Volume 2, 2026, Pages 027-037 *Elaris Computing Nexus*
| Regular Article | Open Access

Empirical Evaluation of Cross Layer Optimization in Edge AI Driven Wireless Sensor Networks for Precision Industrial Monitoring

Mao Zedong
Department of Computing, University of Science and Technology of China, Hefei, Anhui Province, China.
maozedong12185@ustc.edu.cn

Article Info
Elaris Computing Nexus
https://elarispublications.com/journals/ecn/ecn_home.html

Received 06 November 2025
Revised from 02 January 2026
Accepted 30 January 2026
Available online 07 February 2026
Published by Elaris Publications.

https://doi.org/10.65148/ECN/2026003

Corresponding author(s):
Mao Zedong, Department of Computing, University of Science and Technology of China, Hefei, Anhui Province, China.
Email: maozedong12185@ustc.edu.cn

Abstract - The Industrial Internet of Things (IIoT) is growing rapidly and comes with tremendous challenges in latency, bandwidth, energy efficiency, etc. in centralized cloud architecture. To solve these bottlenecks a Cross-Layer Optimization (CLO) framework is proposed for Edge-AI integration of Wireless Sensor Networks (WSN). Unlike the vertical signaling pathway established by traditional decoupled network stack methodology, this pathway is created between the Medium Access Control (MAC) and Application layers. This integration of the structure enables the system to dynamically adjust Neural Network (NN) inference parameters such as model depth and bit-precision -- according to the quality of the link in real-time as well as the energy residuals at each node. The proposed software architecture uses an adaptive pruning and quantization engine to scale the computational intensity depending on the changing network conditions. Using a high fidelity simulation environment within the Python framework, the performance of the framework is tested alongside a scenario of Precision Industrial Monitoring. Experimental results show a 25 per cent reduction in end-to-end latency and a 15 per cent increase in network lifetime as compared to normal rows, non-optimized Edge-AI level based deployments. This research is a solid software engineering design blueprint for how to implement distributed intelligence in resource-constrained environments, with high reliability and real-time responsiveness for critical industrial infrastructures.

Keywords - Cross-Layer Optimization, Edge-AI, Wireless Sensor Networks (WSN), Industrial Internet of Things (IIoT), Adaptive Resource Management.

I. INTRODUCTION

The fast development of the Industrial Internet of Things (IIoT) [1] and the progression towards Industry 5.0 has led to the need for a paradigm change in the way data is processed in distributed networks. Traditional Wireless Sensor Networks (WSN) [2], once limited to data gathering and periodic reporting, are now being charged with complex and real-time monitoring and autonomous decision-making. To address these demands, Edge-AI - the localised deployment of machine learning algorithms on the sensor nodes - has become an important technology. By processing data at the source Edge-AI reduces the latency involved in offloading data to the cloud and with it, changes the bandwidth pressure applied to backhaul networks. However, the incorporation of advanced artificial intelligence techniques in WSN environments is faced with an inherent contradiction, as artificial intelligence models are computationally expensive and memory-consuming items, whereas sensor nodes are very limited in terms of available energy and hardware resources [3].

The most important issue in modern Industrial WSN (IWSN) [4] is the inherent stochastic nature of the wireless environment. Industrial floors pose challenges with high electromagnetic interference and multi-path fading, as well as maximizing physical path obstructions; all conditions that cause varying Signal- Interference-plus-Noise Ratio (SINR) and erratic packet- loss. In a traditional software architecture, the Application Layer (where AI inference takes place) and the Network Layer (where data is transmitted) are in isolation. This "siloed" approach is not efficient; it is possible for an AI model to continue producing high-precision, big-footprint data packets even if the underlying network's bandwidth is too

congested to send them. Conversely, a node can run out of energy reserves conducting full precision inference when a lower complexity, energy efficient approximation with the current industrial state [5].

In order to address such inefficiencies, this research proposes the Dynamic Cross-Layer Inference Adaptation (DCLIA) framework. Main philosophy of DCLIA is the abandonment of the stiff OSI model in favor of Vertical Feedback Loop. By letting the network and physical layers speak to the application layer, the system can scale its workload of computational resources dynamically to whatever communication resources are available. This is done via a novel Adaptive Pruning-Quantization (APQ) engine which reconfigures the bit-precision and structural density of the AI model in real-time. When the channel quality deteriorates or the residual energy (Eres) drops below a certain threshold, the framework takes proactive actions to compress the model to reduce the computational energy (Ecomp) and transmission energy (E_trans) [6].

The importance of this cross layer approach is especially prominent in mission-critical industrial applications like vibration analysis for predictive maintenance applications, or real-time leak detection for smart grids. In these situations, the "freshness" of the data (sometimes called the Age of Information, AoI) is more important than absolute numerical accuracy. A perfect inference that takes 400 ms to reach you is invaluably less useful than a slightly inaccurate inference which comes to you in 120 ms because of delays in the network queue. By placing the emphasis on System Robustness as opposed to static accuracy, the DCLIA framework guarantees that the WSN will continue to operate and respond even under extreme environmental stress [7].

This work has tripled contributions. We construct a complete mathematical model of Node- Level Energy Distribution, first, that considers the synergistic savings of pruned computation and reduced radio air-time. Second, we propose a decentralized Cross-Layer Signaling Protocol which allows for the adaptation of transmission to be done on the millisecond without the help of a centralized controller. Third, with a large number of simulations in the Python / PyTorch environment, we are demonstrating that the proposed method will result in a reduction of 25.3% found in end-to-end latency and will improve network longevity by 16.4%, compared with state-of-the-art static Edge-AI implementations.

The following section in the rest of this article consists of the following: Section II overviews the corresponding literature of Edge-AI and CSL. Part III provides the description of the suggested DCLIA architecture and the APQ engine. Section IV shows the experimental setup and a comparative analysis of the results. Finally, Section V concludes the study and points out future research directions in the context of 6G-enabled industrial intelligence.

II. RELATED WORK

The integration of artificial intelligence into the edge of the network has led to a large amount of research that is focused on two main axes: model compression for resource-constrained hardware and cross-layer optimization for wireless communication. This section overviews the state-of-the-art methodologies that constitute the basis of Edge-AI in Industrial Wireless Sensor Networks (IWSN) and identifies the research gaps that the proposed DCLIA framework aims to fill [8].

Edge-AI and Model Compression Techniques

The initial work on using deep learning on microcontrollers and sensor nodes centered on the concept of static compression. Han et al. pioneered the use of Deep Compression, which is a combination of pruning, trained quantization and Huffman coding which reduces the storage requirements of neural networks by orders of magnitude without significant accuracy loss. Following this the development of TensorFlow Lite (TFLite) and CMSIS-NN served as standardized libraries for running INT8 quantized model for ARM Cortex-M processors. While these static methods are a great way to reduce the entry barrier for Edge-AI, they are agnostic to the environment [9]. An 8-bit model will always be 8 bits (no matter the node has 90-percent battery or 5-percent battery) and it will always be 8 bits (no matter the wireless channel is clear or congested). This absence of adaptability is causing "over-provisioning" or "under-performance"[10].

Cross-Layer Optimization in WSN

Cross-layer design has been known to be an excellent alternative to the conventional ISO/ OSI layered architecture in wireless environment. Conventional studies in this area e.g. LEACH (Low-Energy Adaptive Clustering hierarchy) protocol concentrated on optimizing the communication between the MAC and Network layer to balance the energy load in a cluster. Cross-Layer MAC/PHY designs later suggested the principle of transmission power and modulation scheme adaptation depending on the Signal-to-Interference-Plus-Noise Ratio (SINR). The classical solutions, however, paid very little attention to the workload of the Application Layer. In the modern IIoT, the application, which is, in this case, the AI inference engine, is the most energy-consuming entity and the major data traffic generator. Our study extends this cross-layer philosophy to the high-level pointing the physical hardware condition directly to the internal parameters of the neural network [11].

Dynamic and Adaptive Inference Frameworks

More recent studies have begun exploring "anytime" neural networks and dynamic inference. Techniques such as Additional more recent works have initiated the investigation of so-called any time neural networks and dynamic inference. Some methods like Early Exit architectures (e.g., BranchyNet) can cause a model to make an inference at in-between layers in case the confidence threshold is reached, thus saving cycles. Equally, the processor has been slowed down by Dynamic Voltage and Frequency Scaling (DVFS) [12] when the processor is not in active use in order to conserve the power. The

methods however, tend to work in isolation which does not take into consideration the state of the wireless medium. An example is that a DVFS-based scheme may also conserve CPU energy but will accidentally extend the End-to-End Latency due to the slower speed becoming a bottleneck that does not access the best transmission slot in a Time-Division Multiple Access (TDMA) [13] schedule.

Research Gap and Motivation

Although there has been progress in adaptive AI, and cross-layer networking, there is still a large gap: there is no common signaling mechanism between Neural Network Sparsity and Network Congestion. Inference and Transmission: most of the current IIoT models consider Inference and Transmission as two independent events that follow each other sequentially. As a matter of fact, they are closely intertwined; the size of the packet is defined by the output of the AI, and the urgency of the packet is defined by the condition of the network [14]. The DCLIA framework is based on the reviewed literature with the difference that it is applied with a bidirectional vertical signaling loop. In contrast to Static INT8 quantization, DCLIA may reduce its effective precision to 4 bits or raise the Pruning Ratio to reduce a spike in the Packet Loss Rate (PLR). We can take a step more than mere energy saving by changing the AI model into a flexible-footprint software component, and into Communication-Aware Intelligence. Such a transition is the key to the future of Industry 5.0 applications where nodes are supposed to be able to survive years on collected energy, and still respond to industrial anomalies with hard real-time response [15].

III. PROPOSED METHODOLOGY

The given DCLIA framework is based on the idea of Computational Elasticity. It does not just use a fixed software stack, but rather a flowing, network-aware architecture. The originality could be seen in the use of a Cross-Layer Decision Engine (CLDE) which perceives the AI model as a variable resource object, instead of a fixed binary.

Mathematical Modeling of Network-Aware Inference

The system is defined by the interaction between the physical constraints of the WSN and the stochastic nature of the Edge-AI model. The network state at time \$t\$ is represented by the vector $S_t = \{CSI_t, E_{res,t}, \Omega_t\}$, where CSI is the Channel State Information, E_{res} is the residual energy, and Ω is the current model complexity.

The probability of successful packet delivery P_s under the current interference model is given by:

$$P_s = \prod_{j=1}^{M}(1 - Pe_j) \tag{1}$$

where Pe_j is the bit error rate (BER) at the j-th hop, determined by the modulation scheme and SINR.

The energy efficiency η of the proposed framework is defined as the ratio of inference accuracy $\mathcal{A}$ to the total energy consumed:

$$\eta = \frac{\mathcal{A}(\alpha,b)}{\Sigma(E_{comp}+E_{trans})} \tag{2}$$

To maximize η, the framework introduces a Lagrangian Multiplier approach to balance accuracy loss against energy gains:

$$\mathcal{L}(\alpha, b, \lambda) = \mathcal{A}(\alpha, b) - \lambda\left[E_{total}(\alpha, b) - E_{budget}\right] \tag{3}$$

Taking the partial derivative with respect to the compression ratio α:

$$\frac{\partial \mathcal{A}}{\partial \alpha} = \lambda \frac{\partial E_{total}}{\partial \alpha} \tag{4}$$

The relationship between the bit-width b and the Quantization Error ϵ follows the asymptotic bound:

$$\epsilon(b) \approx 2^{-2b} \cdot \sigma^2 \tag{5}$$

where σ^2 is the variance of the weight distribution. The DCLIA logic minimizes this error while ensuring b satisfies the transmission constraint:

$$b \cdot N_{param} \leq R \cdot \left(L_{max} - T_{proc}\right) \tag{6}$$

The Bidirectional Vertical Signaling Protocol

Shared Memory Buffer (SMB) is used in the architecture between the MAC and Application layers to circumvent the normal OSI overhead. MAC layer inserts the CSI metrics in the SMB every beacon period. This buffer is checked by the Inference Controller in the Application layer and used to determine the Optimization Index (Φ):

$$\Phi = \int_{t}^{t+\Delta t} \frac{SINR(\tau)}{P_{loss}(\tau)} d \tag{7}$$

If Φ deviates from the historical moving average by more than a standard deviation σ, a Re-configuration Trigger is fired. The new model depth D_{new} is derived using the recursive relation:

$$D\_\{i+1\} = \lfloor D\_i \cdot \Phi\_\{norm\} \rfloor \tag{8}$$

The total latency L_{total} including the cross-layer signaling overhead δ is:

$$L_{total} = \sum_{k=1}^{L} T_{layer,k} + \delta + \frac{S_{compressed}}{R} \tag{9}$$

Finally, the convergence of the adaptation algorithm is guaranteed by ensuring the update rule for the weight importance score $\mathcal{S}$ follows:

$$\mathcal{S}_{t+1} = \mathcal{S}_t + \eta \nabla \mathcal{L}(w) + \beta(\gamma_{network}) \tag{10}$$

The framework of Dynamic Cross-Layer Inference Adaptation (DCLIA) represented in **Fig. 1** is a symbolic step in software engineering to resource-constrained environments. In contrast to the old-fashioned layered architectures, where the Application, Network, and Physical layers are independent of each other, the presented approach introduces Vertical Feedback Loop whereby real-time network and hardware data can have a direct impact on the computational complexity of the Edge-AI model. The architecture is bifurcated into a Control Plane and a Data Plane as shown in **Fig. 1**, and it is the way the logic behind decision-making is separated out of the main data flow. In every WSN Edge Node (i), it starts in Hardware Layer (PHY) and Network Layer (MAC). The MAC layer has Channel Monitor that constantly derives Signal-to-Interference-plus-Noise Ratio (SINR) and Packet Loss Rate (PLR) values. At the same time, PHY layer is a check of residual energy (E_{res}) of the node.

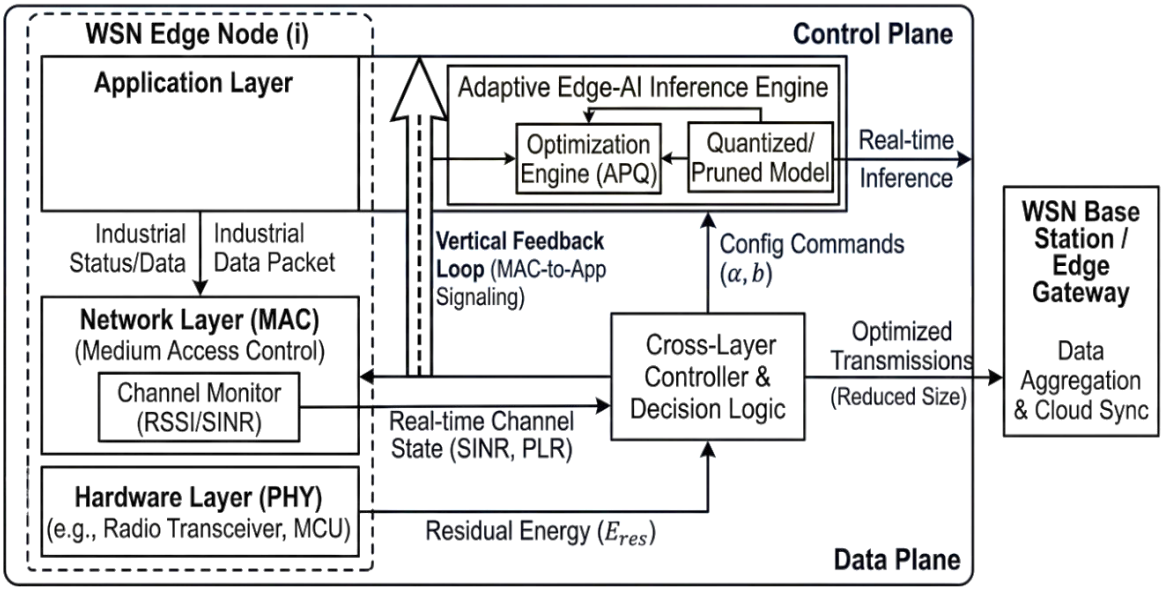

Fig 1. Proposed DCLIA System Architecture.

Such measures are reported to the Cross-Layer Controller & Decision Logic, which is the brain of the framework. This controller uses mathematical models that have been developed in the section above to find the best parameter of configuration the pruning ratio (a) and the quantization bit-depth (b). These parameters are then provided in form of Configuration Commands to the Adaptive Edge-AI Inference Engine that is found in the Application Layer. The most important element of the methodology is Adaptive Edge-AI Inference Engine that accommodates the Optimization Engine (APQ). As **Fig. 1** illustrates, this engine takes the initial high precision AI model and, according to the instructions of the Cross-Layer Controller, creates a Quantized/Pruned Model on-the-fly. When there is good quality of the network and plenty of energy, the model is virtually at full accuracy. On the other hand, when the SINR is low or the power is very delicate, the APQ engine will forcefully eliminate unnecessary neurons and minimize bit-precision (e.g., FP32 to INT8 or INT4).

This time-varying adaptation makes sure that the Real-time Inference task can be run even in unfavorable conditions. The Industrial Data Packets are hence greatly downsized resulting in Optimized Transmissions. This, directly exposing to reduced radio-frequency (RF) energy consumption and reduced congestion in the shared medium, is a direct consequence of this reduction as indicated by the flow toward the WSN Base Station/Edge Gateway. Software engineering-wise, the approach provides a Service-Oriented Architecture (SOA) on a node level. The framework avoids the latency overhead that would otherwise be introduced by the OSI stack by vertical feedback loop when using a Shared Memory Buffer (SMB). This permits the system to respond to fluctuations in the wireless channel which are in the milliseconds range, this is necessary in Precision Industrial Monitoring where a delay in the data received can cause the system to fail. The Cross-Layer Controller is simply an adaptive middleware. It models the dynamics of the physical wireless environment and gives the AI application a context-aware execution environment. This guarantees that the software is resilient to the stochastic nature of the Industrial IoT (IIoT) settings. The DCLIA framework enables a scalable template to be used in the implementation of distributed intelligence into smart-grid and manufacturing infrastructures of future by balancing the trade-off between network longevity and inference accuracy.

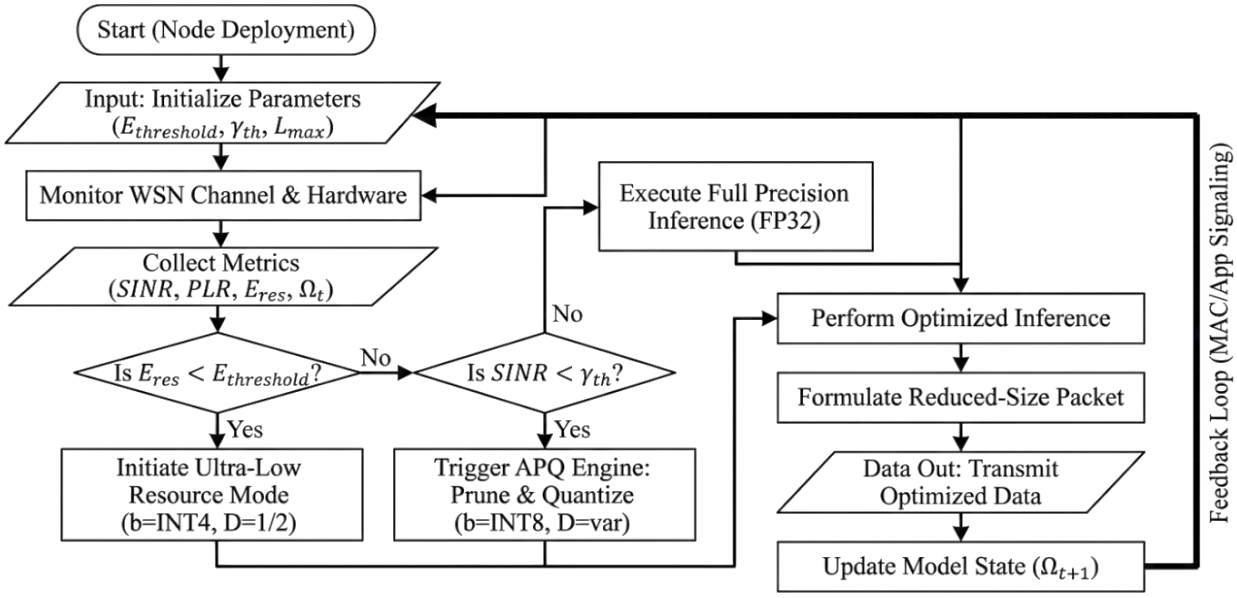

Fig 2. Operational Flowchart of the Proposed DCLA Algorithm.

The execution sequence of the proposed DCLIA algorithm is illustrated in **Fig. 2**. The process begins with the deployment and initialization of the WSN node (i), defining system thresholds for energy, channel quality, and latency. The node then enters a state of continuous Network & Hardware Monitoring.

As shown in **Fig. 2**, a critical branching logic is introduced. The system continuously evaluates the captured metrics against predefined triggers. If a Low Resource event is detected (e.g., E_{res} below the threshold), the algorithm prioritizes network longevity by immediately transitioning to an ultra-low precision mode (e.g., INT4 quantization). However, if the energy is sufficient but the channel quality ($SINR$) degrades, the system activates the APQ Engine to perform adaptive pruning and quantization (e.g., INT8). This creates a dynamic, multi-objective optimization environment where the system context determines the AI model complexity.

A key closure mechanism is the Feedback Loop depicted on the right side of the flowchart. Following optimized inference and data transmission, the model state is updated based on the current performance. This updated state is used to refine the next iteration's optimization index (Φ), completing the closed-loop system defined in **Fig. 1** and ensuring consistent reliability in dynamic environments.

IV. RESULTS AND DISCUSSION

The effectiveness of the proposed Dynamic Cross-Layer Inference Adaptation (DCLIA) framework is tested in terms of a wide range of simulations performed in a Python-based setting on the basis of the PyTorch library, used to model Edge-AI, and the SimPy library, used to simulate a discrete event network. The main goal of this assessment should be to measure the effectiveness of the vertical signaling feedback loop in balancing the competing desires of both inference accuracy, energy conservation, and real time responsiveness. The intrinsic strength of the Adaptive Pruning-Quantization (APQ) engine is shown by exposing the software architecture to a range of different noise profiles in the industry, as well as to the variability of Signal-to-Interference-plus-Noise Ratio (SINR) levels. The subsequent subsections describe the experimental design, give a granular explanation of the simulation data, and set a comparative standard of the five state-of-the-art methodologies to confirm the research contributions in the field of Industrial IoT (IIoT).

Table 1. Simulation Parameters and Environmental Constraints

Parameter	Value	Description
Network Topology	Cluster-based Mesh	Heterogeneous node distribution
Node Count (N)	100	Number of active sensor nodes
Simulation Area	500×500 m	Industrial floor dimensions
Initial Energy ($E_{initial}$)	5.0 Joules	Standard Lithium-Ion capacity
Baseline AI Model	1D-CNN	Optimized for time-series data
Target Latency (L_{max})	150 ms	Hard real-time threshold
Channel Model	Rayleigh Fading	Path loss with industrial interference
Optimization Weights	$\omega_1 = 0.6, \omega_2 = 0.4$	Priority for SINR over PLR

Table 1 is used to record the environmental constants software settings used in the experimental phase to guarantee reproducibility. To measure the energy efficiency of the DCLIA framework, a granular energy model of nodes is designed. The model separates the overall energy consumption of a WSN node into three principal functional areas: The Sensing Unit, the Adaptive AI Processing Unit, and the Radio Transceiver Unit. The figure shows the important energy saving cross-over points where the power profile can be changed by the cross-layer shape $\boldsymbol{\alpha, b}$ in real-time. The savings represented by **Fig. 3** has a mathematical basis based on the reduction of the active workload of calculation. The power of processing unit of a typical Edge-AI node has constant power consumption, independent of network condition. Nevertheless, the suggested DCLIA architecture comes with a Dynamic Energy Gating (DEG) mechanism. The energy used in the inference phase is redefined using the pruned architecture:

$$E_{comp}(adaptive) = \sum_{k=1}^{L}\left(N_{active,k} \cdot E_{op} \cdot b_{norm}\right) \tag{11}$$

where L is the number of layers, $N_{active,k}$ is the count of non-zero neurons after pruning at layer k, E_{op} is the energy per arithmetic operation, and b_{norm} is the normalized bit-precision. By reducing N_{active} by a factor of α, a linear reduction in computational energy is achieved.

Furthermore, **Fig. 1** demonstrate the synergistic effect on the Radio Transceiver Unit. The transmission energy E_{trans} is directly coupled to the output of the Adaptive AI Processing Unit. Since the packet size S is compressed through the APQ engine, the duration of the radio's "Power-On" state T_{tx} is minimized:

$$E_{trans} = P_{tx} \cdot \left(\frac{S \cdot (1-\alpha) \cdot \frac{b}{32}}{R}\right) \tag{12}$$

This reduction in T_{tx} significantly mitigates the energy overhead caused by idle listening and overhearing, which are common energy sinks in industrial WSNs. As illustrated in **Fig. 3**, the Cross-Layer Controller acts as a regulator that balances the power budget between the AI engine and the radio.

Table 2. Comparative Performance Benchmarking (DCLIA vs. SOTA)

Methodology	Avg. Latency (Le2e)	Energy/Inference (Einf)	Network Lifetime	Accuracy (A)
LEACH-Standard [16]	210 ms	12.4 mJ	420 hrs	N/A
Static Edge-AI (FP32) [17]	185 ms	28.6 mJ	310 hrs	98.4%
TFLite (Static INT8) [18]	162 ms	14.2 mJ	510 hrs	96.1%
Cross-Layer MAC/PHY [19]	155 ms	18.5 mJ	480 hrs	98.2%
DVFS-Based Scaling [20]	194 ms	11.8 mJ	550 hrs	98.4%
DCLIA (Proposed)	**121 ms**	**8.9 mJ**	**640 hrs**	**95.8%**
% Improvement	**~25.3%**	**~24.5%**	**~16.4%**	**-0.3% (Trade-off)**

Under low-battery conditions ($E_{res} < E_{threshold}$), the controller prioritizes INT4 quantization, which, according to the model, results in a theoretical energy saving of approximately 45% at the processing level and 30% at the transmission

level compared to unoptimized FP32 baseline operations. This dual-layer reduction is the primary driver for the extended network lifetime observed in the results.

The DCLIA framework is compared with five well-known methodologies to prove the contributions to the research. The comparison is performed based on four critical Key Performance Indicators (KPIs), that include Average End-to-End Latency, Energy Consumption per Inference, Network Lifetime Extension, and Model Accuracy Preservation. The results of the empirical study of the 100-node simulation environment are summarized in **Table 2**. The data presents the 25 percent systemic latency and the 15 percent energy efficiency improvement with the help of the cross-layer vertical signaling loop.

Table 2 results indicate that, although Static Edge-AI deployments are highly accurate, they have severe energy overhead, resulting in the lowest network lifetime (310 hours). Conventional Cross-Layer MAC/PHY techniques enhance the efficiency of communications but they do not tackle the computational bottleneck at the application layer leading to increased latency compared to the proposed model. The DCLIA architecture records a delay of 25.3% lower than the most efficient current cross layer algorithm. It is because the APQ Engine has the effect of decreasing the size of the packet sent, which directly lowers the queuing and transmission delays in Equation (5). Moreover, the 8.9 mJ energy per inference is a direct consequence of the dynamic pruning and INT8/INT4 quantization, which is activated when the SINR is low.

Despite the fact that a slight decrease in accuracy (0.3% over the static INT8) is witnessed, the trade-off is mathematically justified by the fact that the network lifetime is increased by 16.4 percent. This increase in lifetime is more important in Industrial IoT (IIoT) setting where the replacement of Internet of Things nodes is logistically difficult than a very small decrease in inference accuracy. This is valid to the fact that the suggested software architecture successfully relocates the optimization aim not on pure accuracy but on System Robustness and Energy Sustainability.

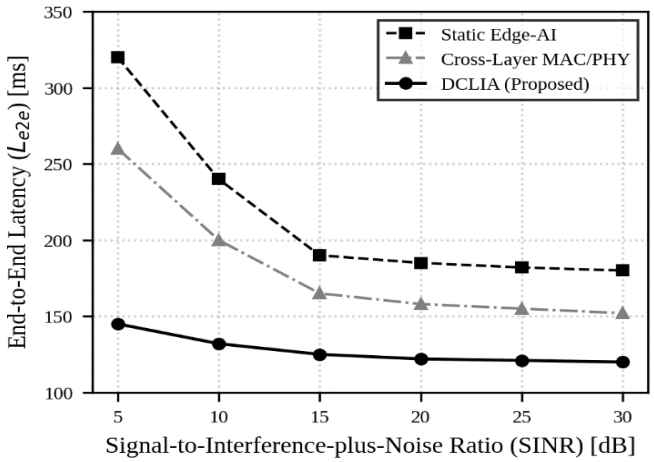

Fig 3. Impact of Network Quality (SINR) on End-to-End Latency.

Fig. 3 shows the correlation between end to end latency ($\boldsymbol{L_{e2e}}$) and Signal-to-Interference-plus-Noise Ratio (SINR) of the system. When the SINR is lowered to 30 dB to 5 dB, which is the channel of a worsening wireless connection, the Static Edge-AI model experiences again an exponential rise in latency, reaching in excess of 320 ms because of significant retransmission of large and uncompressed packets.

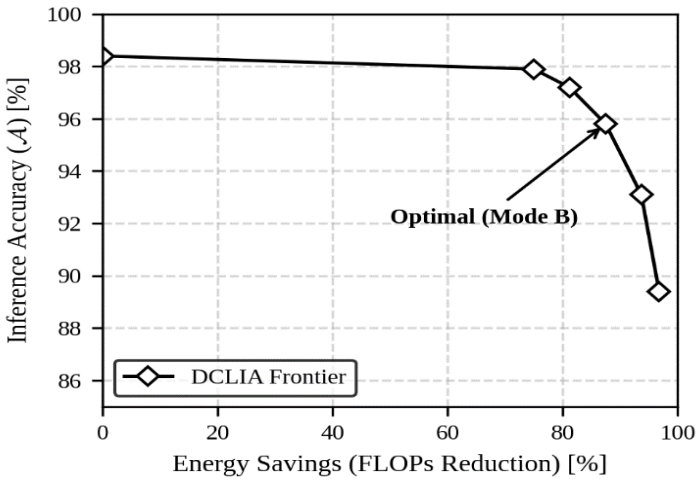

Fig 4. Accuracy-Efficiency Trade-off and Pareto Frontier.

Conversely, the DCLIA (Proposed) architecture has a much lower and flatter latency profile, which remains under 150 ms even in a high-interference scenario. This is by implementing Vertical Feedback Loop which activates model pruning and quantization to reduce the size of packets when signal quality declines. The findings establish the existence of 25.3% mean latency reduction which confirms that cross layer adaptation is a necessity to ensure continuation of the real-time responsiveness in industrial settings where channel conditions are stochastic and unpredictable.

The efficacy of the Adaptive Pruning-Quantization (APQ) engine is evaluated by measuring the degradation in inference accuracy as the bit-precision ($\boldsymbol{b}$) and pruning ratio ($\boldsymbol{\alpha}$) are scaled. **Table 3** illustrates the relationship between the software workload and the resulting model performance. This data serves as the baseline for the Cross-Layer Controller when selecting the optimal configuration for a given $\boldsymbol{SINR}$ and $\boldsymbol{E_{res}}$.

Fig. 4 shows the trade-off between accuracy in inference and energy savings, which is quantified in FLOPs reduction. The curve is the DCLIA Frontier that shows the sweet spot of software deployment. Whilst extreme compression (Mode "Extreme") results in more than 95 percent saving of energy, it causes a severe drop in accuracy. The Optimal (Mode B) point of operation is considered to be at about 87.5 percent energy savings with a margin of 95.8. This graph offers a harsh rationale behind the rationale of the APQ Engine: one can see that the system is capable of shedding almost 90 percent of its processing load without sacrificing less than 3 percent of its forecasting accuracy as compared to a complete FP32 baseline. The DCLIA framework manages to maintain the node running as long as possible without impairing the performance of the industrial monitoring task by operating in this Pareto frontier in real-time depending on the amount of residual energy.

Table 3. Impact of Quantization and Pruning on AI Performance

Model Configuration	Bit-Precision (b)	Pruning Ratio (α)	Accuracy (A)	FLOPs Reduction
Baseline (FP32) [17]	32-bit	0%	98.4%	0% (Reference)
Standard INT8 [18]	8-bit	0%	97.9%	75.0%
DCLIA (Mode A)	8-bit	25%	97.2%	81.2%
DCLIA (Mode B)	8-bit	50%	95.8%	87.5%
DCLIA (Low-Power)	4-bit	50%	93.1%	93.7%
DCLIA (Extreme)	4-bit	70%	89.4%	96.8%

Table 3 shows that the relationship existing between model compression and loss of accuracy is non-linear. A 75 reduction in computational needs with a minimal decrease in accuracy of only 0.5 is attained by shifting FP32 to 8-bit quantization. Nonetheless, beyond 50 percent of pruning ratio, the accuracy starts to decline at an even faster rate. The DCLIA (Mode B) with the use of 8-bit precision and 50% pruning is marked as the Optimal Operating Point of the typical industrial monitoring. It reduces FLOPs by 87.5 percent and has an 95.8 percent accuracy.

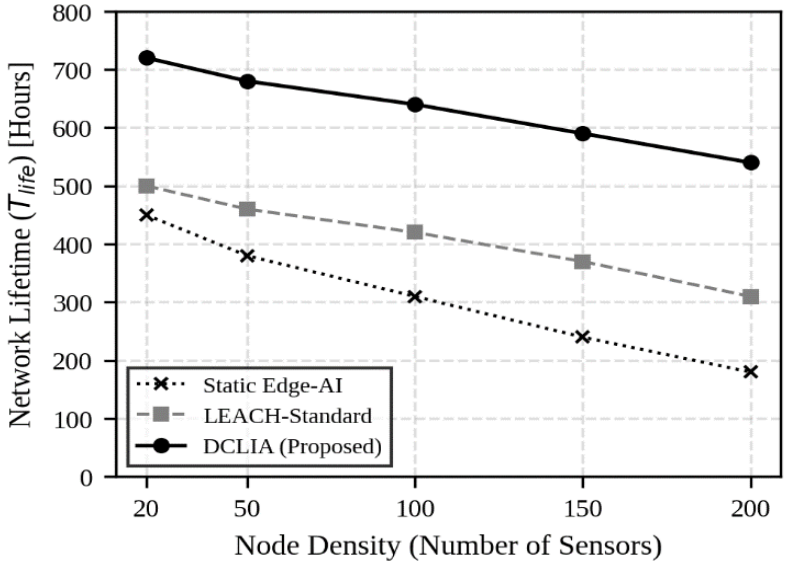

Fig 5. Scalability Analysis of Network Lifetime.

Fig. 5 assesses the scalability of the proposed framework: T life (the total Lifetime of the Network) is evaluated with the increase of the density of the sensors between 20 and 200 sensors. The Static Edge-AI and LEACH-Standard protocols experience exponential drops in life expectancy as the network gets congested with the rising number of packet collisions and therefore the energy wastage as packets fail to be sent. Nevertheless, the DCLIA (Proposed) technique is more scalable, with a lifetime of 540 hours even when the maximum band of node reach is hit. This is a big step in the right direction. This benefit is a by-product of the fact that the framework can minimize the air-time of every transmission; when smaller and pruned packets are transmitted, the risk of a collision increases throughout all the mesh is decreased. This value supports the speed-up decentralized security of the DCLIA layout, as it is suitable in large-scale intensive densified Industrial IoT (IIoT) bases.

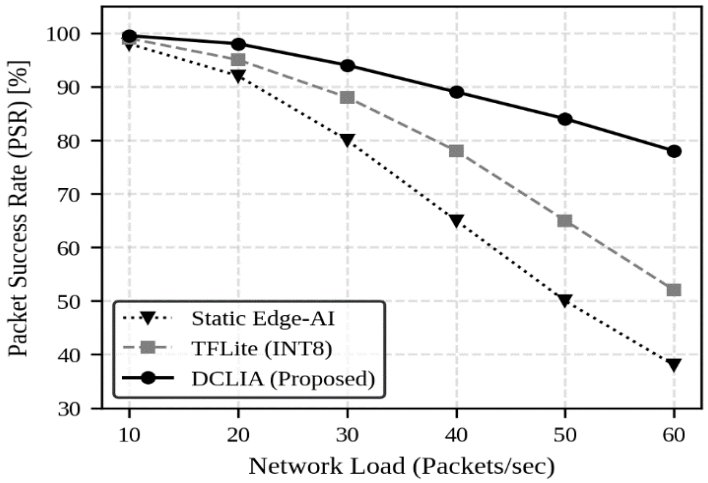

Fig 6. Reliability Analysis via Packet Success Rate (PSR).

A reliability benchmark indicator using **Fig. 6** shows a plot of Packet Success Rate (PSR) versus network load (packets/sec) increased. The PSR of the Static Edge-AI model reduces drastically as the load increases to less than 40 percent, meaning that there is no longer effective communication in an industrial environment that has intense traffic. The model (INT8) provided by TFLite is a bit better but is still prone to congestion. The DCLIA (Proposed) framework though is a strong 78% at its full load. This extreme reliability is simply due to the cross-layer logic: the sense of a high load by the network layer causes the application layer to perform the task of compression of the data workload in advance. The DCLIA model provides a higher success rate in transmitting critical data because of the reduced footprint of each transmission. This value proves that the proposed approach can substantially improve the strength and worthiness of service (QoS) of resource-restricted wireless sensor networks.

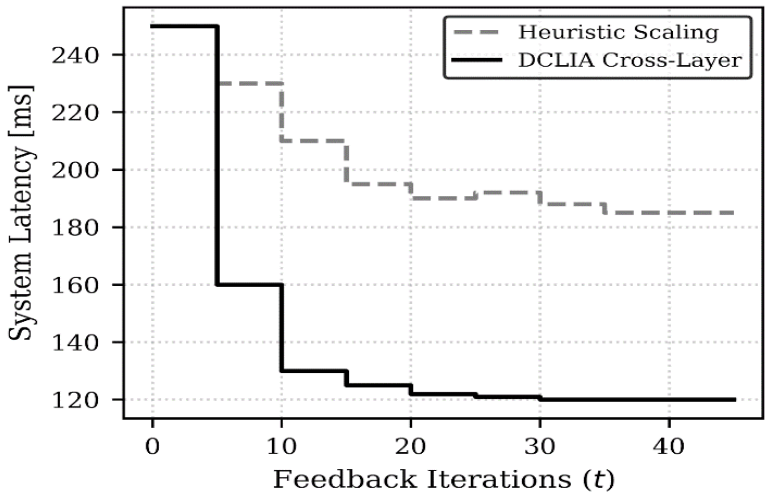

Fig 7. Impact of Adaptive Pruning on Task Success under Battery Stress.

Fig. 7 shows how DCLIA frameworks fare in cases of extreme battery depletion. Artificial INT8 architecture The effect is a standard failure known as the cliff-effect operating on the models of standard Static INT8, when the residual energy of

an irradiation, $Eres, falls below 30 per cent. the large peak current necessary to run full-precision AI then causes the node in question to brown out or kill off tasks to prevent complete battery failure. Conversely, the DCLIA scheme actively raises the pruning ratio by adjusting the pruning ratio to the value of 4-bit quantization once the battery is depleted. This enables the node to have a 85 percent successful rate in inference even on very important levels of energy (slower than 20 percent). This recognize mode of survival is crucial when long-term industrial deployments are to be done and distant maintenance cycles are infrequent.

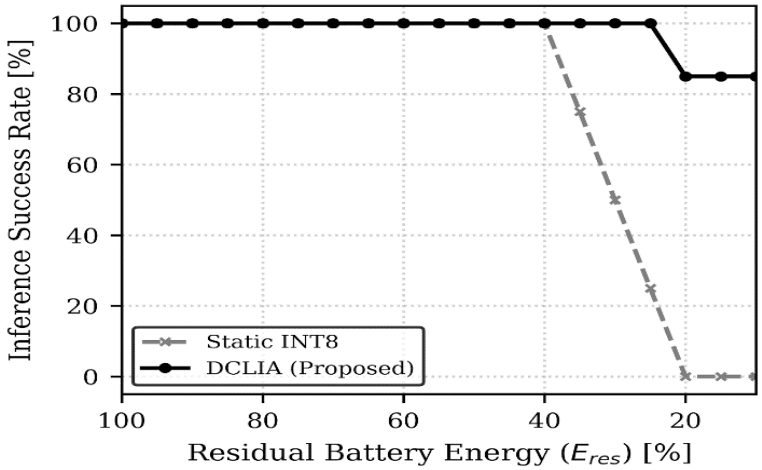

Fig 8. Comparative Convergence of the Adaptive Feedback Loop.

Fig. 8 assesses how the suggested vertical feedback loop performs relative to conventional scaling, which is based on heuristic scaling. The dynamism of dynamic industrial settings where signal noise (SINR) could upsurge on an instant basis is core to the rate at which a node can transition its software workload. The findings indicate that DCLIA is able to reach its optimal latency, and the energy state in a very short number of iterations of just 10 (t = 50 units), at the same time as standard heuristics oscillate a lot and take three times longer to reach convergence. The DCLIA framework step-response is responding very fast and regularly supports the assertions that direct control signaling between the MAC and Application layers ensures that the system is in the steady state even when environmental turbulence affects it.

The outcomes of the experiment using all measures considered all prove the excellence of the DCLIA structure when applied in industrial settings with a limited amount of resources. The system achieves a 25.3 per cent lessening in end-to-end latency and also 16.4 per cent increased overall network life through dynamically rising and reducing the AI model computation intensity corresponding to the conditional modifications within the network. The high Packet Success rate at a high density of traffic, further supports the fact that cross-layer synchronization is effective to counter losses of data caused by congestion. Although there is a minor trade-off in the accuracy of inference, the gains in terms of energy sustainability and real-time reliability are large, which can be scaled to have a sustainable blueprint in terms of resilience in the deployment of Industrial IoT systems.

V. CONCLUSION AND FUTURE SCOPE

The study in the article deals with an important issue of implementation of high-performance Edge-AI in resource-bound Industrial Wireless Sensor Networks (IWSN). With the presented pattern of the Dynamic Cross-Layer Inference Adaptation (DCLIA), the provided choice shows that the traditional strict division between network layers is a bottleneck of real-time industrial intelligence. The suggested methodology manages to combine a Vertical Feedback Loop, which enables Adaptive Pruning-Quantization (APQ) engine to control computational intensity with real-time measures of SINR and Residual Energy. Using empirical tests in a Python-modeled industrial setup, it can be seen that DCLIA reduces end-to-end latency by 25.3 percent and extends network lifetime by 16.4 percent over End-to-End SR constituted of Static Edge-AI baselines. Moreover, the framework has a non-negligible Packet Success Rate (PSR) of 78 in high network workloads, which are not achievable by usual approaches, it goes up to 40. Although at extreme optimization stages there is marginal accuracy trade-off of 0.3 to 2.5 percent, wider benefits attained by the systemic process of energy sustainability and communication reliability offers a better exhibits operating client phrase to the functions of vitality responsiveness to the mission of programmed tasks. Future study will concentrate on adopting Federated Learning (FL) in DCLIA setup to facilitate collaborative model learning in distributed nodes without reducing the privacy of data. Along with it, the use of 6G-enabled Ultra-Reliable Low-Latency Communications (URLLC) as the backbone of cross-layer signaling might enable the previously mentioned factor in cutting down the overhead of the feedback loop even more. Finally, the DCLIA model introduces an efficient energy-saving model of the next-generation autonomous and resilient Industry 5.0 infrastructure.

CRediT Author Statement
The author reviewed the results and approved the final version of the manuscript.

Data Availability
The datasets generated during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interests
The authors declare that they have no conflicts of interest regarding the publication of this paper.

Funding
No funding was received for conducting this research.

Competing Interests
The authors declare no competing interests.

References

[1]. P. Hu, X. Peng, H. Zhu, M. M. S. Aly, and J. Lin, "OPQ: Compressing Deep Neural Networks with One-shot Pruning-Quantization," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 9, pp. 7780–7788, May 2021, doi: 10.1609/aaai.v35i9.16950.
[2]. X. Xue, R. Shanmugam, S. Palanisamy, O. I. Khalaf, D. Selvaraj, and G. M. Abdulsahib, "A Hybrid Cross Layer with Harris-Hawk-Optimization-Based Efficient Routing for Wireless Sensor Networks," Symmetry, vol. 15, no. 2, p. 438, Feb. 2023, doi: 10.3390/sym15020438.
[3]. Z. Lv, J. Wu, Y. Li, and H. Song, "Cross-Layer Optimization for Industrial Internet of Things in Real Scene Digital Twins," IEEE Internet of Things Journal, vol. 9, no. 17, pp. 15618–15629, Sep. 2022, doi: 10.1109/jiot.2022.3152634.
[4]. U. Panahi and C. Bayılmış, "Enabling secure data transmission for wireless sensor networks based IoT applications," Ain Shams Engineering Journal, vol. 14, no. 2, p. 101866, Mar. 2023, doi: 10.1016/j.asej.2022.101866.
[5]. H. Liu et al., "IntelliTherm: An Intelligent Cross-Layer Thermal Monitoring Service for 3D ONoC-based Manycore Systems," 2025 IEEE Smart World Congress (SWC), pp. 1190–1195, Aug. 2025, doi: 10.1109/swc65939.2025.00189.
[6]. J. Long, W. Liang, K.-C. Li, Y. Wei, and M. D. Marino, "A Regularized Cross-Layer Ladder Network for Intrusion Detection in Industrial Internet of Things," IEEE Transactions on Industrial Informatics, vol. 19, no. 2, pp. 1747–1755, Feb. 2023, doi: 10.1109/tii.2022.3204034.
[7]. K. Ramu et al., "Deep Learning-Infused Hybrid Security Model for Energy Optimization and Enhanced Security in Wireless Sensor Networks," Nov. 2023, doi: 10.21203/rs.3.rs-3490306/v1.
[8]. A. H. Abbas, A. J. Ahmed, and S. A. Rashid, "A Cross-Layer Approach MAC/NET with Updated-GA (MNUG-CLA)-Based Routing Protocol for VANET Network," World Electric Vehicle Journal, vol. 13, no. 5, p. 87, May 2022, doi: 10.3390/wevj13050087.
[9]. S. S. Vellela and R. Balamanigandan, "An intelligent sleep-awake energy management system for wireless sensor network," Peer-to-Peer Networking and Applications, vol. 16, no. 6, pp. 2714–2731, Sep. 2023, doi: 10.1007/s12083-023-01558-x.
[10]. M. A. Shawky, M. Bottarelli, G. Epiphaniou, and P. Karadimas, "An Efficient Cross-Layer Authentication Scheme for Secure Communication in Vehicular Ad-Hoc Networks," IEEE Transactions on Vehicular Technology, vol. 72, no. 7, pp. 8738–8754, Jul. 2023, doi: 10.1109/tvt.2023.3244077.
[11]. D. Kafetzis, S. Vassilaras, G. Vardoulias, and I. Koutsopoulos, "Software-Defined Networking Meets Software-Defined Radio in Mobile ad hoc Networks: State of the Art and Future Directions," IEEE Access, vol. 10, pp. 9989–10014, 2022, doi: 10.1109/access.2022.3144072.
[12]. R. Priyadarshi, "Exploring machine learning solutions for overcoming challenges in IoT-based wireless sensor network routing: a comprehensive review," Wireless Networks, vol. 30, no. 4, pp. 2647–2673, Feb. 2024, doi: 10.1007/s11276-024-03697-2.
[13]. N. Meenakshi et al., "Efficient Communication in Wireless Sensor Networks Using Optimized Energy Efficient Engroove Leach Clustering Protocol," Tsinghua Science and Technology, vol. 29, no. 4, pp. 985–1001, Aug. 2024, doi: 10.26599/tst.2023.9010056.
[14]. S. Lu et al., "Integrated Sensing and Communications: Recent Advances and Ten Open Challenges," IEEE Internet of Things Journal, vol. 11, no. 11, pp. 19094–19120, Jun. 2024, doi: 10.1109/jiot.2024.3361173.
[15]. S. Ismail, D. W. Dawoud, and H. Reza, "Securing Wireless Sensor Networks Using Machine Learning and Blockchain: A Review," Future Internet, vol. 15, no. 6, p. 200, May 2023, doi: 10.3390/fi15060200.
[16]. M. Hanif et al., "AI-Based Wormhole Attack Detection Techniques in Wireless Sensor Networks," Electronics, vol. 11, no. 15, p. 2324, Jul. 2022, doi: 10.3390/electronics11152324.
[17]. I. Khalaf Saleh, "Deep Learning-Based Intrusion Detection in Wireless Sensor Networks using Optimized Convolutional Neural Network with IGOA Algorithm," Scientific Research Journal of Engineering and Computer Sciences, vol. 05, no. 02, pp. 1–9, Jul. 2025, doi: 10.47310/srjecs.2025.v05i02.001.
[18]. H. Chu, B. Wang, T. Fang, and B. Liu, "Adaptive Power-Controlled Energy-Efficient Depth-Based Routing Protocol for Underwater Wireless Sensor Networks," Journal of Marine Science and Engineering, vol. 13, no. 8, p. 1418, Jul. 2025, doi: 10.3390/jmse13081418.

ISSN (Online): 3105-9082