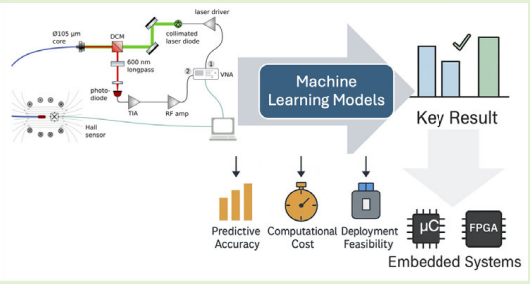
Enhancing All-Optical NV Center Magnetometry With Machine Learning: Model Selection for Efficient and Deployable Quantum Sensors

José Luis Ávila-Jiménez, Ann-Sophie Bülter, Ludwig Horsthemke, Francisco J. Rodriguez-Lozano, Manuel Ortiz-Lopez, and Peter Glösekötter, *Member, IEEE*

***Abstract*—Negatively charged nitrogen–vacancy (NV) centers in diamond are promising platforms for magnetic field sensing under ambient conditions, yet practical deployment is often limited by hardware complexity, power consumption, and computational cost. All-optical approaches based on fluorescence lifetime measurements avoid microwave excitation and enable simpler, potentially more scalable sensor designs, but the nonlinear response to magnetic and thermal fields requires advanced modeling. In this study, a systematic comparison of 18 regression algorithms—including linear models, ensemble methods, support vector machines, and symbolic regression—was conducted for predicting magnetic field strength from NV fluorescence lifetime data. A dataset of 22 758 measurements was employed, and models were assessed in terms of both predictive accuracy and computational footprint (training time, model complexity, and inference latency). The results indicate that ensemble techniques such as LightGBM and XGBoost achieve near-perfect accuracy while maintaining subsecond inference times, making them suitable for real-time embedded deployment. In contrast, symbolic regression provides negligible inference cost but suffers from significantly lower accuracy. These findings offer quantitative guidance for the selection of machine learning models that balance precision and efficiency, paving the way toward compact, low-power, all-optical NV magnetometers for resource-constrained platforms.**



***Index Terms*— Machine learning, nitrogen–vacancy (NV) sensor, symbolic regression.**

I. Introduction

NEGATIVELY charged nitrogen–vacancy (NV) centers in diamond have become an attractive quantum sensing element for magnetic fields, owing to their exceptional spin coherence and optical addressability at room temperature. In most NV-based magnetometers, microwaves are used to manipulate the spin sublevels and read out magnetic resonance, yielding high sensitivity and fine spatial resolution [1], [2], [3]. However, using microwaves complicates the sensor design by increasing size, power consumption, and the need for RF shielding, while also risking unwanted environmental interactions.

An alternative strategy bypasses microwaves altogether by exploiting changes in the NV fluorescence lifetime under varying magnetic fields. As the local field modifies the spin-state mixing, the temporal decay of the emission carries information about the field strength. In addition to magnetic fields, the fluorescence lifetime is influenced by thermal effects, producing a coupled, nonlinear response that resists separation via standard analytical formulas [4], [5]. To disentangle these effects, recent work has turned to data-driven modeling [6]. The result is a step toward a practical, all-optical magnetometer suitable for industrial applications.

When dealing with small magnetic fields of less than 10 mT, it is not possible to directly determine the magnetic field from the fluorescence intensity using all-optical measurements

Received 5 October 2025; accepted 26 October 2025. Date of publication 10 November 2025; date of current version 15 December 2025. This work was supported in part by the Spanish Ministry of Science, Innovation and Universities; in part by the State Research Agency; in part by the European Regional Development Fund through MCIU/AEI/10.13039/501100011033/FEDER, EU under Project PID2023-148396NB-I00; and in part by QuantumIRES Project. The associate editor coordinating the review of this article and approving it for publication was Dr. Satyendra K. Mishra. *(Corresponding author: José Luis Ávila-Jiménez.)*

José Luis Ávila-Jiménez, Francisco J. Rodriguez-Lozano, and Manuel Ortiz-Lopez are with the Department of Electronic and Computer Engineering, Universidad de Córdoba, 14071 Córdoba, Spain (e-mail: jlavila@uco.es; fj.rodriguez@uco.es; mortiz@uco.es).

Ann-Sophie Bülter, Ludwig Horsthemke, and Peter Glösekötter are with the Department of Electrical Engineering and Computer Science, FH Münster—University of Applied Sciences, 48565 Steinfurt, Germany (e-mail: annsophie.buelter@fh-muenster.de; l.horsthemke@fh-muenster.de; peter.gloesekoetter@fh-muenster.de).

Digital Object Identifier 10.1109/JSEN.2025.3628661

with NV-rich diamonds [7]. The motivation for this study is to solve this issue by using various symbolic regression methods, which aim to discover mathematical expressions linking NV fluorescence lifetimes to magnetic field strength, and compare the methods. Beyond raw prediction accuracy, each method's computational footprint—training time, model complexity, and inference cost—is critically assessed, since the ultimate goal is deployment on resource-constrained platforms such as microcontrollers or FPGAs. By comparing various symbolic regression algorithms under identical datasets and cross-validation protocols, those that strike the best balance between predictive performance and efficiency are identified, paving the way toward compact, low-power, all-optical NV magnetometers.

II. State of the Art

In the context of NV center-based magnetometry at ambient conditions, all-optical detection schemes are gaining traction to simplify hardware by eliminating microwave excitation [8].

However, approaches based on optically detected magnetic resonance (ODMR) have primarily been used in combination with machine learning for optimized sensor design. Tsukamoto et al. [9] demonstrated that Gaussian process regression applied to NV ensemble spectra can predict magnetic fields with sub-2 μT accuracy, outperforming traditional ODMR-based analytical models. Homrighausen et al. [10] ran a neural inference model on an ESP32 microcontroller, demonstrating real-time magnetic field estimation from NV center spectra with microtesla-level precision.

Recent work on NV centers has turned to data-driven modeling. For example, Horsthemke et al. [6] trained a neural network to map the frequency-domain fluorescence response directly to the magnetic field strength, learning the complex physical relationship without relying on an explicit spin Hamiltonian model. The fluorescence signal is recorded across a range of excitation frequencies, capturing both amplitude and phase information, which vary with the magnetic field strength. Instead of relying on predefined physical models, the network is trained to directly predict the magnetic field from the spectral data. This approach enables a more robust and accurate magnetometric system, featuring a high magnetic bandwidth from zero field onward.

Concurrently, recent NV-based demonstrations illustrate how all-optical or minimally complex optical paths can be paired with efficient inference to approach embedded, real-time operation. On the one hand, an edge–machine-learning NV magnetometer has been run on a low-cost microcontroller, evidencing the feasibility of on-device inference but also the importance of choosing models with favorable latency/accuracy tradeoffs [10]. On the other hand, fiber-bundle implementations with nanodiamonds have achieved remote read-out of magnetic fields and gradients under ambient conditions, showcasing distributed sensing with relatively simple optical hardware while leaving room to optimize computational cost for deployment [11]. Together with the broader commercialization and integrated-photonics trends [12], [13], [14], [15], these examples motivate a systematic assessment of regression algorithms that jointly optimize predictive performance and inference efficiency for resource-constrained platforms.

Deploying machine learning algorithms on edge devices, such as microcontrollers and FPGAs, poses significant challenges due to limited computational and energy resources [16]. While frameworks like TensorFlow Lite Micro enable small neural networks to run on low-power microcontrollers [17], [18], symbolic regression provides an attractive alternative by generating closed-form expressions that require only basic arithmetic operations, significantly reducing memory usage and computational cost. In the context of FPGAs, Tsoi et al. [16] integrated symbolic expressions generated by genetic programming into the hls4ml toolchain, achieving a 13× speedup in inference latency compared to a three-layer neural network, while maintaining over 90% of the original accuracy. These findings underscore the potential of interpretable models, such as symbolic expressions and small decision trees, for efficient deployment in resource-constrained environments [19].

Symbolic regression can be a powerful technique for discovering explicit relationships in optical sensor data, offering interpretable models that reflect the underlying physics.

In spectroscopy and fluorescence sensing, where the number of variables is often moderate, symbolic regression can uncover closed-form expressions linking sensor outputs to target properties [19], [20]. For instance, Chong et al. [20] applied symbolic regression combined with physics-informed spectral descriptors to predict adsorption energies and charge-transfer properties from IR and Raman spectra with high robustness. Similarly, Sierra-Vélez et al. used symbolic regression to model reflectance spectra of bio-inspired photonic structures, obtaining dimensionally consistent formulas that accurately reproduce the optical response [21].

In industrial process control, Kay et al. [22] demonstrated the construction of an interpretable soft sensor for product quality assessment by combining dimensionality reduction with symbolic regression, achieving performance comparable to conventional machine learning models while providing physical insight. These studies illustrate the growing adoption of symbolic regression in optical sensing domains, where model transparency and expression compactness are highly valued [19].

On top of that, in scenarios with limited data, symbolic regression often outperforms deep learning models by avoiding overfitting and requiring fewer parameters. Höfler [23] compared symbolic regression, deep neural networks, and partial least-squares on gas sensor array data, finding that symbolic models achieved higher accuracy on small datasets. However, symbolic regression typically entails higher offline training costs due to the expansive search over function spaces.

Conversely, ensemble methods such as gradient boosting (e.g., XGBoost [24] and LightGBM [25]) can provide near-perfect predictive performance and favorable training and prediction time tradeoffs in data-rich regimes, albeit with reduced interpretability. Ultimately, the choice between symbolic regression and other machine learning techniques depends on the desired balance between model interpretability, predictive accuracy, and computational constraints [19].

As symbolic regression has been shown to provide promising results, this study aims to assess different symbolic regression techniques, analyzing their tradeoffs between accuracy and computational resource requirements, to identify models suitable for real-time deployment in embedded sensing applications based on quantum sensors using NV centers.

III. Materials and Methods

This section outlines the methodology, including a brief description of the experimental setup, data acquisition, dataset structure, regression models, and evaluation metrics. Fluorescence measurements are acquired from NV-rich microdiamonds using a custom optical setup under controlled magnetic fields to generate the dataset. The resulting data are then structured to enable model training with a diverse range of regression algorithms, whose performance is systematically compared. Finally, both predictive accuracy and computational efficiency are assessed to quantify the suitability of each approach for deployment in resource-constrained environments.

A. Experimental Setup

The experimental setup shown in Fig. 1 is configured to study the fluorescence behavior of NV-rich microdiamonds under laser excitation and controlled magnetic fields. The measurement uses a frequency-domain approach, a generalized variant of the basic all-optical method where the frequency response of amplitude and phase is recorded. This approach provides more information about the system, which can then be used as input data for the symbolic regression methods.

A 520 nm collimated laser diode (ams-OSRAM PLT5 520B [26]) with a mean optical output power of 10 mW serves as the excitation source. The laser is modulated by a driver circuit featuring a comparator that switches the diode on and off in synchronization with the positive and negative half-cycles of the incoming waveform. The laser light passes through a dichroic mirror (Thorlabs DMPS567R [27]) and is then coupled into an optical fiber with a core diameter of 105 μm, whose end facet is placed near NV-rich microdiamonds within a glass cuvette, arranged to completely cover the fiber tip. The fluorescence emitted by the NV centers under laser excitation is collected through the same fiber and passed through a long-pass filter (Thorlabs FELH600 & LA1951-AB [28]) to separate the signal from the excitation light before being focused onto a photodiode, where a photocurrent is generated. This current is amplified by a transimpedance amplifier (12 kΩ gain, 75 MHz bandwidth) and analyzed using a Vector Network Analyzer (VNA). The VNA generates a frequency-modulated stimulus to drive the laser diode while simultaneously measuring the response of the fluorescence signal in terms of amplitude and phase across a frequency range of 1–100 MHz at 401 equidistant discrete points.

An electromagnet capable of generating magnetic fields up to 80 mT was used to apply external fields to the NV centers. The magnetic field strength was continuously monitored using a Hall effect sensor.

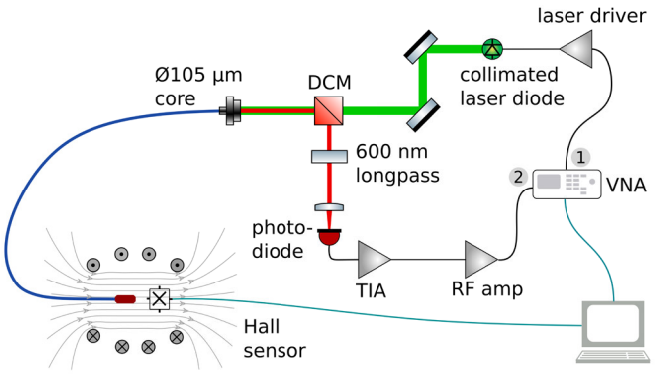


Fig. 1. Frequency-domain measurement setup.

B. Dataset Description

The dataset employed in this study comprises a total of 22 758 individual data points, each corresponding to a fluorescence lifetime measurement. For every data point, a set of 802 features is included, reflecting both the magnitude and phase responses of the fluorescence emission signal. The behavior of this fluorescence response can be described as the convolution of the excitation signal with the decay dynamics, functioning as a low-pass filter. As the frequency increases, the low-pass filter effect results in a decrease in fluorescence amplitude, while the phase of the fluorescence signal shifts from 0° to 90° [29]. In the context of a bi-exponential fluorescence decay, the transfer function is given by the sum of two first-order low-pass filters, resulting in

$$H(s) = \frac{a_1}{s + \frac{1}{\tau_1}} + \frac{a_2}{s + \frac{1}{\tau_2}} \tag{1}$$

with the time constants τ_1 and τ_2 and their amplitudes a_1 and a_2. The VNA is calibrated using the system response $H(s)|_0$ at the reference condition $B = 0$. This reference response accounts for the transmission characteristics of all optical and electrical components involved in the measurement, as well as the intrinsic decay dynamics of the fluorescence. The VNA then compensates for $H(s)|_0$, yielding the normalized measurement

$$H_r(s) = \frac{H(s)|_B}{H(s)|_0} \tag{2}$$

which has unit magnitude and zero phase at $B = 0$. Hence, only the variations induced by the parameter B on the material are recorded.

All measurements were systematically collected under a variety of controlled magnetic field conditions, using the previously described frequency-domain measurement setup. Since the focus of this study is on the unambiguous determination of the magnetic field from the fluorescence lifetime for low magnetic fields, the dataset only contains measurements up to a magnetic field of 18 mT.

To ensure the robustness and reliability of the dataset, each recorded observation does not represent a single instantaneous measurement but rather the average response computed over five repeated frequency sweeps. The resulting averaging duration corresponds to the minimum in the Allan deviations of magnitude or phase, which ensures that the measurement

is performed at the lowest noise level. Shorter averaging times would increase the contribution from photon shot noise and other fast fluctuations, whereas longer averaging times would increase the influence of long-term drift and 1/f noise. Furthermore, before their use in regression modeling, all input features were systematically normalized to standardize scales and improve consistency across variables. Standardization was performed independently for each feature, such that the training set features have zero mean and unit variance. The scaling parameters were computed exclusively from the training set to avoid data leakage, and then applied to both the training and test sets. This normalization step was essential to facilitate effective model training and to ensure fair comparisons among different regression algorithms, ultimately promoting better predictive accuracy and model generalizability [30].

Following data preprocessing, the dataset was partitioned into ten distinct subsets to enable a 10-fold cross-validation procedure. This approach involves repeatedly training and validating models on different data partitions, providing a comprehensive assessment of each model's predictive performance, stability, and generalization capability [31].

While long-term drift of the NV centers was not explicitly modeled, the use of 10-fold cross-validation across the entire dataset provides resilience against slow variations in the measurement conditions. As a result, the dataset retains realistic variability while suppressing spurious fluctuations, enabling a fair evaluation of regression models under practical noise conditions.

C. Algorithms Description

To effectively model the relationship between fluorescence lifetime measurements and physical parameters, a broad set of regression algorithms was evaluated. These include a wide range of machine learning paradigms, enabling a thorough assessment of different regression strategies. The selected models encompass both parametric and nonparametric approaches, including linear models, ensemble methods, and symbolic regression techniques. This selection ensures a comprehensive comparison across various learning paradigms, balancing interpretability, predictive performance, and computational efficiency.

1) *Decision Tree Regression [32]:* A nonparametric model that splits data recursively into subsets using feature thresholds, forming a tree structure. It is widely used due to its interpretability and efficiency.
2) *Elastic Net [33]:* A linear regression model that combines L1 (Lasso) and L2 (Ridge) penalties to balance feature selection and regularization. It is particularly useful when features are highly correlated.
3) *Gradient Boosting Regression [34]:* An ensemble learning method that builds decision trees sequentially, correcting errors from previous iterations to improve predictive accuracy.
4) *K-Nearest Neighbors (KNNs) Regression [35]:* A nonparametric algorithm that predicts values based on the average of the k-nearest training examples. It is useful for capturing local patterns in data.
5) *Lasso Regression [36]:* A linear regression model with L1 regularization that shrinks some coefficients to zero, effectively performing feature selection.
6) *LightGBM Regression [25]:* A gradient boosting framework optimized for speed and efficiency by using histogram-based learning. It is well-suited for large datasets.
7) *Linear Regression [37]:* A simple parametric approach modeling the relationship between a dependent variable and one or more independent variables.
8) *Random Forest Regression [38]:* An ensemble method that constructs multiple decision trees and averages their predictions to reduce overfitting and improve generalization.
9) *Support Vector Regression (SVR) [39]:* A regression technique using support vector machines (SVMs) [40] that aims to fit data within a specified margin while minimizing error.
10) *GP Symbolic Regressor [41]:* An approach that searches for mathematical expressions representing data relationships using genetic programming.
11) *XGBoost Regression [24]:* An efficient and scalable implementation of gradient boosting with regularization and parallel computing, widely used in machine learning competitions and real-world applications.

D. Evaluation Metrics

The evaluation of regression models has been conducted using classical performance metrics, including mean squared error (mse), root-mean-squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). These metrics provide a comprehensive assessment of model accuracy and reliability [31].

mse measures the average squared difference between actual and predicted values, where lower values indicate better performance. RMSE, as the square root of MSE, expresses errors in the same unit as the target variable and penalizes large deviations more heavily than MAE. MAE, on the other hand, calculates the average absolute difference between actual and predicted values, making it less sensitive to outliers. Lastly, R^2 represents the proportion of variance in the target variable that is explained by the model, with values ranging from 0 to 1, where a higher value indicates better explanatory power.

In addition to assessing model performance, the computational cost associated with model execution was also considered. Specifically, the time required for both the training and testing processes, measured in seconds, was evaluated. Training time represents the total duration needed to train the model, while prediction time quantifies the time required to generate predictions on the test dataset. These metrics are crucial for evaluating the feasibility of deploying models on embedded devices or using them in real-time applications.

IV. Results and Discussion

The performance results, as summarized in Table I and represented in Fig. 2, reveal a broadly high level of predictive accuracy across most models, with several methods achieving

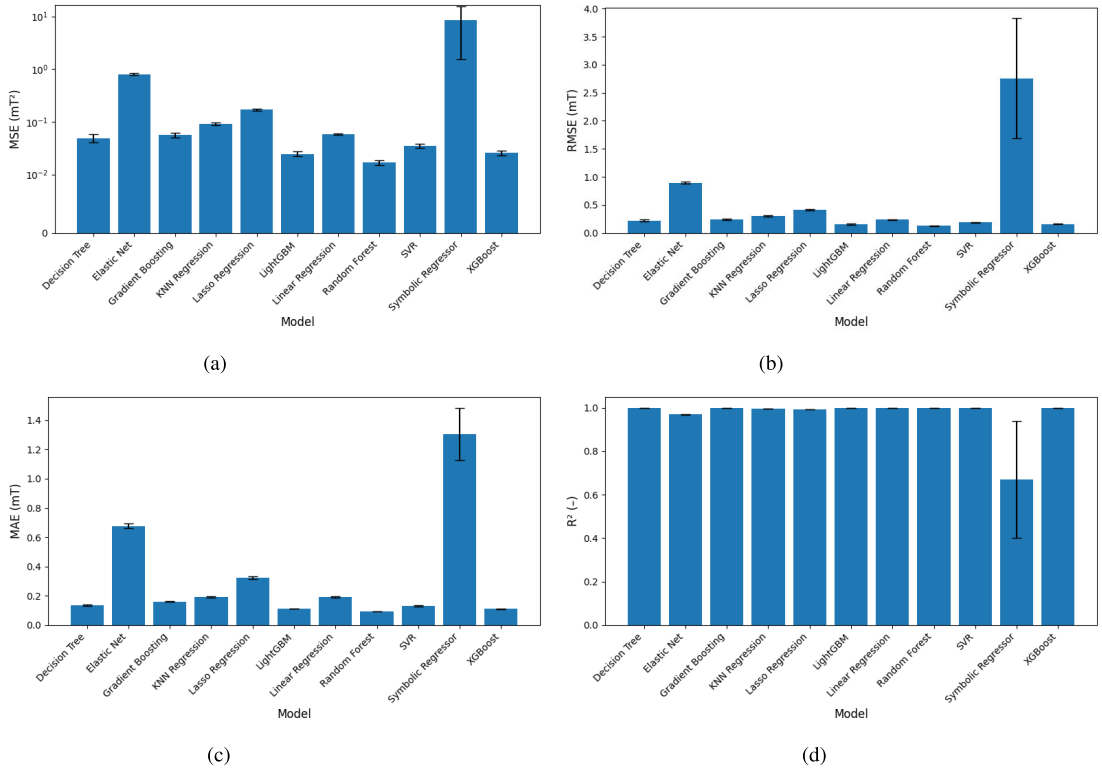


Fig. 2. Performance metrics for each regression model (mean and standard deviation across partitions). (a) mse, (b) RMSE, (c) MAE, and (d) coefficient of determination (R^2).

near-perfect fits. In particular, ensemble-based models such as LightGBM, random forest, and XGBoost, as well as SVR, exhibit exceptionally low MSEs — on the order of 0.017–0.026 mT^2, and correspondingly low root-mean-squared errors (RMSEs) and mean absolute errors (MAEs). These models also record R^2 values that are effectively 1 (ranging from 0.998 to 0.999), indicating that they explain nearly all the variance in the data. Such performance suggests that these models are able to identify the underlying relationships in the data with high fidelity, making them strong candidates for applications where predictive precision is a primary goal.

In contrast, simpler models like decision tree and linear regression also achieve commendable performance, with MSEs of approximately 0.049 and 0.058 mT^2, respectively, and R^2 values around 0.998. These results imply that even models with relatively straightforward structures can perform well on the given dataset, provided that the underlying data relationships are sufficiently linear or can be effectively partitioned by decision rules.

Similarly, regularized linear models such as elastic net and lasso regression deliver acceptable performance, although their error metrics are somewhat higher compared to the best-performing ensemble methods. This slight degradation in accuracy may be attributed to the inherent bias introduced by regularization, which, while aiding in the prevention of overfitting, can sometimes limit the model's ability to capture complex patterns in the data. This phenomenon is a manifestation of the bias-variance tradeoff, where the regularization term increases bias but reduces variance, resulting in more robust but potentially less accurate models when the underlying relationships in the data are highly nonlinear or complex [42].

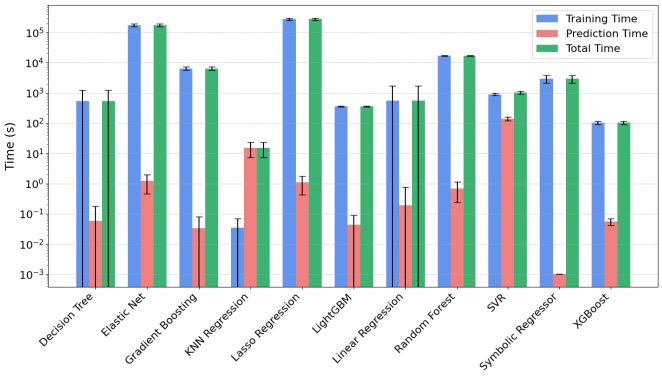


Fig. 3. Training, prediction, and total execution times for each regression model. Bars indicate mean values, with error bars representing standard deviation across partitions. Time is represented on a logarithmic scale.

The GP symbolic regression approach, with an MSE of 8.634 mT^2 and an R^2 of only 0.67, performs substantially worse than all other models considered. The high standard deviations accompanying these metrics further suggest that its predictions are highly variable, indicating instability in its estimation process.

The execution time analysis reveals substantial differences among the various models, underscoring the importance of evaluating computational cost alongside predictive accuracy. As shown in Table II and illustrated in Fig. 3, both training

TABLE I
MEAN AND STANDARD DEVIATION OF REGRESSION METRICS ACROSS ALL PARTITIONS

Model	MSE (mT^2)		RMSE (mT)		MAE (mT)		R^2	
	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s
Decision Tree	0.049	0.009	0.220	0.020	0.137	0.004	0.998	0.000
Elastic Net	0.805	0.039	0.897	0.022	0.676	0.017	0.969	0.001
Gradient Boosting	0.056	0.006	0.237	0.012	0.161	0.004	0.998	0.000
KNN Regression	0.092	0.006	0.303	0.010	0.193	0.004	0.996	0.000
Lasso Regression	0.171	0.008	0.413	0.010	0.323	0.007	0.993	0.000
LightGBM	0.025	0.003	0.157	0.008	0.111	0.002	0.999	0.000
Linear Regression	0.058	0.002	0.240	0.005	0.192	0.003	0.998	0.000
Random Forest	0.017	0.002	0.129	0.006	0.093	0.001	0.999	0.000
SVR	0.035	0.003	0.187	0.009	0.131	0.005	0.999	0.000
GP Symbolic Regressor	8.634	7.087	2.758	1.069	1.305	0.180	0.670	0.268
XGBoost	0.026	0.002	0.160	0.007	0.111	0.002	0.999	0.000

TABLE II
MEAN AND STANDARD DEVIATION OF TRAINING AND PREDICTION TIME ACROSS ALL PARTITIONS

Model	Training Time (s)		Prediction Time (s)		Total Time (s)	
	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s
Decision Tree	535.407	690.812	0.058	0.121	535.465	690.812
Elastic Net	175 670.859	19 172.181	1.216	0.751	175 672.075	19 171.931
Gradient Boosting	6424.911	797.430	0.033	0.048	6424.943	797.422
KNN Regression	0.034	0.034	15.129	7.888	15.163	7.889
Lasso Regression	278 207.511	23 848.242	1.098	0.669	278 208.610	23 848.511
LightGBM	358.226	17.167	0.043	0.045	358.268	17.171
Linear Regression	548.921	1132.622	0.194	0.573	549.115	1133.053
Random Forest	16 901.524	675.645	0.687	0.448	16 902.212	675.429
SVR	902.328	90.361	140.059	19.782	1042.387	109.270
GP Symbolic Regressor	2979.194	851.527	0.001	0.000	2979.195	851.527
XGBoost	102.000	12.946	0.055	0.014	102.054	12.946

and prediction times vary considerably across models, further highlighting the tradeoffs involved in selecting an appropriate regression algorithm. For instance, tree-based ensemble methods, like XGBoost and LightGBM, stand out with their rapid training and prediction times—XGBoost completes training in roughly 102 s and prediction in about 0.055 s, while LightGBM requires approximately 358 s for training and 0.043 s for prediction. These models' low prediction times make them particularly attractive for applications where both real-time predictions and frequent model updates are necessary.

Elastic net and lasso regression incur high training times (175 671 and 278 208 s, respectively), even though their prediction times remain relatively low. Such prolonged training duration may be justified in scenarios where the slight improvement in accuracy is critical, yet it can be a limiting factor in time-sensitive environments or when computational resources are constrained.

Another observation arises with KNN regression. Unlike most other models, KNN regression requires almost negligible time for training (around 0.034 s) because it essentially stores the training data without learning an explicit model. However, this benefit is offset by its prediction phase, which is considerably slower (approximately 15.129 s on average). This tradeoff is characteristic of lazy learning algorithms and highlights the fact that a low training time does not necessarily translate into overall computational efficiency, especially when rapid predictions are needed [43].

Finally, models such as decision trees, random forests, and SVR present more balanced execution profiles. Their training times, while higher than those of XGBoost or LightGBM, are still within a reasonable range (with decision tree training around 535 s and random forest around 16 901 s), and their prediction times remain low. This balance makes them versatile candidates for a variety of applications where moderate computational costs can be tolerated in exchange for robust performance.

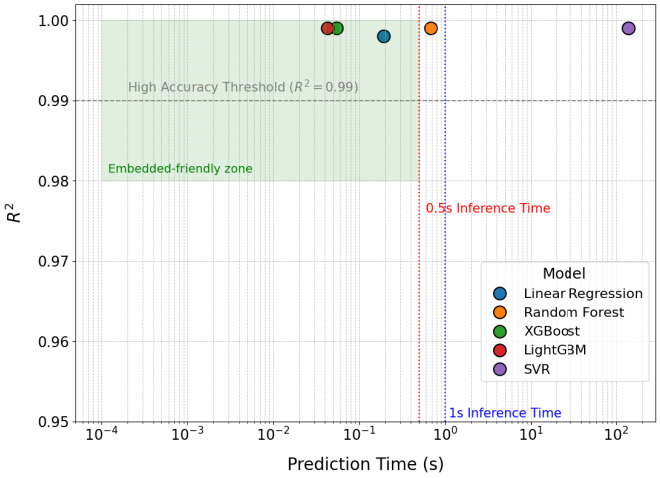


Fig. 4. Tradeoff between R^2 and prediction time for methods with $R^2 > 0.95$. Prediction time is plotted on a logarithmic scale.

Fig. 4 shows the tradeoff between accuracy and prediction time for methods with $R^2 > 0.95$. Reference lines at 0.5 and 1 s illustrate practical thresholds for real-time inference in embedded systems. While 1 s is often cited as a practical

upper limit [44], stricter requirements, such as response times below 500 ms are required for interactive or safety-critical applications [45]. The shaded area highlights models that are both accurate and fast enough for deployment in resource-constrained environments.

While ensemble methods such as LightGBM and XGBoost demonstrate excellent predictive performance ($R^2 \approx 0.999$) with subsecond inference times, other methods like SVR, despite achieving similar accuracy, exhibit significantly higher latency. In contrast, the symbolic regressor offers ultrafast inference but at the cost of lower accuracy.

In comparison with prior work, Tsukamoto et al. [9] achieved sub-2 μT accuracy using Gaussian process regression on ODMR spectra, and Homrighausenet al. [10] demonstrated microtesla-level precision on an ESP32 with ODMR-based readout. By contrast, the all-optical fluorescence lifetime approach yields mean absolute errors of 90–130 μT. This reduced sensitivity is explained by the weaker magnetic contrast of lifetime signals and the broader field range explored. Nevertheless, a complementary contribution is provided through a systematic benchmark of regression models under a microwave-free modality, in which hardware complexity is reduced and energy-efficient deployment in resource-constrained platforms is facilitated.

It should be noted that although averaging and feature standardization reduce short-term noise contributions, long-term drift was not explicitly addressed in this study. Future work will focus on incorporating drift compensation strategies to further enhance robustness in extended measurement experiments.

While the present evaluation demonstrates robust intrasetup performance, the generalization of models trained on a specific NV configuration to other experimental conditions (e.g., varying defect concentrations, temperature ranges, or noise levels) remains an open question. Addressing this challenge could require systematic validation and may benefit from transfer learning or domain adaptation strategies, which it is identified as an important direction for future work.

A. Statistical Analysis

The statistical significance of the performance differences among the regression models was evaluated with the Friedman test, a nonparametric, rank-based counterpart of the repeated-measures ANOVA [46]. For every fold of the 10-fold cross-validation, the models were first ranked from best to worst with respect to each performance measure (MSE, RMSE, MAE, R^2, training time, and prediction time). The Friedman procedure then tests the null hypothesis H_0 that all models have identical average ranks, that is, any observed differences arise purely from sampling variation. Because it relies on ordinal ranks rather than the raw metric values, the test does not require normality or homoscedasticity, making it well-suited to the heteroscedastic error distributions and highly skewed execution times observed in the study.

The resulting χ_F^2 statistics and associated p-values (Table III) are extremely large and far below the conventional $\alpha = 0.05$ threshold for every metric. Consequently, H_0 is rejected in all cases, indicating that the choice of model has a statistically significant impact on both predictive accuracy and computational cost. Owing to this global significance, a Nemenyi posthoc analysis [46] was performed to identify which specific pairs of models differ, as summarized below.

TABLE III
FRIEDMAN TEST RESULTS FOR ALL METRICS, INCLUDING INDICATION OF STATISTICAL SIGNIFICANCE AT $\alpha = 0.05$

Metric	Friedman Statistic	p-value	Significant?
MSE (mT^2)	98.64	1.02×10^{-16}	Yes
RMSE (mT)	98.64	1.02×10^{-16}	Yes
MAE (mT)	98.84	9.31×10^{-17}	Yes
R^2	98.64	1.02×10^{-16}	Yes
Training Time (s)	98.78	9.55×10^{-17}	Yes
Predict time (s)	88.53	1.05×10^{-14}	Yes

TABLE IV
NUMBER OF PAIRWISE COMPARISONS (MODEL PAIRS) THAT ARE STATISTICALLY SIGNIFICANT ($p < 0.05$) ACCORDING TO THE NEMENYI POSTHOC TEST

Metric	Significant Pairs
MSE	20
RMSE	20
MAE	21
R^2	20
Training Time	18
Prediction Time	16

Pairwise comparisons were further examined using the Nemenyi test ($\alpha = 0.05$). Table IV summarizes the number of model pairs that exhibit statistically significant differences for each metric. Accuracy metrics (MSE, RMSE, and R^2) show 20–21 significant contrasts, mainly involving the symbolic regressor and elastic net versus the tree-based ensembles. For computational cost, 18 contrasts are significant for training time and 16 for prediction time, highlighting SVR as the slowest and KNN/Lasso as the fastest approaches.

V. CONCLUSION

Given the increasing demand for energy-efficient, real-time quantum sensors, this work set out to quantitatively guide the selection of machine learning models suitable for embedded NV center-based magnetometry. To this end, a statistically robust comparison of eleven regression algorithms was conducted, using a dataset of 22 758 fluorescence lifetime measurements obtained under controlled magnetic fields.

Using the Friedman test followed by Nemenyi posthoc comparisons, it was confirmed that the differences among models are highly significant for every accuracy metric: MSE, RMSE, MAE, and R^2, as well as for both training and prediction times, with all p-values well below 10^{-14}. These results rule out the possibility that the observed performance gaps arise from random fold-to-fold fluctuations, demonstrating that algorithm choice has a material impact on sensing accuracy and computational burden.

Boosted tree ensembles, in particular LightGBM and XGBoost, consistently deliver the highest accuracy ($R^2 \approx 0.999$) while maintaining subsecond inference latency. Their favorable average ranks place them on the Pareto frontier for applications requiring both high precision and

real-time operation on embedded hardware. In contrast, simpler models—including decision tree, KNN, Lasso, and ordinary least-squares linear regression—sacrifice only a small amount of accuracy yet offer the shortest training and prediction times, making them compelling when memory, energy, or on-device retraining constraints dominate.

SVR achieves excellent predictive quality but is statistically the slowest at inference, whereas the symbolic regressor is virtually instantaneous to evaluate, yet performs markedly worse than all other models in every accuracy metric. These extremes illustrate that neither theoretical model complexity nor strict interpretability alone guarantees superior results. Overall, the analysis underscores the need to balance predictive power against computational cost and supplies quantitative guidance for choosing regression techniques suitable for low-power, real-time NV magnetometers.

REFERENCES

[1] J.-P. Tetienne et al., "Magnetic-field-dependent photodynamics of single NV defects in diamond: An application to qualitative all-optical magnetic imaging," *New J. Phys.*, vol. 14, no. 10, Oct. 2012, Art. no. 103033, doi: 10.1088/1367-2630/14/10/103033.

[2] G. Balasubramanian et al., "Nanoscale imaging magnetometry with diamond spins under ambient conditions," *Nature*, vol. 455, no. 7213, pp. 648–651, Oct. 2008, doi: 10.1038/nature07278.

[3] Y. Xie et al., "A hybrid magnetometer towards femtotesla sensitivity under ambient conditions," *Sci. Bull.*, vol. 66, no. 2, pp. 127–132, Jan. 2021, doi: 10.1016/j.scib.2020.08.001.

[4] F. A. Pedroza-Montero, K. J. Santacruz-Gomez, R. Meléndrez-Amavizca, and M. Barboza-Flores, "Commercial nanodiamonds for precise fluorescence-based temperature sensing," *Appl. Phys. Lett.*, vol. 125, no. 7, Aug. 2024, Art. no. 073701, doi: 10.1063/5.0219532.

[5] D. K. Bommidi and A. D. Pickel, "Temperature-dependent excited state lifetimes of nitrogen vacancy centers in individual nanodiamonds," *Appl. Phys. Lett.*, vol. 119, no. 25, Dec. 2021, Art. no. 254103, doi: 10.1063/5.0072357.

[6] L. Horsthemke et al., "Towards resolving the ambiguity in low-field, all-optical magnetic field sensing with high NV-density diamonds," *Eng. Proc.*, vol. 68, no. 1, p. 8, 2024. [Online]. Available: https://www.mdpi.com/2673-4591/68/1/8

[7] R. Wunderlich, R. Staacke, W. Knolle, B. Abel, and J. Meijer, "Magnetic field and angle-dependent photoluminescence of a fiber-coupled nitrogen vacancy rich diamond," *J. Appl. Phys.*, vol. 130, no. 12, Sep. 2021, Art. no. 124901, doi: 10.1063/5.0059330.

[8] S. Sengottuvel et al., "Microwave-free imaging magnetometry with nitrogen-vacancy centers in nanodiamonds at near-zero field," *Phys. Rev. Appl.*, vol. 23, no. 3, Mar. 2025, Art. no. 034001, doi: 10.1103/PhysRevApplied.23.034001.

[9] M. Tsukamoto et al., "Accurate magnetic field imaging using nanodiamond quantum sensors enhanced by machine learning," *Sci. Rep.*, vol. 12, 2022, Art. no. 13942, doi: 10.1038/s41598-022-18115-w.

[10] J. Homrighausen, L. Horsthemke, J. Pogorzelski, S. Trinschek, P. Glösekötter, and M. Gregor, "Edge-machine-learning-assisted robust magnetometer based on randomly oriented NV-ensembles in diamond," *Sensors*, vol. 23, no. 3, p. 1119, Jan. 2023, doi: 10.3390/s23031119.

[11] M. Jani, P. Czarnecka, Z. Orzechowska, M. Mrózek, W. Gawlik, and A. M. Wojciechowski, "Sensing of magnetic-field gradients with nanodiamonds on optical glass-fiber facets," *ACS Appl. Nano Mater.*, vol. 6, no. 13, pp. 11077–11084, Jul. 2023, doi: 10.1021/acsanm.3c00887.

[12] E. Oh et al., "Perspective on quantum sensors from basic research to commercial applications," *AIAA J.*, vol. 62, no. 11, pp. 4029–4053, Nov. 2024.

[13] S. E. Crawford et al., "Quantum sensing for energy applications: Review and perspective," *Adv. Quantum Technol.*, vol. 4, no. 8, Aug. 2021, Art. no. 2100049.

[14] T. Dutta et al., "Advances in integrated quantum photonics for quantum sensing and communication," *J. Mater. Chem. C*, vol. 13, no. 23, pp. 11521–11561, 2025, doi: 10.1039/d4tc05290k.

[15] S. Devendrababu, S. Ganguly, and K. Hemachandran, "Mapping quantum industry demands to education: A critical analysis of skills, qualifications, and modalities," *EPJ Quantum Technol.*, vol. 12, no. 1, p. 105, Dec. 2025, doi: 10.1140/epjqt/s40507-025-00406-6.

[16] H. F. Tsoi et al., "Symbolic regression on FPGAs for fast machine learning inference," *Proc. EPJ Web Conf.*, vol. 295, Apr. 2024, Art. no. 09036, doi: 10.1051/epjconf/202429509036.

[17] R. Immonen and T. Hämäläinen, "Tiny machine learning for resource-constrained microcontrollers," *J. Sensors*, vol. 2022, pp. 1–11, Nov. 2022, doi: 10.1155/2022/7437023.

[18] R. David et al., "Tensorflow lite micro: Embedded machine learning for tinyml systems," in *Proc. Mach. Learn. Syst.*, vol. 3, 2021, pp. 800–811. [Online]. Available: https://proceedings.mlsys.org/paper_files/paper/2021

[19] N. Makke and S. Chawla, "Interpretable scientific discovery with symbolic regression: A review," *Artif. Intell. Rev.*, vol. 57, p. 2, 2024, doi: 10.1007/s10462-023-10622-0.

[20] Y. Chong et al., "Machine learning of spectra-property relationship for imperfect and small chemistry data," *Proc. Nat. Acad. Sci. USA*, vol. 120, no. 20, May 2023, Art. no. 2220789120, doi: 10.1073/pnas.2220789120.

[21] J. Sierra-Velez, A. Vial, D. Macías, M. Inchaussandague, and D. Skigin, "Modeling the optical properties of biological structures using symbolic regression," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 112, no. 3, Sep. 2025, Art. no. 034404, doi: 10.1103/w8jn-qs2z.

[22] H. Kay et al., "Constructing a symbolic regression-based interpretable soft sensor for industrial data analytics and product quality control," *Ind. Eng. Chem. Res.*, vol. 63, no. 9, pp. 4083–4092, Mar. 2024, doi: 10.1021/acs.iecr.3c04021.

[23] L. Höfler, "Good results from sensor data: Performance of machine learning algorithms for regression problems in chemical sensors," *Sens. Actuators B, Chem.*, vol. 421, Dec. 2024, Art. no. 136528, doi: 10.1016/j.snb.2024.136528.

[24] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[25] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3146–3154. [Online]. Available: https://papers.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[26] AMS OSRAM AG. (2025). *OSRAM Metal Can TO56, PLT5-520B Green Laser Diode*. Accessed: Jul. 22, 2025. [Online]. Available: https://ams-osram.com/products/lasers/color-lasers-eel/osram-metal-can-to56-plt5-520b

[27] I. Thorlabs. (2018). *DMSP567R 25 mm × 36 Mm Shortpass Dichroic Mirror, 567 Nm Cutoff*. Accessed: Jul. 22, 2025. [Online]. Available: https://www.thorlabs.com/thorproduct.cfm?partnumber=DMSP567R

[28] (2012). *FELH0600 25 mm Longpass Filter, Cut-On Wavelength: 600 nm*. Accessed: Jul. 22, 2025. [Online]. Available: https://www.thorlabs.com/thorproduct.cfm?partnumber=FELH0600

[29] J. R. Lakowicz, *Principles of Fluorescence Spectroscopy*. Cham, Switzerland: Springer, 2006.

[30] A. K. Jain, P. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[31] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Cham, Switzerland: Springer, 2009.

[32] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984. [Online]. Available: https://www.routledge.com/Classification-and-Regression-Trees/Breiman-Friedman-Olshen-Stone/p/book/9780412048418

[33] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[34] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.

[35] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *Int. Stat. Rev. / Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, Dec. 1989, doi: 10.2307/1403797.

[36] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. Ser. B: Stat. Methodology*, vol. 58, no. 1, pp. 267–288, Jan. 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[37] F. Galton, "Regression towards mediocrity in hereditary stature," *J. Anthropological Inst. Great Britain Ireland*, vol. 15, pp. 246–263, Mar. 1886, doi: 10.2307/2841583.
[38] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/a:1010933404324.
[39] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 9, 1997, pp. 155–161. [Online]. Available: https://papers.nips.cc/paper/1996/hash/d38901788c533e7b50e175e3bc9c5303-Abstract.html
[40] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[41] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press, 1992. [Online]. Available: https://mitpress.mit.edu/9780262111706/genetic-programming/
[42] C. M. Bishop, *Pattern Recognition and Machine Learning*. Cham, Switzerland: Springer, 2006.
[43] J. Singh, "Computational complexity and analysis of supervised machine learning algorithms," in *Next Generation of Internet of Things*. Singapore: Springer, 2023, pp. 195–206.
[44] S. Mittal, "A survey of techniques for improving energy efficiency in embedded computing systems," *Int. J. Comput. Aided Eng. Technol.*, vol. 6, no. 4, pp. 440–459, 2014.
[45] S. Heydari and Q. H. Mahmoud, "Tiny machine learning and on-device inference: A survey of applications, challenges, and future directions," *Sensors*, vol. 25, no. 10, p. 3191, May 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/10/3191
[46] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jul. 2006. [Online]. Available: http://www.jmlr.org/papers/v7/demsar06a.html

José Luis Ávila-Jiménez received the B.Sc. degree in software engineering from the University of Córdoba, Córdoba, Spain, in 2001 and 2013, respectively, and the M.Sc. degree in computer engineering from UNED, Madrid, Spain, in 2006, and the Ph.D. degree from the University of Córdoba, in 2013.

He has been with the Department of Electronics and Computer Engineering, University of Córdoba, since 2021. He is a co-author of many articles in international journals and conferences. He has participated in several research projects. His current research interests include deep learning and data mining.

Ann-Sophie Bülter received the M.Sc. degree in electrical engineering from FH Münster—University of Applied Sciences, Steinfurt, Germany, in 2024. She is currently pursuing the Ph.D. degree with the University of Granada, Granada, Spain.

Her current research interests include magnetometry and current sensing with NV centers and edge machine learning.

Ludwig Horsthemke received the M.Sc. degree in electrical engineering from FH Münster, Steinfurt, Germany, in 2021, and the Ph.D. degree in information and communication technology from the University of Granada, Granada, Spain, in 2025.

His research focuses on quantum sensing and embedded electronics.

Francisco J. Rodriguez-Lozano received the B.Sc. degree in computer science, the M.Sc. degree in distributed renewable energies, and the Ph.D. degree in computer engineering from the University of Córdoba, Córdoba, Spain, in 2016, 2017, and 2020, respectively.

He has been with the Department of Electronics and Computer Engineering, University of Córdoba, since 2018, where he is currently a member. He has been the main author of several articles in high-ranking international journals. He has participated in different conferences and research projects. His research interests include computer vision, machine learning, and embedded systems.

Manuel Ortiz-Lopez received the M.Sc. degree in physics electronics from the University of Granada, Granada, Spain, in 1987, and the Ph.D. degree from the University of Córdoba, Córdoba, Spain, in 2013.

From 1987 to 1996, he was with the Research and Development Department of several companies (TECOSA, Fujitsu, and Telefonica). He has been with the Department of Electronic and Computer Engineering, University of Córdoba, since 1996. He has authored or co-authored several articles covering topics on embedded systems and wired and wireless networks applied to communications for energy management systems and applications.

Peter Glösekötter (Member, IEEE) received the Dipl.-Ing. and Ph.D. degrees in electrical engineering from the University of Dortmund, Dortmund, Germany, in 1997 and 2002, respectively.

Since 2006, he has been a Professor at Münster University of Applied Sciences, Steinfurt, Germany. His research interests include embedded systems, circuit architectures for emerging technologies, and quantum sensing.

Dr. Glösekötter is a member of the IEEE Solid-State Circuits Society and IEEE Computer Society.