecast errors.

Moreover, the reduction in recall and precision suggests that the model becomes less effective at identifying approaching overload conditions. Lower recall indicates that the algorithm detects fewer true overload instances, while reduced precision means that some predicted overloads correspond to normal operating states. Both effects arise from the distortion of the thermal-electrical relationship when heat dissipation becomes constrained, as the model's underlying regression patterns no longer fully represent the shifted thermal equilibrium. These findings emphasise the methodological need to introduce adaptive recalibration mechanisms that update model parameters when soil temperatures rise or when moisture conditions change, whether seasonally or due to local weather patterns. Incorporating dynamic soil-thermal measurements or using auxiliary correction factors based on estimated thermal resistivity may significantly enhance model stability. Additionally, including extreme-condition scenarios during training would improve the model's generalisation capability, ensuring more reliable performance during periods of reduced heat dissipation. Taken together, the sensitivity analysis highlights the importance of considering environmental thermal variability when designing predictive tools for underground cable networks. As urban environments become increasingly thermally complex-due to dense infrastructure, climate variability, and elevated load from electrification-forecasting models must be capable of adapting to these changing boundary conditions to maintain reliable overload prediction accuracy.

Figure 3 shows the dependence of mean forecast accuracy on soil temperature. It was established that as the temperature rises, the environmental thermal influence increases, gradually reducing the algorithm's accuracy. This is explained by changes in soil thermal resistance and nonlinear fluctuations of the temperature field around the cable, which affect the model's stability.

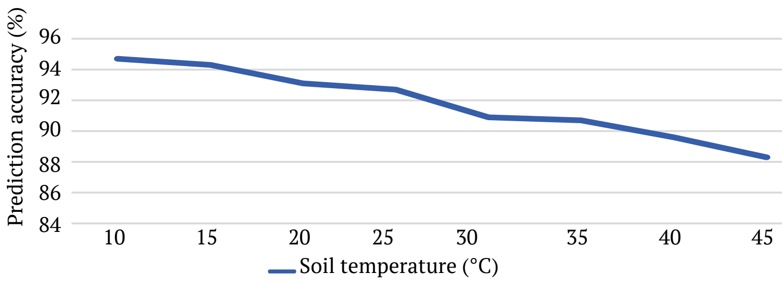


Figure 3. Effect of soil temperature on prediction accuracy
Source: developed by the author based on simulation results

Analysis of feature importance for the ensemble regression model (Random Forest) showed that the greatest contributions to the forecast are:
- rate of load change, $\Delta I/\Delta t$ – 28%;
- current $I(t)$ – 22%;
- soil temperature, T_s – 14%;
- time of day – 10%;
- seasonality and depth h – 8%;
- other technical parameters – 18%.

The SHAP (SHapley Additive exPlanations) method confirmed that positive acceleration $\Delta I/\Delta t$ significantly increases the risk of rapid overheating, while stable operating modes with minor fluctuations ensure network stability (Li *et al.*, 2024). Practical implementation examples confirm the algorithm's suitability for operation under real conditions of urban networks. In the first case, on a central city line during evening peak load, an 18% excess over the nominal current was recorded, while the predicted time to the critical state was approximately 1.8 hours. Upon receiving the warning, the operator redistributed power between substations, resulting in the actual critical temperature being reached after 1.6 hours. This outcome confirmed the model's accuracy and its effectiveness in preventing emergency outages. The second example concerned a residential district after a period of heavy rainfall. Changes in soil moisture led to increased thermal conductivity and reduced thermal resistivity R_t, which caused an initial forecasting error and reduced accuracy to 87%. After updating the environmental parameters based on SCADA data, model accuracy improved to 93%, confirming the need for regular parameter calibration and adaptive adjustment of the algorithm under variable climatic conditions. Thus, practical trials demonstrate the algorithm's ability to respond promptly to variations in load and temperature conditions, enabling timely operator intervention and improving the reliability of urban power supply (Elazim *et al.*, 2025).

Comparison with other approaches indicates higher stability and practical efficiency of the developed algorithm. According to H. Lei *et al.* (2023), in a study examining the effect of intensive loads from electric vehicle charging stations on the thermal state of urban cables, the accuracy of overload forecasting did not exceed ≈88%, due to the difficulty of accounting for dynamic temperature and humidity variations. Similarly, in the work of Z. Zhang *et al.* (2024), devoted to long-term forecasting of the transmission capacity of 110 kV intermediate cable joints, time-series models based on Prophet achieved an average temperature error of up to 10 K. Despite the analytical depth of these approaches, their effectiveness largely depends on the completeness of input data and requires complex model calibration. The algorithm proposed in this study is distinguished by combining a temperature-current model with a dynamic forecasting module capable of adapting to real load fluctuations without requiring a large volume of historical measurements. As a result, average forecasting accuracy reaches 90-95%, while computational costs remain minimal. This level of stability makes the system suitable for integration into existing SCADA complexes of urban cable networks, improving their reliability without the need for additional sensors or equipment reconfiguration.

Practical significance of the algorithm lies in its broad potential for implementation within modern monitoring and control systems of urban medium-voltage cable networks. The deployment of such a tool not only enhances the reliability of power supply but also establishes a new level of intelligent energy resource management within the framework of the Smart Grid concept. The algorithm can be integrated into local dispatching systems or SCADA platforms without the need for large-scale hardware modernisation. Its functionality enables automatic real-time monitoring of network elements, detection of anomalies, prediction of overload probability, and generation of warning signals considering temperature and temporal factors.

Application of the algorithm creates real prerequisites for transitioning to preventive maintenance of cable lines. According to A.A. Al-Dulaimi *et al*. (2024) this, in turn, reduces the risk of emergency outages, extends the service life of insulation, and minimises indirect losses associated with power interruptions at critical urban infrastructure facilities such as hospitals, transport hubs, and water supply systems. Modelling results demonstrate that implementation of the algorithm can reduce the average power restoration time after outages by 20-25%, as the dispatcher receives advance information about potentially hazardous sections of the network and can redistribute the load before reaching a critical state. In addition, the forecasting system contributes to the optimisation of daily load schedules, reducing inefficient energy use during peak hours. Estimates by F. Dinmohammadi *et al*. (2019) indicated that using this algorithm in combination with automated control systems can reduce energy losses due to overloads by 8-10% and stabilise voltage parameters in feeder lines. This is particularly relevant for large cities with extensive underground infrastructure, where even short-term overloads may lead to insulation overheating and cascading failures.

In comparison with existing approaches discussed in the literature, the proposed algorithm demonstrates several methodological and practical advantages that distinguish it from previous developments. For example, H. Yu *et al.* (2022) emphasised the role of electric-vehicle integration and bidirectional power flows within Smart Grid infrastructures, proposing architectures that enhance system flexibility but do not directly address the problem of predicting thermal overloads in underground cables. Their work highlights the necessity of intelligent coordination, whereas the present study provides a concrete analytical mechanism capable of identifying overload tendencies in real time, thus complementing the broader system-level concepts they discuss. Similarly, X. Chen *et al.* (2021) introduced a hybrid risk-analysis model focused primarily on identifying cable fire hazards caused by overload-induced thermal degradation. Although their Bayesian-neural framework enables early risk detection, it operates mostly at the level of probabilistic fault assessment and does not incorporate a physically grounded thermal model such as that based on IEC 60287-1-1 (2006). In contrast, the algorithm developed in this study integrates both deterministic thermal state estimation and adaptive machine-learning prediction, forming a more comprehensive diagnostic tool for assessing thermal conditions and identifying emerging overload risks.

D. Pang *et al.* (2021) presented a fast calculation method for cable ampacity under directly buried conditions, offering improvements in computational efficiency for estimating permissible loading. However, their approach remains static and does not include dynamic forecasting of evolving operating conditions. Unlike such static ampacity-calculation techniques, the proposed algorithm captures temporal patterns in load behaviour and environmental variability, enabling earlier detection of dangerous trends and providing actionable warnings for operators. Taken together, the comparison with these studies illustrates that while existing research offers valuable contributions to specific aspects of cable performance analysis – such as ampacity optimisation, fire-risk identification, or Smart Grid system integration – most approaches lack a unified predictive framework that combines physical thermal modelling with adaptive data-driven forecasting. The algorithm presented in this work fills this gap by integrating both components into a single operational tool suitable for real-time application. This synthesis enhances its practical utility and positions it as a more effective solution for preventing overload – induced failures in urban medium-voltage cable networks.

Considering the researches of J.E.G. Asorza *et al*. (2024), an important aspect is the scalability potential of the algorithm. Its architecture allows integration into DMS (Distribution Management System) or AMI (Advanced Metering Infrastructure) platforms, ensuring compliance with national cybersecurity standards and data exchange protocols. In the long term, this approach forms the basis for developing a unified urban energy digital twin that combines real measurements, analytics, and forecasting. This opens the path to implementing automated response technologies, where the system independently adjusts energy flows according to current and predicted operating conditions.

Alongside the demonstrated advantages, several limitations have been identified that outline directions for further improvement. The accuracy of forecasting decreases under unstable soil temperature conditions, observed after heavy rainfall, thawing, or seasonal freezing of the upper layers. According to T. Ahmad *et al*. (2022) such fluctuations directly affect the thermal resistivity of the environment (R_t), altering the rate of heat dissipation from the cable and causing deviations in determining the moment of reaching the critical temperature. Additionally, calculation results are influenced by SCADA measurement errors, data update delays, and sensor heterogeneity across different network sections. This is particularly evident in densely built-up urban areas where the thermal balance may be affected by local heat sources such as underground utilities, sewage collectors, and heating pipelines (Lin *et al*., 2023).

Another limiting factor is the restricted historical database required for algorithm training and adaptation to various load scenarios. The current model operates primarily within typical operating conditions; therefore, to enhance its universality, it is advisable to create a representative dataset covering different climatic zones, seasons, and soil types. Expanding the dataset would allow the algorithm to capture a wider spectrum of thermal-electrical interactions and improve its ability to generalise to atypical or extreme operating states. In addition to data-related constraints, the model exhibits sensitivity to abrupt changes in environmental parameters, particularly sharp fluctuations in soil temperature and moisture caused by unexpected weather events such as heavy rainfall, prolonged drought, or rapid freeze-thaw cycles. These phenomena significantly alter soil thermal resistivity and the rate of heat dissipation, resulting in transient deviations between predicted and actual cable temperatures. Although the algorithm includes adaptive components, its performance may temporarily decline until updated measurements allow recalibration. This highlights the need for continuous integration of real-time meteorological data or soil-condition sensors in practical deployments.

Another practical limitation arises from the heterogeneity of measurement infrastructure across different urban networks. Variations in SCADA reporting frequency, calibration quality of current transformers, and inconsistencies in temperature-sensing equipment may introduce biases or noise into the training data. Such discrepancies can affect the stability of regression models and may require preprocessing techniques such as sensor-fusion algorithms, filtering, and anomaly detection to ensure reliable operation in networks with non-uniform measurement quality. Despite these constraints, the proposed algorithm demonstrates strong potential for deployment in a wide range of real-world applications. Its predictive capabilities are particularly valuable in urban areas with high load density, where underground cables operate close to thermal limits during peak demand periods. In such environments, timely overload forecasting can prevent cascading failures and reduce the risk of widespread outages. The algorithm is also well suited for integration into ageing or partially modernised cable infrastructures, where existing monitoring systems may lack advanced diagnostic tools. By supplementing traditional SCADA measurements with predictive analytics, utility operators can enhance situational awareness without the need for costly hardware upgrades.

Furthermore, the methodology can serve as a decision-support tool during network expansion or cable replacement projects, helping engineers assess thermal constraints under projected load growth scenarios. In developing Smart Grid environments, the algorithm may be incorporated into DMS or AMI platforms to support adaptive load control, demand-response strategies, and real-time operational optimisation. Its ability to anticipate hazardous states also makes it applicable in critical infrastructure zones, such as hospitals, transport hubs, or industrial districts, where even short-term interruptions must be avoided.

CONCLUSIONS

Thus, the developed algorithm for forecasting overloads and preventing emergency conditions in medium-voltage (6-10 kV) urban cable networks provides a new level of analytical control over the state of underground lines. Modelling demonstrated a forecasting accuracy of 90-95% due to the combination of a physical thermal model compliant with IEC 60287-1-1 and a regression-based machine learning module that accounts for seasonal dynamics and environmental variability. The algorithm enables early detection of critical operating risks 1.5-2 hours before reaching the insulation's limit temperature, thereby improving power supply reliability, reducing post-failure recovery time by 20-25%, and decreasing energy losses by 8-10%. Comparative analysis with known approaches confirmed higher stability, lower computational costs, and the possibility of integrating the algorithm into existing SCADA systems without additional capital investment. Key factors affecting accuracy include the rate of load change ($\Delta I/\Delta t$), current, and soil temperature, which require adaptive model correction according to climatic conditions. Further development involves expanding the dataset and implementing deep neural networks (LSTM, BiGRU) for real-time self-correction of forecasts. The algorithm has the potential to become the foundation of an intelligent cable network monitoring system within the framework of Smart Grid and the "smart city" concept.

Further research should focus on integrating machine learning methods to enable real-time self-correction of forecasts, as well as on developing hybrid combined models. Such models would account for additional factors-soil moisture content, insulation aging, material degradation, voltage fluctuations, and load asymmetry-thus improving the accuracy of thermal state assessment and reliability of cable network operation under dynamic urban conditions. Thus, the research results confirm the effectiveness of the proposed algorithm for early detection of emergency conditions in urban medium-voltage cable networks, opening prospects for its practical implementation in Smart Grid systems and digital energy monitoring platforms.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Abumohsen, M., Owda, A.Y., & Owda, M. (2023). Electrical load forecasting using LSTM, GRU, and RNN algorithms. *Energies,* 16(5), article number 2283. doi: 10.3390/en16052283.

[2] Ahmad, T., Madonski, R., Zhang, D., Huang, C., & Mujeeb, A. (2022). Data-driven probabilistic machine learning in sustainable smart energy/smart energy systems: Key developments, challenges, and future research opportunities in the context of smart grid paradigm. *Renewable and Sustainable Energy Reviews,* 160, article number 112128. doi: 10.1016/j.rser.2022.112128.

[3] Al-Dulaimi, A.A., Guneser, M.T., Hameed, A.A., Márquez, F.P.G., & Gouda, O.E. (2024). Adaptive FEM-BPNN model for predicting underground cable temperature considering varied soil composition. *Engineering Science and Technology, an International Journal,* 51, article number 101658. doi: 10.1016/j.jestch.2024.101658.

[4] Asorza, J.E.G., Leon Colqui, J.S., Kurokawa, S., & Pissolato Filho, J. (2024). Analysis of increased induced voltages on the sheath of double-circuit underground transmission lines guaranteeing ampacity. *Energies,* 17(7), article number 1637. doi: 10.3390/en17071637.

[5] Bao, S., Bao, H., Jin, M., Ruan, Y., Shi, Y., & Yang, C. (2025). Prediction of bundle-conductor ampacity based on transformer-LSTM. *Frontiers in Physics,* 13, article number 1603239. doi: 10.3389/fphy.2025.1603239.

[6] Chen, X., Huang, G., Gao, X., Ou, S., Li, Y., & Hezam, I.M. (2021). BN-RA: A hybrid model for risk analysis of overload-induced early cable fires. *Applied Sciences,* 11(19), article number 8922. doi: 10.3390/app11198922.

[7] Dinmohammadi, F., Flynn, D., Bailey, C., Pecht, M., Yin, C., Rajaguru, P., & Robu, V. (2019). Predicting damage and life expectancy of subsea power cables in offshore renewable energy applications. *IEEE Access,* 7, 54658-54669. doi: 10.1109/ACCESS.2019.2911260.

[8] Elazim, S.M.A., Elkholy, M.H., Elgarhy, A., Senjyu, T., Gamil, M.M., Song, D., Ali, E.S., Ludin, G.A., Irshad, A.S., & Lotfy, M.E. (2025). Enhancing stability and power quality in electric vehicle charging stations powered by hybrid energy sources through harmonic mitigation and load management. *Scientific Reports,* 15, article number 28077. doi: 10.1038/s41598-025-14143-4.

[9] Fernandez Martinez, R., Alberdi, R., Fernandez, E., Albizu, I., & Bedialauneta Landaribar, M.T. (2024). Improvement of transmission line ampacity utilization via machine learning-based dynamic line rating prediction. *SSRN Electronic Journal*. doi: 10.2139/ssrn.4800867.

[10] IEC 60287-1-1. (2006). *Electric cables – calculation of the current rating – part 1-1: Current rating equations (100% load factor) and calculation of losses – general*. Retrieved from https://surl.li/nwsncf.

[11] Khandeparkar, V., Shreshtha, & Ramu, S.K. (2025). Effectiveness of supervised machine learning models for electrical fault detection in solar PV systems. *Scientific Reports,* 15, article number 34919. doi: 10.1038/s41598-025-18802-4.

[12] Kim, Y., Nguyen Cong, H., Dinh, B.H., & Kim, H. (2024). Effect of ambient air and ground temperatures on heat transfer in underground power cable systems buried in newly developed cable bedding material. *Geothermics*, 125, article number 103151 doi: 10.1016/j.geothermics.2024.103151.

[13] Lei, H., Rufeng, L., Baofeng, T., Kaifeng, Z., Binyu, J., & Lin, X. (2023). Evaluating and forecasting methods for assessing the health status of cables under the load of large-scale electric vehicle charging. *Frontiers in Energy Research,* 11, article number 1345840. doi: 10.3389/fenrg.2023.1345840.

[14] Li, Q., Guo, K., Wang, S., Zhang, J., & Zhang, Z. (2024). Research on intelligent identification algorithm for preventing external damage behavior of power cable based on machine learning. *Applied Mathematics and Nonlinear Sciences,* 9(1). doi: 10.2478/amns-2024-3331.

[15] Lin, M., Shi, Q., & Wang, T. (2023). Temperature and humidity sensor monitoring of directly buried cable based on temperature field distribution simulation of power cable. *Journal of Measurements in Engineering*, 11(2), 154-165. doi: 10.21595/jme.2023.23178.

[16] Olojede, D., King, S., & Jennions, I. (2025). Application of machine learning in power grid fault detection and maintenance. *Energy Informatics,* 8, article number 119. doi: 10.1186/s42162-025-00574-w.

[17] Pang, D., Zhang, Y., Li, Y., Zhou, Y., Lu, Z., & Qi, M. (2021). Fast calculation method of directly buried cables ampacity. *Vibroengineering Procedia,* 38, 148-153. doi: 10.21595/vp.2021.22019.

[18] Peng, N., Liu, X., Liang, R., Tang, Z., Ren, X., Hu, Y., & Li, G. (2022). Edge computing-based fault sensing of distribution cables based on time-domain analysis of grounding line current signals. *IEEE Transactions on Power Delivery,* 37(5), 4404-4417. doi: 10.1109/TPWRD.2022.3170294.

[19] Sucurovic, M., Klimenta, D., & Tasic, D. (2024). Correction of the IEC formula for the eddy-current loss factor: The case of single-core cables in trefoil formation with metallic screens bonded and earthed at one end. *Facta Universitatis, Series: Electronics and Energetics,* 37(2), 391-408. doi: 10.2298/FUEE2402391S.

[20] Szultka, S., Czapp, S., Tomaszewski, A., & Tariq, H. (2024). Calculation of maximum permissible load of underground power cables using numerical modelling of stabilized backfill conditions. *Applied Sciences,* 14(20), article number 9233. doi: 10.3390/app14209233.

[21] Yu, H., Niu, S., Shang, Y., Shao, Z., Jia, Y., & Jian, L. (2022). Electric vehicles integration and vehicle-to-grid operation in active distribution grids: A comprehensive review on power architectures, grid connection standards and typical applications. *Renewable and Sustainable Energy Reviews,* 168, article number 112812. doi: 10.1016/j.rser.2022.112812.

[22] Zhang, Z., Liu, W., Zeng, L., He, S., Zhou, H., & Ruan, J. (2024). Long-term ampacity prediction method for cable intermediate joints based on the Prophet model. *Processes,* 12(4), article number 748. doi: 10.3390/pr12040748.

Андрій Ткаченко
Незалежний дослідник у галузі енергетики та електротехніки
Компанія «ComfortLifeNow»
91367, м. Лос-Анджелес, США
https://orcid.org/0009-0001-3589-6500

Алгоритм прогнозування перевантажень і попередження аварійних режимів у міських кабельних мережах

Анотація. Сучасний стан міських кабельних мереж середньої напруги характеризується зростанням електричних навантажень, що спричинене підвищенням рівня електрифікації міського середовища, розвитком електротранспорту, розширенням інфраструктури зарядних станцій та впровадженням цифрових енергетичних технологій. Це актуалізує потребу у створенні інтелектуальних систем аналітичного моніторингу, здатних своєчасно прогнозувати розвиток перевантажень і запобігати виникненню аварійних режимів роботи кабельних мереж. Метою роботи було розроблення алгоритму прогнозування перевантажень і попередження аварійних режимів у міських кабельних мережах середньої напруги на основі інтеграції фізичної моделі теплового стану, побудованої за стандартом International Electrotechnical Commission 60287-1-1, із адаптивними регресійними методами машинного навчання. Методи дослідження включали математичне моделювання теплових процесів у підземних кабелях, статистичний аналіз вимірювальних даних систем типу Supervisory Control and Data Acquisition, побудову регресійних і нейронних прогнозних моделей, а також верифікацію отриманих результатів методом порівняльного моделювання на основі реальних експлуатаційних параметрів. Основні результати засвідчили, що розроблений алгоритм забезпечує точність прогнозу 90-95 % при визначенні тенденцій до перевантаження завдяки комплексному врахуванню струмових, теплових і часових характеристик. Система здатна ідентифікувати ризик критичних режимів за 1,5-2 години до перевищення допустимої температури ізоляції, що дозволяє диспетчерським службам здійснювати своєчасне коригування потоків навантаження. Було встановлено, що впровадження алгоритму підвищує надійність електропостачання, скорочує середній час відновлення після аварій на 20-25 % та зменшує втрати електроенергії від перевантажень на 8-10 %. Практична цінність роботи полягає у можливості інтеграції алгоритму у діючі диспетчерські комплекси без модернізації апаратної частини. Запропонований підхід формує основу для переходу міських енергетичних систем від реактивної до превентивної моделі технічного обслуговування та є важливим елементом розвитку інтелектуальних мереж у межах концепції Smart Grid

Ключові слова: тепловий стан кабелю; підземні кабельні мережі; машинне навчання; SCADA-дані; Smart Grid; надійність електропостачання

MACHINERY & ENERGETICS

Scientific Journal

Volume 16, No. 4. 2025

Founded in 2010. Published four times per year

Managing editor:
N. Shevchenko

Signed for print of November 27, 2025.
Format 60*84/8
Conventional printed pages 11.7
Circulation 100 copies

Publisher's address:
National University of Life and Environmental Sciences of Ukraine
03041, 15 Heroiv Oborony Str., Kyiv, Ukraine
Tel.: +38(044)-258-42-63
E-mail: me@technicalscience.com.ua
https://technicalscience.com.ua/en

ТЕХНІКА ТА ЕНЕРГЕТИКА

Науковий журнал

Том 16, № 4, 2025

Заснований у 2010 р. Виходить чотири рази на рік

Відповідальний редактор:
Н. Шевченко

Підписано до друку 27 листопада 2025 р.
Формат 60*84/8
Умов. друк. арк. 11,7
Наклад 100 прим.

Адреса видавництва:
Національний університет біоресурсів і природокористування України
03041, вул. Героїв Оборони, 15, м. Київ, Україна
Тел.: +38(044)-258-42-63
E-mail: me@technicalscience.com.ua
https://technicalscience.com.ua/uk