Journal of Defence and Security Industries: Strategy and Technology, 2026, 1(2)

Science & Technology Section

ML-Based RF Calibration of Military T/R Modules

Alperen M. Ünal[1], Yusuf Yiğitbaşı[1], Akil Kutlu[1], Fatma Hatipoğlu[1,*]

Abstract

In this work, Machine Learning (ML) techniques are incorporated into the Radio Frequency (RF) calibration workflow with the objective of reducing associated costs and enhancing operational efficiency for the military grade Transmitter/Receiver (T/R) modules. RF calibration entails a series of measurements that verify the module's performance at certain temperatures, power levels and frequencies; when performed with conventional procedures, it is both time-consuming and expensive. In this study for a single unit, the traditional RF calibration process requires approximately 17 h and involves ten dedicated test and calibration stations as well as seven operators. Ensuring RF calibration, seven distinct Received-Signal-Strength-Indicator (RSSI) values are recorded; in this study focuses on one of these parameters, denoted RSSI F, which is essential for the correct functioning of the target module. In validation, the ML predicted RSSI F values are directly benchmarked against measured references under a ±12 mV error envelope; the comparison yielded a 66.80% success rate.

Keywords: Machine learning, artificial intelligence, RF calibration, regression, XGBoost, LightGBM, and TabNet.

Received: 12.03.2026 Accepted: 21.05.2026 Research Article

[1] ASELSAN Inc., Ankara, Türkiye
* Corresponding author.
✉ fatmakart1407@gmail.com

Cite this article as:
https://doi.org/10.65834/jdsi.12.52
Ünal, A.M., Yiğitbaşı, Y., Kutlu, A. & Hatipoğlu, F. (2026). ML-Based RF Calibration of Military T/R Modules. *Journal of Defence and Security Industries: Strategy and Technology 1(2)*, 110-128.

1. Introduction

Communication is a vital nerve network for the army's command-control, intelligence sharing and unit coordination; without timely and secure information flow, decisions are delayed, synchronization between units deteriorates and operational effectiveness is significantly reduced. Communication between every unit has an impact on the strength of operability in the army. Considering that many units sending and receiving voice and digital data, linearity is critical. The army is obliged to fulfil its duties in different places, under different climatic conditions for long periods of service time. Therefore, military technology Radio Frequency (RF) devices must be capable of delivering high performance across the operating temperature range from -40°C to +55°C. The system shall operate without performance degradation across specified temperature extremes and therefore must successfully satisfy the high-temperature and low-temperature qualification tests defined in MIL-STD-810H, 501.7 and 502.7 (United States Department of Defence [DOD], 2019). In view of RF Transceiver device as a system, expecting multiple subsystem, such as power supply, Power Amplifier (PA) and filters have to perform proportionally.

The basis of the transferring an information in the environment, carrying the frequency and amplifying of signals have high level of linearity. Amplified the signals according to varied required signal levels, bandwidths, waveform have to meet the performance metrics. These metrics are 1 dB compression point P_{1dB}, power gain, Power-Added Efficiency (PAE), Adjacent Channel Power Ratio (ACPR), Third-Order Intercept Point (IP_3), Peak-to-Average Power Ratio (PAPR). Managing broad bandwidths and performance metrics while preserving the signal integrity and efficiency is a significant issue.

Modern communications systems cover High Frequency (HF), Very High Frequency (VHF) and Ultra High Frequency (UHF) communication transmitters. The communication units generate signals in baseband and amplify these signals to the medium by carrying the frequencies. According to the technical standards and specs, PAs amplify the generated signals with varying amplitude and phase components. This amplifying process must preserve signal integrity, minimize interference, be accurate and linear. Therefore, the PAs are of critical importance by influencing the power efficiency of radio transmitters and receivers. (Ortega-González, Pño-Gómez, & Madueño-Pulido, 2025).

RF calibration is the process of setting specific parameters into digital units so that a device or module operates stably and efficiently. The RF calibration process can be performed in three different approaches. The first approach relies on iterative testing and exhaustive search techniques, which are characterized by high cost and significant time consumption. To overcome these limitations, recent research emphasizes a shift toward Machine Learning (ML) and fully integrated self-calibration strategies. Unlike manual calibration that necessitates extensive data collection, ML-based techniques can attain near-optimal performance settings using only a fraction of the measurements required by exhaustive searches. Bayesian optimization with both global and local search algorithms is used to predict control voltages of circuits (Chang, Wen, Hong, Padmanabhan, Franzon, & Floyd, 2025).

In the work of Dikmese et al., modern ML methods are proposed to predict the dynamic nonlinear behaviour of wideband RF PAs. Neural networks, k-nearest neighbours and several tree-based ML algorithms are adapted to handle complex-valued signals and applied to PA modelling. Evaluation with measured data from two commercial base-station PAs shows that gradient boosting performs best, yielding comparable or better performance than memory-polynomial reference models in terms of Normalized Mean-Squared Error (NMSE) and Adjacent-Channel Error-Power Ratio (ACEPR). The authors suggest that applying techniques for Digital Pre-Distortion (DPD) linearization could significantly reduce predictor computational complexity by developing a single predictor capable of modelling a PA across different output-power levels (Dikmese et al., 2019).

Sriram and Tuffy present an ML-based accelerated optimization framework for RF PA design that utilizes Max-Min Latin Hypercube Sampling with CatBoost gradient boosting to intelligently explore multidimensional parameter spaces. This approach reduces simulation requirements by 65% while maintaining ±0.4 dBm accuracy for most modes. The trained CatBoost model predicts P2dB performance across the entire design space, transforming PA design from exhaustive search to intelligent exploration and achieving a 58.24%–77.78% reduction in total simulation time with an average R^2 of 0.901 across 15 PA operating modes, enabling design cycles twice as fast without compromising accuracy (Sriram & Tuffy, 2025).

Tong et al., propose a measurement-driven automated PA design approach that leverages load-pull measurement data and ML methods for the design of harmonically controlled PAs. By introducing a two-stage optimization strategy using the Particle-Swarm Optimization (PSO) algorithm, accurate matching of harmonic impedances is achieved. The automated flow demonstrates high performance in a 2.4 GHz PA, attaining a maximum Power-Added Efficiency (PAE) of 73.4% (or 73.6%) and a maximum output power of 40 dBm, confirming that the process is able to avoid multiple simulation-tool iterations (Tong et al., 2024).

Cai et al., present a dynamic behavioural-modelling technique for multi-device RF PAs that often exhibit strongly nonlinear behaviour with long-term memory effects. The method combines a real-valued decomposed-piecewise approach with Support-Vector Regression (SVR) to accurately model distinct characteristics of the multi-device system across different power levels. Experimental results show that the decomposed piecewise SVR model provides significant performance improvements over standard SVR models, achieving approximately 10 dB-14 dB reduction in Normalized Mean Square Error (NMSE) for a single-transistor PA and an NMSE of 50 dB for a multi-device Doherty PA (Cai et al., 2020).

Uğur used a dataset consisting of measurements from 200 power-converter modules obtained during prototype and pilot production at three power levels, three temperatures and 128 frequency points. The features included driver-bias voltage, output-bias voltage, target output power, temperature, gain and the current drawn from the main power source. Hyper-parameter optimization is performed automatically with the third-party AX platform, which explored the hyper-parameter space using Sobol sampling to identify the configuration yielding the best performance. Among the evaluated models, a Multilayer Perceptron (MLP) achieved the lowest prediction error 1.07% Mean Absolute Error (MAE) on the second dataset, outperforming both a Convolutional Neural Network (CNN) and a linear-regression baseline. The inclusion of the primary-current feature in the second dataset further improved the MLP's gain-prediction accuracy, confirming the relevance of current-related variables for Printed Circuit Board (PCB) characterization. Conversely, the CNN's performance deteriorated when the larger, current-augmented dataset is used and its training cost is markedly higher than that of the MLP, indicating that CNNs are less suited to high-dimensional tabular data in this context (Uğur, 2023).

Artificial Intelligence (AI) can optimize non-linearity. Military tech communication devices and systems need optimization to prevent any performance decreases. Climatic conditions, supply current and voltages, waveforms, aging, design limitations cause non-linearity. Although the hardware is the same design, units identified by different component which has different serial number may differ slightly in their test results. To remove non-linearity, RF calibration is essential and highly recommended.

An ML-based prediction of the RF calibration via digital control operations is presented in this paper. While ML techniques have been previously explored for RF PA modelling and calibration-related tasks, their specific application to predicting digital control parameter values for T/R-modules using large-scale production data remains limited (Cai et al., 2020; Chang et al., 2025; Dikmese et al., 2019; Sriram et al., 2025; Uğur et al., 2023). In this study, we address this application-specific gap by generating a ML-based model that predicts digital control parameter values of the RF circuit of the settings with a negligible error margin, replacing conventional and self-test calibration methods. To achieve this, thousands of data rows are processed, interpolated and combined to create benchmark comparisons across multiple models, resulting in a prediction success rate of 66.80% within the defined error bounds. Although ML techniques have been previously used for RF PA behavioural modelling and circuit design optimization, their application to predicting digital control knob parameters for military-grade T/R modules directly from large-scale production test data without requiring post-fabrication calibration runs remains an unexplored domain. Among the seven Received-Signal-Strength-Indicator (RSSI) parameters recorded during RF calibration, RSSI F is identified by the RF designers as the most critical parameter governing stable device operation and is therefore selected as the sole prediction target. The RSSI F value of fabricated T/R PA modules is predicted using the LightGBM algorithm, which significantly reduces the cost and labour force associated with iterative calibration approaches. The methodology consists of large-scale data acquisition, interpolation to fill missing entries, training and benchmarking of several regression algorithms (including XGBoost, LightGBM, and TabNet) with hyper-parameters tuned via Bayesian Optimization and selection of the best performing LightGBM based on error metrics such as Root Mean Squared Error (RMSE), Root Mean Square Percentage Error (RMSPE) and Mean Absolute Percentage Error (MAPE). The proposed framework offers high repeatability and robustness to process variations; however, it depends on the availability of a comprehensive measurement database and its accuracy may degrade with limited data, while the current implementation addresses only a single performance metric, leaving multi-objective calibration as a future challenge. In conclusion, the study confirms that ML-driven calibration specifically gradient-boosting regression can reliably predict digital control parameter values of T/R module with negligible error, streamlining the calibration workflow.

The key contributions of this work are:

- An ML-based RF calibration framework that reduces calibration time from 17 hours to minutes.
- Early application of gradient-boosting-based ML methods for RF PA calibration in production-scale military T/R modules.
- Systematic comparison of XGBoost, LightGBM, and TabNet on harmonized datasets.
- 66.80% prediction accuracy within ±12 mV error bounds using LightGBM regression.

The remainder of this paper is organized as follows: Section II describes the methodology and RF calibration workflow. Section III presents the dataset characteristics and pre-processing steps. Section IV details the ML models and performance metrics. Section V discusses the experimental results and model comparisons. Section VI concludes the paper with summary remarks and future research directions.

2. Methodology

RF calibration is the process of setting specific parameters so that a device operates stably and efficiently. With the current method, calibrating the output power of a single device takes about 17 h and incurs additional costs such as labour and test-setup expenses. A regression-based ML approach is developed to shorten the RF calibration time and make it more cost-effective. The model is built in Python, using NumPy and Pandas for data pre-processing and manipulation and Scikit-Learn for applying the ML algorithms. The algorithms employed are XGBoost (using the XGBoost library), LightGBM (using the LightGBM library), and TabNet (implemented with the PyTorch/Torch library).

During model training, the data are partitioned into training, validation and test sets with proportions of 70%, 15% and 15% respectively and 5-fold cross-validation is employed to mitigate overfitting. The ML models are trained on an Intel Xeon Gold 6254 CPU and an NVIDIA Grid T4-4Q GPU.

2.1. Dataset & Data Pre-processing

The calibrated measurements including frequency, power level, temperature and waveform labels are retrieved from a Structured Query Language (SQL) database. The database includes 187 PA modules records. For 167 of these modules, the data are partitioned into training, validation and test sets, while the remaining 20 modules are never exposed to the model and used for only testing the model to obtain the performance metrics of the model.

The experiment employed two distinct datasets. The first is referred to as the calibration-dataset, includes four fundamental measurement attributes: waveform, temperature, power level and frequency. The second is designated as the module-test dataset and contains the same four attributes and additionally, five RSSI values obtained during module-level testing to characterize module-specific behaviour.

The module-test dataset, an independent source of T/R-module measurements, comprises a single temperature level, one waveform, two distinct power levels and 40 frequency points. In contrast, the calibration-dataset comprises three temperature levels, two waveforms, five power levels and 64 frequencies. To harmonize the two datasets, the 40 frequencies present in the module-test data are linearly interpolated to match the 64 frequencies of the RF calibration values and the two power levels are interpolated to five power levels. To capture the intrinsic characteristics of each module, a comprehensive attribute set consisting of 41 distinctive features is derived using an interpolated dataset.

2.2. Feature Engineering

To effectively capture the complex RF dependencies and module-specific signatures embedded in the T/R-module measurements, a systematic feature engineering strategy is employed on the harmonized dataset. The raw module-test records provide five distinct Received-Signal-Strength-Indicator (RSSI) channels denoted RSSI A, RSSI B, RSSI D, RSSI F, and RSSI G each representing an independent receiver path. Rather than relying solely on these absolute values, the relative interactions between paths are encoded by constructing pairwise features across four physically meaningful antenna pairings: A–B, A–D, B–D, and F–D. For each pair, eight mathematical transformations are computed, yielding 32 derived descriptors.

These transformations are deliberately chosen to expose complementary aspects of the RF behaviour. The arithmetic mean and sum quantify common-mode signal strength and aggregate energy content, whereas the difference highlights relative imbalances that may indicate differential phase shifts or asymmetric attenuation between paths. The product captures multiplicative channel interactions, while the ratio and inverse ratio furnish scale-invariant metrics that suppress absolute power fluctuations and emphasize proportional disparities. The geometric mean characterizes multiplicative central tendency, and the harmonic mean provides an appropriate averaging measure for rate-like quantities such as signal-to-noise contributions. To ensure numerical stability particularly under weak-signal or near-null conditions a small regularization constant ($\varepsilon = 1\times10^{-8}$) is incorporated into all ratio and harmonic mean computations, thereby preventing division-by-zero artefacts.

This set of 32 engineered interaction features is augmented with the original five RSSI measurements and the four environmental/operational covariates: ambient temperature, waveform type, transmit power level, and operating frequency. The resulting 41-dimensional feature vector constitutes a rich, informative representation that enables the learning algorithm to discern intricate nonlinear mappings between the module's operating context and the target RSSI F calibration response. By explicitly modelling both intra-channel coupling phenomena and extrinsic operational conditions, this feature space enhances discriminative capacity and supports robust generalization, particularly for unseen modules during inference.

2.3. ML Models

This study addresses a regression problem based on tabular dataset. In such datasets, variables typically exhibit non-linear and highly complex interactions, which often renders classical linear models inadequate. Consequently, tree-based gradient boosting algorithms tend to demonstrate superior prediction performance under these conditions. In this study, XGBoost and LightGBM are used as base gradient boosting models. Additionally, the TabNet algorithm, based on deep learning is evaluated to provide a comparative analysis between traditional boosting methods and a neural network-based approach.

2.3.1. XGBoost

XGBoost, proposed by Chen and Guestrin, is an advanced gradient boosting algorithm known for its high scalability. This method uses second-order, i.e., Hessian gradient information and a regularized objective function containing both L_1 and L_2 penalties. These features provide fast and stable performance on large datasets, reducing overfitting (Chen and Guestrin, 2016).

2.3.2. LightGBM

LightGBM, proposed by Ke et al., is a gradient-boosting framework that achieves high training speed and predictive accuracy on large-scale datasets. It reduces computational burden substantially through two innovative techniques: Gradient-based One-Side Sampling, which retains instances with large gradients while randomly discarding a large portion of instances with small gradients; and Exclusive Feature Bundling, which merges mutually exclusive sparse features into a single composite feature, thereby decreasing dimensionality without loss of information. Consequently, LightGBM delivers fast, memory-efficient learning while maintaining state-of-the-art performance (Ke et al., 2017).

2.3.3. TabNet

TabNet, proposed by Arik and Pfister, is a deep learning architecture for tabular data that performs feature selection step-by-step through an attention-based mechanism. Thus, it provides interpretability by nature. By applying a sparse mask to determine the most important features at each decision step, the model achieves high prediction accuracy while additionally offering strong explainability (Arik and Pfister, 2019).

2.4. Performance Metrics

The study employed MAPE, RMSE and RMSPE as performance metrics to evaluate the models from different perspectives. RMSE measures the absolute error in the unit of the target variable. It is sensitive to large residuals, which can cause scale effects when the dataset is extensive and non-uniform. MAPE expresses the error as a percentage, offering a simpler and more interpretable measure, yet it becomes unstable when the true values approach zero due to its sensitivity to small magnitudes. RMSPE combines percentage-based error measurement with sensitivity to relative errors, thereby providing a more stable and comparable performance assessment for datasets with a wide range of values.

2.4.1. RMSE

As emphasized by Chai and Draxler and Willmott and Matsuura, RMSE is a regression performance metric defined as the square root of the mean of the squared errors; it penalizes large deviations more heavily and, because it shares the same unit as the target variable, is straightforward to interpret (Chai and Draxler, 2014; Matsuura, 2005).

This formulation is presented in Equation (1).

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}} \tag{1}$$

y_i: real value
$\hat{y}_i$: predicted value
n: number of values

2.4.2. MAPE

As emphasized by Hynman&Koehler, and Armstrong&Collopy, MAPE expresses forecast errors in percentage terms, making it easy to interpret, but it tends to inflate errors when the true value approaches zero (Hynman and Koehler, 2006; Armstrong and Collopy, 1992).

The MAPE formula is presented in Equation (2).

$$MAPE = \frac{\%100}{n} \sum_{i=0}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{2}$$

y_i: real value
$\hat{y}_i$: predicted value
n: number of values

2.4.3. RMSPE

As highlighted by De Myttenaere et al., RMSPE is a regression performance metric that computes the square root of the mean of squared percentage errors, thereby imposing heavier penalties on large relative deviations. It is especially advantageous in tabular-data problems where scale disparities are evident (De Myttenaere et al., 2016).

The RMSPE formula is presented in Equation (3).

$$RMSPE = \sqrt{\frac{1}{n} \sum_{i=0}^{n} \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2} \tag{3}$$

y_i: real value
$\hat{y}_i$: predicted value
n: number of values

2.5. Statistical Validation via Cluster Bootstrap

To address the inherent uncertainty associated with reporting performance metrics from a fixed test set of 20 modules, a cluster bootstrap procedure is employed at the module level. This approach preserves the within-module correlation of the 320 measurement points per module, providing a statistically valid estimate of metric variability. The 20 unseen test modules are resampled with replacement to generate B = 5000 bootstrap samples. For each resample, the success rate (within ±12 mV tolerance), outlier count, RMSE, MAPE, and RMSPE are recomputed for all four models. The 95% confidence intervals are derived using the percentile method, taking the 2.5th and 97.5th quantiles of the bootstrap distributions. This procedure quantifies the sensitivity of the reported metrics to the specific composition of the test set.

2.6. Hyperparameter Optimization

Hyperparameter tuning is carried out using a Bayesian Optimization framework. This probabilistic search strategy builds a surrogate model of the objective function and explores the hyperparameter space efficiently, thereby locating near-optimal configurations with a reduced number of evaluations. During the optimization, RMSE served as the primary performance metric for all models. For XGBoost, however, both RMSE and the RMSPE are employed, allowing the model's error to be assessed in absolute as well as relative (percentage) terms and enabling a direct comparison of the two evaluation criteria. This dual-metric scheme highlights how the choice of performance metric influences model selection while demonstrating the capability of Bayesian Optimization to discover high performance hyperparameter settings efficiently. Consequently, optimal hyperparameter sets are identified for each algorithm, and the overall predictive accuracy is maximized.

Table 1 lists the hyperparameter settings that achieved the best performance for XGBoost in the context in this study. The table shows the configurations obtained when the model is separately with respect to RMSE and RMSPE, indicating that these are the optimal XGBoost hyperparameters for this work.

Table 1. The optimal hyperparameters for XGBoost.

Model	Hyperparameter	Value	Hyperparameter	Value
	RMSE		RMSPE	
XGBoost	n Estimators	1414	n Estimators	1998
	Max Depth	12	Max Depth	12
	Learning Rate	0.0580	Learning Rate	0.03527
	Min Child Weight	3	Min Child Weight	9
	Subsample	0.7659	Subsample	0.7847
	Colsample Bytree	0.8824	Colsample Bytree	0.8670
	Reg Alpha	0.4147	Reg Alpha	3.1873
	Reg Lambda	38.0753	Reg Lambda	36.0197
	Gamma	0.9294	Gamma	0.8727
	Max Bin	249	Max Bin	-

Table **2** lists the hyperparameters of LightGBM that are obtained through the optimization process and represent the best-performing configuration for this study.

Table 2. The optimal hyperparameters for LightGBM.

Model	Hyperparameter	Value
	RMSE	
LightGBM	n Estimators	2465
	Max Depth	16
	Learning Rate	0.0589
	Num Leaves	305
	Min Child Samples	14
	Subsample	0.7956
	Colsample Bytree	0.9755
	Reg Alpha	0.7870
	Reg Lambda	44.2262
	Max Bin	265

Table 3 lists the hyper-parameters of TabNet that are obtained through the optimization process and represent the best-performing configuration for this study.

Table 3. The optimal hyperparameters for TabNet.

Model	Hyperparameter	Value
	RMSE	
TabNet	n d	36
	n a	26
	n Steps	4
	Gamma	1.4718
	Lambda Sparse	0.0020
	Batch Size	256
	Virtual Batch Size	128

3. Results

The predictive performance is assessed on the forecasts of 20 modules, which together comprise 6400 individual data points. For the outlier detection analysis, the 12mV tolerance threshold is defined based on domain expertise provided by the RF system designers, who identified this value as the critical limit for ensuring stable module operation; any prediction deviating from the reference value by more than this limit is classified as an outlier.

Four regression models are trained for the forecasting task, two XGBoost models are built, each with its hyperparameters tuned separately for the RMSE and RMSPE loss functions. The LightGBM and TabNet models are tuned with respect to RMSE. All hyperparameter searches are carried out using Bayesian Optimization to ensure systematic and reproducible selection of the optimal configurations.

As illustrated in Table 4, the four models are compared across three error metrics MAPE, RMSE and RMSPE. In this evaluation the LightGBM model achieved the best overall performance, followed by XGBoost model whose hyperparameters are optimized for RMSPE. The TabNet model exhibited the poorest performance among the four approaches.

Table 4. Model performance evaluation.

	RMSE (mV)	MAPE (%)	RMSPE (%)
XGBoost-RMSE	13.152	0.605	0.761
XGBoost-RMSPE	13.151	0.598	0.754
LightGBM	12.292	0.570	0.714
TabNet	15.742	0.712	0.925

Figure 1 and Table 5, the XGBoost model optimized for RMSPE identified 2,215 outliers and had a maximum deviation of 48.52 mV, whereas the model optimized for RMSE detected 2,254 outliers with a maximum deviation of 43.47 mV; thus, the RMSE-optimized model exhibits a lower maximum deviation. The RMSE-optimized XGBoost model is less effective at limiting the total number of outliers and detects more outliers with larger deviations. In contrast, the RMSPE-optimized model finds fewer outliers, and the deviations of those outliers are smaller.

Table 5. Outlier count of models.

	Outlier Count
XGBoost-RMSE	2254
XGBoost-RMSPE	2215
LightGBM	2124
TabNet	2704

LightGBM and XGBoost exhibit divergent performance depending on the evaluation metric. As illustrated in Table 4 the LightGBM attains lower overall error metrics (e.g., MAPE, RMSE, RMSPE) than XGBoost model whose hyperparameters are tuned for RMSE. However, when the outlier detection results shown in Figure 1 and Table 5 are considered, LightGBM identifies 2,124 outlier fewer than the 2,254 detected by the RMSE-optimized XGBoost and its mean absolute error is 18.32 mV versus 19.45 mV for XGBoost. However, LightGBM maximum deviation is 50.25 mV, which exceeds the 43.47 mV maximum deviation of the RMSE-optimized XGBoost. Consequently, LightGBM is superior in outlier count and mean deviation, whereas XGBoost-RMSE performs better in terms of maximum deviation; the preferred model thus depends on the primary evaluation criterion.

In this study, in addition to the gradient boosting algorithms, the deep-learning-based TabNet model is also evaluated and its results are obtained. As shown in Figure 1 and Table 5, TabNet produced the highest number of outliers among the tested models and are not able to achieve a performance comparable to that of the tree-based gradient boosting algorithms.

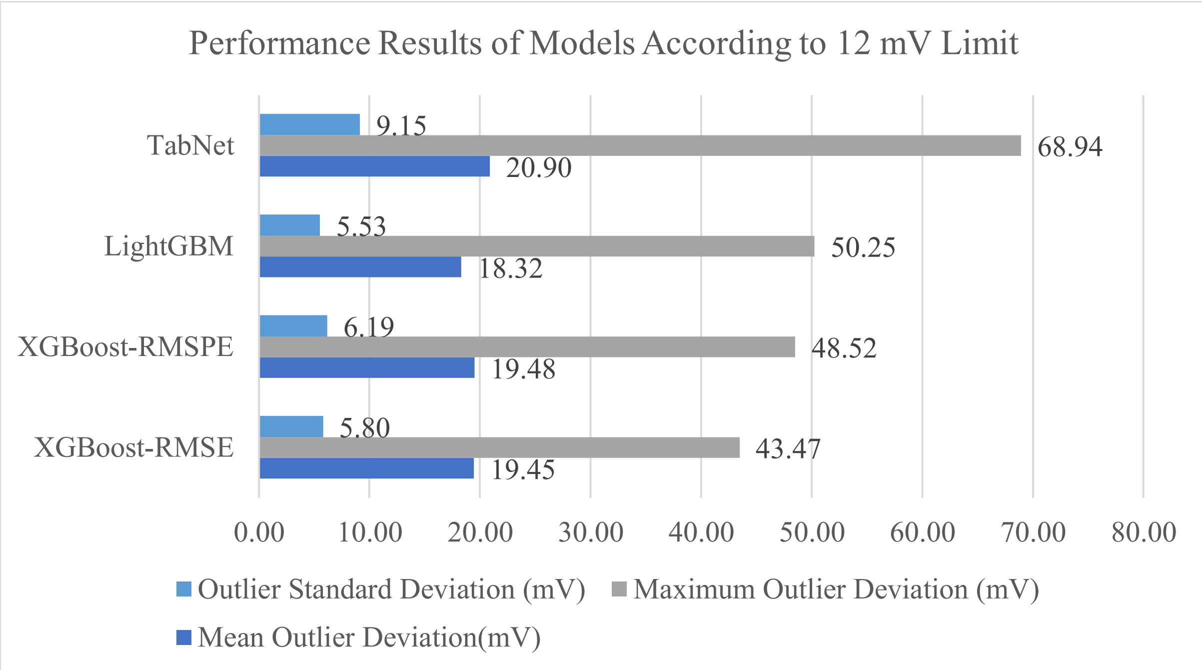

Figure 1. Results of models according to 12 mV limit.

Table 6. Bootstrap-based performance metrics with%95 confidence intervals (B=5000).

Model	Success Rate (%)	RMSE (mV)	MAPE (%)	RMSPE (%)
XGBoost-RMSE	64.77 (63.63-65.90)	13.13 (12.93-13.34)	0.60 (0.60 -0.62)	0.76 (0.74-0.78)
XGBoost-RMSPE	65.38 (64.25-66.51)	13.15 (12.92-13.38)	0.60 (0.59-0.61)	0.75 (0.74-0.77)
LightGBM	66.80 (65.66-67.95)	12.29 (12.08-12.50)	0.57 (0.59-0.58)	0.71 (0.70-0.72)
TabNet	57.74 (56.52-58.95)	15.74 (15.40-16.08)	0.71 (0.56-0.58)	0.92 (0.90-0.94)

Table 6 shows that the three gradient-boosting models (XGBoost-RMSE, XGBoost-RMSPE, and LightGBM) consistently outperform TabNet. Their success rates are higher (64.8%–66.8% versus 57.7%), their RMSE values are lower (12.3–13.2 mV versus 15.7 mV), and their percentage errors are smaller (MAPE ≤ 0.60%, RMSPE ≤ 0.76%). The near-identical RMSE values between XGBoost-RMSE and XGBoost-RMSPE (13.15 mV for both) stem from the fact that the target variable F lies in the range of approximately 1200–1500 mV, while the prediction errors are around 13 mV. At this scale, percentage errors (~0.6–0.8%) are nearly proportional to absolute errors, so optimizing for RMSE and optimizing for RMSPE yield very similar models and performance. Nevertheless, XGBoost-RMSPE achieves a slightly higher success rate (65.38% versus 64.77%) and a marginally lower RMSPE (0.75% versus 0.76%), confirming that the relative-error objective provides a small but measurable advantage. Moreover, the 95% confidence intervals for these metrics are narrow (e.g., success-rate width ≈ 2 pp, RMSE width ≈ 0.4 mV), indicating that performance is stable across different test splits. In contrast, TabNet has wider intervals and larger errors, confirming that tree-based boosting methods provide both greater accuracy and higher reliability for this regression task.

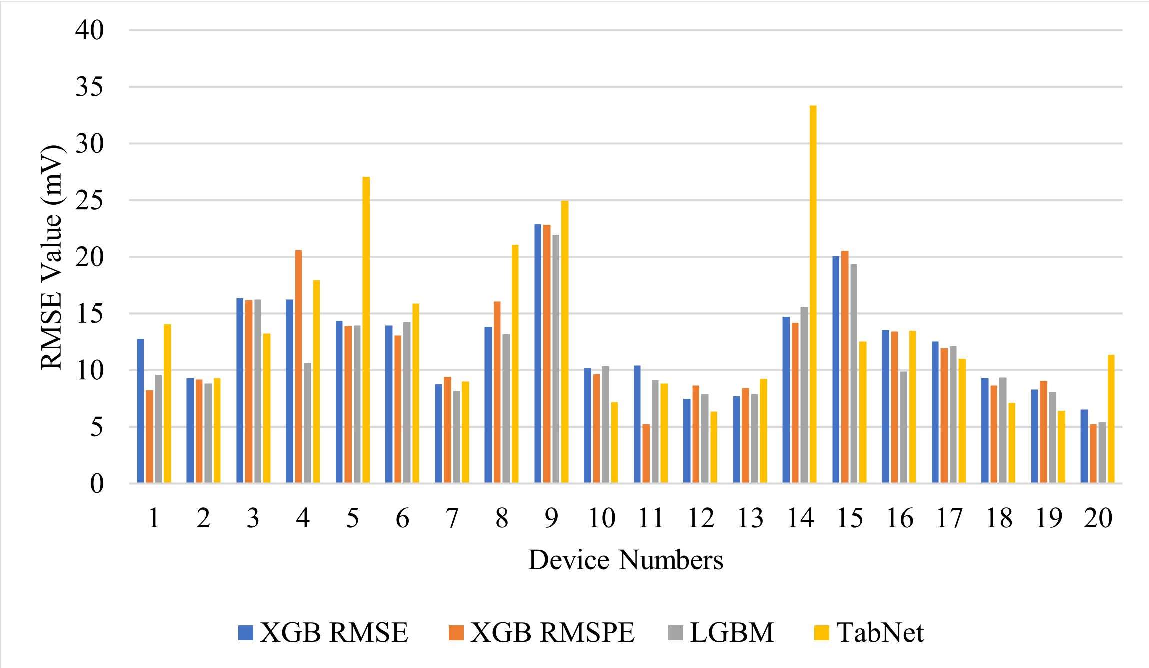

Figure 2. Per-Module RMSE Comparison Across Four Models for the 20 Test Modules.

The per-module RMSE values in Figure 2 demonstrate that the LightGBM ensemble provides the most consistently low error across the 20 test modules; it achieves the smallest RMSE for 7 of the 20 units and its overall mean RMSE (≈ 9.6 mV) is lower than those of the two XGBoost variants (mean ≈ 12.0 mV) and TabNet (mean ≈ 14.8 mV). XGBoost-RMSE and XGBoost-RMSPE show comparable average performance but exhibit larger variability, with several modules (e.g. 1, 8, 9, 15) producing RMSEs above 20 mV, whereas LightGBM never exceeds 22 mV. TabNet attains the highest RMSE on several modules (14) and yields the greatest spread of errors, confirming its inferior predictive accuracy for this dataset. Consequently, the table indicates that gradient-boosting models particularly LightGBM are more accurate and more robust to module-to-module variations than the TabNet architecture.

The per module MAPE values in Figure 3 show that the LightGBM model consistently delivers the lowest errors for the majority of the 20 test modules (13 out of 20), attaining a mean MAPE of ≈ 0.49%, which is lower than the averages of XGBoost-RMSE (≈ 0.62%), XGBoost-RMSPE (≈ 0.66%) and TabNet (≈ 0.73%). XGBoost-RMSE and XGBoost-RMSPE produce comparable performance, but their errors are more dispersed, with several modules (9, 15) exceeding 1% MAPE. TabNet exhibits the poorest consistency, recording the highest MAPE on several modules (5 and 14) while only marginally outperforming the boosting models on a few easy cases (7179 and 7185). These observations indicate that tree-based gradient-boosting ensembles particularly LightGBM provide more accurate and robust percentage-error predictions across heterogeneous modules than the TabNet architecture.

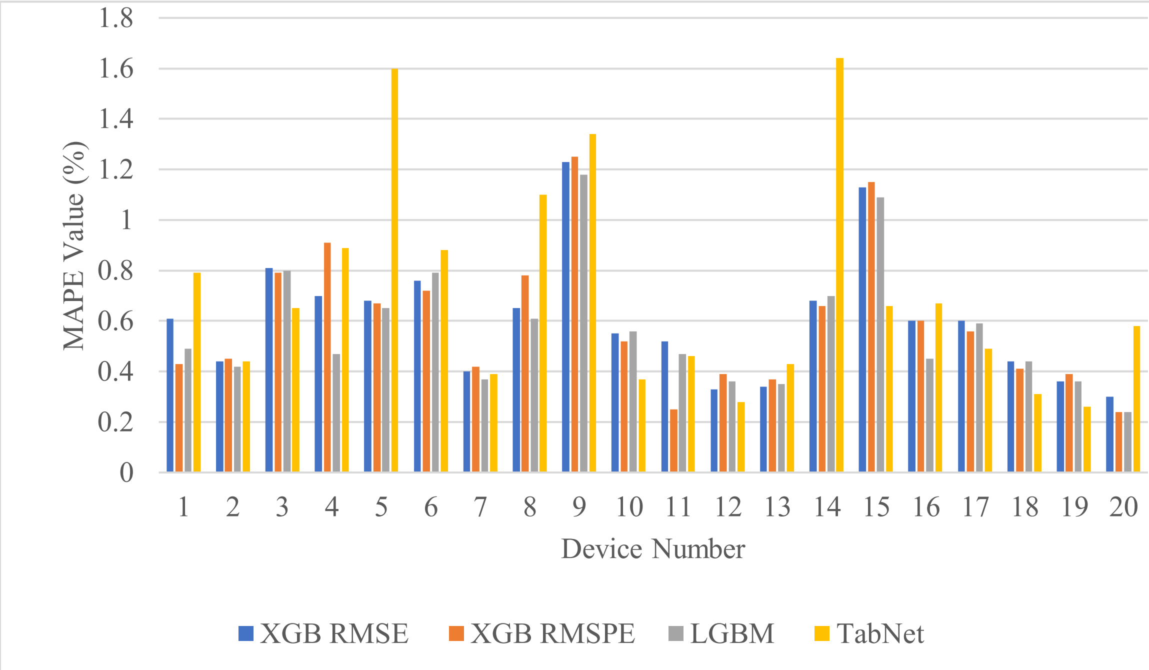

Figure 3. Per-Module MAPE Comparison Across Four Models for the 20 Test Modules.

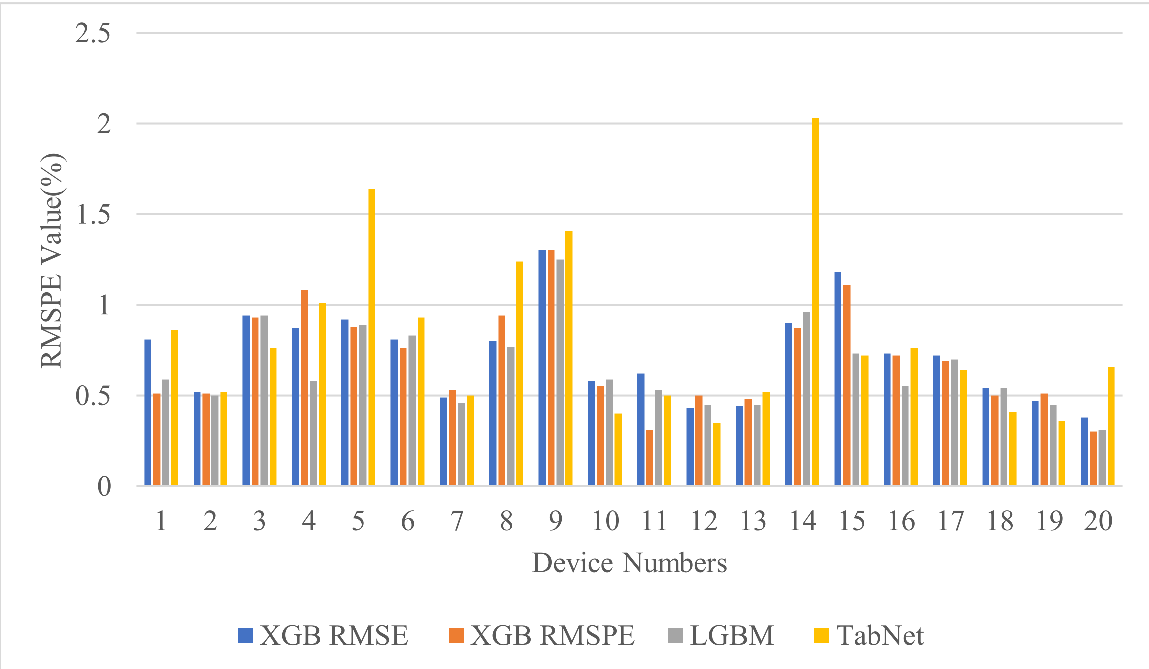

Figure 4. Per-Module RMSPE Comparison Across Four Models for the 20 Test Modules.

The per-module RMSPE values in Figure 4 demonstrate that the LightGBM ensemble yields the most accurate and stable percentage-error performance across the 20 test units. LightGBM attains the lowest RMSPE on 13 of the 20 modules, with a mean RMSPE of ≈ 0.65%, whereas the two XGBoost variants achieve comparable average errors (≈ 0.72 % for XGB-RMSE and ≈ 0.73% for XGB-RMSPE) but display a broader spread, reaching values above 1 % on several modules (6 and 15). TabNet exhibits the poorest overall consistency, presenting the highest RMSPE on multiple hard cases (14 and 5) and a mean RMSPE of ≈ 0.%. These observations confirm that the tree-based gradient-boosting approach, particularly LightGBM, provides superior and more reliable RMSPE predictions compared with both XGBoost configurations and the TabNet architecture.

The predictive performance is evaluated on the forecasts of twenty modules, corresponding to a total of 6400 measurement points. Four regression models XGBoost optimized for RMSE, XGBoost optimized for RMSPE, LightGBM and TabNet are assessed on outlier metrics (outlier count, outlier percentage, and mean absolute error). The LightGBM model achieved the lowest outlier occurrence (outlier count = 2124, outlier percentage = 33.19%, mean absolute error = 18.32 mV), followed by the XGBoost model optimized for RMSPE (outlier count = 2215, outlier percentage = 34.61%, mean absolute error = 19.48 mV). The XGBoost model tuned for RMSE yielded outlier count = 2254, outlier percentage = 35.22%, mean absolute error = 19.45 mV, whereas TabNet performed the worst (outlier count = 2704, outlier percentage = 42.25%, mean absolute error = 20.90 mV).

Outlier detection is carried out over the 6400 measurement points. The number of outliers identified by each model is 2124 for LightGBM (33.19%), 2215 for XGBoost–RMSPE (34.61%), 2254 for XGBoost–RMSE (35.22%) and 2704 for TabNet (42.25%). Accordingly, the proportion of predictions flagged as outliers is lowest for LightGBM, followed by the XGBoost configurations, and highest for TabNet. Among the outliers, LightGBM exhibits the lowest mean absolute error (18.32 mV) and the tightest dispersion (standard deviation = 5.53 mV), whereas XGBoost–RMSE and XGBoost–RMSPE record comparable mean values (19.45 mV and 19.48 mV, respectively) with standard deviations of 5.80 mV and 6.19 mV; TabNet, meanwhile, shows the highest mean absolute error (20.90 mV), the largest maximum deviation (68.94 mV) and the widest spread (standard deviation = 9.15 mV). These results indicate that LightGBM provides the most reliable performance overall, achieving the fewest outliers and the smallest error magnitude among them. The XGBoost models also deliver strong control, with the RMSPE-optimized variant incurring a slightly lower outlier count than the RMSE-optimized one, while the RMSE-optimized model records a marginally lower mean absolute error and maximum deviation on the outlier subset. In contrast, TabNet exhibits a markedly larger outlier count and higher error statistics, suggesting the deep-learning-based tabular approach is less suitable for this dataset.

The study therefore demonstrates that, for the examined tabular forecasting problem, gradient-boosting methods (particularly LightGBM and the XGBoost variants) outperform the deep-learning-based TabNet both in terms of outlier frequency and outlier magnitude reduction. Moreover, among the XGBoost models, hyperparameter tuning with RMSPE as the objective yields a slightly lower outlier percentage than tuning with RMSE, whereas the

RMSE-optimized model achieves a marginally better mean absolute error on the identified outliers. These findings support the preference for gradient boosting algorithms when addressing tabular prediction tasks in this domain, where interpretability, training efficiency and outlier control are jointly critical.

4. Discussion

The results show that the proposed ML-based approach can predict RF calibration parameters with a notable degree of accuracy. Within a ±12 mV tolerance envelope, 66.80% of the predicted values agree with the reference measurements. This performance is attributable to the selected features, which directly capture the operating dynamics of the module. Importantly, the ML predictor achieves this agreement using calibration and module-test data that are substantially less extensive than the measurement sequences required by conventional calibration procedures. The 17 h manual calibration routine is reduced to a matter of minutes, yielding measurable savings in labour and turnaround time.

Prior work in RF PA design has similarly highlighted the time-intensive nature of conventional measurement-based calibration. Chang et al. (2025) report that traditional measurement-weighted methods for PA efficiency and linearity characterization, while accurate, require lengthy measurement sequences. Sriram and Tuffy (2025) address a related challenge in PA design, employing a CatBoost model with Max-Min Latin Hypercube Sampling to reduce simulation requirements by 65% and achieve ±0.4 dBm prediction accuracy. While the tasks and metrics differ from those in the cited work, the present study shares the same motivation: replacing time-consuming measurement campaigns with data-driven prediction.

The results reveal a clear performance–complexity trade-off among the evaluated models. TabNet, despite its greater architectural complexity, yielded the weakest predictive performance and the highest outlier count, indicating that deep learning complexity does not confer an advantage for this particular tabular dataset. LightGBM and the RMSE-optimized XGBoost, by contrast, delivered superior accuracy with lower computational demands, consistent with the broader literature on gradient boosting versus deep learning for tabular data. A notable limitation is that model performance may degrade when applied to production batches with module characteristics substantially different from those in the training set. This generalizability concern underscores the need for future work to evaluate the models on larger and more diverse production datasets.

5. Conclusion

In this work a data-driven calibration framework is presented in which the raw calibration measurements and module-test results are directly employed as inputs to a ML predictor. This approach achieved a prediction success rate of 66.80% with an error bounded to ±12 mV, demonstrating that the extracted features capture the intrinsic behaviour of the RF PA without the need for extensive additional measurement. By replacing the traditional 17 h manual calibration routine with an automated process that completes within a few minutes, the method substantially reduces module related expenses, labour effort and overall turnaround time. The outcome proves that high-accuracy, rapid calibration is feasible for rugged military communication equipment operating under demanding environmental constraints.

In this study, four regression approaches are compared for tabular RF calibration forecasting and found that gradient boosting methods clearly outperform the deep-learning-based TabNet. Among the evaluated models, LightGBM achieved the most reliable outlier performance, while the XGBoost variant tuned for RMSE provided a robust overall error profile. Consequently, gradient boosting algorithms are the preferred solution for this type of task and future research should investigate ensemble techniques that combine the algorithms complementary strengths.

While the current results are promising, several approaches are able to be explored to further improve the solution. Expanding the training database with more data and diverse ML algorithms will increase the model's success. More characteristic feature extraction from the module-test data is able to provide richer representations and improve prediction accuracy, especially in tight tolerance windows. To address the slight performance degradation observed at ultra-low tolerances, alternative learning paradigms such as deep neural networks, ensemble methods and hybrid models are able to be investigated. Finally, integrating an online adaptation mechanism that continuously updates the predictor with field-collected calibration results will enable long term performance monitoring and further reduce maintenance costs of deployed systems.

Conflict of Interest

The authors declare no conflicts of interest.

Funding

No funding has been received for this research.

Acknowledgment

The authors gratefully acknowledge ASELSAN A.Ş. for providing the computational resources and data used in this study, as well as for its support during the experimental phase. Also, the authors would like to thank to all individuals who contributed to the execution of this study and provided technical assistance during the measurement and testing procedures particularly the ASELSAN Communication and Information Technologies Business Sector, Department of Hardware Test and Verification Engineering.

Ethical Statements

The study is performed solely on electronic hardware and software; no ethical committee approval is required.

Data and Code Availability

The calibration data, feature definitions and software used in this work are not publicly available. The all data is provided by ASELSAN Communication and Information Technologies Business Sector.

Supplementary Materials

Not applicable.

References

Arik, S. O., & Pfister, T. (2019). TabNet: Attentive interpretable tabular learning. arXiv preprint arXiv:1908.07442.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8(1), 69–80.

Cai, J., King, J. B., Yu, C., Pan, B., Sun, L., & Liu, J. (2020). Dynamic behavioural modelling of RF power amplifiers based on decomposed piecewise machine learning technique. *International Journal of Microwave and Wireless Technologies*, 1-7. https://doi.org/10.1017/S1759078720001208

Chang, Y., Wen, Y., Hong, Z., Padmanabhan, B., Franzon, P. D., & Floyd, B. A. (2025). A 27–30 GHz T/R module with reflection-type phase shifting and machine-learned calibration. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 72(9), 4649-4660. https://doi.org/10.1109/TCSI.2024.3524280

Chai, T., & Draxler, R. R. (2014). Root Mean Square Error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247–1250.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

De Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing*, *192*, 38-48.

Dikmese, S., Anttila, L., Campo, P. P., Valkama, M., & Renfors, M. (2019). Behavioral modeling of power amplifiers with modern machine learning techniques. In *2019 IEEE MTT-S International Microwave Conference on Hardware and Systems for 5G and*

Beyond (IMC-5G) (pp. 1-3). IEEE. https://doi.org/10.1109/IMC-5G47857.2019.9160381

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Ortega-González, F. J., García, J. A., Patiño-Gómez, M., & Madueño-Pulido, D. (2025). High-Efficiency Transmitters Operating at HF, VHF, and UHF Frequencies. *IEEE Microwave Magazine*. https://doi.org/10.1109/MMM.2025.3592159

Sriram, A., & Tuffy, N. (2025). Accelerating RF power amplifier design via intelligent sampling and ML-based parameter tuning. arXiv. https://arxiv.org/abs/2507.11928

Tong, Y., Wang, J., Fu, T., Su, J., & Liu, J. (2024). Measurement-Driven Automated Design of High-Efficiency Power Amplifiers. In *2024 9th International Conference on Integrated Circuits and Microsystems (ICICM)* (pp. 129-133). IEEE. https://doi.org/10.1109/ICICM63644.2024.10814353

Uğur, A. B. (2023). *Derin öğrenme tabanlı güç yükseltici modülü kalibrasyonu* (Master's thesis, Gazi University).

U.S. Department of Defense. (2019). *MIL-STD-810H: Environmental engineering considerations and laboratory tests*. http://everyspec.com/MIL-STD/MIL-STD-0800-0899/MIL-STD-810H_55998/

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*(1), 79-82.