ISSN
2822-4450

AINTELIA SCIENCE NOTES

Volume 3, Issue 2
December 2024

2024
Volume: 3
Issue: 2

AIntelia

AIntelia Sciences Notes

Volume 3, Issue 2

December , 2024

Editor

İlyas Özer

Bandırma Onyedi Eylül University

Bandırma, Balıkesir, Turkey

More information about this series at https://sciencenotes.com.tr/

Contents

Digital Twin Framework for Real-Time Structural Health Monitoring and Flood Resilience in Urban Smart Infrastructure

Wahbiyah Fatinah Kassab[1], Mihran Rushd Gerges[2]

[1] *National Engineering School of Gabes, Zrig Eddakhlania, Tunisia, wahbiyah.kassab@enig.rnu.tn, ORCID: 0009-0004-6224-5297*

[2*] *Minia University, Minia, Egypt, mihran.gerges@mu.edu.eg, ORCID: 0009-0004-4256-1452*

Urban infrastructure systems face compounding threats from structural deterioration and climate-driven flood events. This paper proposes an integrated Digital Twin (DT) framework that couples real-time Structural Health Monitoring (SHM) with urban flood resilience management for smart city environments. The framework leverages a heterogeneous Internet of Things (IoT) sensor network — comprising accelerometers, strain gauges, piezometers, and water-level sensors — feeding continuously into a cloud-hosted digital replica of urban built assets. A hybrid machine learning pipeline combining Long Short-Term Memory (LSTM) autoencoders for anomaly detection with ensemble Kalman filtering for real-time model state updating is presented. The flood resilience module integrates hydrodynamic simulation with sensor-driven early-warning logic, enabling proactive emergency response. Architecture components, data flows, computational requirements, and implementation challenges are systematically described. The proposed framework provides a replicable, scalable foundation for data-driven urban resilience, directly applicable to bridges, drainage networks, and critical urban assets. Key contributions include a unified ontology linking structural and hydrological digital twin layers, a latency-optimised edge-cloud processing pipeline, and a multi-criteria decision support module for infrastructure managers.

Keywords: *digital twin; flood resilience; smart city; urban infrastructure*

© 2024 Published by *AIntelia*

1. Introduction

The convergence of accelerating urbanisation, ageing civil infrastructure, and the intensification of extreme weather events driven by climate change has created a critical imperative for intelligent, adaptive infrastructure management systems. Urban areas concentrate population, economic activity, and interdependent critical systems in ways that amplify both the probability and consequences of structural failures and flood events (Alibrandi, 2022). Traditional inspection-based approaches — relying on periodic visual assessment and post-hoc damage surveys — are increasingly inadequate for managing the dynamic risks facing modern cities.

Structural Health Monitoring (SHM) has evolved substantially over the past two decades from point-sensor campaigns to continuously operating, data-rich sensor networks capable of delivering real-time structural performance data (Malekloo et al., 2022). In parallel, digital twin (DT) technology has emerged as a powerful paradigm for creating live, continuously updated virtual replicas of physical systems, integrating sensor data, computational models, and decision-support logic into a unified cyber-physical architecture (Liu et al., 2023). The application of DTs to civil infrastructure remains, however, fragmented: structural monitoring and flood risk management have largely developed as separate domains with separate data architectures and stakeholder communities.

Flood events represent one of the most damaging and frequently occurring natural hazards globally. According to the 2020 global natural disaster assessment, floods were the most frequent disaster type, affecting more than 200 countries (Tanoue et al., 2023). Urban drainage infrastructure, frequently under-dimensioned against projected future precipitation intensities, constitutes a critical vulnerability node. The integration of flood early-warning systems with real-time structural assessment within a shared digital twin architecture offers significant untapped potential for urban resilience.

This paper addresses this gap by proposing a unified DT framework that integrates real-time SHM and urban flood resilience management. The framework is designed for scalability across heterogeneous urban asset types and for interoperability with existing smart city data platforms. The contributions of this work are threefold: (i) a formal architectural specification for a hybrid structural-hydrological digital twin; (ii) a machine learning pipeline for real-time anomaly detection and model state updating; and (iii) a multi-criteria decision support module linking structural condition assessment with flood risk scenarios.

The remainder of the paper is organised as follows. Section 2 reviews the relevant literature. Section 3 presents the proposed framework architecture. Section 4 describes the machine learning and data assimilation components. Section 5 discusses the flood resilience integration. Section 6 addresses implementation challenges and scalability considerations. Section 7 concludes with directions for future research.

2. Literature Review

The concept of the digital twin, originally formulated in the context of aerospace manufacturing by Grieves (2014), has been progressively adapted to the built environment over the past decade. In civil infrastructure applications, a DT is defined as a dynamic, bi-directional digital representation of a physical asset that is continuously synchronised with its physical counterpart through sensor data streams, enabling real-time condition assessment, predictive analytics, and decision support (Liu et al., 2023).

Building Information Modelling (BIM) has provided the geometric and semantic foundation for many DT implementations in the construction domain. Baghalzadeh Shishehgarkhaneh et al. (2022) conducted a bibliometric review of IoT, BIM, and DT integration in the construction industry, identifying accelerating research output and highlighting the growing importance of interoperability standards. Li et al. (2022) reviewed advanced information technologies in civil infrastructure construction and maintenance, emphasising the role of sensor fusion and data analytics platforms. The integration of BIM with IoT sensing and structural analysis into a coherent DT framework for SHM was demonstrated by Hu et al. (2024), who developed a wireless sensing system coupled to a finite-element updating pipeline.

A key enabler of operational DTs is the availability of standards for data exchange and model interoperability. Liu et al. (2023) reviewed DT technologies for civil infrastructure through the lens of the Journal of Infrastructure Intelligence and Resilience, documenting the fragmented standards landscape and calling for unified ontologies. Kong and Hucks (2023) presented a photogrammetry-based DT framework specifically tailored to historic structures, demonstrating the extension of DT principles beyond conventional engineered systems.

Contemporary SHM systems deploy diverse sensor modalities — accelerometers, strain gauges, displacement sensors, fibre Bragg gratings, and increasingly vision-based systems — in permanently installed configurations that generate continuous data streams (Malekloo et al., 2022). The assimilation of these data into computational models for damage detection, localisation, and severity estimation has been the central research challenge in the field for two decades.

Machine learning approaches have transformed SHM practice, enabling the extraction of damage-sensitive features from high-dimensional sensor data without explicit physical model assumptions. Malekloo et al. (2022) provided a comprehensive overview of machine learning in SHM, documenting the shift from supervised classification approaches requiring labelled damage data towards unsupervised anomaly detection methods suitable for operational infrastructure where damage examples are rarely available. LSTM neural networks have demonstrated particular efficacy for temporal anomaly detection in continuously monitored structures due to their capacity to learn complex temporal dependencies in vibration and strain response data (Rizvi et al., 2023).

Autoencoders — neural networks trained to reconstruct their input from a compressed latent representation — have become a dominant paradigm for unsupervised SHM anomaly detection. Reconstruction errors elevated above a learned baseline threshold provide a principled, data-driven damage indicator (Pan et al., 2023). The integration of railway bridge SHM into a digital twin architecture using low-cost wireless IoT sensors and clustering-based machine learning was demonstrated in a real operational deployment, confirming the practical viability of this approach (Bado et al., 2024). Sakr and Sadhu (2023) further demonstrated the visualisation of SHM information within BIM-integrated IoT platforms.

The application of digital twin principles to urban flood management has accelerated substantially in response to growing recognition of urban flood risk as a climate adaptation priority (Tanoue et al., 2023). Urban Digital Twins for flood applications typically integrate hydrodynamic simulation models with real-time sensor networks, satellite remote sensing, and meteorological forecast data to enable dynamic flood inundation mapping and early-warning functionality (Ferré-Bigorra et al., 2022).

The DestinEarth initiative launched by the European Commission in 2022 represents the most ambitious publicly funded effort to develop a high-resolution digital twin of the Earth system for climate resilience applications, including urban flood simulation. Tanoue et al. (2023) explored the potential of DT, 3D city models, and sensor networks for climate resilience in smart cities, conducting a systematic review of 68 articles across three databases and documenting the maturation of early-warning system development. Backhaus and Stamm (2023) presented a city-scale urban DT for flood risk management in Dresden, Germany, integrating a 3D web application with a comprehensive urban data platform supporting automated generation of 2D hydro-numerical models.

The integration of machine learning with physics-based hydrodynamic simulation has emerged as a key approach for balancing computational efficiency with predictive accuracy (Ferré-Bigorra et al., 2022). Ensemble Kalman filtering provides a mathematically rigorous framework for assimilating real-time sensor observations into running simulation models, enabling continuous state estimation under uncertainty (Lei et al., 2023). Yuan et al. (2022) reviewed community-scale big data sources — including infrastructure sensor data, mobility data, and crowdsourced feeds — as inputs to smart flood resilience systems, identifying machine learning and deep learning as key analytical technologies.

Despite parallel advances in structural DTs and flood risk DTs, the integration of these two domains within a unified cyber-physical architecture remains largely unexplored. Structural monitoring systems typically focus on load effects and material degradation without considering hydrological boundary conditions, while flood management platforms rarely incorporate structural condition data. This decoupling misses critical interactions: flood events exert dynamic hydrostatic and hydrodynamic loads on bridges, retaining walls, and drainage infrastructure; conversely, structural deterioration may reduce flood-critical assets below design performance thresholds. A unified DT framework that simultaneously tracks structural condition and flood risk, sharing sensor infrastructure and computational platforms, would significantly enhance the analytical and decision-support capability available to urban infrastructure managers.

3. Methodology

The proposed framework, designated Urban Infrastructure Digital Twin (UIDT), consists of four principal layers: (i) a Physical Asset Layer comprising the monitored infrastructure and its embedded sensor network; (ii) a Data Acquisition and Communication Layer handling real-time data ingestion and edge processing; (iii) a Digital Twin Core comprising the structural simulation, hydrological model, and machine learning pipeline; and (iv) a Decision Support and Visualisation Layer providing actionable outputs to stakeholders. Figure 1 (conceptual) illustrates the inter-layer data flows and principal computational components.

The UIDT architecture follows a cyber-physical systems design principle (Ferré-Bigorra et al., 2022), wherein bidirectional coupling between the physical and virtual domains is maintained continuously. Physical sensor observations update the digital twin state; DT-generated predictions and alerts are fed back to operational decisions governing physical asset management. This closed-loop architecture distinguishes operational digital twins from conventional asset models.

The physical asset layer encompasses the instrumented infrastructure — bridges, retaining walls, drainage tunnels, and surface water control structures — along with the embedded and deployed sensor network. The sensor suite is designed as a heterogeneous array combining complementary measurement modalities to capture the full spectrum of structural and hydrological state variables relevant to both SHM and flood resilience assessment.

Structural sensors include triaxial micro-electro-mechanical systems (MEMS) accelerometers deployed at key structural nodes for continuous vibration monitoring; resistive and fibre Bragg grating (FBG) strain gauges for direct stress measurement at fatigue-critical cross-sections; linear displacement sensors monitoring joint openings and deflections; and corrosion potential probes for reinforced concrete durability assessment. Environmental sensors include temperature and humidity transducers, precipitation gauges, and barometric pressure sensors whose measurements are essential for separating environmental and structural contributions to monitored responses (Malekloo et al., 2022).

The flood resilience sensor array comprises ultrasonic and pressure-transducer water-level sensors in drainage channels, culverts, and at-risk urban watercourses; soil moisture sensors in embankments and ground adjacent to retaining structures; flow velocity meters in combined sewer overflow structures; and optical and radar-based remote sensing feeds from low-Earth orbit satellites providing areal precipitation and inundation extent data. All sensor nodes are equipped with IoT communication modules transmitting data over LoRaWAN for low-power, long-range telemetry or over cellular 4G/5G links for higher-bandwidth accelerometer and FBG channels (Sakr and Sadhu, 2023).

Raw sensor streams are received by gateway nodes deployed at each monitored asset. Edge processing, implemented on low-power system-on-chip platforms, performs local signal conditioning — anti-aliasing filtering, spike rejection, and timestamp synchronisation — before transmitting compressed data packages to the cloud platform. Edge processing capacity is further used to compute first-order feature statistics (root mean square acceleration, peak strain, water level rate-of-change) that enable low-latency local alerting independent of cloud connectivity, ensuring fail-safe operation under network outage conditions.

The communication architecture follows a three-tier topology: sensor node to local edge gateway (short-range IEEE 802.15.4 or Bluetooth Low Energy); edge gateway to cloud ingestion service (LoRaWAN or cellular); cloud ingestion service to DT processing pipeline (MQTT publish-subscribe protocol with QoS Level 1 delivery guarantee). Data latency from sensor measurement to DT state update is maintained below 5 seconds for water-level and accelerometer channels designated as safety-critical, consistent with operational requirements for urban flood early-warning systems.

The DT Core comprises three tightly coupled computational modules: the Structural Finite Element (FE) Model, the Hydrodynamic Flood Model, and the Machine Learning Inference Engine. The FE model is an as-built, validated representation of each monitored structural asset, implemented in an open-source FE framework and parameterised using material properties derived from construction records, non-destructive testing, and BIM data (Liu et al., 2023; Kong and Hucks, 2023). The FE model is updated continuously using sensor-derived modal parameters through an ensemble Kalman filter assimilation algorithm.

The Hydrodynamic Flood Model implements a two-dimensional shallow-water equation solver for the urban drainage and surface water network, discretised on a high-resolution computational mesh derived from airborne LiDAR survey data and BIM-linked infrastructure geometry. Real-time boundary conditions are provided by the water-level and flow sensor array, with meteorological forecast forcing ingested from national hydrological services. The model produces probabilistic inundation maps at 15-minute forecast horizons, updated at each sensor data assimilation cycle (Backhaus and Stamm, 2023).

The Machine Learning Inference Engine hosts the LSTM autoencoder ensemble responsible for structural anomaly detection, the gradient-boosted regression models for flood impact prediction, and the multi-criteria decision analysis module. All ML models are retrained periodically on accumulated operational data to maintain representativeness of current structural condition and seasonal hydrological regimes.

4. Results and Discussion

The structural anomaly detection pipeline is based on an ensemble of LSTM autoencoders trained on extended baseline periods of sensor data recorded under normal structural conditions. Each autoencoder in the ensemble consists of a bidirectional LSTM encoder that compresses the multivariate sensor time series into a fixed-length latent vector representation, followed by an LSTM decoder that reconstructs the original sequence. The reconstruction error, measured as the mean absolute error between input and reconstructed sequences over a sliding window, constitutes the primary anomaly indicator.

The ensemble approach, combining N = 20 independently initialised autoencoders trained on bootstrap resamples of the baseline dataset, provides uncertainty estimates on the anomaly score in addition to point estimates. Anomaly thresholds are set at the 99th percentile of the reconstruction error distribution observed across the baseline period, separately for each sensor channel and environmental condition stratum. This stratification is essential to prevent false positives arising from temperature-driven reversible changes in structural stiffness — a well-documented confounding factor in SHM (Malekloo et al., 2022; Pan et al., 2023).

Encoder architecture comprises three stacked bidirectional LSTM layers of dimensions [64, 32, 16] with layer normalisation and dropout regularisation (rate 0.2) applied between layers. The bottleneck latent dimension is set to 8 for single-structure deployments. Decoder architecture mirrors the encoder. Training employs the Adam optimiser with an initial learning rate of 1e-3 and cosine annealing scheduling over 200 epochs. Sequence length is set to 1024 samples, corresponding to approximately 10 minutes at 100 Hz sampling for accelerometer channels.

Continuous synchronisation of the FE model with the physical structure is achieved through an ensemble Kalman filter (EnKF) implementation. The EnKF maintains a probabilistic ensemble of N = 100 FE model realisations, each perturbed by samples from the prior uncertainty distribution over structural parameters — primarily elastic moduli, boundary condition stiffnesses, and mass distributions. At each assimilation step, the ensemble is propagated forward in time using the FE model, and the ensemble spread is updated using the Kalman gain matrix computed from the cross-covariance between model-predicted and observed sensor responses.

The EnKF implementation follows the square-root ensemble formulation for numerical stability under small ensemble sizes. Localisation is applied to manage spurious long-range correlations between sensors and model parameters that may arise in large-dimensional model spaces. The computational cost of the EnKF, dominated by N parallel FE model evaluations per assimilation step, is managed through model reduction — the FE model is reduced to a modal subspace representation preserving the lowest 50 vibration modes — enabling near-real-time updating at 1-minute assimilation intervals on commodity cloud computing hardware.

Flood impact on monitored structural assets is predicted through a hybrid physics-machine learning approach. The hydrodynamic model provides probabilistic water-level and flow velocity fields at each asset location at 15-minute forecast horizons. These fields are converted to structural load demands — hydrostatic lateral pressures, scour potential at bridge foundations, and overtopping risk for retaining walls — through empirically calibrated transfer functions derived from physical model tests and historical flood event data.

A gradient-boosted regression model (XGBoost) is trained on historical paired records of flood conditions and observed structural responses to predict the structural response increment expected under forecast flood loading. The predicted structural response is superimposed on the current DT state estimate to produce a conditional anomaly probability — the probability that the structural response will exceed the anomaly detection threshold given the forecast flood loading. This conditional probability drives the flood-structural risk score reported to decision-makers (Yuan et al., 2022; Lei et al., 2023). The flood resilience module integrates the real-time hydrodynamic model outputs with a rule-based early warning logic to generate tiered alert levels for urban infrastructure managers and emergency responders. The alert hierarchy comprises four levels: Level 1 (Watch) triggered by precipitation forecast exceedance of a site-specific threshold; Level 2 (Advisory) triggered by water-level sensor readings exceeding the 2-year return period discharge threshold; Level 3 (Warning) triggered by forecast inundation extent intersecting critical infrastructure; and Level 4 (Emergency) triggered by observed structural anomaly detection coinciding with Level 2 or Level 3 hydrological conditions.

The tiered alert architecture is designed to support progressive mobilisation of emergency resources, reducing both the cost of false alarms and the response latency under genuine threat conditions. Alert dissemination is achieved through integration with national emergency communication systems and direct notification to infrastructure asset managers via the UIDT dashboard application. Alert metadata — including the spatial extent of forecast inundation, the identity and condition score of at-risk structural assets, and the recommended inspection or operational response actions — are included in all alerts above Level 2.

The multi-criteria decision support module synthesises structural condition scores, flood risk indicators, and asset criticality data to generate prioritised intervention recommendations for infrastructure managers. The decision logic implements a weighted scoring approach in which each monitored asset receives a composite risk score as a linear combination of: (i) the current structural anomaly probability (weight 0.40); (ii) the 24-hour flood impact probability (weight 0.35); (iii) the asset criticality index reflecting traffic volume, redundancy, and consequence of failure (weight 0.25).

Assets with composite risk scores exceeding pre-defined intervention thresholds are flagged for immediate inspection, operational restriction, or pre-emptive protective measures. The decision support outputs are presented through an interactive geospatial dashboard displaying the real-time condition state of all monitored assets overlaid on the forecast inundation maps, enabling spatial reasoning about the co-location of structural vulnerability and flood hazard. Dashboard functionality includes time-series visualisation of individual sensor channels, FE model animation, and scenario comparison tools (Ferré-Bigorra et al., 2022; Backhaus and Stamm, 2023).

5. Discussion

A primary challenge in operationalising the UIDT framework is the integration of heterogeneous data sources generated by different asset owners, management systems, and sensor vendors operating under different data standards and communication protocols. Urban infrastructure is typically owned and managed by multiple entities — national highway authorities, local government drainage departments, utility companies — each maintaining separate asset management systems with limited interoperability. The UIDT framework addresses this through a standardised data ingestion layer that implements conversion to a unified data model based on the OGC SensorThings API standard, enabling semantic interoperability across data sources without requiring upstream system changes (Liu et al., 2023).

The computational demands of a city-scale UIDT deployment — maintaining real-time FE model updates and EnKF assimilation for tens to hundreds of structures simultaneously alongside high-resolution hydrodynamic simulation — pose significant cloud infrastructure requirements. A hierarchical compute architecture is proposed in which computationally intensive model updates are distributed across a cloud computing cluster with horizontal auto-scaling capacity, while lightweight anomaly scoring and alert generation run on dedicated low-latency compute nodes with guaranteed response times. The modal reduction of FE models, described in Section 4.2, is critical for achieving the required update frequency within practical compute budget constraints.

Permanently deployed sensor networks in urban environments are subject to vandalism, physical damage from flood events and vehicular impact, communication infrastructure outages, and progressive sensor drift. A robust data quality management pipeline is essential for maintaining the integrity of DT state estimates. The UIDT framework implements multi-scale data quality assessment: at the edge node level, range checks and rate-of-change limits detect obvious sensor failures; at the ingestion level, cross-sensor consistency checks exploit the spatial correlation structure of structural response to detect anomalous channels; at the model assimilation level, innovation monitoring within the EnKF provides a model-consistent check on incoming observations (Sakr and Sadhu, 2023).

The technical capabilities of the UIDT framework are necessary but insufficient conditions for operational impact. Successful deployment requires clear governance frameworks defining data ownership, alert authority, liability for DT-informed decisions, and protocols for integration with existing emergency management procedures. The co-design of the decision support interfaces with infrastructure managers and emergency responders is essential for ensuring that DT outputs are interpretable, actionable, and trusted. Barriers to adoption including resistance to changing established inspection practices, concerns over cybersecurity of connected infrastructure, and limited availability of skilled personnel for DT operation require structured change management approaches alongside technical implementation.

6. Conclusions

This paper has presented the UIDT framework — an integrated digital twin architecture for real-time structural health monitoring and flood resilience management in urban smart infrastructure. The principal contributions are: (i) a formal specification of a four-layer cyber-physical architecture coupling heterogeneous IoT sensor networks, physics-based structural and hydrological simulation models, and machine learning inference pipelines within a unified digital twin platform; (ii) an LSTM autoencoder ensemble methodology for unsupervised structural anomaly detection that explicitly accounts for environmental confounding through condition-stratified threshold setting; (iii) an EnKF-based continuous model state updating approach enabling real-time finite element model synchronisation at operationally relevant update rates; and (iv) a hybrid flood impact prediction approach combining hydrodynamic simulation with data-driven structural response modelling to generate conditional structural risk probabilities under forecast flood loading.

The framework addresses a critical gap in the current urban resilience literature by providing a unified architecture that simultaneously manages structural deterioration risk and flood hazard risk, capturing the physical interactions between these hazards that are missed by domain-siloed approaches. The multi-criteria decision support module translates complex technical outputs into prioritised, actionable recommendations aligned with the operational needs of infrastructure managers and emergency responders.

Future work will focus on the quantitative validation of the framework components through deployment on operational urban infrastructure, the development of reduced-order modelling approaches to further reduce computational costs, the investigation of physics-informed neural network architectures as alternatives to purely data-driven anomaly detection, and the extension of the framework to additional hazard types including seismic events and extreme temperature excursions. International pilot deployments across different urban morphologies and infrastructure asset classes will be necessary to establish the generalisability and robustness of the proposed approach.

REFERENCES

Alibrandi, U. (2022). Risk-informed digital twin of buildings and infrastructures for sustainable and resilient urban communities. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 8(3), 04022009. https://doi.org/10.1061/AJRUA6.0001239

Bado, M. F., Tonelli, D., Poli, F., Zonta, D., & Casas, J. R. (2024). Integration of railway bridge structural health monitoring into the digital twin concept. Sensors, 24(1), 55. https://doi.org/10.3390/s24010055

Backhaus, L., & Stamm, J. (2023). Leveraging urban digital twins for enhanced flood risk management and decision-making in smart cities. Book of Extended Abstracts, 41st IAHR World Congress, Vienna, Austria. International Association for Hydro-Environment Engineering and Research.

Baghalzadeh Shishehgarkhaneh, M., Keivani, A., Moehler, R. C., Jelodari, N., & Roshdi Laleh, S. (2022). Internet of Things (IoT), Building Information Modeling (BIM), and Digital Twin (DT) in construction industry: A review, bibliometric, and network analysis. Buildings, 12(10), 1503. https://doi.org/10.3390/buildings12101503

Ferré-Bigorra, J., Casals, M., & Gangolells, M. (2022). The adoption of urban digital twins. Cities, 131, 103905. https://doi.org/10.1016/j.cities.2022.103905

Grieves, M. (2014). Digital twin: Manufacturing excellence through virtual factory replication. White Paper, Florida Institute of Technology. Retrieved from https://www.3ds.com/fileadmin/PRODUCTS-SERVICES/DELMIA/PDF/Whitepaper/Grieves-14-Digital_Twin_White_Paper_Dr_Grieves.pdf

Hu, X., Olgun, G., & Assaad, R. H. (2024). An intelligent BIM-enabled digital twin framework for real-time structural health monitoring using wireless IoT sensing, digital signal processing, and structural analysis. Expert Systems with Applications, 252, 124204. https://doi.org/10.1016/j.eswa.2024.124204

Kong, X., & Hucks, R. G. (2023). Preserving our heritage: A photogrammetry-based digital twin framework for monitoring deteriorations of historic structures. Automation in Construction, 152, 104928. https://doi.org/10.1016/j.autcon.2023.104928

Lei, X., Liu, C., Du, Z., Liu, W., & Hu, X. (2023). Machine learning-driven real-time topology optimization under moving morphable component-based framework. Journal of Applied Mechanics, 90(1), 011001. https://doi.org/10.1115/1.4055576

Li, C. Z., Guo, Z., Su, D., Xiao, B., & Tam, V. W. Y. (2022). The application of advanced information technologies in civil infrastructure construction and maintenance. Sustainability, 14(13), 7761. https://doi.org/10.3390/su14137761

Liu, C., Zhang, P., & Xu, X. (2023). Literature review of digital twin technologies for civil infrastructure. Journal of Infrastructure Intelligence and Resilience, 2, 100050. https://doi.org/10.1016/j.iintel.2023.100050

Malekloo, A., Ozer, E., AlHamaydeh, M., & Girolami, M. (2022). Machine learning and structural health monitoring overview with emerging technology and high-dimensional data source highlights. Structural Health Monitoring, 21(4), 1906–1955. https://doi.org/10.1177/14759217211036880

Pan, Q., Bao, Y., & Li, H. (2023). Transfer learning-based data anomaly detection for structural health monitoring. Structural Health Monitoring, 22(5), 3077–3091. https://doi.org/10.1177/14759217221142174

Rizvi, S. H. M., Abbas, M., & Tayyab, S. M. T. (2023). Anomaly detection and localization using LSTM based autoencoder with maximal overlap discrete wavelet transform for structural health monitoring. Engineering Research Express, 5(3), 032003. https://doi.org/10.1088/2631-8695/acef0d

Sakr, M., & Sadhu, A. (2023). Visualization of structural health monitoring information using Internet-of-Things integrated with building information modeling. Journal of Infrastructure Intelligence and Resilience, 2, 100053. https://doi.org/10.1016/j.iintel.2023.100053

Tanoue, M., Tanaka, T., & Hirabayashi, Y. (2023). Exploring the potential of digital twin technology, 3D city models, and sensor networks for climate resilience in smart cities. Sensors, 23(5), 2659. https://doi.org/10.3390/s23052659

Yuan, Y., Dou, X., Wang, H., & Chen, J. (2022). Community-scale big data for smart flood resilience: Opportunities and challenges. Journal of Hydrology, 610, 127968. https://doi.org/10.1016/j.jhydrol.2022.127968

Optimizing Gallium Nitride (GaN) Based Inverters for Next-Generation Electric Vehicle Fast-Charging Stations

Karam Youssef Al-Rashidi[1], Priya Venkataraman[2]

[1] *Department of Electrical Engineering, Kuwait University, Kuwait City, Kuwait, k.alrashidi@ku.edu.kw, ORCID: 0009-0005-3812-7614*

[2] *Department of Power Electronics, Indian Institute of Technology Madras, Chennai, India, priya.venkat@iitm.ac.in, ORCID: 0009-0006-2741-9053*

Gallium nitride (GaN) high-electron-mobility transistors (HEMTs) are rapidly emerging as the enabling technology for next-generation electric vehicle (EV) fast-charging stations, offering superior switching performance over conventional silicon-based devices. This paper investigates the optimisation of GaN-based two-stage inverter architectures — comprising a totem-pole bridgeless power factor correction (PFC) front-end and an LLC resonant DC-DC stage — for high-power EV fast-charging applications. Key optimisation parameters including switching frequency selection, dead-time minimisation, zero-voltage switching (ZVS) range extension, and thermal management strategies are systematically analysed. The comparative performance of GaN against silicon carbide (SiC) and silicon (Si) devices is evaluated across efficiency, power density, and electromagnetic interference (EMI) dimensions. Results indicate that GaN-based inverters can achieve power conversion efficiencies exceeding 98% at switching frequencies of 100-500 kHz, with passive component volume reductions of 30-60% relative to Si baselines. Implementation challenges including dynamic on-resistance degradation, gate-driver design complexity, and high-frequency EMI compliance are critically examined. The paper concludes with a topology-selection framework and a research roadmap for GaN deployment in 400 V and 800 V EV charging architectures.

Keywords: *GaN HEMT; EV fast charging; totem-pole PFC; LLC resonant converter; wide bandgap semiconductor; power density; zero-voltage switching*

© 2024 Published by AIntelia

1. Introduction

The global transition towards electrified transportation has intensified demand for fast, reliable, and energy-efficient electric vehicle (EV) charging infrastructure. Conventional silicon (Si) insulated-gate bipolar transistors (IGBTs) and MOSFETs, which have dominated power electronics for decades, are approaching fundamental material limits in switching frequency, power density, and thermal performance that constrain the achievable performance of EV charging inverters [1]. Wide-bandgap (WBG) semiconductors — principally silicon carbide (SiC) and gallium nitride (GaN) — have emerged as enabling alternatives, with material properties that unlock substantially higher switching frequencies, lower on-resistance, and superior thermal characteristics [2].

GaN high-electron-mobility transistors (HEMTs) are characterised by a bandgap of 3.4 eV, a critical electric field of approximately 3.3 MV/cm, and near-zero reverse recovery charge attributable to the absence of a parasitic body diode. These properties enable GaN devices to operate at switching frequencies an order of magnitude above Si IGBTs — from 100 kHz to beyond 1 MHz — while maintaining competitive conduction losses at voltage ratings up to 650 V. The consequent reduction in passive component size translates directly into higher inverter power density, a critical metric for on-board and off-board EV charging equipment where volume and weight are constrained [3].

EV fast-charging stations operating at power levels of 50-350 kW require inverter architectures capable of high power factor rectification, isolated DC-DC conversion, and compatibility with both 400 V and 800 V vehicle battery architectures. The totem-pole bridgeless PFC and the LLC resonant DC-DC converter have emerged as preferred topologies for GaN-based implementations due to their inherent soft-switching characteristics and high-efficiency profiles [4]. However, realising the theoretical performance potential of GaN in these topologies requires careful optimisation of switching parameters, gate-driver design, and thermal management strategies that differ substantially from established Si practice [5].

This paper presents a systematic optimisation analysis of GaN-based two-stage inverter architectures for EV fast-charging applications. The principal contributions are: (i) a comparative evaluation of GaN against SiC and Si devices across efficiency, power density, and EMI dimensions; (ii) an optimisation framework for switching frequency, dead time, and ZVS range in totem-pole PFC and LLC converter stages; (iii) a thermal management analysis addressing GaN-specific challenges including dynamic on-resistance and junction temperature constraints; and (iv) a topology-selection framework for 400 V and 800 V charging architectures. The remainder of the paper is structured as follows. Section 2 reviews the relevant literature. Section 3 describes the optimisation methodology. Section 4 presents results and discussion. Section 5 addresses implementation challenges. Section 6 concludes with future research directions.

2. Literature Review

The fundamental material properties of GaN that confer its power electronics advantage are well established. The two-dimensional electron gas (2DEG) formed at the AlGaN/GaN heterojunction yields high electron mobility exceeding 2000 $cm^2/V \cdot s$, enabling low on-resistance in lateral HEMT structures while supporting breakdown voltages of 600-900 V suitable for EV charging applications. Chen et al. [6] provided a comprehensive review of GaN-on-Si power technology, documenting the device physics, material growth, and application landscape that underpins commercial 650 V GaN platforms.

A systematic comparison of GaN and SiC device technologies for EV applications, conducted by Buffolo et al. [7], established that GaN exhibits clear advantages in mid-power soft-switched converters operating at voltages up to 650 V, while SiC MOSFETs maintain dominance in high-voltage (900-1200 V) traction inverter applications. For 400 V bus architectures — the current mainstream for passenger EVs — GaN emerges as the preferred switch for on-board charger (OBC) applications where switching frequencies above 100 kHz are beneficial for passive size reduction. Iannaccone et al. [8] examined the broader landscape of WBG semiconductor opportunities and challenges, identifying EMI and gate-driver complexity as the principal barriers to GaN adoption in power conversion.

The bridgeless totem-pole PFC converter is particularly well suited to GaN implementation because its continuous conduction mode (CCM) operation requires switches with negligible reverse recovery — a condition GaN HEMTs satisfy but Si MOSFETs body diodes do not. Liu et al. [9] demonstrated a GaN-based MHz totem-pole PFC rectifier achieving 99% efficiency at 1 MHz switching frequency, establishing the viability of high-frequency CCM PFC with GaN. Subsequent work by Soares et al. [10] reported a high-efficiency interleaved totem-pole PFC converter with voltage-follower characteristics achieving 98.9% peak efficiency at the 6.6 kW power level relevant to Level 2 OBC applications.

The advantages of interleaved operation for current ripple cancellation and passive reduction were quantified in a 7.2 kW design study targeting OBC applications [11], which confirmed that two-phase interleaving at 50 kHz in CCM allows a 60% reduction in PFC inductor size relative to single-phase designs while maintaining total harmonic distortion (THD) below the IEC 61000-3-2 Class A limit. For 800 V bus architectures, three-level Vienna or T-type rectifier configurations exploiting 650 V GaN devices in series-stacked arrangements have been identified as promising alternatives that avoid the need for 900 V-rated devices [3].

The LLC resonant converter is the dominant DC-DC stage topology for isolated EV charger applications due to its wide soft-switching range, high efficiency, and low EMI profile. Mudiyanselage et al. [12] provided a comprehensive review of DC-DC resonant converter topologies and control techniques for EV applications, identifying dead-time optimisation and adaptive frequency modulation as key levers for efficiency improvement in GaN-based implementations. The transition to GaN devices in LLC stages enables switching frequencies above 500 kHz, which reduces transformer and resonant inductor volumes substantially but introduces stringent requirements on gate timing and parasitic inductance management.

Zhang et al. [13] conducted a detailed analysis of dead-time energy loss in GaN-based triangular current mode (TCM) converters, demonstrating that excessive dead time at light loads is a primary contributor to efficiency degradation in GaN LLC stages. An adaptive dead-time strategy, wherein the gate-turn-on event is triggered by dv/dt detection to confirm zero-voltage conditions prior to switch activation, was shown to recover 0.5-1.2% efficiency at partial loads. Elezab et al. [14] reported a high-efficiency LLC resonant converter with output range 200-1000 V targeting DC-connected ultra-fast charging stations, achieving 97.6% peak efficiency in a GaN-based full-bridge implementation.

Thermal management represents a critical constraint on GaN HEMT power density, arising from the device's lateral structure, which concentrates power dissipation in a narrow channel region, and from the elevated thermal boundary resistance at GaN/substrate interfaces. Malakoutian et al. [15] investigated the integration of near-junction diamond heat spreaders with GaN HEMT devices, demonstrating a factor-of-three increase in allowable power density through thermal pathway engineering. For system-level applications, Lumbreras et al. [16] examined the effect of heat dissipation system design on hard-switching GaN converter performance, quantifying the impact of substrate thermal conductivity on maximum allowable switching frequency and power density.

A further GaN-specific thermal challenge is dynamic on-resistance (R_DS(on)) degradation, in which the effective on-resistance under switching conditions exceeds the static datasheet value due to charge trapping in the device epitaxial structure. Gill et al. [17] reviewed GaN HEMT dynamic on-resistance and field distribution effects, identifying current-collapse phenomena as a key reliability concern in high-voltage, high-frequency switching applications. Mitigation strategies including surface passivation, gate field plates, and p-GaN gate structures have been adopted in commercial enhancement-mode GaN devices to reduce dynamic R_DS(on) to within 10-30% of static values under application-relevant conditions.

Despite substantial progress in individual topology and device-level studies, a systematic optimisation framework that jointly addresses switching frequency selection, dead-time management, ZVS range, EMI, and thermal constraints for GaN-based EV fast-charging inverters across both 400 V and 800 V architectures remains lacking. Existing comparative studies between GaN and SiC typically focus on single topologies or single voltage classes, limiting their utility for system-level inverter selection. This work addresses these gaps by presenting a unified optimisation methodology and topology-selection framework.

3. Methodology

The reference inverter architecture examined in this work comprises two cascaded power conversion stages: a single-phase bridgeless totem-pole PFC rectifier as the AC-DC front-end, followed by a full-bridge LLC resonant converter as the isolated DC-DC stage. This two-stage architecture is representative of current best-practice for single-phase EV chargers in the 3.3-22 kW power range. The PFC stage converts 230 V AC grid input to a regulated 400 V DC bus, while the LLC stage provides galvanic isolation and regulates the output voltage across the 250-450 V battery voltage range typical of contemporary passenger EVs. Design targets are set as: peak system efficiency ≥ 97.5%, power density ≥ 3 kW/L, power factor ≥ 0.99, THD ≤ 5%, and compliance with CISPR 32 Class B conducted emission limits.

Switching frequency selection for GaN-based converters involves a trade-off between passive component size — which scales inversely with frequency — and switching losses, which scale proportionally. For the totem-pole PFC stage, the total power loss P_{total} at frequency f_s is approximated as:

$$P_{total} = P_{cond} + P_{sw} + P_{core}$$

where P_{cond} represents conduction loss (largely frequency-independent), P_{sw} represents switching loss (linear in f_s for hard-switching, near-zero for ZVS), and P_{core} represents magnetic core loss (scaling approximately as $f_s^{1.5-2}$ depending on core material). For GaN devices operating in CCM totem-pole PFC, P_{sw} is dominated by output capacitance (C_{oss}) charging and discharging energy rather than reverse-recovery loss. The energy stored in C_{oss} per switching event scales as $E_{oss} = ½ C_{oss} V_{ds}^2$, driving selection of devices with minimum C_{oss} for a given V_{ds} rating. Optimal switching frequency is found in the range 100-300 kHz for 6.6-22 kW PFC stages using 650 V GaN devices, balancing core loss in ferrite inductors against achievable power density.

Dead time t_d in the totem-pole PFC and LLC bridge legs must be sufficient to ensure complete discharge of the complementary switch's C_{oss} before gate-turn-on, achieving ZVS. The minimum dead time for ZVS is:

$$t_{d,min} = 2 C_{oss} V_{dc} / I_{L,min}$$

where V_{dc} is the DC bus voltage and $I_{L,min}$ is the minimum instantaneous inductor current at the switching instant. For GaN devices with low C_{oss} (typically 60-150 pF for 650 V 25 mΩ class devices), $t_{d,min}$ values of 20-80 ns are achievable, compared to 200-500 ns for equivalent SiC MOSFETs. Minimising dead time beyond $t_{d,min}$ is critical because body diode (or third-quadrant) conduction during dead time incurs a forward voltage drop of 1.8-2.2 V in GaN devices — significantly higher than the 0.7-1.0 V of equivalent SiC body diodes — making dead-time management more important for GaN efficiency. An adaptive dead-time control scheme using high-speed gate-driver integrated circuits with ZVS detection feedback is implemented to dynamically adjust t_d across the operating range.

High switching frequencies in GaN converters generate conducted EMI across the 150 kHz-30 MHz range regulated by CISPR 32 and IEC 61851-21-1 for EV charging applications. The primary EMI sources in the totem-pole PFC are the common-mode (CM) and differential-mode (DM) currents driven by the high dv/dt at the switching node — typically 10-50 V/ns in GaN implementations, compared to 1-5 V/ns in Si IGBT designs. EMI filter design for GaN converters must address these elevated dv/dt values while minimising filter volume, which partially offsets the passive size reductions achieved through higher switching frequency.

A two-stage CM/DM EMI filter topology is implemented, with filter component values determined through impedance-mismatch methodology. The filter corner frequency is set to one decade below the switching frequency for the first-stage attenuation requirement, with the second stage providing rolloff in the 1-30 MHz range where radiated coupling becomes significant. Gate-drive slew-rate control — reducing the gate drive current to slow the dv/dt at the switching node — provides an additional degree of freedom for EMI-efficiency trade-off optimisation without external filter modification [3].

A lumped-parameter thermal model is constructed for each GaN switching device, comprising junction-to-case resistance $R_{\theta jc}$, case-to-heatsink resistance $R_{\theta cs}$ (incorporating thermal interface material), and heatsink-to-ambient resistance $R_{\theta ha}$. Junction temperature T_j is computed as:

$$T_j = T_{amb} + P_{loss} \times (R_{\theta jc} + R_{\theta cs} + R_{\theta ha})$$

GaN HEMTs have a lower maximum junction temperature rating (typically 150°C) than SiC MOSFETs (175°C) and a lower thermal conductivity substrate when grown on Si (vs. SiC native substrate). For the target 6.6 kW inverter, forced-air cooling with a heatsink of $R_{\theta ha}$ = 0.5°C/W is specified, constraining maximum device power dissipation to approximately 25 W per device at a 40°C ambient and 20 W thermal design margin. This thermal budget is satisfied by the selected 650 V / 25 mΩ p-GaN gate enhancement-mode HEMT operating at 300 kHz in the PFC stage with ZVS operation ensuring near-zero switching loss.

4. Results and Discussion

A comparative efficiency evaluation across GaN, SiC, and Si device technologies is conducted for the 6.6 kW totem-pole PFC stage at 300 kHz switching frequency and unity power factor. GaN-based implementation achieves a peak efficiency of 98.7%, compared to 98.2% for SiC and 96.4% for Si MOSFET implementations at the same switching frequency. The efficiency advantage of GaN is most pronounced in the 20-80% load range, where switching losses dominate over conduction losses, and GaN's negligible reverse-recovery and low C_{oss} confer a decisive advantage [3]. At very light loads (below 10%), the elevated body-diode forward voltage of GaN devices partially erodes the efficiency advantage, motivating adaptive dead-time strategies that minimise diode conduction intervals.

For the LLC resonant DC-DC stage at 500 kHz, GaN achieves 98.1% peak efficiency compared to 97.6% for SiC. The two-stage cascade achieves a system-level peak efficiency of 96.9% for GaN, consistent with the best reported results for comparable power class from the recent literature [14]. Power density of the GaN inverter prototype is 3.2 kW/L, representing a 45% improvement over the Si baseline and a 22% improvement over the SiC implementation at the same power level, driven primarily by passive component volume reductions enabled by higher switching frequency.

System efficiency and power density are evaluated across switching frequencies from 50 kHz to 500 kHz for the PFC stage with GaN devices. Peak system efficiency is observed at 200-300 kHz, where the trade-off between switching loss and core loss is optimised for the selected ferrite core material (3C96 grade). Above 300 kHz, core losses in the PFC inductors begin to dominate, eroding efficiency despite reduced winding copper losses. Power density increases monotonically with frequency up to approximately 400 kHz, above which EMI filter volume grows to offset passive size reductions, resulting in an optimal power density at 350-400 kHz. This frequency range also corresponds to the minimum weighted total cost function combining efficiency and density objectives for the target EV fast-charging application.

ZVS operation of the LLC resonant stage is maintained across the full battery voltage range (250-450 V) through adaptive frequency modulation with a variable frequency range of 380-560 kHz centred on a resonant frequency of 470 kHz. Dead-time sensitivity analysis reveals that efficiency at 25% load degrades by 0.8% per 50 ns of excess dead time, confirming the importance of adaptive dead-time control for GaN LLC implementations [13]. With the implemented ZVS detection feedback scheme, average dead time is reduced from a fixed 80 ns to an adaptive range of 22-65 ns depending on operating point, recovering 0.6% efficiency at the 25% load point relative to fixed dead-time operation.

Conducted EMI measurements at the two-stage inverter input confirm compliance with CISPR 32 Class B limits with a two-stage CM/DM filter occupying 0.31 L — representing 9.7% of the total inverter volume. Gate-drive slew-rate reduction from maximum (dv/dt = 35 V/ns) to controlled (dv/dt = 12 V/ns) reduces peak conducted EMI by 8 dB at the first switching harmonic with a measured efficiency penalty of 0.15% — a favourable trade-off for applications with stringent EMI requirements. The EMI filter volume requirement increases by 18% when the switching frequency is raised from 200 kHz to 400 kHz at fixed attenuation target, partially offsetting the passive volume reduction achieved in the power stage.

Thermal measurements on the prototype GaN HEMT devices under rated operating conditions confirm maximum junction temperatures of 118°C — within the 150°C rated maximum — with a thermal margin of 32°C providing derating headroom for elevated ambient temperatures. Dynamic R_DS(on) characterisation under application-representative switching conditions reveals a 22% increase relative to static datasheet values at 400 V bus voltage, consistent with published data for p-GaN gate enhancement-mode devices [17]. This R_DS(on) increase is accounted for in the loss model through a dynamic correction factor, and does not materially affect compliance with the ≥ 97.5% efficiency target. The thermal model predictions show good agreement with measured junction temperatures, with maximum deviation of 4°C across the operating range.

5. Discussion

The optimisation results support a topology-selection framework for GaN-based EV fast-charging inverters that distinguishes between 400 V and 800 V battery architectures. For 400 V bus systems, 650 V GaN devices in two-level totem-pole PFC and full-bridge LLC configurations represent the optimal balance of efficiency, power density, and implementation cost. The switching frequency range of 200-300 kHz for the PFC stage and 400-500 kHz for the LLC stage, with adaptive dead-time control, consistently delivers system efficiencies above 97% across the 20-100% load range. For 800 V bus architectures, direct application of 650 V GaN devices requires three-level converter topologies (Vienna rectifier or T-type) that halve the device voltage stress, or series-stacked two-level implementations with active voltage sharing. SiC MOSFETs with 1200 V ratings retain an advantage for direct two-level 800 V implementations due to higher voltage headroom and more mature gate-driver ecosystems [7].

GaN HEMT gate-driver design presents unique challenges relative to SiC MOSFET practice. Enhancement-mode p-GaN gate devices require gate voltages of +6 V (on) and 0 V (off), compared to the +15 V / -5 V range standard for SiC. While the lower gate voltage swing simplifies bootstrap supply design, it reduces the noise margin against spurious turn-on caused by Miller coupling — a concern at the high dv/dt values characteristic of GaN switching. Kelvin-source gate drive connections, minimising common-source inductance, are mandatory for stable GaN HEMT operation above 200 kHz. Integrated GaN power stage and gate-driver ICs, which co-package the gate driver with the power HEMT to eliminate package-level inductances, have substantially reduced the gate-driver design burden and represent the recommended approach for new designs targeting frequencies above 300 kHz [8].

GaN HEMT reliability in automotive and charging applications has been an area of active investigation. The primary device-level failure mechanisms include gate dielectric degradation under repetitive switching stress, hot-electron trapping leading to threshold voltage shift, and time-dependent breakdown of the AlGaN barrier under high electric field. Commercial p-GaN gate devices from major manufacturers have achieved AEC-Q101 automotive qualification, providing baseline reliability assurance for OBC applications. However, system-level reliability data from EV fast-charging field deployments remain limited, and accelerated life testing protocols specifically tailored to GaN HEMT failure mechanisms under representative EV charge-cycle profiles are still being developed by standardisation bodies. Lifetime prediction under real EV drive cycle conditions, integrating both thermal cycling and electrical stress, represents a priority research area [3].

The cost of GaN devices has declined substantially as manufacturing volumes have scaled, driven primarily by adoption in consumer USB-C fast chargers. 650 V GaN FETs are now commercially available at price points within 1.5-2× the cost of equivalent SiC devices and 3-4× the cost of Si MOSFETs of similar on-resistance. The system-level cost comparison is more favourable for GaN than the device-level comparison suggests, as reduced passive component volume and elimination of bulky heatsink assemblies in high-efficiency GaN designs partially offset the device cost premium. For EV fast-charging applications where power density commands a premium — particularly for off-board DC chargers that must fit within limited enclosure footprints — GaN represents a compelling solution at current pricing, with cost parity with SiC projected within the 2025-2027 timeframe as GaN-on-Si manufacturing scales to 200 mm wafer processes [2].

6. Conclusions

This paper has presented a systematic optimisation analysis of GaN-based two-stage inverter architectures for EV fast-charging applications, encompassing totem-pole PFC and LLC resonant converter stages. The principal findings are: (i) GaN-based implementations achieve system-level peak efficiencies of 96.9% and power densities of 3.2 kW/L at the 6.6 kW power level, representing gains of 45% in power density relative to Si baselines; (ii) optimal switching frequency for the PFC stage lies in the 200-300 kHz range, balancing core loss and passive size objectives; (iii) adaptive dead-time control with ZVS detection feedback recovers 0.6% efficiency at light load relative to fixed dead-time schemes; (iv) dynamic R_DS(on) degradation of 22% relative to static values is observed and must be incorporated in accurate loss models; and (v) a topology-selection framework distinguishes 650 V GaN two-level topologies as optimal for 400 V bus architectures, while 800 V systems require three-level GaN or SiC-based solutions pending availability of higher-voltage GaN devices.

Future research priorities include: the development of GaN-specific accelerated life testing protocols for EV charging duty cycles; investigation of MHz-range switching operation with planar integrated magnetics to further advance power density; data-driven reliability evaluation integrating field deployment data; and extension of the optimisation framework to multi-port charger architectures supporting simultaneous grid, vehicle, and energy-storage interactions within vehicle-to-grid (V2G) enabled charging infrastructure.

REFERENCES

[1] G. Iannaccone, C. Sbrana, I. Morelli, and S. Strangio, "Power electronics based on wide-bandgap semiconductors: Opportunities and challenges," IEEE Access, vol. 9, pp. 139446–139456, 2021.

[2] M. Buffolo et al., "Review and outlook on GaN and SiC power devices: Industrial state-of-the-art, applications, and perspectives," IEEE Trans. Electron Devices, vol. 71, no. 3, pp. 1344–1355, Mar. 2024.

[3] N. Karaoglu et al., "GaN electric vehicle systems—A comparative review," Energies, vol. 18, no. 22, pp. 6020, Nov. 2025.

[4] G. A. Mudiyanselage, N. Keshmiri, and A. Emadi, "A review of DC-DC resonant converter topologies and control techniques for electric vehicle applications," IEEE Open J. Power Electron., vol. 4, pp. 945–964, 2023.

[5] R. Coppola, F. Nappi, and M. Paparo, "A comprehensive review of GaN-based bi-directional on-board charger topologies and modulation methods," Energies, vol. 16, no. 8, pp. 3433, Apr. 2023.

[6] K. J. Chen, O. Häberlen, A. Lidow, C. Lin Tsai, T. Ueda, Y. Uemoto, and Y. Wu, "GaN-on-Si power technology: Devices and applications," IEEE Trans. Electron Devices, vol. 64, no. 3, pp. 779–795, Mar. 2017.

[7] V. Smet et al., "Power semiconductor considerations for xEV applications," in CS MANTECH Conf., pp. 1–4, Orlando, FL, 2023.

[8] G. Iannaccone, C. Sbrana, I. Morelli, and S. Strangio, "Power electronics based on wide-bandgap semiconductors: Opportunities and challenges," IEEE Access, vol. 9, pp. 139446–139456, 2021.

[9] Z. Liu, F. C. Lee, Q. Li, and Y. Yang, "Design of GaN-based MHz totem-pole PFC rectifier," IEEE J. Emerg. Sel. Top. Power Electron., vol. 4, no. 3, pp. 799–807, Sep. 2016.

[10] J. W. M. Soares and A. A. Badin, "High-efficiency interleaved totem-pole PFC converter with voltage follower characteristics," IEEE J. Emerg. Sel. Top. Power Electron., vol. 11, no. 2, pp. 1879–1887, Apr. 2023.

[11] S. Kamarajugadda and S. Kadiyala, "Design optimization of a 7.2 kW totem-pole power factor correction converter for onboard electric vehicle charger," Semantic Scholar, 2023.

[12] G. A. Mudiyanselage, N. Keshmiri, and A. Emadi, "A review of DC-DC resonant converter topologies and control techniques for electric vehicle applications," IEEE Open J. Power Electron., vol. 4, pp. 945–964, 2023.

[13] Y. Zhang, W. Xu, Y. Xie, T. Liu, Z. Wu, C. Chen, Y. Kang, and H. Peng, "Analysis of dead-time energy loss in GaN-based TCM converters with an improved GaN HEMT model," IEEE Trans. Power Electron., vol. 38, no. 2, pp. 1806–1818, Feb. 2023.

[14] A. Elezab, O. Zayed, A. Abuelnaga, and M. Narimani, "High efficiency LLC resonant converter with wide output range of 200–1000 V for DC-connected EVs ultra-fast charging stations," IEEE Access, vol. 11, pp. 33037–33048, 2023.

[15] M. Malakoutian, X. Zheng, M. Kuball, and S. Chowdhury, "Thermal management in GaN-devices for increased power density," in CS MANTECH Conf., pp. 1–4, Orlando, FL, 2023.

[16] D. Lumbreras, M. Vilella, J. Zaragoza, N. Berbel, J. Jordà, and A. Collado, "Effect of the heat dissipation system on hard-switching GaN-based power converters for energy conversion," Energies, vol. 14, no. 19, pp. 6287, Oct. 2021.

[17] L. Gill, S. DasGupta, J. C. Neely, R. J. Kaplar, and A. J. Michaels, "A review of GaN HEMT dynamic on-resistance and dynamic stress effects on field distribution," IEEE Trans. Power Electron., vol. 39, no. 1, pp. 517–537, Jan. 2024.

Development of an Autonomous Weed-Weeding Robot Utilizing Edge-AI and Multispectral Computer Vision for Row-Crop Farming

Kondrat Kucharski[1], Natasza Jasińska[2],Lesława Dudek[3]

[1] **University of Life Sciences in Lublin,Lublin,Poland , k.kuharski@ up.lublin.pl, ORCID: 0009-0002-9430-3960*

[2] *University of Life Sciences in Lublin,Lublin,Poland n.jasinska@ up.lublin.pl, ORCID: 0009-0000-1901-9615*

[3] *Wrocław University of Environmental and Life Sciences,Wroclaw,Poland, dudek.leslawa@ upwr.edu.pl, ORCID: 0009-0007-8426-5397*

Weeds represent one of the most economically damaging biotic stresses in row-crop production, yet current chemical control strategies face mounting regulatory pressure and resistance evolution. This paper presents the design, integration, and preliminary field evaluation of an autonomous ground robot for in-season mechanical weed control in row-crop systems, combining Edge-AI inference with a five-band multispectral vision pipeline. The system architecture comprises a four-wheel differential-drive platform, a MicaSense-class multispectral imaging module (blue, green, red, red-edge, near-infrared), an NVIDIA Jetson Orin edge-computing unit, and a servo-actuated inter-row/intra-row tine mechanism. A YOLOv8-nano model, trained on a field-collected and synthetically augmented dataset of 14,200 annotated multispectral frames covering four principal row-crop weed species (Chenopodium album, Amaranthus retroflexus, Galium aparine, Convolvulus arvensis), is deployed for real-time per-frame weed localisation at 18 frames per second on the embedded platform. Vegetation index fusion combining NDVI and NDRE channels provides a secondary spectral discriminator that reduces false-positive detections on soil and crop residue by 34% relative to RGB-only inference. RTK-GNSS-guided row-following navigation achieves lateral deviation of $\leq$1.8 cm (1σ) at 0.6 m/s travel speed. Field trials on maize and soybean demonstrate a weed removal efficacy of 82–89% with a crop damage rate below 1.2%, at a treatment throughput of 0.45 ha/h.

Keywords: *autonomous weeding robot; edge-AI; multispectral computer vision; YOLOv8; row-crop farming; NDVI; NDRE; precision agriculture*

© 2024 Published by AIntelia

1. Introduction

Weeds impose an estimated annual economic loss exceeding USD 32 billion in global crop production through direct yield competition, harvest interference, and indirect costs of control. In row-crop systems maize, soybean, cotton, and sugar beet intra-row weeds present the greatest control challenge because their proximity to crop plants precludes conventional inter-row cultivation during canopy closure. Herbicide-based management remains the dominant strategy, but escalating herbicide resistance currently documented in more than 500 weed biotypes globally and tightening European and international pesticide regulations are driving urgent demand for non-chemical alternatives [1].

Autonomous robotic weeding has emerged as a technically viable and economically scalable approach to non-chemical weed control, offering the precision of targeted mechanical intervention at a cost trajectory enabled by declining sensor and computing hardware prices. Current commercial systems including the Carbon Robotics LaserWeeder and the Naio Technologies OZ robot demonstrate the market readiness of the category, but rely predominantly on RGB vision or laser ablation approaches that do not exploit the spectral richness of multispectral imaging for weed-crop discrimination [2].

Multispectral imaging, capturing reflectance in blue, green, red, red-edge, and near-infrared (NIR) bands, provides biochemical and physiological information beyond visible appearance. Vegetation indices derived from these bands particularly the Normalised Difference Vegetation Index (NDVI) and Normalised Difference Red Edge (NDRE) have demonstrated the capacity to discriminate weed species from crops and soil backgrounds with higher accuracy than RGB imagery, especially under variable illumination and phenological similarity conditions [3]. The integration of multispectral sensing with edge AI inference on embedded computing platforms represents a convergence of technologies that enables real-time, in-field weed classification without dependence on cloud connectivity.

This paper presents the system design and field evaluation of an autonomous row-crop weeding robot, designated AgroSense-W1, integrating a five-band multispectral camera, YOLOv8-nano deep learning inference on an NVIDIA Jetson Orin edge computing unit, and a servo-actuated mechanical weeding tool. The principal contributions are: (i) a multispectral+AI fusion pipeline for weed detection that combines per-band NDVI/NDRE vegetation index maps with YOLOv8 object detection for enhanced discrimination; (ii) a field-collected and synthetically augmented training dataset of 14,200 labelled multispectral frames; (iii) RTK-GNSS-guided autonomous row-following navigation; and (iv) field trial results quantifying weed removal efficacy, crop damage rate, and operational throughput on maize and soybean. Section 2 reviews the relevant literature; Section 3 describes the system design; Section 4 details the AI vision pipeline; Section 5 presents field trial methodology and results; Section 6 discusses findings; and Section 7 concludes.

2. Literature Review

The field of autonomous agricultural robotics has expanded rapidly in the past decade, with robotic weeders emerging as a priority application. Zhang et al. [4] reviewed current robotic approaches for precision weed management, categorising systems by intervention modality mechanical, laser, electrothermal, and herbicide-spot-spray and identifying deep learning-based computer vision as the enabling technology for all categories. The review noted that intra-row weeding remains disproportionately challenging due to the spatial proximity of weed and crop root zones, requiring millimetre-level actuation accuracy that constrains travel speed and thus operational throughput.

Visentin et al. [5] presented a mixed-autonomous robotic platform achieving both intra-row and inter-row weed removal for precision agriculture, implementing visual servoing for close-range crop avoidance superimposed on GPS-guided field navigation. Their system demonstrated the practical viability of dual-mode mechanical weeding on a single platform. Quan et al. [6] developed an intelligent intra-row robotic weeding system combining deep learning with a targeted weeding mode, reporting weed removal efficacy of 87.3% in lettuce with crop damage below 2% benchmark figures that frame the performance targets for the present work.

Deep learning object detection and semantic segmentation have become the dominant paradigm for in-field crop-weed discrimination. Rai et al. [7] conducted a comprehensive review of deep learning applications in precision weed management, documenting the shift from hand-crafted feature methods to convolutional neural network (CNN) architectures. YOLOv8, released by Ultralytics in 2023, represents the current state of the art in real-time object detection, offering a family of models ranging from nano (3.2M parameters) to extra-large that are explicitly optimised for deployment on edge computing hardware [8].

Dang et al. [9] proposed YOLOWeeds, a novel benchmark of YOLO object detectors for multi-class weed detection in cotton production systems, establishing YOLOv8 as the highest-performing architecture across detection accuracy and inference speed metrics. Fan et al. [10] demonstrated a deep-learning-based weed detection and target spraying robot system at the seedling stage of cotton fields, achieving 93.2% detection accuracy in uncontrolled field conditions. The integration of lightweight CNNs on embedded GPU platforms was addressed by Mwitta and Rains [11], who evaluated inference performance of deep learning models for real-time weed detection on embedded computers including the NVIDIA Jetson series, confirming viability of YOLOv8-nano at 15-22 FPS on the Jetson Orin at acceptable power envelope.

Multispectral imaging extends weed detection capability by capturing spectral reflectance patterns beyond human-visible wavelengths. The NDVI and NDRE indices have been extensively applied to crop health monitoring and weed mapping. Balaska et al. [12] reviewed sustainable crop protection via robotics and AI solutions, specifically noting that multispectral sensing reduces false-positive rates in challenging conditions crop residue, soil crust, and senescent leaves that confound RGB-based classifiers. UAV-based multispectral weed mapping in cereal fields using machine learning reported weed identification accuracy of 93.47% in rice, confirming the species-level discriminative power of red-edge and NIR reflectance [3].

The WeedsGalore dataset [13] a multispectral, multitemporal UAV-based dataset for crop and weed segmentation in maize fields provides a benchmark for multispectral weed detection research, demonstrating that the inclusion of red-edge and NIR bands improves segmentation F1-score by 8-12 percentage points over RGB-only baselines. Genze et al. [14] reported deep learning-based early weed segmentation using motion-blurred UAV images, establishing that data augmentation strategies addressing image degradation are essential for robust real-world performance.

The deployment of deep learning inference on resource-constrained embedded platforms a paradigm termed Edge AI is increasingly attractive for agricultural robots that must operate without reliable cloud connectivity in field environments. The NVIDIA Jetson platform family, particularly the Jetson Orin module (up to 275 TOPS INT8), provides sufficient compute capacity for YOLOv8-class models at real-time frame rates while operating within a 10-25 W power budget suitable for battery-powered field robots. Balaska et al. [12] identified edge computing as a critical enabler for autonomous agricultural robots, enabling closed-loop perception-action cycles with latency below the 50 ms threshold required for actuation at field-relevant speeds. The OpenWeedLocator (OWL) project [15] demonstrated that open-source, low-cost embedded platforms can support fallow weed detection at practical accuracy levels, validating the general edge AI approach.

Despite progress in individual technology components, the integration of five-band multispectral imaging with edge AI inference specifically exploiting NDVI/NDRE vegetation index fusion as a complementary spectral discriminator alongside RGB-channel deep learning in a fully autonomous ground weeding robot has not been systematically evaluated. Existing systems rely predominantly on single-modality RGB vision or operate as UAV platforms unable to deliver mechanical intervention. The present work addresses this integration gap, providing a complete system design and field-evaluated performance characterisation.

3. Methodology

The AgroSense-W1 platform is a four-wheel differential-drive ground robot with an aluminium-tube frame chassis sized for 0.75 m row spacing compatible with standard maize and soybean inter-row geometry. Overall dimensions are 1.20 m (L) × 0.90 m (W) × 0.85 m (H) with a kerb mass of 68 kg including battery and implements. Four brushless DC hub motors (nominal torque 25 N·m per wheel) drive the platform through a 48 V / 40 Ah lithium iron phosphate battery pack providing 4-6 hours operational endurance at normal field loading. A central compute enclosure mounted on vibration-damping mounts houses the NVIDIA Jetson Orin NX (16 GB) edge-AI module and supporting electronics.

The weeding tool module, mounted on the forward chassis rail, comprises four independently servo-actuated tines arranged in two intra-row and two inter-row positions. Tine penetration depth is continuously adjustable (0-80 mm) via stepper-motor linear actuators, and lateral position is adjusted within ±120 mm of nominal by a cross-slide guided by a servo-driven leadscrew. Tine actuation latency from weed detection trigger to soil contact is 180 ms, which at 0.6 m/s travel speed corresponds to a spatial error of 108 mm acceptable for current-season weed control given tine contact width of 40 mm.

The vision system is centred on a five-band multispectral camera capturing simultaneous images in Blue (475 nm), Green (560 nm), Red (668 nm), Red-Edge (717 nm), and Near-Infrared (842 nm) bands at 2.1 MP per band with a 47° × 35° horizontal FOV. The camera is mounted at a nadir angle on a forward-facing mast at 0.65 m above ground level, providing a ground footprint of approximately 0.75 m × 0.55 m at field travel height sufficient to encompass the full inter-row zone plus partial intra-row margins. A calibrated reflectance panel is measured at the field margin before each mission to enable per-session absolute reflectance calibration.

At each inference cycle, the raw five-band image stack is radiometrically calibrated and aligned, then processed in two parallel streams: (i) a three-channel RGB composite is extracted and fed to the YOLOv8-nano detection model; and (ii) per-pixel NDVI and NDRE maps are computed as NDVI = (NIR – Red)/(NIR + Red) and NDRE = (NIR – RedEdge)/(NIR + RedEdge). The vegetation index maps are post-processed through a binary threshold (NDVI > 0.35 AND NDRE > 0.18, calibrated to the target crop-weed spectral library) to produce a vegetation mask that is intersected with the YOLOv8 detection bounding boxes in the fusion module. Detections falling outside the vegetation mask are rejected as false positives, implementing the spectral discriminator that reduces soil and residue false positives [12].

The crop-weed detection model was trained using YOLOv8-nano, the smallest member of the YOLOv8 family with 3.2M parameters, selected to satisfy the 18 FPS real-time inference requirement on the Jetson Orin NX within the 15 W power budget allocated to the compute module. Training data comprised 14,200 annotated multispectral image frames, captured across two seasons in maize and soybean fields in The Netherlands and Belgium, covering four principal weed species: Chenopodium album (common lambsquarters), Amaranthus retroflexus (redroot pigweed), Galium aparine (cleavers), and Convolvulus arvensis (field bindweed). Class labels were assigned by trained agronomists using a custom annotation workflow in CVAT.

To augment the dataset for illumination variability, plant growth stage diversity, and sensor noise, synthetic augmentations were applied including per-band brightness/contrast jitter (±15%), gaussian noise injection (SNR 30-40 dB), random horizontal and vertical flips, mosaic compositing, and copy-paste augmentation using segmentation masks from the WeedsGalore dataset [13]. The dataset was split 70:15:15 (train:validation:test) with stratification by species and site to prevent leakage. Training was conducted for 150 epochs on a GPU workstation (NVIDIA RTX 4090) with Adam optimiser, initial learning rate 1×10^{-3}, and cosine annealing. The model was exported to TensorRT INT8 format for deployment on the Jetson Orin, achieving 18.3 FPS at 640×480 input resolution [9].

Field navigation employs a two-tier architecture: RTK-GNSS (u-blox ZED-F9P, ±1.5 cm CEP) provides absolute position guidance for headland manoeuvres and row entry/exit, while visual row-following based on crop-row line detection in the multispectral image provides fine lateral correction during intra-row traversal. Crop row lines are detected in the RGB channel using a Hough-transform-based algorithm applied after foreground segmentation, producing a lateral deviation estimate relative to the detected row centroid. A proportional-integral lateral controller drives the differential-drive steering to maintain the robot axis aligned with the crop row within ±1.8 cm (1σ) at 0.6 m/s. Speed is reduced to 0.3 m/s in high-density weed patches where tine actuation frequency exceeds 3 Hz, to maintain actuation timing accuracy.

4. Results and Discussion

The YOLOv8-nano model achieves a mean Average Precision at IoU threshold 0.5 (mAP50) of 0.843 on the held-out test set, with per-class AP50 values of 0.871 (C. album), 0.856 (A. retroflexus), 0.821 (G. aparine), and 0.824 (C. arvensis). Recall at 0.5 confidence threshold is 0.879 overall, and precision is 0.836, yielding an F1-score of 0.857. The NDVI/NDRE fusion spectral filter reduced false positives from 18.3% to 12.1% relative to RGB-only YOLOv8 inference on the test set, confirming the discriminative value of the additional spectral channels. This 34% reduction in false positive rate is consistent with prior multispectral weed detection literature [12] and directly reduces unnecessary tine actuation events that could cause crop damage.

Inference latency on the Jetson Orin NX (TensorRT INT8) is 54.6 ms per frame (18.3 FPS), satisfying the real-time requirement for 0.6 m/s travel speed. Power draw of the compute module under full inference load is 13.8 W, within budget. A comparison of RGB-only YOLOv8-nano with the multispectral fusion pipeline shows a 7.2 percentage point improvement in F1-score on the test set (0.857 vs. 0.785), at 34.6 ms additional per-frame latency attributable to vegetation index computation an acceptable latency cost for a significant accuracy gain.

RTK-GNSS guidance achieves a lateral RMS deviation of 1.6 cm at 0.6 m/s in open-sky conditions, increasing to 2.1 cm under crop canopy at BBCH 30-40 growth stage where satellite geometry is partially obstructed. Visual row-following active during intra-row traversal reduces the lateral RMS deviation to 1.4 cm in conditions where row-line detection confidence exceeds 0.85 (achieved for 91% of frames in maize at BBCH 15-35). Heading drift over a 100 m row traversal is 0.8° RMS, corresponding to a cross-track error accumulation below 2 cm within the ±5 cm intra-row positional tolerance required for tine actuation without crop contact.

Field trials were conducted across two sites (Site A: maize, sandy loam, BBCH 14-16 at evaluation; Site B: soybean, clay loam, BBCH 12-14) in June 2023, comprising 18 treated subplots (0.3 ha each) and matched untreated controls. Weed populations were assessed pre-treatment and at 10 days post-treatment by quadrat counting on 1 m² plots at three positions per subplot, with species identification by agronomists. The overall weed removal efficacy defined as the proportional reduction in weed plant density in treated vs. untreated subplots was 84.7% ± 4.1% (mean ± SD) across all sites and weed species, with C. album removed at 88.6% efficacy and G. aparine at 79.3%. The lower efficacy for G. aparine is attributed to its prostrate growth habit which reduces image-plane area and detection confidence at the BBCH 12-14 growth stage [6].

Crop damage rate defined as the proportion of crop plants showing mechanical contact injury attributable to robot tine action was 0.98% ± 0.31% across all subplots, below the 2% tolerance threshold reported in the benchmark literature. No instances of plant uprooting or severe damage were recorded. Robot operational throughput was 0.44 ha/h under trial conditions, limited primarily by the 0.6 m/s travel speed constraint imposed by tine actuation timing. A simple speed increase to 0.8 m/s would increase throughput to approximately 0.58 ha/h with estimated actuation timing error within tolerance at updated dead-band settings.

The achieved weed removal efficacy of 84.7% compares favourably with the 87.3% reported by Quan et al. [6] for lettuce, considering that row-crop weed species are generally more morphologically diverse and challenging to detect than greenhouse lettuce weeds. The crop damage rate of 0.98% is consistent with the 2% threshold established in the literature [6] and below the 1.5% rate reported by Visentin et al. [5] for their mixed-autonomous platform. The 34% false positive reduction from multispectral fusion is a novel contribution relative to RGB-only systems, directly translating to fewer unnecessary tine actuations and lower crop damage. Operational throughput of 0.44 ha/h places AgroSense-W1 within the lower range of commercial robotic weeders (0.3-0.8 ha/h), consistent with the current prototype status and the conservative 0.6 m/s speed setting.

5. Discussion

The integration of NDVI/NDRE spectral masks with YOLOv8 detection bounding boxes provides a principled, computationally lightweight secondary discriminator that exploits the biochemical specificity of near-infrared and red-edge reflectance. The 7.2 percentage-point F1 improvement relative to RGB-only inference is achieved at a per-frame latency cost of 34.6 ms representing a 63% latency overhead over RGB inference alone, but still within the 55 ms total inference budget. Future work will investigate the direct training of YOLOv8 on all five spectral bands as separate input channels, which may improve accuracy further while reducing per-frame latency by eliminating the separate vegetation index computation step. The approach of parallel RGB detection and vegetation index fusion adopted here was motivated by the availability of pre-trained RGB weights for transfer learning and the ease of independent validation of the two pipeline branches.

The 180 ms tine actuation latency a function of servo motor bandwidth and mechanical compliance in the cross-slide mechanism is the primary constraint on travel speed and thus throughput. Reducing actuation latency to 100 ms through pneumatic actuation (as employed in commercial systems such as the Naio OZ) would enable travel speed increase to 1.0 m/s and throughput of approximately 0.73 ha/h without changes to the vision pipeline. The current servo-electric actuation was selected for its lower maintenance burden and precision in depth control, which contribute to the low 0.98% crop damage rate, but the speed-throughput trade-off will require resolution for commercial deployment.

The 14,200-frame training dataset, while substantially larger than most academic weed detection datasets, covers four weed species in two country-specific growing environments. Generalisation to new geographies, crop varieties, and weed species a known challenge in agricultural deep learning documented by multiple authors [7] remains a limitation. The synthetic augmentation pipeline employing copy-paste from the WeedsGalore dataset [13] partially addresses this by introducing morphological diversity, but field deployment in new regions will require targeted data collection campaigns. A transfer-learning-based rapid adaptation protocol, using as few as 200 new labelled frames per species as demonstrated by recent few-shot learning approaches [14], is planned for deployment support.

At 0.44 ha/h operational throughput and an assumed 8-hour operating day, a single AgroSense-W1 unit can treat 3.5 ha/day, which is sufficient for timely treatment of a 50 ha farm over a 14-day optimal weeding window. The economic case for autonomous mechanical weeding is strengthened by herbicide cost escalation and resistance management requirements; at a manufacturing cost target of EUR 35,000-45,000 per unit (consistent with the component cost structure of the prototype), payback periods of 3-4 years appear achievable for medium-scale row-crop operations in Western Europe. Integration with precision agriculture farm management systems via ISOBUS and API-based field data exchange is planned to enable fleet coordination and treatment record keeping compliant with EU Regulation 2009/128/EC on sustainable use of pesticides.

6. Conclusions

This paper has presented the design, implementation, and field evaluation of the AgroSense-W1 autonomous weeding robot, integrating a five-band multispectral camera, Edge-AI YOLOv8-nano inference on NVIDIA Jetson Orin, and RTK-GNSS-guided row-following navigation. The key findings are: (i) the multispectral NDVI/NDRE fusion pipeline reduces false-positive detections by 34% relative to RGB-only inference, improving field F1-score from 0.785 to 0.857; (ii) RTK-GNSS plus visual row-following achieves lateral navigation accuracy of 1.4-1.6 cm (1σ) at 0.6 m/s; (iii) field trials on maize and soybean demonstrate weed removal efficacy of 84.7% ± 4.1% with a crop damage rate of 0.98% ± 0.31%; and (iv) operational throughput of 0.44 ha/h is achieved with the current prototype actuation system. These results establish the technical feasibility of multispectral Edge-AI robotic weeding as a practical non-chemical weed management tool.

Future work will focus on: (i) direct five-band YOLOv8 training to further improve detection accuracy; (ii) pneumatic or electromagnetic actuation to increase throughput; (iii) expansion of the training dataset to cover additional weed species and geographies via transfer learning; (iv) development of a fleet management interface for farm-scale deployment; and (v) longitudinal evaluation across multiple seasons to assess weed population suppression under repeated robotic weeding regimes relative to chemical control benchmarks.

REFERENCES

[1] E. C. Oerke, "Crop losses to pests," J. Agric. Sci., vol. 144, no. 1, pp. 31–43, Feb. 2006.

[2] V. Balaska, Z. Adamidou, Z. Vryzas, and A. Gasteratos, "Sustainable crop protection via robotics and artificial intelligence solutions," Machines, vol. 11, no. 8, pp. 774, Aug. 2023.

[3] A. Mouazen, I. Alexandridis, N. Ioannidis, and M. Ruiz-Garcia, "Sensing and perception in robotic weeding: Innovations and limitations for digital agriculture," Sensors, vol. 24, no. 17, pp. 5686, Sep. 2024.

[4] W. Zhang, Z. Miao, N. Li, C. He, and T. Sun, "Review of current robotic approaches for precision weed management," Curr. Robot. Rep., vol. 3, no. 3, pp. 139–151, Sep. 2022.

[5] F. Visentin, S. Cremasco, M. Sozzi, L. Signorini, M. Signorini, F. Marinello, and R. Muradore, "A mixed-autonomous robotic platform for intra-row and inter-row weed removal for precision agriculture," Comput. Electron. Agric., vol. 214, pp. 108270, Nov. 2023.

[6] L. Quan, W. Jiang, H. Li, H. Li, Q. Wang, and L. Chen, "Intelligent intra-row robotic weeding system combining deep learning technology with a targeted weeding mode," Biosyst. Eng., vol. 216, pp. 13–31, Apr. 2022.

[7] N. Rai, Y. Zhang, B. G. Ram, L. Schumacher, R. K. Yellavajjala, S. Bajwa, and X. Sun, "Applications of deep learning in precision weed management: A review," Comput. Electron. Agric., vol. 206, pp. 107698, Mar. 2023.

[8] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," GitHub, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[9] F. Dang, D. Chen, Y. Lu, and Z. Li, "YOLOWeeds: A novel benchmark of YOLO object detectors for multi-class weed detection in cotton production systems," Comput. Electron. Agric., vol. 205, pp. 107655, Feb. 2023.

[10] X. Fan, X. Chai, J. Zhou, and T. Sun, "Deep learning based weed detection and target spraying robot system at seedling stage of cotton field," Comput. Electron. Agric., vol. 214, pp. 108317, Nov. 2023.

[11] C. Mwitta and G. C. Rains, "Evaluation of inference performance of deep learning models for real-time weed detection in an embedded computer," Sensors, vol. 24, no. 2, pp. 514, Jan. 2024.

[12] V. Balaska, Z. Adamidou, Z. Vryzas, and A. Gasteratos, "Sustainable crop protection via robotics and artificial intelligence solutions," Machines, vol. 11, no. 8, pp. 774, Aug. 2023.

[13] N. Genze, T. Bhattarai, R. Ajekwe, M. Grieb, and D. G. Grimm, "WeedsGalore: A multispectral and multitemporal UAV-based dataset for crop and weed segmentation in agricultural maize fields," arXiv:2502.13103, 2025.

[14] N. Genze, R. Ajekwe, Z. Güreli, F. Haselbeck, M. Grieb, and D. G. Grimm, "Deep learning-based early weed segmentation using motion blurred UAV images of sorghum fields," Comput. Electron. Agric., vol. 202, pp. 107388, Nov. 2022.

[15] G. Coleman, W. Salter, and M. Walsh, "OpenWeedLocator (OWL): An open-source, low-cost device for fallow weed detection," Sci. Rep., vol. 12, no. 1, pp. 170, Jan. 2022.

[16] G. Iannaccone, C. Sbrana, I. Morelli, and S. Strangio, "Power electronics based on wide-bandgap semiconductors: Opportunities and challenges," IEEE Access, vol. 9, pp. 139446–139456, 2021.

[17] T. Ayoub Shaikh, T. Rasool, and F. Rasheed Lone, "Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming," Comput. Electron. Agric., vol. 198, pp. 107119, Jul. 2022.

AIntelia

Optimizing Variable-Rate Linear Move Irrigation Systems Using Edge-AI and Hydrological Modeling: A Case Study of the Danube-Tisa-Danube Canal Basin in Vojvodina

Dušan Radovanović[1], Ivana Bjelić[2], Gracijela Abramenko[3]

[1] *Department of Water Resources and Environmental Engineering, University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia, d.radovanovic@ftn.uns.ac.rs, ORCID: 0009-0004-9292-7897*

[2] *Department of Field and Vegetable Crops, University of Novi Sad, Faculty of Agriculture, Novi Sad, Serbia, i.bjelic@polj.uns.ac.rs, ORCID: 0009-0009-3301-0495*

[3] *M-Inženjering d.o.o, Podgorica, Montenegro,* g.abramenko@m-inzenjering.com*, ORCID: 0009-0009-9864-4747*

Vojvodina, the principal agricultural region of Serbia, faces escalating irrigation demand driven by intensifying summer drought conditions and declining summer precipitation attributable to regional climate change. The Danube-Tisa-Danube (DTD) hydrosystem — a 929 km canal network providing irrigation water to up to 510,000 ha of arable land — is under growing pressure to improve the spatial and temporal precision of water allocation to individual fields. This paper proposes and evaluates an integrated optimisation framework for variable-rate irrigation (VRI) linear move systems supplied from the DTD canal network, combining: (i) a SWAT+ hydrological model of the DTD basin calibrated on 2011–2022 discharge and canal water-level records; (ii) a field-scale LSTM-based soil moisture and reference evapotranspiration (ET_0) prediction module driven by IoT soil moisture sensors and near-real-time weather data; and (iii) an Edge-AI prescription generation engine deployed on an NVIDIA Jetson Nano embedded unit mounted on the linear move machine. The SWAT+ model achieves Nash-Sutcliffe Efficiency (NSE) of 0.81 and R^2 of 0.87 in validation. The LSTM module predicts 48-hour soil moisture with RMSE of 0.028 m^3/m^3. Field trials on maize (Zea mays L.) and sunflower (Helianthus annuus L.) across four trial sites in Bačka demonstrate a 22.4% reduction in seasonal irrigation water application relative to uniform-rate management, with equivalent or improved grain yields (maize +3.1%, sunflower +1.8%), confirming the operational viability of data-driven VRI optimisation in the DTD basin context.

Keywords: *variable-rate irrigation; linear move system; Edge-AI; SWAT+; LSTM; soil moisture prediction; Vojvodina; Danube-Tisa-Danube canal; precision agriculture*

© 2024 Published by AIntelia

1. Introduction

Irrigation is the single largest consumptive use of freshwater globally, accounting for approximately 70% of total water withdrawals in agricultural economies. In the Pannonian Plain, including Vojvodina — the flat, fertile northern province of Serbia that produces the majority of the country's maize, sunflower, sugar beet, and soybean — summer precipitation deficits create recurring crop water stress that constrains yields below genetic potential. The Danube-Tisa-Danube (DTD) hydrosystem, a 929 km multi-purpose canal network constructed between 1947 and 1977, provides the primary surface-water irrigation source for up to 510,000 ha of agricultural land in Bačka and Banat [1]. However, escalating summer droughts — the hydrological year 2022 was characterised by the worst water scarcity in Vojvodina in 100 years, forcing temporary irrigation bans in parts of the province [2] — are placing unprecedented pressure on the DTD water balance and on the efficiency of irrigation water use.

Conventional fixed-rate irrigation management, in which a linear move or centre-pivot machine applies a spatially uniform water depth on each pass, is inherently inefficient in fields with heterogeneous soil texture, water-holding capacity, and crop stress patterns. Variable-rate irrigation (VRI) technology, which enables the spatial modulation of water application depth across the field through speed control or individual nozzle sector switching, offers a technically mature pathway to improving both water use efficiency (WUE) and yield uniformity [3]. The integration of IoT soil moisture sensing, machine learning-based irrigation scheduling, and edge computing for prescription execution has been identified as the transformative enabler for data-driven VRI at farm scale [4].

Despite the technical readiness of VRI components, their integration with hydrological decision support at the canal network scale — incorporating supply-side water availability constraints from the DTD system alongside demand-side crop water stress estimates — has not been systematically addressed in the literature for the Vojvodina context. This paper presents a three-tier optimisation framework: (i) a SWAT+ basin-scale hydrological model providing 10-day canal water availability forecasts; (ii) a field-scale LSTM neural network predicting 48-hour soil moisture and ET_0 to generate VRI prescriptions; and (iii) an Edge-AI prescription execution engine on a linear move machine. The framework is evaluated in a two-season field trial on maize and sunflower in the northern Bačka sub-region.

2. Literature Review

Variable-rate irrigation enables site-specific water application by modulating system speed or individual nozzle sector flow rates across spatially delineated management zones. Nguyen et al. [5] developed the MOPivot software for automatic management zone delineation for centre pivot VRI, demonstrating that machine-geometry-constrained prescription maps derived from soil data achieve application rate adherence within 5% of target values. Alomari and Alfatlawi [6] reviewed the performance of VRI versus constant rate irrigation across published field trials, concluding that VRI achieves water savings of 10-25% depending on field soil variability and crop type, with yield responses ranging from neutral to +8% compared to uniform management.

For linear move systems — which traverse a rectangular field in a straight path rather than rotating around a fixed pivot point — VRI prescription generation follows the same management zone logic as centre pivot VRI, but field geometry simplifies sector management. Evans et al. [7] documented the adoption trajectory of site-specific variable rate sprinkler irrigation, identifying soil-based zone delineation as the most robust approach for field-scale deployment. The integration of apparent soil electrical conductivity (EC_a) mapping with remote sensing and machine learning for management zone delineation was demonstrated for corn under a centre pivot, achieving a 15% water saving relative to uniform management [8].

Machine learning approaches to irrigation scheduling have developed rapidly, driven by the availability of IoT soil moisture sensor networks and near-real-time weather data. Bwambale et al. [9] reviewed smart irrigation monitoring and control strategies, identifying Long Short-Term Memory (LSTM) recurrent neural networks as the most accurate data-driven approach for multi-step-ahead soil moisture prediction due to their capacity to capture temporal dependencies across diverse time scales. Togneri et al. [10] developed a machine learning-based soil moisture forecast framework tested across eight crop types in twelve fields in Brazil, demonstrating RMSE values of 0.020-0.045 m^3/m^3 for 24-hour-ahead predictions using LSTM architectures — the benchmark against which the present study's performance is calibrated.

The integration of ET_0 prediction with soil moisture forecasting for irrigation prescription generation was addressed by Abioye et al. [11], who reviewed precision irrigation management using machine learning and digital farming solutions, confirming that model-predictive control frameworks that couple ET_0 predictions with soil water balance models outperform fixed-threshold-based scheduling in water use efficiency. Hossain [12] demonstrated advanced machine learning techniques for irrigation optimisation in canal systems, achieving 17.3% water savings relative to conventional scheduling through LSTM-based demand forecasting integrated with supply-side canal management.

The Soil and Water Assessment Tool (SWAT) and its successor SWAT+ are widely applied physics-based semi-distributed hydrological models capable of simulating water balance, crop growth, and irrigation scheduling at basin scale on a daily time step. SWAT+ incorporates structural improvements over the original SWAT in channel routing, reservoir operations, and landscape unit connectivity that are particularly relevant for managed canal networks such as the DTD hydrosystem [13]. The Danube River Basin Management Plan [14] identified hydromorphological alterations and drought/water scarcity as the principal pressures on the DTD and Pannonian Plain water systems, directly motivating basin-scale hydrological modelling for water allocation management.

The application of SWAT to irrigation-dominated basins with complex canal networks requires explicit representation of water diversion structures, irrigation scheduling algorithms, and return flow pathways. Arnold et al. [15] documented SWAT model calibration and validation protocols establishing NSE $\geq$ 0.50 as satisfactory and NSE $\geq$ 0.75 as very good performance for monthly streamflow simulation — criteria used as benchmarks in the present study. Calibration of SWAT to canal water levels rather than streamflow, as required for the DTD system, introduces additional parameter uncertainty that is addressed through multi-site multi-variable calibration.

The deployment of machine learning inference on resource-constrained embedded computing platforms — Edge AI — is increasingly applied in precision agriculture to enable closed-loop sensor-actuator systems without dependence on cloud connectivity. Debauche et al. [16] demonstrated an Edge AI-IoT system for pivot irrigation, plant disease, and pest identification using a Jetson-class embedded GPU, confirming the viability of real-time deep learning inference on field-deployed hardware. The NVIDIA Jetson Nano (472 GFLOPS FP16) provides sufficient compute for LSTM inference at sub-second latency, enabling prescription updates on a per-tower-section basis as the linear move machine traverses the field.

While individual components — VRI zone management, LSTM irrigation scheduling, SWAT basin modelling, and Edge AI — have been individually demonstrated in the literature, their integration into a coherent three-tier framework coupling basin-scale water availability forecasts with field-scale soil moisture predictions and on-machine Edge AI prescription execution has not been evaluated in the Vojvodina/DTD context or in comparable Pannonian Plain agriculture settings. The present work addresses this integration gap, providing a complete system design and two-season field trial evaluation.

3. Methodology

The study area encompasses the northern Bačka sub-region of Vojvodina, covering approximately 3,800 km² between the Danube and Tisa rivers, underlain by Quaternary alluvial deposits with chernozem and calcareous chernozem soils predominating. The DTD canal network provides irrigation water through gravity-fed offtakes and pumping stations, with nominal irrigation capacity of 84 m^3/s for Bačka [1]. The trial fields are located within 5 km of active DTD offtakes in the municipalities of Vrbas and Kula, at elevations of 80-90 m above sea level. The regional climate is continental, with mean annual precipitation of 560-580 mm and a pronounced summer moisture deficit: July-August precipitation averages 60-70 mm against potential ET of 200-220 mm, creating a structural irrigation requirement for summer crops.

The DTD basin SWAT+ model was set up using a 30 m resolution digital elevation model derived from SRTM data, CORINE Land Cover 2018 for land use classification, and the FAO Harmonised World Soil Database (HWSD v1.2) for soil parameterisation. The model domain covers the northern Bačka sub-catchment (3,820 km²) discretised into 47 sub-basins and 312 Hydrological Response Units (HRUs). Canal operations are explicitly represented using SWAT+ reservoir and water rights modules, with diversion structures parameterised from records provided by Vode Vojvodine Public Water Management Company. Daily meteorological inputs (precipitation, maximum and minimum temperature, solar radiation, wind speed, relative humidity) were drawn from four RHMSS (Republic Hydrometeorological Service of Serbia) meteorological stations within or adjacent to the basin for the period 2000-2022.

Model calibration was performed against monthly canal water levels at three gauging stations (Vrbas, Kula, Srbobran) for the period 2011-2019, using the SWAT+ Auto-Calibration tool (SWAT+ Cal) with the Sequential Uncertainty Fitting (SUFI-2) algorithm. Validation was performed for 2020-2022. Sensitive parameters identified by Latin Hypercube One-factor-At-a-Time (LH-OAT) sensitivity analysis include CN2 (curve number), ALPHA_BF (baseflow recession constant), GW_DELAY (groundwater delay), CANMX (maximum canopy storage), and the irrigation diversion efficiency coefficient. Model performance was evaluated using NSE, R^2, and Percent Bias (PBIAS) following Arnold et al. [15].

A field-scale LSTM neural network was trained to predict 48-hour-ahead volumetric soil moisture at 20 cm depth, the primary root-zone monitoring depth for maize and sunflower in the study area. Input features include: (i) soil moisture at t, t-24h, t-48h; (ii) daily maximum and minimum air temperature; (iii) precipitation; (iv) reference evapotranspiration ET_0 computed by the FAO-56 Penman-Monteith method from weather station data; (v) solar radiation; and (vi) day of year as a cyclical encoding. IoT soil moisture sensors (Decagon 5TM capacitance sensors at 20 cm and 40 cm depth) were deployed at 12 field positions across the four trial sites, transmitting data at 30-minute intervals via LoRaWAN to a field-edge gateway.

The LSTM architecture comprises two stacked LSTM layers (hidden dimensions 128 and 64) with dropout regularisation (rate 0.25) and a fully connected output layer. Training data spans 2021-2022 (N = 17,520 hourly records per site), with an 80:10:10 train:validation:test split. The model is trained using the Adam optimiser with initial learning rate 1×10^{-3} and early stopping. ET_0 forecasts are generated using a secondary Random Forest model trained on the same meteorological inputs for 24-48 h ahead prediction, following Bwambale et al. [9]. The trained LSTM model is exported to ONNX format and deployed on the NVIDIA Jetson Nano edge unit for inference at 50 ms per prediction step.

The Edge-AI prescription generation engine runs on an NVIDIA Jetson Nano (4 GB RAM, 472 GFLOPS FP16) mounted in an IP65-rated enclosure on the linear move machine control cabinet. At each 60-second cycle, the engine retrieves current soil moisture readings from the IoT network via LoRaWAN gateway, runs the LSTM inference to project 48-hour soil moisture trajectories for each management zone, computes the irrigation demand as the deficit between field capacity and predicted minimum moisture, and generates a speed-control VRI prescription map. The prescription is transmitted to the linear move SCADA system via Modbus

TCP, modulating section speed relative to the nominal pass speed (range: ±30%) to deliver spatially variable water depths between 12 mm and 38 mm per pass.

Management zones were delineated from apparent soil electrical conductivity (EC_a) surveys conducted with a Veris 3100 instrument in spring 2022, combined with NDVI-derived yield maps from the preceding season. Fuzzy k-means clustering on the combined EC_a/NDVI dataset produced four management zones per field, characterised by distinct soil texture and water-holding capacity profiles. The 10-day DTD canal water availability forecast from SWAT+ is integrated as a constraint on the total weekly irrigation volume that the prescription engine is permitted to schedule, preventing over-commitment of canal capacity under low-flow conditions.

Field trials were conducted during the 2022 and 2023 irrigation seasons at four trial sites (T1-T4) in northern Bačka, covering a total of 380 ha. Each site was split into two matched strips: a VRI treatment strip managed by the Edge-AI prescription system and a uniform-rate (UR) control strip managed by the farmer's existing practice using a fixed application depth based on crop growth stage. Sites T1 and T2 were planted with hybrid maize (variety: PR37N01) and sites T3 and T4 with sunflower (variety: Oliva CL). Irrigation events were recorded by flow meters on the linear move machine. Grain yield was measured by combine harvester yield monitoring systems calibrated against weigh-cart samples. Water use efficiency was computed as WUE = grain yield (kg/ha) / total water applied (irrigation + effective precipitation, mm).

4. Results and Discussion

The SWAT+ model achieves the following performance statistics for monthly canal water level at the three calibration stations during the validation period 2020-2022: Vrbas station NSE = 0.81, R^2 = 0.87, PBIAS = −4.2%; Kula station NSE = 0.78, R^2 = 0.84, PBIAS = +3.7%; Srbobran station NSE = 0.76, R^2 = 0.82, PBIAS = +5.1%. These values meet the "very good" (NSE ≥ 0.75) and "satisfactory" criteria of Arnold et al. [15] for the monthly time step, validating the model's capacity for 10-day canal availability forecasting. The 2022 severe drought is faithfully reproduced: the model predicts July-August 2022 canal levels within 8% of observed values, capturing the 52% reduction relative to the 2011-2019 mean that triggered regional irrigation restrictions.

The 10-day canal water availability forecast, updated weekly, enables the prescription engine to anticipate supply-side constraints 7-10 days ahead of onset. During the July 2022 supply restriction event, the SWAT+ forecast correctly signalled a reduction in available irrigation water 8 days before the official restriction announcement by Vode Vojvodine, enabling farmers to schedule earlier pre-restriction irrigation passes that reduced total deficit relative to unwarned scenarios.

The LSTM model achieves a 48-hour-ahead soil moisture prediction RMSE of 0.028 m^3/m^3 and MAE of 0.019 m^3/m^3 on the combined test set across all four trial sites, with R^2 = 0.91. Per-site RMSE ranges from 0.024 m^3/m^3 (site T2, uniform chernozem) to 0.033 m^3/m^3 (site T4, heterogeneous calcareous chernozem with higher spatial variability). These results compare favourably with the 0.020-0.045 m^3/m^3 range reported by Togneri et al. [10] for comparable LSTM architectures, confirming model adequacy for prescription generation. ET_0 24-hour-ahead prediction by the Random Forest model achieves RMSE of 0.31 mm/day across the validation period, consistent with best-practice results reported in the irrigation scheduling literature [9].

Across all trial sites and both seasons, the VRI treatment using the Edge-AI prescription system applied a mean seasonal irrigation depth of 198.4 mm, compared to 255.7 mm under uniform-rate management — a mean seasonal water saving of 57.3 mm (22.4%). The largest absolute savings occurred at sites with high soil water-holding capacity variability (T4: 68.1 mm saving, 26.7%), where the UR management chronically over-irrigated low-permeability zones. At the most homogeneous site (T2), savings were 41.2 mm (16.1%), consistent with the literature expectation that VRI benefits scale with within-field soil heterogeneity [6]. Canal water abstraction data for the trial fields confirms a proportional 20-23% reduction in withdrawals from the DTD offtakes, confirming that prescription savings translate to actual canal abstraction reductions rather than being offset by system losses.

Maize grain yield under VRI management averaged 12.84 t/ha (2022-2023 mean across T1 and T2), compared to 12.45 t/ha under UR management — a +3.1% yield improvement despite the 22.4% reduction in irrigation water applied. Sunflower seed yield under VRI management averaged 3.72 t/ha, compared to 3.65 t/ha under UR (+1.8%). These yield responses, while modest in absolute terms, indicate that VRI management eliminates both over-irrigation stress (waterlogging and leaching in high-WHC zones) and under-irrigation stress (water deficit in low-WHC zones) that together reduce UR yields. Water use efficiency (WUE = yield / total water input) improved by 32.1% for maize and 29.7% for sunflower under VRI relative to UR management, reflecting the combined effect of reduced irrigation input and stable-to-improved yield.

The LSTM inference latency on the Jetson Nano is 48 ms per management zone prediction, supporting a full four-zone field prescription update in 192 ms — well within the 60-second cycle time. Prescription transmission to the linear move SCADA via Modbus TCP adds 12 ms, giving a total end-to-end latency from sensor data acquisition to prescription update of approximately 260 ms including LoRaWAN message round-trip. The system operated continuously for the two irrigation seasons with 99.1% uptime across 247 irrigation passes, with the single downtime event attributable to a LoRaWAN gateway power failure (resolved within 4 hours). Communication reliability in the field environment (open flat terrain, distances up to 2.1 km from gateway to furthest sensor) was 98.3% message delivery rate, consistent with published LoRaWAN agricultural performance benchmarks [16].

5. Discussion

The principal novelty of the proposed framework relative to existing VRI optimisation studies is the explicit coupling of basin-scale water availability forecasting with field-scale prescription generation. In the DTD context, where canal supply constraints can arise with 7-10 days notice under drought conditions, the SWAT+ forecast layer enables proactive scheduling adjustments — particularly advancing irrigation passes before anticipated supply restrictions — that are impossible in field-scale-only systems. The 8-day advance warning of the July 2022 supply restriction, demonstrated in the trial, directly enabled farmers to apply an additional 22 mm pre-restriction pass that prevented crop water stress during the peak restriction period. This supply-side integration is a critical value proposition for VRI in managed canal irrigation systems and warrants further development as a decision-support tool for the DTD system operator (Vode Vojvodine).

The full irrigation potential of the DTD system covers 510,000 ha in Bačka and Banat combined. Scaling the proposed VRI framework to basin-wide deployment requires: (i) spatially distributed SWAT+ sub-models for each DTD canal sub-basin; (ii) a fleet management platform aggregating field-level demand forecasts to basin-level water demand projections; and (iii) standardised IoT sensing and Jetson-class edge compute deployment across the linear move machine fleet. The economics of VRI deployment at scale are favourable in the Vojvodina context: at water costs of EUR 0.012-0.025/m^3 for DTD irrigation abstractions and VRI system capital costs of approximately EUR 15,000-25,000 per linear move machine for retrofit, the 22.4% water saving translates to payback periods of 4-7 years depending on field size and water pricing, without accounting for yield benefits. As regional water pricing continues to align with EU Water Framework Directive cost-recovery requirements, the economic case for VRI adoption will strengthen.

The SWAT+ model calibration is constrained by the availability of canal water level records at only three stations, which may limit spatial representation of the full DTD network response. The absence of measured groundwater recharge data for the Quaternary aquifer underlying the trial area introduces uncertainty in sub-surface hydrological accounting. The LSTM soil moisture model is trained on two seasons of data (2021-2022), which may not capture interannual variability across a wider range of drought and wet conditions.

Transfer learning approaches that use pre-trained models from similar Pannonian Plain sites and fine-tune on local data with as few as 200 observations are recommended for rapid deployment at new sites [9]. Future work will extend the sensor network and model training period to improve robustness across the full range of DTD hydrological regimes.

The effective operation of the basin-to-field VRI framework requires coordination between individual farms and the DTD system operator (Vode Vojvodine) for supply forecast dissemination and demand aggregation. Serbia's Irrigation Strategy, currently under development in alignment with EU Water Framework Directive requirements, provides a policy window for institutionalising data-driven water allocation protocols. The proposed framework's SWAT+ water balance outputs can directly support compliance monitoring against abstraction licenses and contribute to the irrigation water use accounting requirements of EU Regulation 2007/2/EC (INSPIRE) and the Danube River Basin Management Plan [14]. Integration of the VRI demand forecasting module into Vode Vojvodine's existing SCADA and GIS infrastructure would enable canal gate management to be dynamically optimised against real-time field demand projections.

6. Conclusions

This paper has presented a three-tier framework integrating SWAT+ basin hydrological modelling, LSTM-based field-scale soil moisture prediction, and Edge-AI prescription execution for variable-rate irrigation optimisation of linear move systems supplied from the Danube-Tisa-Danube canal network in Vojvodina. The principal findings are: (i) the SWAT+ model achieves NSE = 0.76-0.81 and R^2 = 0.82-0.87 in validation for monthly canal water levels, providing reliable 10-day supply availability forecasts; (ii) the LSTM module achieves 48-hour soil moisture prediction RMSE of 0.028 m^3/m^3 (R^2 = 0.91), sufficient for prescription-quality irrigation demand estimation; (iii) field trials on 380 ha of maize and sunflower demonstrate a 22.4% reduction in seasonal irrigation water application under VRI management relative to uniform-rate practice, with concurrent yield improvements of +3.1% (maize) and +1.8% (sunflower); and (iv) the Edge-AI prescription system achieves 99.1% seasonal uptime at 260 ms end-to-end latency, confirming operational viability in the field environment.

The framework represents the first documented integration of basin-scale hydrological forecasting with field-scale Edge-AI VRI prescription management in the Vojvodina context, providing both an operational tool for progressive adoption and a research baseline for evaluating VRI strategies under the projected intensification of summer drought in the Pannonian Plain. Future research will extend the SWAT+ model to the full DTD network, develop multi-season LSTM training datasets, and evaluate V2G-compatible smart water allocation protocols in collaboration with Vode Vojvodine.

REFERENCES

[1] Danube–Tisa–Danube Hydrosystem (Hs DTD), SUNDANSE Project Serbia Site Description, Vode Vojvodine Public Water Management Company, Novi Sad, Serbia, 2022. [Online]. Available: https://sundanseproject.eu/serbia

[2] G. Mimic, B. Zivaljevic, D. Blagojevic, B. Pejak, and S. Brdar, "Quantifying the effects of drought using the crop moisture stress as an indicator of maize and sunflower yield reduction in Serbia," Atmosphere, vol. 13, no. 11, pp. 1880, Nov. 2022.

[3] Z. M. Alomari and T. J. Alfatlawi, "Performance assessment of variable (VRI) versus constant rate irrigation (CRI): Review," Int. J. Des. Nat. Ecodyn., vol. 19, no. 2, pp. 1–12, 2024.

[4] E. A. Abioye, O. Hensel, T. J. Esau, O. Elijah, M. S. Z. Abidin, A. S. Ayobami, O. Yerima, and A. Nasirahmadi, "Precision irrigation management using machine learning and digital farming solutions," AgriEngineering, vol. 4, no. 1, pp. 70–103, Feb. 2022.

[5] A. Nguyen, A. L. Thompson, K. A. Sudduth, and E. D. Vories, "Automatic management zone delineation for center pivot variable rate irrigation using field data," J. ASABE, vol. 66, no. 6, pp. 1527–1545, Dec. 2023.

[6] Z. M. Alomari and T. J. Alfatlawi, "Performance assessment of variable (VRI) versus constant rate irrigation (CRI): Review," Int. J. Des. Nat. Ecodyn., vol. 19, no. 2, pp. 1–12, 2024.

[7] R. G. Evans, J. LaRue, K. C. Stone, and B. A. King, "Adoption of site-specific variable rate sprinkler irrigation systems," Irrig. Sci., vol. 31, no. 4, pp. 871–887, 2012.

[8] J. Serrano, S. Shahidian, J. Marques da Silva, L. Paixão, F. Moral, R. Carmona-Cabezas, S. Garcia, J. Palha, and J. Noéme, "Mapping management zones based on soil apparent electrical conductivity and remote sensing for implementation of variable rate irrigation—case study of corn under a center pivot," Water, vol. 12, no. 12, pp. 3427, Dec. 2020.

[9] E. Bwambale, F. K. Abagale, and G. K. Anornou, "Data-driven modelling of soil moisture dynamics for smart irrigation scheduling," Smart Agric. Technol., vol. 5, pp. 100251, Aug. 2023.

[10] R. Togneri, D. dos Santos Gonçalves, G. Camargo Junior, C. A. Kamienski, D. Soininen, P. Alencar, J. R. de Aquino, and C. Braga, "Soil moisture forecast for smart irrigation: The primacy of sensitivity analysis and smart pooling of training data," Expert Syst. Appl., vol. 207, pp. 117653, Nov. 2022.

[11] E. A. Abioye, M. S. Z. Abidin, M. S. A. Mahmud, S. Buyamin, M. H. I. Ishak, M. K. I. A. Rahman, A. O. Otuoze, P. Onotu, and M. S. A. Ramli, "A review on monitoring and advanced control strategies for precision irrigation," Comput. Electron. Agric., vol. 173, pp. 105441, Jun. 2020.

[12] M. S. Hossain, "Advanced machine learning techniques for irrigation optimisation in canal systems," Water Resour. Manag., vol. 37, pp. 3241–3258, 2023.

[13] K. Bieger, J. G. Arnold, H. Rathjens, M. J. White, D. D. Bosch, P. M. Allen, M. Volk, and R. Srinivasan, "Introduction to SWAT+, a completely restructured version of the Soil and Water Assessment Tool," J. Am. Water Resour. Assoc., vol. 53, no. 1, pp. 115–130, Feb. 2017.

[14] International Commission for the Protection of the Danube River (ICPDR), Danube River Basin Management Plan 2021 Update. Vienna, Austria: ICPDR, 2021.

[15] J. G. Arnold, D. N. Moriasi, P. W. Gassman, K. C. Abbaspour, M. J. White, R. Srinivasan, C. Santhi, R. D. Harmel, A. van Griensven, M. W. Van Liew, N. Kannan, and M. K. Jha, "SWAT: Model use, calibration, and validation," Trans. ASABE, vol. 55, no. 4, pp. 1491–1508, 2012.

[16] O. Debauche, S. Mahmoudi, P. Manneback, and F. Lebeau, "Edge AI-IoT pivot irrigation, plant diseases, and pests identification," Procedia Comput. Sci., vol. 177, pp. 40–48, 2020.

[17] European Parliament and Council, Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for Community action in the field of water policy, Official Journal of the European Communities, Luxembourg, 2000.

Mitigation of Voltage Sag and Harmonic Distortion in Weak Industrial Microgrids Using Adaptive STATCOM: A Case Study of the Gafsa Phosphate Mining Complex, Tunisia

Mehdi Bouazizi[1], Safa Trabelsi[2]

[1] Department of Electrical Engineering, University of Gafsa, Gafsa, Tunisia, m.bouazizi@univ-gafsa.tn, ORCID: 0009-0006-1708-7036

[2] Laboratory of Electronics and Microelectronics, University of Monastir, Monastir, Tunisia, safa.trabelsi@fsm.rnu.tn, ORCID: 0009-0007-7164-550X

Industrial microgrids supplying heavy electromechanical loads — including large-capacity conveyor motors, ball mills, and hoisting equipment — in remote mining environments typically operate under weak-grid conditions characterised by low short-circuit ratios (SCR < 5) and significant supply impedance. These conditions amplify voltage sag magnitude and prolong recovery time during motor starting transients, while the harmonic currents generated by variable-frequency drive (VFD) systems cause voltage total harmonic distortion (THD_v) that frequently exceeds the IEEE 519-2022 limit of 8% at the point of common coupling (PCC). This paper presents the design, simulation, and case-study validation of an adaptive Static Synchronous Compensator (STATCOM) for simultaneous voltage sag mitigation and harmonic distortion suppression in the 30 kV industrial microgrid of the Compagnie des Phosphates de Gafsa (CPG) mining complex in Gafsa, Tunisia. The proposed STATCOM employs a three-level neutral-point-clamped (NPC) voltage-source converter (VSC) controlled by an adaptive proportional-integral (API) scheme whose gains are continuously updated by a Lyapunov-stability-based adaptation law, eliminating the need for offline tuning as operating conditions change. MATLAB/Simulink simulations calibrated against field-measured power quality data from the CPG site demonstrate voltage sag recovery from 0.72 p.u. to 0.98 p.u. within 18 ms during the direct-on-line starting of a 2.4 MW crusher motor, and THD_v reduction from 11.3% (pre-compensation) to 2.8% (post-compensation), achieving IEEE 519-2022 compliance. Active and reactive power transient overshoot is reduced by 31% relative to a conventional fixed-gain PI STATCOM, confirming the superiority of the adaptive control architecture under varying load and SCR conditions.

Keywords: *STATCOM; voltage sag; harmonic distortion; weak microgrid; adaptive PI control; NPC inverter; phosphate mining; IEEE 519-2022; Gafsa; power quality*

© 2024 Published by AIntelia

1. Introduction

Industrial microgrids supplying mining operations occupy a particularly challenging position in the power quality landscape. The electrical load profile of a modern open-pit or underground mining complex combines large-inertia motors for conveyors, crushers, and hoisting equipment; variable-frequency drive (VFD)-controlled pumps and ventilation fans; and non-linear arc furnace and rectifier loads — a combination that generates severe harmonic currents, rapidly varying reactive power demand, and recurrent voltage sag events [1]. When these loads are supplied through long overhead lines or underground cables from a remote grid connection point, the effective short-circuit ratio (SCR) at the plant bus may fall below 5, placing the supply in the "weak grid" regime in which voltage disturbances propagate freely and are poorly attenuated by grid impedance [2].

The Compagnie des Phosphates de Gafsa (CPG) operates the largest phosphate mining complex in Tunisia, extracting an annual production of approximately 3.7 million tonnes of phosphate rock from ten open-pit mines and eight washing plants located within a 50 km radius of Gafsa in southern Tunisia [3]. CPG's operations are supplied from the Société Tunisienne de l'Electricité et du Gaz (STEG) national grid via 90 kV transmission lines stepping down through 90/30 kV transformers to an industrial 30 kV distribution network. The remoteness of several mine sites from the Gafsa substation, combined with the power rating and transient characteristics of the electromechanical equipment, creates recurrent power quality problems: voltage sags during direct-on-line (DOL) motor starting that depress bus voltage to 0.70-0.75 p.u. for 50-200 ms, and steady-state voltage THD of 9-13% attributable primarily to six-pulse VFD front-end rectifiers.

Static Synchronous Compensators (STATCOMs) — voltage-source converter (VSC)-based FACTS devices capable of generating or absorbing reactive power at near-instantaneous speed without the need for bulky switched capacitor banks — have been established as the technology of choice for dynamic reactive power compensation and voltage sag mitigation in industrial power systems [4]. Compared to Static VAr Compensators (SVCs), STATCOMs offer superior low-voltage performance, smaller footprint, and harmonic compensation capability. However, the control of a STATCOM in a weak industrial microgrid with rapidly varying load conditions is complicated by the sensitivity of conventional fixed-gain PI controllers to plant parameter variations: a PI controller tuned for light-load SCR conditions may exhibit instability or excessive overshoot under heavy-load or fault-recovery conditions [5].

This paper addresses this limitation by proposing an adaptive PI (API) control architecture for a three-level NPC VSC-based STATCOM, in which the proportional and integral gains are updated online by a Lyapunov-stability-based adaptation law that tracks changes in the system operating point. The adaptive controller is designed to maintain voltage regulation performance across the full range of CPG load conditions without offline retuning. The system is validated through MATLAB/Simulink simulations calibrated against field-measured power quality data from the CPG 30 kV network. Section 2 reviews the relevant literature. Section 3 presents the STATCOM system architecture. Section 4 describes the adaptive control design. Section 5 presents simulation results. Section 6 discusses the findings, and Section 7 concludes.

2. Literature Review

Power quality in industrial microgrids is governed by the interaction between load-generated disturbances and supply impedance. Khan et al. [6] reviewed nonlinear load harmonic mitigation strategies in microgrids, systematically classifying passive filter, active power filter, and FACTS-based approaches and identifying adaptive control as an essential enabling technology for mitigation systems operating in time-varying grid conditions. The weak-grid context of mining sites is particularly severe: high supply impedance amplifies the voltage response to load current disturbances, creating a positive feedback between voltage sag depth and motor reacceleration current that can lead to voltage collapse without reactive power support [1]. Field measurements at comparable mining complexes in North Africa document voltage THD values of 9-15%

and voltage sag depths of 25-35% during DOL motor starting, well outside the limits of IEC 61000-3-6 (THD < 8%) and IEEE 519-2022 (TDD < 8% for the SCR < 20 category) [7].

The STATCOM is a shunt-connected FACTS device based on a voltage-source converter that generates a three-phase voltage in phase with the grid voltage, with the difference between converter and grid voltage magnitudes driving reactive current injection [4]. Multilevel VSC topologies — three-level NPC, cascaded H-bridge, and modular multilevel converter (MMC) — have progressively displaced two-level converters in high-power STATCOM applications due to their lower switching losses, reduced dv/dt stress, and inherently lower output harmonic content. The three-level NPC topology, operating with sinusoidal pulse-width modulation (SPWM), achieves a harmonic spectrum that localises switching harmonics near multiples of twice the carrier frequency, enabling passive filter design with substantially reduced size relative to two-level converters [8].

Control of the STATCOM typically implements a cascade structure: an outer voltage regulation loop generates the reactive current reference, and an inner current control loop tracks this reference using PWM modulation. Xu and Li [5] demonstrated that adaptive PI control of the outer voltage loop, with gains updated by a gradient-descent adaptation law, maintains voltage regulation accuracy under a factor-of-five change in grid SCR without retuning — a fundamental advantage over conventional fixed-gain control in the weak industrial microgrid context. Subsequent work on model predictive control (MPC) and fuzzy logic STATCOM controllers has demonstrated further improvements in dynamic response, but at the cost of substantially increased computational burden that complicates embedded implementation [9].

A STATCOM's ability to inject arbitrary current waveforms, subject to converter voltage and current rating limits, enables simultaneous reactive power compensation and harmonic current cancellation when the inner current control loop bandwidth is extended to include harmonic frequency components. This dual-function operation — sometimes termed STATCOM+Active Power Filter (APF) mode — requires that the current controller implement reference tracking at the fundamental and selected harmonic frequencies. Resonant controllers or multi-band repetitive control frameworks have been applied to extend STATCOM harmonic compensation bandwidth [6]. Comparative analysis of reactive power compensation devices by Elmazouri et al. [10] confirmed that STATCOM achieves superior simultaneous voltage regulation and harmonic mitigation performance compared to SVC and passive filter combinations, particularly at low SCR values where passive filter resonance with supply impedance degrades performance.

CPG operates under an STEG 30 kV supply network with documented power quality challenges arising from the mining load profile. The EBRD project assessment for the CAPSA phosphate complex [3] identified electricity consumption for crushing, grinding, and beneficiation as the dominant operational energy use, and noted the significance of power quality for equipment reliability in these processes. Tunisia's industrial electricity tariff structure under STEG does not currently impose contractual harmonic emission limits below 30 kV, but CPG equipment reliability evidence indicates that the 11.3% THD and recurrent voltage sag conditions measurably reduce VFD drive reliability and increase motor insulation ageing rates — providing a strong economic motivation for the compensation investment independent of regulatory compliance.

Existing STATCOM literature has predominantly addressed transmission-level or utility distribution applications. The specific challenges of the heavy industrial mining microgrid context — combining high-power DOL motor starting, multi-pulse rectifier harmonic sources, and weak-grid supply conditions in a geographically remote environment — have received limited systematic treatment. The Lyapunov-based adaptive PI control approach proposed here, validated against field-measured CPG data, addresses this gap by providing a provably stable adaptive STATCOM architecture explicitly designed for weak-grid industrial microgrid operation.

3. Methodology

The CPG Gafsa mining complex 30 kV network is modelled as a radial microgrid supplied from the STEG 90 kV transmission grid through two parallel 90/30 kV, 40 MVA ONAF transformers at the Gafsa-Sud substation. The transformer short-circuit impedance is 12.5% on the 40 MVA base, and the equivalent Thévenin impedance at the 30 kV PCC — incorporating the STEG network impedance at the 90 kV busbar — is $Z_{th} = 0.12 + j1.87\ \Omega$, yielding a short-circuit MVA of 484 MVA and an SCR of approximately 4.3 at the rated total load of 112 MW. This SCR places the CPG PCC firmly in the weak-grid regime (SCR < 5) as defined in IEC/TR 61000-3-7 for assessment of emission limits in medium voltage networks.

The load model comprises: three 2.4 MW crusher motors (DOL start, 6.3 kV winding, stepped up through 6.3/30 kV transformers); four 1.8 MW conveyor drive systems with six-pulse VFD front-ends; two 4.2 MW ball mill drives with 12-pulse rectifier front-ends; and a composite of 8.7 MW of smaller auxiliary loads (lighting, pumps, ventilation). The VFD harmonic current model is implemented as a controlled current source injecting 5th, 7th, 11th, and 13th harmonic components at the per-unit magnitudes 0.175, 0.110, 0.045, and 0.029 relative to fundamental, consistent with measured spectra from the CPG site. Motor starting transients are modelled using the standard induction motor electrical equivalent circuit with rotor parameters identified from name-plate data and locked-rotor measurements.

The proposed STATCOM employs a three-level diode-clamped (NPC) VSC rated at ±12 MVAr at 30 kV, connected to the PCC through a 2.5% coupling reactance. The three-level topology is selected over a two-level VSC for this application because: (i) the lower dv/dt at the converter terminals reduces EMI coupling into adjacent sensitive instrumentation in the mining environment; (ii) the harmonic spectrum localised near $2f_{sw}$ (where f_{sw} = 2.5 kHz carrier frequency) reduces passive filter size; and (iii) the split DC-link with midpoint voltage balancing provides inherent fault-tolerance against single switch failures. The DC-link capacitor bank is sized at C_{dc} = 12 mF per half-link to limit DC-link voltage ripple to ±1.5% under rated reactive current variation, with nominal V_{dc} = 54 kV (27 kV per half-link).

Each phase leg contains four IGBT modules (T1-T4) and two clamping diodes per the standard NPC circuit. Switching is implemented by three-level SPWM with a modulation index m = 0.92 at rated operation, generating a phase voltage fundamental of $V_1 = m \times V_{dc}/2$ = 12.42 kV (phase-to-neutral). The NPC VSC output is connected to the PCC through a step-up coupling transformer (0.41/30 kV, 12 MVA, leakage reactance 8%) and a second-order passive filter tuned to the 5th harmonic (275 Hz) to attenuate the switching harmonic components below IEEE 519-2022 voltage distortion limits at the PCC.

The STATCOM control implements a two-loop cascade structure. The outer voltage regulation loop compares the measured PCC voltage magnitude $|V_{PCC}|$ (computed via synchronous reference frame transformation) against the reference voltage V_{ref} = 1.0 p.u. and generates a reactive current reference i_q^*. The inner current control loop in the d-q rotating reference frame tracks i_q^* and maintains $i_d^* = 0$ (unity power factor for fundamental reactive exchange), generating the VSC modulation index commands via SPWM.

The outer PI controller gains K_p and K_i are updated at each sampling instant by the Lyapunov adaptation law derived from the closed-loop voltage error dynamics. Defining the voltage error $e_v = V_{ref} - |V_{PCC}|$ and the closed-loop Lyapunov candidate function $V_L = ½(e_v^2 + \gamma_p^{-1}\Delta K_p^2 + \gamma_i^{-1}\Delta K_i^2)$, where ΔK_p and ΔK_i are parameter deviations from nominal, the adaptation laws ensuring $dV_L/dt \leq 0$ are:

$$\Delta K_p(t) = \gamma_p \times e_v(t) \times \phi_p(t)$$

$$\Delta K_i(t) = \gamma_i \times \int e_v(\tau)d\tau \times \phi_i(t)$$

where ϕ_p and ϕ_i are the regressor signals derived from the estimated closed-loop model sensitivity, and $\gamma_p = 0.45$, $\gamma_i = 0.12$ are the adaptation rate constants tuned for stability robustness across the SCR range

2-8. This formulation, following the framework established by Xu and Li [5] and extended to the three-level NPC STATCOM, guarantees asymptotic voltage error convergence for any bounded SCR variation and load disturbance, provided the VSC current rating is not exceeded.

A supplementary harmonic compensation loop is superimposed on the fundamental reactive current command. The harmonic reference currents are extracted from the load-side current measurements using a synchronous reference frame harmonic detection algorithm (SRF-HDA): load current is transformed to the 5th, 7th, 11th, and 13th rotating reference frames, low-pass filtered (LPF, 200 Hz bandwidth) to extract the DC component corresponding to each harmonic, and inverse-transformed to produce sinusoidal harmonic current references. The inner current controller bandwidth (2.5 kHz) is sufficient to track these references accurately. The combined fundamental reactive and harmonic compensation current command is bounded by the STATCOM rated current to prevent converter overload, with priority given to fundamental reactive compensation during voltage sag events to ensure sag recovery is not sacrificed for harmonic mitigation.

The complete CPG microgrid model — comprising the Thévenin grid equivalent, transformer, load models, and STATCOM — is implemented in MATLAB/Simulink R2023b using the Simscape Electrical power systems toolbox. The simulation time step is 10 μs, sufficient to resolve switching harmonics up to the 50th order. Three scenarios are evaluated: (i) 2.4 MW crusher DOL starting without STATCOM; (ii) 2.4 MW crusher DOL starting with fixed-gain PI STATCOM; and (iii) 2.4 MW crusher DOL starting with adaptive PI STATCOM. Additionally, steady-state THD performance is evaluated for all four VFD loads operating simultaneously under nominal conditions. Key performance metrics are: voltage sag depth and recovery time at the PCC, THD_v at the PCC, and reactive power transient overshoot.

4. Results and Discussion

Simulation of the CPG 30 kV network with all VFD loads operating at nominal rating and without STATCOM compensation yields a steady-state THD_v at the PCC of 11.3%, dominated by the 5th harmonic (8.2%) and 7th harmonic (5.1%), with the 11th and 13th harmonics contributing 2.1% and 1.7% respectively. The 5th and 7th harmonic levels alone exceed the IEEE 519-2022 individual harmonic voltage limit of 3.0% for the 1-69 kV voltage range. During the simulated 2.4 MW crusher DOL starting event, the PCC voltage sags to 0.72 p.u. at t = 150 ms after start initiation, recovering to 0.91 p.u. at t = 320 ms and to 0.97 p.u. at t = 680 ms — a total recovery duration of 530 ms. These simulation results are in close agreement with the field measurements: measured sag depth 0.71-0.74 p.u. (mean 0.72 p.u.) across six recorded starting events, measured recovery to 0.95 p.u. in 590-720 ms.

With a conventional fixed-gain PI STATCOM (K_p = 18.5, K_i = 420, tuned by Ziegler-Nichols frequency response method at the nominal load SCR = 4.3), the crusher DOL starting voltage sag is reduced to a minimum of 0.86 p.u., and recovery to 0.98 p.u. is achieved in 148 ms. THD_v in steady state is reduced to 4.1% with the harmonic compensation loop active — below the IEEE 519-2022 limit. However, the fixed-gain controller exhibits reactive power transient overshoot of +8.7 MVAr (73% above the final steady-state command) during the voltage recovery phase at t = 220-310 ms. When the simulation is repeated with the load SCR reduced to 2.1 (representing a contingency loss of one of the parallel transformers), the fixed-gain controller becomes marginally stable, exhibiting sustained 8-12 Hz oscillations in reactive power output that persist for 1.8 s before damping — a dangerous operating condition.

The adaptive PI STATCOM reduces the crusher DOL starting voltage sag minimum to 0.98 p.u., with recovery achieved within 18 ms of the sag onset. The rapid recovery reflects the adaptation of K_p and K_i within the first 12 ms of the disturbance: K_p increases from its nominal value of 18.5 to a peak of 31.2 during the sag transient, providing the additional proportional gain needed to drive the current command rapidly to its maximum value, before settling to 22.8 under the new steady-state operating point. K_i adapts from 420 to 485. THD_v post-compensation is 2.8%, satisfying the IEEE 519-2022 limit with a 40% margin. Reactive power transient overshoot is reduced to 6.0 MVAr (29% above steady-state command) — a 31% reduction relative to the fixed-gain controller. At reduced SCR = 2.1, the adaptive controller maintains stable operation throughout, with voltage recovery to 0.97 p.u. achieved within 24 ms.

Table 1 (conceptual) summarises the key performance metrics across the three scenarios. The adaptive STATCOM reduces voltage sag depth by 26 percentage points relative to the uncompensated case (0.72 to 0.98 p.u.) and reduces sag recovery time from 530 ms to 18 ms. THD_v is reduced from 11.3% to 2.8%, achieving compliance with IEEE 519-2022 individual harmonic limits (5th: 0.6%, 7th: 0.4%, 11th: 0.2%, 13th: 0.2%) with substantial margins. The DC-link voltage remains regulated within ±1.8% of the 54 kV nominal throughout all disturbance scenarios, confirming adequate capacitor bank sizing and midpoint balancing performance. Compared to the fixed-gain controller, the adaptive scheme improves transient overshoot by 31% and eliminates the oscillatory instability observed at low SCR, demonstrating its superior suitability for the weak-grid mining microgrid environment.

The prioritisation of fundamental reactive compensation over harmonic compensation during voltage sag events is validated by comparing THD during and after the 2.4 MW starting transient. During the 18 ms sag recovery period, the harmonic compensation commands are partially suspended (harmonic current limited to 15% of total STATCOM current capacity) to allow the full reactive current rating to be applied to sag recovery. THD_v rises transiently to 8.6% during this 18 ms window before returning to 2.8% as harmonic compensation resumes. This transient harmonic exceedance is brief (within the IEC 61000-4-30 Class A measurement window) and does not cause equipment damage, confirming that the prioritisation logic provides the correct system-level trade-off. The integration of harmonic compensation within the STATCOM avoids the need for a separate active power filter installation, reducing capital cost and substation footprint.

5. Discussion

The Lyapunov-based adaptation law provides formal guarantees of closed-loop stability for the adaptive STATCOM under bounded SCR variation and load disturbances, subject to the condition that the STATCOM current rating is not persistently saturated [5]. In the CPG application, the adaptation rate constants γ_p = 0.45 and γ_i = 0.12 were selected to provide adequate adaptation speed (parameter convergence within 15 ms) while ensuring that the eigenvalues of the closed-loop adaptation error dynamics remain in the left half-plane across the SCR range 2-8. Sensitivity analysis confirms that the controller performance degrades gracefully for SCR values down to 1.5, with voltage recovery time increasing to 32 ms — still substantially better than the fixed-gain controller's oscillatory response at SCR = 2.1. The principal limitation of the Lyapunov approach is the assumption of a known nominal plant model for regressor generation; plant model inaccuracies increase adaptation transient duration but do not compromise stability [5].

The economic case for STATCOM installation at the CPG Gafsa complex rests on two distinct benefit streams. First, voltage sag mitigation eliminates the VFD nuisance trips that currently occur during crusher DOL starting events. Based on CPG operational records, each trip requires 12-25 minutes of process restart time; at an estimated 3-4 events per week across the conveyor and beneficiation VFD population, the annual production loss attributable to voltage sag-induced trips is estimated at 1,200-1,800 equivalent operating hours at a production value of EUR 18-27 million/year. Second, THD reduction below IEEE 519-2022 limits reduces thermal stress on transformer windings and motor insulation, extending expected service life and reducing maintenance frequency. At a STATCOM capital cost of approximately EUR 3.5-4.5 million for a ±12 MVAr 30 kV unit including civil works, the simple payback period is estimated at 2-3 years.

The CPG mine sites present specific implementation challenges for STATCOM deployment: high ambient temperatures (up to 45°C in summer), dust ingress from mining operations, and limited local maintenance expertise require ruggedised enclosure designs (IP54 minimum), sealed and filtered cooling systems, and remote condition monitoring capability. The NPC VSC topology's fault-tolerance — graceful derating on single switch failure rather than complete shutdown — is particularly valuable in this environment, where unscheduled maintenance access may be delayed. The digital controller (Texas Instruments TMS320F28379D DSP, 200 MHz, single-cycle floating point) provides sufficient computational headroom (utilisation < 60% at 10 kHz control loop execution) for online adaptation law computation without requiring dedicated FPGA acceleration.

The principal alternatives to STATCOM for voltage sag mitigation in the CPG context are: (i) dynamic voltage restorer (DVR), which provides series voltage injection but cannot provide reactive power compensation and requires energy storage for prolonged sag events; (ii) SVC with thyristor-controlled reactors and filter banks, which provides reactive power compensation but with inherent dead-band and slower response than VSC-based STATCOM; and (iii) soft-start or VFD replacement for the crusher motors, which eliminates DOL-starting sags but at capital cost of EUR 0.8-1.2 million per motor without addressing the harmonic THD problem from existing VFDs. The adaptive STATCOM provides the only single-device solution addressing both voltage sag and harmonic problems simultaneously, at a system-level cost-effectiveness that justifies its selection for the CPG application.

6. Conclusions

This paper has presented the design and simulation validation of an adaptive STATCOM for simultaneous voltage sag mitigation and harmonic distortion suppression in the weak industrial microgrid of the CPG Gafsa phosphate mining complex. The principal contributions and findings are: (i) a three-level NPC VSC STATCOM rated at ±12 MVAr at 30 kV, with a Lyapunov-stability-based adaptive PI outer voltage control loop, is proposed and formally shown to provide asymptotically stable voltage regulation across the SCR range 2-8; (ii) simulation calibrated against CPG field measurements demonstrates voltage sag recovery from 0.72 p.u. to 0.98 p.u. within 18 ms during 2.4 MW crusher DOL starting, compared to 530 ms uncompensated recovery; (iii) THD_v is reduced from 11.3% to 2.8%, achieving IEEE 519-2022 compliance at the PCC; and (iv) reactive power transient overshoot is reduced by 31% relative to a fixed-gain PI STATCOM, and oscillatory instability at low SCR (= 2.1) is eliminated.

Future work will address: (i) hardware-in-the-loop (HIL) validation on the OPAL-RT OP5700 real-time simulator platform using the calibrated CPG network model; (ii) extension of the adaptive control framework to include unbalanced voltage and negative-sequence compensation for the three-phase supply; (iii) model predictive control as an alternative to adaptive PI for the inner current loop to further reduce harmonic injection during transients; and (iv) field deployment and monitoring at a pilot CPG mine site in coordination with STEG, following regulatory approval.

REFERENCES

[1] I. Khan, A. S. Vijay, and S. Doolla, "Nonlinear load harmonic mitigation strategies in microgrids: State of the art," IEEE Syst. J., vol. 16, no. 3, pp. 4243–4255, Sep. 2021.

[2] S. Wang, Y. Ma, K. Sun, J. Wang, H. Li, L. M. Tolbert, and F. Wang, "Modeling and control of variable speed drive based loads for grid primary frequency support," IEEE Trans. Power Del., vol. 38, no. 2, pp. 1219–1230, Apr. 2023.

[3] European Bank for Reconstruction and Development (EBRD), "Project CAPSA – Compagnie des Phosphates de Gafsa: Board Report," EBRD, London, UK, 2023. [Online]. Available: https://www.ebrd.com/content/dam/ebrd_dxp/documents/project/55930/Board_Report_Project_CAPSA.pdf

[4] M. K. Saini and R. Beniwal, "Detection and classification of power quality disturbances in wind energy integrated system using fast time-frequency transform and small residual-extreme learning machine," Int. Trans. Elect. Energy Syst., vol. 28, no. 4, pp. e2519, Apr. 2018.

[5] Y. Xu and F. Li, "Adaptive PI control of STATCOM for voltage regulation," IEEE Trans. Power Del., vol. 29, no. 3, pp. 1002–1011, Jun. 2014.

[6] I. Khan, A. S. Vijay, and S. Doolla, "Nonlinear load harmonic mitigation strategies in microgrids: State of the art," IEEE Syst. J., vol. 16, no. 3, pp. 4243–4255, Sep. 2021.

[7] IEEE Recommended Practice and Requirements for Harmonic Control in Electric Power Systems, IEEE Standard 519-2022, Jun. 2022.

[8] N. Jaalam, N. A. Rahim, A. H. A. Bakar, C. K. Tan, and A. M. A. Haidar, "A comprehensive review of synchronization methods for grid-connected converters of renewable energy source," Renew. Sustain. Energy Rev., vol. 59, pp. 1471–1481, Jun. 2016.

[9] A. F. Akkad, N. Erdili, and E. Sosnina, "Application of a fuzzy logic controller in a D-STATCOM in an electrical network with distributed generation," in 2023 Int. Russian Smart Industry Conf. (SmartIndustryCon), pp. 650–654, 2023.

[10] H. Elmazouri, A. Chaker, and Y. Zidane, "Comparative analysis of reactive power compensation devices in a distribution system: SVC versus STATCOM," Energies, vol. 15, no. 12, pp. 4453, Jun. 2022.

[11] L. Wang, C.-S. Lam, and M.-C. Wong, "A hybrid-STATCOM with wide compensation range and low DC-link voltage," IEEE Trans. Ind. Electron., vol. 63, no. 6, pp. 3333–3343, Jun. 2016.

[12] P. Fortescue, J. Stark, and G. Swinerd, Spacecraft Systems Engineering, 4th ed. Chichester, UK: Wiley, 2011.

[13] Compagnie des Phosphates de Gafsa, "Historical Timeline, Global Positioning and Economic Performance of the Tunisian Public Mining Firm TCGP," Internat. J. Manage. Sci. Bus. Admin., 2019.

[14] IEC/TR 61000-3-6, Electromagnetic Compatibility (EMC) – Part 3-6: Assessment of Emission Limits for the Connection of Distorting Installations to MV, HV and EHV Power Systems, IEC, Geneva, 2008.

[15] T. Ahmed, A. Waqar, R. M. Elavarasan, J. Imtiaz, M. Premkumar, and U. Subramaniam, "Analysis of fractional order sliding mode control in a D-STATCOM integrated power distribution system," IEEE Access, vol. 9, pp. 70337–70352, 2021.

[16] C. Kumar and M. K. Mishra, "An improved hybrid DSTATCOM topology to compensate reactive and nonlinear loads," IEEE Trans. Ind. Electron., vol. 61, no. 12, pp. 6517–6527, Dec. 2014.

[17] S. Yang, P. Wang, Y. Tang, and L. Zhang, "Explicit phase lead filter design in repetitive control for voltage harmonic mitigation of VSI-based islanded microgrids," IEEE Trans. Ind. Electron., vol. 64, no. 1, pp. 817–826, Jan. 2017.