IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 June 2025, accepted 11 July 2025, date of publication 16 July 2025, date of current version 22 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3589947

RESEARCH ARTICLE

Ensemble Methods for Peristaltic Pump Accuracy Enhancement

DAVIDE PRIVITERA[1,2], **ALESSANDRO MECOCCI**[1], **AND SANDRO BARTOLINI**[1]
[1]Department of Information Engineering and Mathematics, University of Siena, 53100 Siena, Italy
[2]Pharma Integration, 53100 Siena, Italy
Corresponding author: Davide Privitera (privitera@diism.unisi.it)

ABSTRACT This study investigates how ensemble learning techniques can be employed for enhancing peristaltic pump accuracy in pharmaceutical manufacturing, and demonstrates significant accuracy improvements through the novel E-AR implementation, with gains of up to 53.93% at 0.3 ml volume compared to 47% achievable with single models. To establish the foundation for ensemble methods evaluation, we first conduct a comprehensive validation of traditional Adaptive Dosing Control System (ADCS) across an extended volume range (0.1-2.0 ml), demonstrating base performance improvements. In this investigation, we develop a novel offline performance indicator enabling rapid assessment of compensation strategies without extensive physical testing, showing strong correlation with actual measurements. These premises enable a thorough investigation of various ensemble configurations, revealing volume-dependent performance patterns where different models excel under specific conditions, suggesting that practical applications may benefit from volume-specific model selection. The comparison with a very accurate reference mechanical pump, demonstrates that our ADCS solutions achieve comparable or superior performance across most volumes while maintaining the cost-effectiveness. Statistical validation via a multi-dimensional framework confirms the significance of these improvements through multiple complementary tests: paired t-tests showing significant mean differences with $p \leq 0.001$, Mann-Whitney U tests confirming distributional shifts, Levene tests demonstrating variance modifications with statistics up to 801.65, and mixed linear model analysis with F-statistics ranging from 0.004 to 1497.75 confirming global effects.

INDEX TERMS Adaptive control, dosing accuracy, ensemble, peristaltic pump.

I. INTRODUCTION AND MOTIVATION

Accuracy in drug dosing is critical in the pharmaceutical industry for patient safety, treatment efficacy, and economic results. The U.S. Food and Drug Administration (FDA) has consistently emphasized that even small dosing errors can pose major risks to patient health, particularly with anti-cancer drugs [1], [2].

Beyond patient risk, contamination and cross-contamination represent significant challenges in pharmaceutical manufacturing. Traditional dosing devices like volumetric pumps, while accurate, create contamination risks due to direct product contact with difficult-to-clean mechanical parts.

FDA recall data demonstrates contamination's criticality, with sterility-related issues representing the highest category of pharmaceutical recalls [3]. Contamination-related batch recalls create both health threats and substantial economic damage, as evidenced by major facility closures and regulatory reforms following contamination incidents [4].

Economic implications extend beyond contamination risk. Biopharmaceutical production consumes significant portions of R&D budgets from pre-clinical trials to approval [5]. Given low clinical success rates, improved dosing quality offers substantial economic benefits by minimizing waste and batch rejections.

The associate editor coordinating the review of this manuscript and approving it for publication was Jinquan Xu.

Peristaltic pumps (PP) have gained importance in addressing these challenges. Their operating principle — compressing sterile, disposable tubing with rotating rollers — ensures fluid only contacts the tubing's interior, significantly reducing cross-contamination risk.

PP versatility is another advantage. By adjusting tube size and parameters, these systems handle volumes from 0.1 to 250 ml, suiting diverse biopharmaceutical production needs from high-potency drugs to larger infusions [6].

Despite contamination control and versatility advantages, these systems struggle with optimal accuracy, particularly at low volumes [6], [7], [8]. The same mechanism ensuring contamination resistance — single-use sterile tubing — introduces dosing variability through tube wear and temperature fluctuations.

Cost implications of accuracy improvements are substantial. For high-cost pharmaceuticals like Zolgensma ($1.9 million per dose), even modest 1% improvements in dosing accuracy could yield millions in annual savings [9], while increasing production capacity without additional capital investment.

Traditional PP accuracy improvements have focused on mechanical advancements — more precise roller mechanisms, better tubing materials, and sophisticated designs. While yielding some progress, these approaches often involve high costs with limited effectiveness, especially across diverse pharmaceutical production requirements [6], [8], [10].

Recent research has shifted toward software-based compensation techniques, demonstrating significant performance improvements across various industrial domains [11], [12], [13]. These advancements are emerging in pharmaceutical applications as well, where regulatory compliance is facilitated by their seamless integration with established validation frameworks, while patient safety is maintained through existing In-Process Control systems that implement 100% weight verification. A notable example of these systems is the ARIMA-based ADCS [14] which achieved up to 30% accuracy improvement for 1.2 ml volumes relying on continuous model retraining during operation — an online training paradigm essential for adapting to dynamic pump behavior. Since its linear nature limits complex pattern capture, with degraded performance at lower volumes, subsequent research explored machine learning methods: specifically GRU and LSTM neural networks [15]. This selection over recent approaches was driven by empirical evidence and industrial constraints, particularly limited training data availability. While Transformer models show remarkable capabilities they may not offer inherent advantages in this domain [16], [17] and their substantial data requirements exceed practical industrial dosing limitations [15] particularly since pharmaceutical manufacturers typically restrict production data sharing due to regulatory and competitive concerns. GRU networks, with simpler architecture, achieve good performance with less training data, making them better suited for runtime applications [18], [19]. This selection was also driven by computational efficiency considerations critical for industrial deployments. While Transformer architectures require significant computational resources due to their self-attention mechanisms, GRU and LSTM models offer superior inference speed with minimal hardware requirements allowing continuous model retraining during operation, which would be impractical with computationally intensive models like Transformers.

Though AI-powered methods demonstrated improved pattern recognition, they face limitations: computational intensity and inconsistent performance across volume ranges [15].

Ensemble techniques offer promise in addressing these challenges, improving accuracy, robustness, and complex pattern handling while adapting to changing conditions [20], [21]. In PP contexts, where mechanical components, fluid dynamics, and environmental factors create complex interactions, ensemble methods potentially overcome individual model limitations by combining their strengths, generating more robust results than any single model could provide.

While ensemble methods show significant potential, their practical implementation requires efficient evaluation strategies. To address this need, this study introduces a novel offline performance indicator correlating offline prediction metrics with online compensation effectiveness. This indicator enables rapid evaluation without extensive physical testing, reducing resource requirements while maintaining predictive accuracy within ±10% across diverse conditions. Validated across multiple architectures from statistical (AR) to neural networks (GRU), this advancement addresses a critical bottleneck in compensation model development, where traditional evaluation requires time-consuming physical testing cycles.

Finally, while previous ARIMA and GRU model results for 0.3 ml and 1.2 ml volumes are promising, a comprehensive analysis across a broader range is needed to validate these techniques across modern pharmaceutical manufacturing requirements.

In summary, this study addresses these shortcomings and advances PP accuracy through three contributions:

1. An in-depth analysis of ensemble techniques, combining different models' strengths to improve accuracy and robustness.
2. A new offline indicator allowing a priori assessment of compensation model performance, simplifying development and selection of dosing control strategies.
3. A comprehensive validation of advanced models across 0.1-2.0 ml volumes, including direct comparison with modern mechanical pumps, revealing comparable or superior performance from ADCS-enhanced standard pumps.

This study provides understanding of software-based PP compensation techniques with significant impact on pharmaceutical manufacturing, improving drug production safety and efficiency. As industry demands for precision and reliability increase, these advances represent a key step

TABLE 1. Tubing configurations for different target volumes.

Target Volume (ml)	Nozzle i.d. (mm)	Tubing i.d. (mm)	Y-connector o.d. (mm)
0.1 0.2 0.3	0.6	0.5	2.4
1.2 2.0	1.6	1.2	3.6

toward addressing these challenges and improving patient outcomes.

II. SETUP AND METHODOLOGY

This section outlines our experimental methodology for studying ensemble-based ADCS system for PP, detailing both hardware and software configurations, evaluation metrics, and data collection procedures. Our methodology builds on that introduced in previous works [14], [15] and extends the analysis to a wider range of volumes, introducing new techniques. The following subsections offer, among other things, insights into experimental design, ensuring the reproducibility and validity of our results.

A. HARDWARE AND SOFTWARE SETUP

The experimental setup replicates pharmaceutical filling processes using standard industrial components. This choice ensures real-world relevance and was conducted at a company specializing in robotic filling systems for pharmaceutical applications. The workstation maintains industrial-grade precision across various filling volumes.

Central to our system is a PP Flexicon PD12 system with MC100 control unit (firmware S2) (Watson-Marlow Flexicon, Ringsted, Denmark), selected for its flexibility and wide dosing range. The dosing circuit uses Flexicon Accusil^{TM} pump tubing connected with Y-connectors before and after the pump, with additional tubing connecting to a liquid reservoir and nozzle. The tubing length (approximately two meters) remained constant across all setups (Figure 1). Different configurations of nozzle inner diameter (i.d.), tubing i.d., and Y-connector outer diameter (o.d.) accommodated various dosing volumes (Table 1). These configurations follow the pump manufacturer's specifications for efficient volume handling [22].

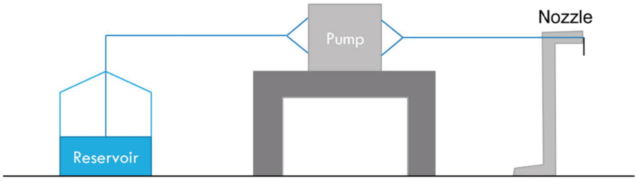


FIGURE 1. Pump tubing setup.

We used a high-precision Wipotec SL-M 250/300 scale (Wipotec, Kaiserslautern, Germany) with Active Vibration Compensation for precise volume measurements even in dynamic industrial environments.

All tests used purified water processed through a Milli-Q® Advantage A10 with Millipak® Express 40 filter, ensuring consistent liquid quality and reducing impurity-related variations.

Before each experiment, we implemented a standardized initialization composed by a purging session consisting of 30 seconds of continuous pump operation to remove air bubbles from the dosing circuit, followed by a calibration for specific target volumes using the MC100 controller's built-in function, ensuring consistent baseline conditions for all experimental runs. This calibration process automatically sets optimal pump parameters based on weight measurements.

Container handling is performed by a Denso VS-050S2 robot (Denso, Kariya, Aichi, Japan), selected for its precision and pharmaceutical compliance, equipped with a Gimatic MPXM gripper (Gimatic, Brescia, Italy).

A Siemens ET 200SP Open Controller (Siemens, Munich, Germany) serves as both PLC and Windows PC-based system, integrating control logic with data processing capabilities.

A key design aspect was treating the dosing system as a closed box by adhering to the pump controller's API interfaces [23], ensuring data and controls remained within operational limits while improving result applicability to different dosing systems.

The software architecture consists of two major modules: a PLC which controls physical components (PP, scale, robot, gripper) and a PC which runs Python applications for prediction, data analysis, and ADCS algorithms, communicating via OPC-UA protocol for robust data exchange.

For model and ensemble implementations, we used:

1. AR models: Statsmodels package [24] for time series analysis.
2. GRU models: Keras package [25] for neural network construction and training.
3. Ensemble models:

- XGBoost: For gradient boosting ensembles [26].
- Scikit-learn's ensemble module: For StackingRegressor implementation [27].

Computation ran on an Intel Xeon Gold 6226 CPU with VGPU T4-2Q for accelerated neural network processing.

This setup allows realistic simulation of filling processes and testing of various ADCS strategies across a wide range of dosing volumes.

B. EXPERIMENTAL PROTOCOL

The experimental protocol consists of the execution of various subsets of 300 fillings based on previous studies, as shown in Figure 2. In this way, we could test the performance of the ADCS under different operational conditions, simulating realistic scenarios in pharmaceutical production.

At the beginning of every subset, the pump is calibrated to ensure the same starting conditions. This is crucial because the calibration process can significantly alter the operating states of the PP [14]. As shown Figure 3 each subset is divided into two distinct phases:

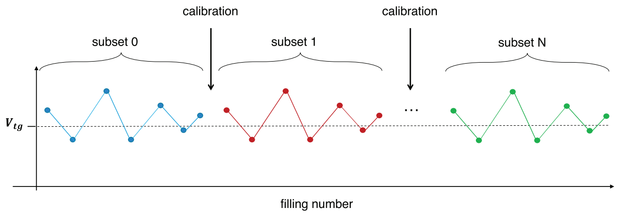


FIGURE 2. **Schematic representation of the experimental filling protocol. Multiple runs (subset 0 to subset N) are performed, each comprising 300 fillings, where weight measurements (colored dots) fluctuate around a target value (V_{tg}, dashed line). Pump calibration procedures (arrows) are implemented between runs to neutralize temperature-related effects and reveal diverse behavioral patterns.**

1. **ADCS OFF**: The first 100 fills are performed without any compensation in order to: a) provide a baseline for pump performance under standard operating conditions and b) generate initial training data for our predictive models. During this phase, the system continuously monitors dosing patterns to initialize the model parameters.
2. **ADCS ON**: The next 200 fills are conducted with ADCS active. This extended phase allows us to: a) evaluate the effectiveness of the compensation system over an extended period, b) observe any long-term trends or adaptations in the system's behavior, and c) study the behavior of the algorithm as it processes its compensated data since, throughout this phase, the system maintains a sliding window of recent measurements, continuously updating model predictions and compensation values to optimize dosing accuracy.

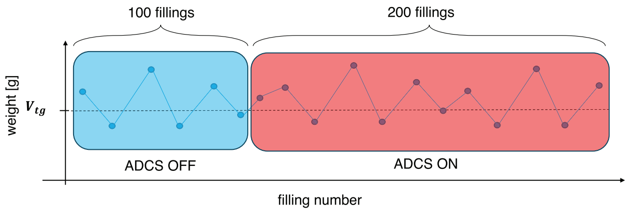


FIGURE 3. **Detailed structure of the experimental protocol phases. Each run consists of two distinct phases: an initial ADCS OFF phase operating without compensation to establish baseline performance and generate training data, followed by an ADCS ON phase where the compensation system is active. The dashed line represents the target weight, while dots indicate individual filling measurements. Pump calibration is performed at each run to ensure consistent initial conditions. This bi-phasic approach enables both system training and evaluation of the compensation algorithm's long-term performance under controlled conditions.**

More specifically, the ADCS implements a continuous predict-and-adjust cycle where the system monitors dosing patterns, predicts upcoming volumes based on historical data, and applies compensatory adjustments at runtime. More details are reported in Section II-D. This adaptive mechanism enables the system to respond dynamically to variations in dosing behavior without requiring prior knowledge of the underlying system model.

After systematic evaluation of various approaches, we prioritized AR and GRU models based on their demonstrated advantages in PP systems through analyses presented in previous works [14], [15]. AR models offer superior interpretability and computational efficiency - crucial requirements for runtime industrial applications where model behavior must be readily understood and validated. GRU networks demonstrate enhanced pattern recognition capabilities for complex behaviors, particularly excelling at micro-volume dosing where precision requirements are most stringent.

The selection of 100 uncompensated fillings as the training window aligns with the analysis presented in [14] and [15], where this configuration demonstrated optimal performance in balancing model convergence with computational overhead. Alternative window sizes were evaluated (50-300 fillings), but larger datasets showed diminishing returns in prediction accuracy while significantly increasing processing time, and smaller ones proved insufficient for reliable pattern recognition.

The experimental protocol's robustness against external factors is ensured through multiple design considerations. The periodic calibration every 300 doses effectively resets any accumulated environmental perturbations, particularly temperature-induced variations, thereby maintaining consistent baseline conditions throughout the experimental sequence. Furthermore, to eliminate any potential accuracy degradation due to tube wear, the complete hydraulic circuit was replaced for each test sequence.

The validation framework for methodological reproducibility incorporates validation of both the measurement system and dosing apparatus. The measurement system features a Wipotec SLM 250/300 high-precision balance characterized by the following specifications: display resolution of 0.002 g, linearity within ±0.001 g, repeatability of ±0.0005 g, and settling time below 120 ms. In accordance with GAMP guidelines, a three-point calibration protocol (0 g, 50 g, 100 g) using certified reference weights is performed prior to each experimental session to ensure measurement precision, with acceptance criteria of ±0.010 g to account for environmental vibrations and measurement noise. The system's reliability is enhanced through AVC technology, which employs a dedicated sensor working in conjunction with the weighing cell to provide real-time compensation for environmental vibrations. The sensors are integrated into a rigid support structure, maintaining a vertical separation of less than 100 mm to optimize vibration compensation while preventing mechanical resonance effects.

Finally for the PP system, as discussed in Section II-A, a standardized initialization procedure is implemented before each experimental session, comprising an initial purging operation to remove air from the tubing, followed by system calibration for the specific target volume using the MC100 controller's built-in function. This calibration process, managed through the pump controller's API, automatically calculates and implements optimal pump parameters based on weight measurements. The experimental validation protocol encompassed extensive testing, enabling statistical averaging of system behavior to mitigate the impact of stochastic variations and establish statistically significant performance metrics.

C. METRICS AND EVALUATION CRITERIA

The appropriate choice of performance metrics is crucial in evaluating ADCS for PP. These metrics should translate dosing system accuracy and precision, serving as capability assessments. Throughout our analysis, we work with three fundamental variables: the target volume (V_{tg}), which represents the desired dosage amount set for the PP; the actual dispensed volume (V_d), which measures what is physically delivered by the system; and the volume predicted by our models (V_p), which represents the system's estimation of the dispensed amount. Using these variables, we can formulate various metrics that provide comprehensive insights into the performance of automatic dosing systems. This section presents the leading metrics and evaluation criteria with their importance and mathematical formulations.

1) STANDARD DEVIATION

To assess system precision, we utilize the Standard Deviation (STD) metric measuring volume dispersion around their mean:

$$\mathrm{STD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(V_d - \mu_d)^2}$$

where μ_d represents the dispensed volumes mean and n refers to dataset sample number.

2) ROOT MEAN SQUARE ERROR

The Root Mean Square Error (RMSE) serves as a primary accuracy measure quantifying average prediction error magnitude, with larger errors penalized more heavily. We employ two distinct RMSE forms:

1. $\mathrm{RMSE}_{\mathrm{Standard}}$: Calculated between dispensed and predicted volumes:

$$\mathrm{RMSE}_{\mathrm{Standard}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(V_d - V_p)^2}$$

2. $RMSE_{PP}$: Modified version directly comparing dispensed to target volumes:

$$\mathrm{RMSE}_{\mathrm{PP}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(V_d - V_{tg})^2}$$

where n refers to dataset sample number.

Figure 4 illustrates $\mathrm{RMSE}_{\mathrm{PP}}$ utility: while traditional STD favors tight precision regardless of target proximity (Scenario A), $\mathrm{RMSE}_{\mathrm{PP}}$ prioritizes achieving correct dosage (Scenario B), making it more suitable for dosing applications where target accuracy outweighs precision around incorrect volumes.

Note that RMSE can represent either $\mathrm{RMSE}_{\mathrm{Standard}}$ or $\mathrm{RMSE}_{\mathrm{PP}}$ depending on context, providing flexibility for consistent metric application throughout analysis.

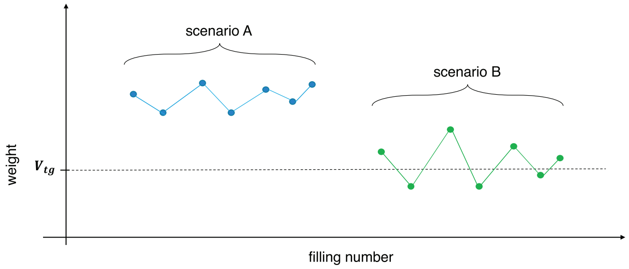


FIGURE 4. Comparison of filling scenarios illustrating difference between precision and accuracy in dosing systems. Scenario A shows high precision but low accuracy, with measurements consistently deviating from target weight (V_{tg}). Scenario B demonstrates lower precision but higher accuracy, with measurements scattered more widely but centered closer to target weight. This comparison highlights why *$RMSE_{PP}$* is more suitable for filling applications, accounting for both measurement dispersion and target proximity.

3) SUBSET-SPECIFIC METRICS

As detailed in Section II-B, our experimental protocol comprises several subsets (Figure 2), each with 300 fillings. For examining the performance across these subsets, we employ two key metrics: RMSE_i which represents the RMSE calculated for each i-th subset and the STD_i which is the STD calculated on each i-th subset. These subset-specific metrics track performance changes across experimental runs, identifying trends or anomalies.

4) AGGREGATE METRICS

For overall performance indication across all subsets, we define:

$$\overline{\mathrm{RMSE}} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{RMSE}_i$$

$$\overline{\mathrm{STD}} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{STD}_i$$

where n is total dataset subset number.

These metrics provide overall system performance view. $\overline{\mathrm{RMSE}}$ measures overall accuracy, while $\overline{\mathrm{STD}}$ indicates dosing system consistency across multiple subsets.

5) COMPENSATED VS. UNCOMPENSATED DATA

We distinguish compensated from uncompensated data using $\mathrm{ADCS}_{\mathrm{ON}}$ and $\mathrm{ADCS}_{\mathrm{OFF}}$ notations as shown in Figure 3, adding these designations to metrics for clear PP system operating mode identification.

For instance, $\overline{\mathrm{RMSE}}_{\mathrm{ADCS_{ON}}}$ denotes average RMSE_i values for compensated dataset subset, effectively evaluating aggregate performance with active compensation while distinguishing from uncompensated measurements.

6) ONLINE VS. OFFLINE ANALYSIS

Our research divides into *online* and *offline* analysis. Online analysis refers to runtime analysis and compensation during pump operation, using RMSE_{PP} to directly measure system ability to hit target volumes. Offline analysis uses previously collected datasets to evaluate models and strategies without continuous physical experiments, employing $\text{RMSE}_{\text{Standard}}$ to evaluate model predictive accuracy by comparing predictions (V_p) against actual dispensed volumes (V_d).

7) ACCURACY

We define relative accuracy measure as RMSE to target volume ratio, expressed in percentage:

$$\text{Acc} = \left(\frac{\overline{\text{RMSE}}}{V_{tg}}\right) \times 100$$

This allows normalized performance comparison across different target volumes. RMSE can be either $\overline{\text{RMSE}}_{\text{Standard}}$ or $\overline{\text{RMSE}}_{\text{PP}}$, depending on focus (prediction accuracy or overall dosing accuracy).

Although PP expresses target volumes in milliliters, we convert to milligrams (mg) using 1000 multiplication factor, based on water density relationship (1 ml equals 1000 mg at standard temperature and pressure).

8) GAIN PERCENTAGE

To quantify ADCS compensation, we use gain percentage, comparing compensated to uncompensated system performance:

$$\text{Gain}_{\text{Metric}} = \left(1 - \frac{\text{Metric}_{\text{ADCS}_{\text{ON}}}}{\text{Metric}_{\text{ADCS}_{\text{OFF}}}}\right) \times 100$$

where Metric can be various performance indicators depending on evaluation aspect.

Unlike other metrics, Gain percentage inherently includes ADCS operational states (ON/OFF) and is used without additional online/offline designations.

9) ERROR CUMULATIVE DISTRIBUTION ANALYSIS

To characterize system error distribution and compensation effectiveness, we employ Error Cumulative Function (ECF) analysis, representing empirical cumulative distribution of absolute dosing errors, providing detailed perspective on error behavior across observed deviations.

For dosing operations sequence, we compute absolute errors in mg:

$$e_i = |V_d^i - V_{tg}| \cdot 1000$$

where V_d^i is i-th dispensed volume in ml, with 1000 multiplication converting errors to mg. After computing errors, we sort in ascending order:

$$e_{(1)} \leq e_{(2)} \leq \cdots \leq e_{(n)}$$

ECF is defined directly in terms of ordered errors:

$$\text{ECF}(e_{(k)}) = \frac{k}{n} \quad \text{for } k = 1, 2, \ldots, n$$

where:

- $e_{(k)}$ is k-th ordered error value (k-th smallest error)
- k represents $e_{(k)}$ index in ordered sequence, indicating how many errors are less than or equal to $e_{(k)}$
- n is total sample number

If particular error value appears multiple times, ECF at that value corresponds to highest index among occurrences, ensuring ECF(e) correctly represents fraction of errors less than or equal to e.

This approach provides key analytical benefits over single-point metrics:

1) **Complete Distribution Analysis**: Unlike aggregate metrics, ECF reveals complete error distribution shape, showing not just magnitude but dosing deviation pattern.
2) **Comparative Performance Assessment**: By plotting ECF curves for compensated ($\text{ECF}_{\text{ADCS}_{\text{ON}}}$) and uncompensated ($\text{ECF}_{\text{ADCS}_{\text{OFF}}}$) operation, we can evaluate compensation effectiveness. Better performance shows steeper rising ECF curve reaching higher values at lower error thresholds, indicating greater small error proportion.

This direct curve comparison clearly shows compensation's effect on dosing error distribution, offering insights beyond simpler statistical measures.

10) STATISTICAL TESTS

To ensure rigorous validation of our results, we implement a statistical framework that allows for objective assessment of the ADCS performance. Our approach implements a two-phase validation methodology, specifically designed to first characterize the baseline behavior of the uncompensated system and then analyze the effectiveness of the compensation mechanism.

Initial phase establishes uncompensated measurement variability statistical significance through variance homogeneity analysis using two complementary tests:

1) **Levene Test (mean-centered)**: Evaluates variance equality across sequential measurements, assessing measurement stability. The test's sensitivity to mean-centered deviations suits detecting systematic precision variations.
2) **Brown-Forsythe Test (median-centered)**: Modified Levene test using median-centering, offering enhanced robustness against non-normality, crucial for validating potentially non-Gaussian measurement results.

These tests apply globally across all sequences and pairwise between consecutive subsets, enabling macro-level stability assessment and detailed sequential variation analysis.

Second phase implements four complementary statistical tests evaluating compensation impact:

1) **Paired t-test**: Evaluates the statistical significance of mean differences between corresponding uncompensated and compensated sequences. This parametric test compares the means of each subset pair, taking into account the intrinsic coupling of the

experimental design, allowing direct assessment of compensation-induced improvements while controlling for sequence-specific variations.

2) **Mann-Whitney U test**: Provides a non-parametric assessment of distributional differences, applied individually to each uncompensated-compensated subset pair. Results are combined through meta-analysis using Fisher's method, maintaining the hierarchical data structure and respecting observation independence. The percentage of subset pairs showing significant differences ($p < 0.05$) reveals how consistent this effect is across different calibration conditions.

3) **Levene test**: Examines variance homogeneity separately for each uncompensated-compensated subset pair, quantifying changes in process variability induced by compensation without violating independence assumptions. Meta-analysis of these paired comparisons, combined with the proportion of subset pairs showing significant variance differences, provides a detailed assessment of the compensation impact on dosing consistency across different operational conditions.

4) **Mixed linear model analysis**: Implements a hierarchical statistical model that explicitly accounts for the nested data structure, treating compensation (ON/OFF) as a fixed effect and subset identity as a random effect. This approach properly addresses within-subset and between-subset variability components, providing a robust evaluation of compensation effectiveness while controlling for calibration-induced variations across experimental sequences.

All tests use significance levels $\alpha = 0.05$. This multi-faceted approach was carefully selected to serve dual purposes: performance evaluation and data characterization. The ECF analysis provides direct visualization of error distributions and their characteristics, while the statistical framework effectively characterizes central tendencies, distributions, and variance structures in the data.

D. EXPERIMENTAL FRAMEWORK AND DATA ACQUISITION

This section provides an overview of our ADCS methodology, focusing specifically on the online training approach and data collection procedures. We detail the foundational experimental framework that underpins our ensemble-based compensation strategies, encompassing both the iterative learning mechanisms employed during runtime operation and the extensive experimental campaigns conducted across multiple volume ranges. The methodology presented herein establishes the baseline against which our ensemble approaches are subsequently evaluated.

1) ADCS IMPLEMENTATION: ONLINE TRAINING METHODOLOGY

In each subset collected with the methodology described in Section II-B, we apply a compensation method called Online Training which proceeds as follows:

1) **Initial Data Acquisition:** Collection of fixed amount of dispensed volumes without compensation (Training Window, TW = 100), corresponding to ADCS_{OFF} phase (Figure 5a).

2) **Model Training:** Training our predicting model using TW. The model can be any predictive algorithm, not restricted to AR model, GRU, or ensemble methods (Figure 5a).

3) **Prediction and Compensation:** The trained model predicts next filling volume given p previous observations (Figure 5b). Based on this prediction, we compute compensation volume using:

$$V = V_{tg} \cdot \frac{V_{tg}}{V_p}$$

This formula adjusts dispensed volume inversely proportional to predicted target deviation.

4) **Dispensing and Data Update:** We apply compensation to pump controller to adjust dispensed quantity and trigger filling (Figure 5c). After dispensing, we update our dataset by removing oldest value in TW and inserting new compensated value (Figure 5d).

5) **Iteration:** The model updates by repeating step 2 with new data. This iterative step corresponds to ADCS_{ON} phase and runs until acquiring desired filling number - 200 in our tests. Upon completion, calibration is performed, and entire cycle resets and repeats from step 1, starting new data collection under ADCS_{OFF}.

To simulate continuous container flow, we use the cyclic process presented in [14] where a single container is filled and emptied in loop. This includes placing container on weighing scale, recording gross weight, moving under needle for filling, commanding PP to dispense, returning to WS for net weight recording, and finally emptying container back into reservoir once it's full. This procedure simulates continuous production using one container instead of multiple.

2) DATASET COLLECTION ACROSS MULTIPLE VOLUMES

Building on previous discussions, we collected data across a wide dosing volume range to establish a foundation for ADCS performance evaluation. This extensive collection effort, comprising thousands of fillings per volume, provided both a solid basis for ensemble method research and insights into ADCS behavior across different volumes.

Our study covered five specific dosing volumes: 0.1 ml, 0.2 ml, 0.3 ml, 1.2 ml, and 2.0 ml. This range addresses micro-dosing challenges while remaining applicable to conventional pharmaceutical applications, expanding previous work [15] that examined only 0.3 ml and 1.2 ml.

Micro-volumes (0.1-0.3 ml) are critical for precise dosing applications where minor variations significantly impact efficacy or cost. Standard volumes (1.2-2.0 ml) correspond to typical pharmaceutical dosing quantities.

This range allows us to analyze system behavior, identify challenges at low volumes where precision is paramount, evaluate performance in standard pharmaceutical scenarios,

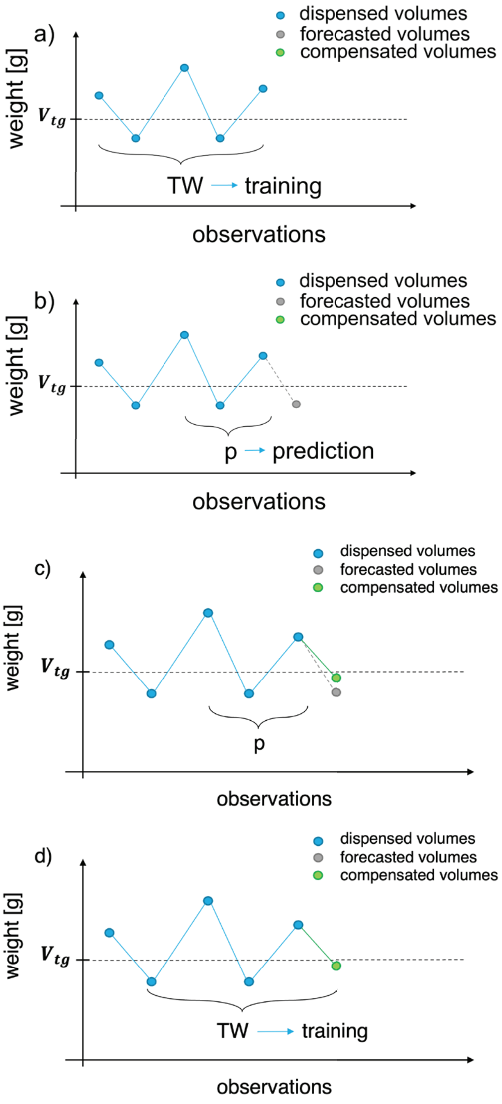


FIGURE 5. ADCS with Online Training representation. a) *TW* dispensed volumes are collected and model trained. b) Considering last *p* doses, forthcoming fill is forecasted. c) PP is compensated according to forecasted volume and next fill triggered resulting in dispensed volume closer to V_{tg}. d) *TW* moves to consider last compensated volume and model is trained again.

and detect volume-dependent trends or limitations in our compensation strategies.

3) BASELINE ACCURACY DETERMINATION

Before developing our ensemble-based ADCS, determining absolute PP accuracy across volumes under realistic conditions was essential, as pump manufacturers typically provide only generic specifications. Our manufacturer claims accuracy of ±0.5% for volumes above 0.5 ml and ±1.0% for volumes below 0.5 ml [28]. However, these specifications inadequately represent performance under varied operating conditions, necessitating our investigation.

We collected uncompensated datasets by operating the pump without compensation for extended fill sequences across each volume, performing calibrations every 300 fillings to replicate our protocol.

Following our statistical framework (Section II-C10), we initiated baseline characterization through variance homogeneity analysis using complementary mean-centered Levene and median-centered Brown-Forsythe tests, establishing robust foundation for assessing measurement stability while analyzing accuracy per formula $\mathrm{Acc}_{\mathrm{ADCS}_{\mathrm{OFF}}}$ presented in Section II-C. Results are reported in Table 2.

Statistical analysis revealed significant variance heterogeneity across all volumes, with both tests yielding p-values $< 1e^{-5}$, demonstrating substantial sequence-to-sequence variability in uncompensated measurements.

Pairwise sequential analysis, quantified through Sequential Differences percentage, revealed an increasing trend with volume (50.98%-64.15%). This metric, derived from pairwise Levene tests ($p < 0.05$), indicates proportion of consecutive sequence pairs exhibiting statistically significant variance differences. For example, at 0.3 ml, 60.38% of consecutive pairs demonstrated significant variance differences (32 out of 53 pairs). This baseline characterization, encompassing 25-53 sequences of 300 fills per volume, establishes robust foundation for subsequent compensation effectiveness evaluation, aligning with our statistical framework's first phase.

4) ADCS PERFORMANCE DATA COLLECTION

Based on baseline accuracy calculation and following the experimental protocol in Section II-D1, we designed a systematic data collection procedure to assess ADCS performance. For each volume, we implemented both AR(10) (autoregressive model with 10-dose window) and GRU(20,6) (single-layer GRU network with 6 neurons processing 20-dose sequences) as described in previous works [14], [15].

For each subset, we calculated online accuracy measures $\mathrm{Acc}^{\mathrm{online}}_{\mathrm{ADCS}_{\mathrm{OFF}}}$ and $\mathrm{Acc}^{\mathrm{online}}_{\mathrm{ADCS}_{\mathrm{ON}}}$, with corresponding $\mathrm{Gain}_{\mathrm{Acc}}$ to assess performance improvements between uncompensated and compensated conditions.

Our study collected almost 80,000 individual dosing data points across all volumes, including valuable information from previous studies for 0.3 ml and 1.2 ml volumes, particularly integrating data from earlier work on Pre-trained approaches [15].

To establish statistical validity, we implemented the multi-dimensional framework discussed in Section II-C10 and summarized in Table 3. Due to GRU(20,6) model's computational complexity requiring longer execution times compared to AR(10), collected subset numbers vary between models (14-31 for GRU vs. 28-38 for AR), while maintaining statistical significance across key metrics.

Mean-level effects demonstrate strong volume dependency, with highest significance at lower volumes ($t = -12.609$, $p < 0.001$ for AR(10) at 0.3 ml; $t = -6.285$, $p < 0.001$ for GRU(20,6) at 0.2 ml). This pattern weakens at larger volumes, notably at 1.2 ml where AR(10) shows non-significant mean-level changes ($t = -0.166$, $p > 0.05$).

TABLE 2. Baseline accuracy results and statistical validation metrics.

Volume (ml)	$Acc_{ADCS_{OFF}}$ (%)	Subsets (n)	Levene Stat	B-F Stat	Sequential Diff (%)
0.1	4.0	32	20.07	15.53	67.74
0.2	2.5	51	12.32	7.87	50.98
0.3	2.0	53	86.90	59.86	60.38
1.2	0.5	53	66.76	62.90	64.15
2.0	0.5	25	39.34	37.75	64.00

TABLE 3. Multi-dimensional statistical validation framework results.

Volume (ml)	Model	Subsets (n)	Mean-level Analysis	Distribution Analysis	Variance Analysis	Global Effects
			t-stat	MW-stat	Levene-stat	Mixed model stat
0.1	AR(10)	32	-3.125^{**}	300.76^{***} (50.0%)	414.96^{***} (68.8%)	46.90^{***}
	GRU(20,6)	18	-3.893^{**}	33.62^{**} (25.0%)	318.55^{***} (87.5%)	27.67^{***}
0.2	AR(10)	38	-4.101^{***}	985.15^{***} (63.2%)	420.79^{***} (60.5%)	223.25^{***}
	GRU(20,6)	31	-6.285^{***}	473.29^{***} (67.7%)	416.87^{***} (76.8%)	158.45^{***}
0.3	AR(10)	31	-12.609^{***}	1934.48^{***} (93.5%)	375.52^{***} (58.1%)	1497.75^{***}
	GRU(20,6)	19	-0.657	457.53^{***} (57.9%)	801.65^{***} (89.5%)	10.40^{**}
1.2	AR(10)	28	-0.166	715.76^{***} (71.4%)	447.55^{***} (75.0%)	0.004
	GRU(20,6)	16	-2.416^{*}	551.58^{***} (93.8%)	108.57^{***} (31.2%)	83.69^{***}
2.0	AR(10)	32	-2.870^{**}	740.51^{***} (65.6%)	370.91^{***} (53.1%)	105.43^{***}
	GRU(20,6)	14	-2.654^{**}	495.34^{***} (65.6%)	232.76^{***} (53.1%)	82.43^{***}

Note: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$
Percentages in parentheses indicate the proportion of subset pairs with significant differences

Distributional analyses through Mann-Whitney U statistics confirm significant transformations across most conditions, with robust effects at smaller volumes—particularly for AR(10) at 0.3 ml (1934.48, $p < 0.001$) with 93.5% of subset pairs showing significant differences. GRU(20,6) demonstrates its strongest distributional transformation effect at 1.2 ml volume (551.58, $p < 0.001$) with 93.8% of subset pairs exhibiting significant differences. Lower consistency at micro-volumes is observed for GRU(20,6) at 0.1 ml, where only 25.0% of subset pairs show significant distributional differences despite an overall significant meta-analysis statistic.

Variance homogeneity analysis through Levene tests reveals significant modifications ($p < 0.001$) across all volume-model combinations, with notably high statistics for GRU(20,6) at 0.3 ml (801.65) with 89.5% of subset pairs showing variance differences, and AR(10) at 1.2 ml (447.55) with 75.0% significant subset pairs. These high percentages indicate consistent variance modification effects across multiple calibration conditions, suggesting robust improvement in dosing precision.

The framework's global validation through mixed linear models demonstrates varied but predominantly significant effects. AR(10) shows exceptionally strong effects at 0.3 ml (1497.75, $p < 0.001$), but non-significant global effects at 1.2 ml (0.004, $p > 0.05$) — a notable contrast with its significant distributional and variance modifications at this volume. GRU(20,6) maintains significant global effects across all volumes, with the strongest effect at 0.2 ml (158.45, $p < 0.001$), providing robust evidence for systematic compensation effects across most volume-model configurations.

TABLE 4. Performance metrics across volumes and models.

Volume (ml)	Model	$Acc^{online}_{ADCS_{OFF}}$ (%)	$Acc^{online}_{ADCS_{ON}}$ (%)	$Gain_{Acc}$ (%)	Time (s)

0.1	AR(10)
	GRU(20,6)
		4.0	2.57
			2.49
				35.75
				37.75
					0.89
					10.85

0.2	AR(10)
	GRU(20,6)
		2.5	1.66
			1.48
				33.6
				40.8
					0.90
					10.92

0.3	AR(10)
	GRU(20,6)
		2.0	1.17
			1.06
				41.5
				47.0
					0.88
					10.87

1.2	AR(10)
	GRU(20,6)
		0.5	0.37
			0.35
				26.0
				28.0
					0.91
					10.94

2.0	AR(10)
	GRU(20,6)
		0.5	0.253
			0.259
				49.4
				48.2
					0.90
					10.89

Table 4 and Figure 6 provide complete representation of accuracy improvements across volumes for both models.

Both AR(10) and GRU(20,6) models achieved significant dosing accuracy improvements across all studied volumes. For smaller volumes (0.1-0.3 ml) with higher initial uncompensated accuracy errors (4.0-2.0%), the compensation system proved highly effective. At 0.1 ml, GRU(20,6) achieved 37.75% gain (2.49% accuracy), while AR(10) reached 35.75% gain (2.57% accuracy). GRU(20,6) provided superior results throughout smaller volumes, with distinctly better gains at 0.2 ml (40.8% vs. 33.6%) and 0.3 ml (47.0% vs. 41.5%). For larger volumes (1.2-2.0 ml) with better initial accuracy (0.5%), both models maintained very good performance gains. At 2.0 ml, AR(10) slightly outperformed GRU(20,6) with 49.4% gain (0.253% accuracy) against

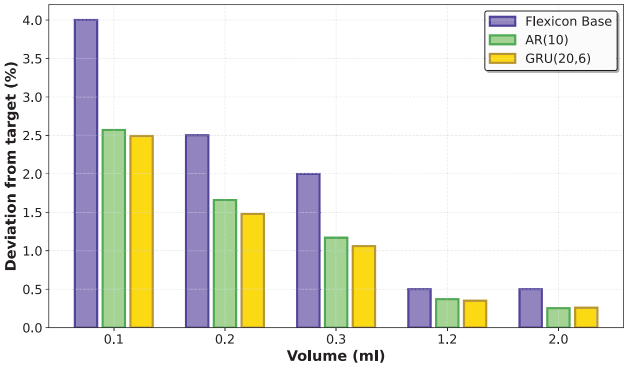


FIGURE 6. Accuracy improvements across volumes and ADCS models. Lower bars indicate higher accuracy, as they represent lower error percentages.

48.2% gain (0.259% accuracy). This demonstrates ADCS capability for large improvements regardless of initial accuracy conditions.

The Time column in Table 4 represents the average time in seconds required for the complete model training and prediction cycle, measured from the start of model training to the generation of the next dose prediction. These timing measurements were conducted on the hardware configuration described in Section II-A. As clearly demonstrated by the results, AR(10) models consistently maintain execution times below 1 second across all volumes. In contrast, GRU(20,6) models require substantially more computational resources, with execution times around 11 seconds. This significant performance difference aligns with previous findings [15] and highlights an important practical consideration for industrial implementation, where cycle time requirements often dictate the feasibility of runtime compensation strategies.

To provide deeper insights into compensation effectiveness, we analyze ECF for representative volumes 0.1 ml and 1.2 ml. The ECF analysis (Section II-C, Figures 7 and 8) reveals distinct improvement patterns across different operational regimes. For 0.1 ml, where initial accuracy challenges are more pronounced, both models demonstrate substantial error distribution improvements. GRU(20,6) shows particularly effective error reduction, evidenced by steeper compensated curve rise in lower error range, indicating higher concentration of measurements with reduced errors. AR(10) exhibits similar improvement patterns with slightly different error distribution characteristics. At 1.2 ml, with better baseline accuracy, compensation effect manifests through systematic error distribution shift. Both models achieve notable improvements, with compensated curves showing consistently higher cumulative fractions at lower error values. The pronounced leftward shift indicates systematic dosing error reduction across the entire distribution.

III. EXPERIMENTS AND RESULTS

Building on previous sections, we now present the core of our research: development, evaluation, and application of ensemble-based ADCS models in real-world settings.

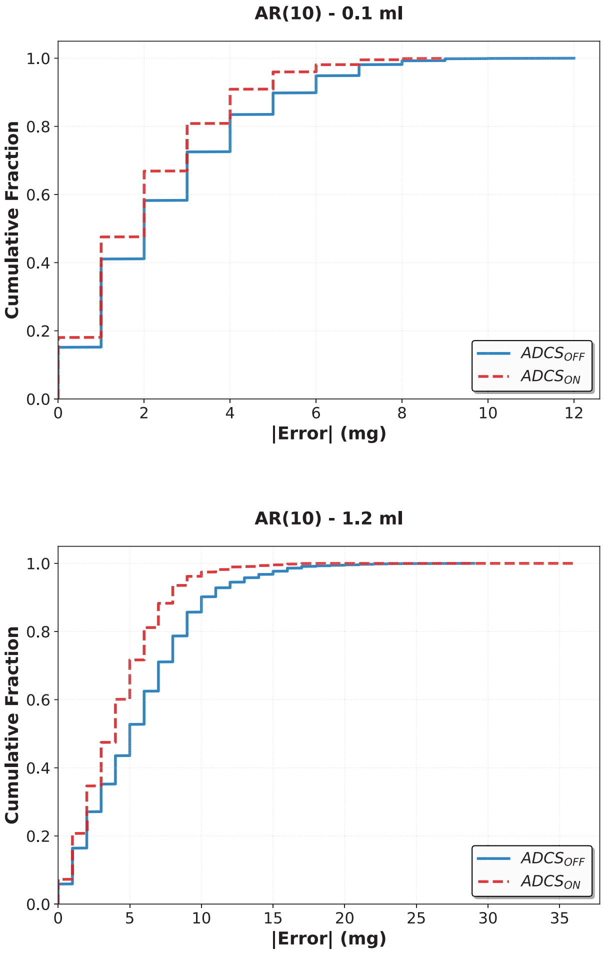


FIGURE 7. Cumulative error distributions for AR(10) model at different doses. Error distributions are shown for both ADCS_{OFF} (blue) and ADCS_{ON} (green) conditions.

We begin by introducing ensemble methods for PP control systems in Section III-A, followed by a novel offline indicator that assesses and selects between ensemble models in Section III-B. This indicator, verified using our extensive dataset, streamlined our research process and reliably predicts real-world performance.

In Section III-C we explore various ensemble strategies with their theoretical foundations and provide comprehensive performance analysis across different dosing volumes. Our results show substantial improvements over traditional single-model approaches, particularly in micro-dosing scenarios.

To validate our ensemble-based ADCS, we conducted real-world testing using a model identified by our offline indicator. These tests, discussed in Section III-D, demonstrate both method effectiveness and practical significance in pharmaceutical manufacturing.

Finally, in the same section, we present comparative analysis benchmarking our ensemble-based ADCS against classical ADCS implementations and state-of-the-art

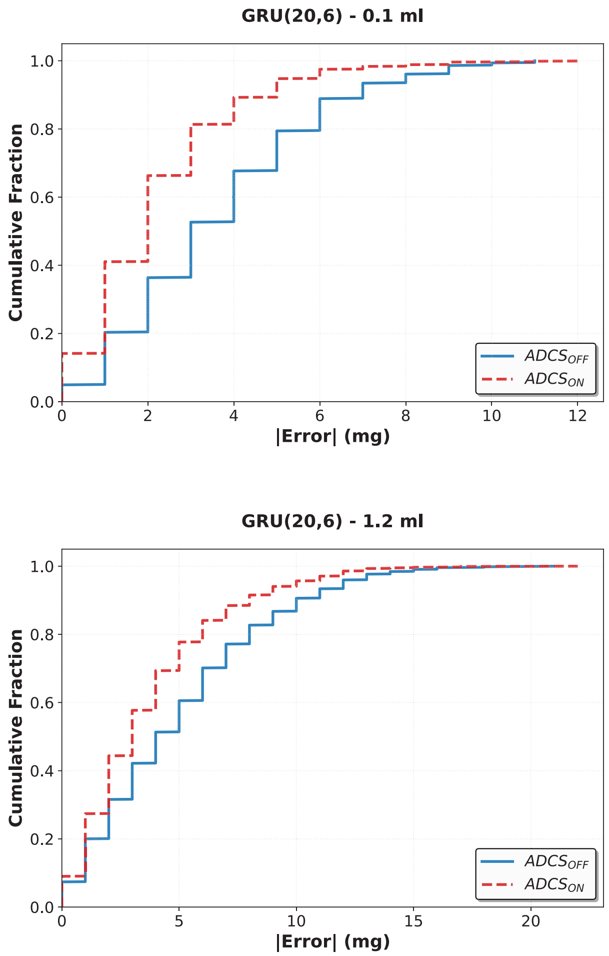


FIGURE 8. Cumulative error distributions for GRU(20,6) model at different doses. Error distributions are shown for both ADCS$_{OFF}$ (blue) and ADCS$_{ON}$ (green) conditions.

mechanical pumps, underscoring advancements achieved and positioning our work within broader dosing technology innovations.

A. PRINCIPLE AND BENEFITS OF ENSEMBLE METHODS

Ensemble methods combine predictions from several models to achieve better performance than any individual model [29]. This approach enhances accuracy and robustness by aggregating different predictive algorithms, effectively averaging predictions to cancel random errors while minimizing outlier influence.

The versatility and effectiveness of ensemble methods have led to widespread adoption across domains [30]. In finance, ensemble techniques predict stock prices and assess risk by capturing complex market dynamics [31], [32]. In meteorology, ensemble forecasting is standard practice, improving weather prediction accuracy by considering outputs from numerous atmospheric models [33], [34], [35]. In medical diagnostics, ensemble methods integrate different expert opinions or test results to enhance diagnosis accuracy and reliability [36], [37]. Common techniques include Bagging, Boosting, Stacking, and specialized approaches, all showing significant improvements over single-model implementations [38], [39], [40], [41].

For ADCS, these benefits translate into more accurate and consistent dosing across various volumes and conditions. ADCS can better adapt to complex PP system behaviors by combining predictions from multiple models, potentially improving compensation for factors like tube elasticity changes, temperature fluctuations, or fluid viscosity variations affecting dosing accuracy. Recent research in other domains has shown similar benefits from dynamic collaborative approaches. For instance, Wang et al. [42] demonstrated that combining dynamic adaptation capabilities with multi-source collaboration significantly improved fault diagnosis accuracy in rotating machinery across variable operating conditions. Their work highlights how dynamic multi-source systems can outperform static models by adapting to changing task requirements, a principle that aligns with our findings.

B. OFFLINE INDICATOR FOR MODEL EVALUATION

Ensemble methods' promising potential faces a major challenge: exhaustive testing of diverse model configurations would be too resource-intensive in real-world settings. We need an efficient evaluation tool to quickly assess model performance without extensive machine testing while reliably indicating real-world performance.

To overcome efficient model evaluation challenges, we developed an offline indicator, serving as a real-world performance proxy allowing efficient evaluation and comparison of different models, including ensemble methods.

1) OFFLINE INDICATOR IMPLEMENTATION

To develop a reliable offline evaluation methodology, we established our approach using models whose real-world performance had been extensively characterized through physical testing (AR, GRU). This strategy provided the essential ground truth dataset needed to validate whether our offline methodology could reliably predict actual compensation effectiveness. Once this correlation was confirmed with reference models, we applied it in the evaluation of new configurations, skipping the physical validation. This approach can be very interesting for achieving affordable coarse design-space-exploration and tuning indications that otherwise would need extensive experiments on the actual machine with the specific algorithm/tuning implemented.

Our analysis utilizes 80,000 measurements across five volumes (reported in Table 3), including historical datasets from previous research campaigns [15]. We apply each candidate model m across all available datasets regardless of original compensation method, separating compensated (ADCS$_{ON}$) from uncompensated (ADCS$_{OFF}$) portions to enable independent evaluation under both operational conditions.

The offline indicator implements this validated approach through a systematic three-step methodology applied indipendently for each model m:

1. **Sliding window-based volume prediction:** As shown in Figure 9, the offline indicator adopts a technique similar to Section II-D1. This sliding window procedure is applied separately to both compensated and uncompensated data. Considering Figure 9a:
   a) Select TW successive doses from given subset (dots in yellow Training Window)
   b) Train selected model m on these TW doses
   c) Model predicts next dose (red cross)
   d) Keep forecasted value for later comparison with actual value (blue dots)

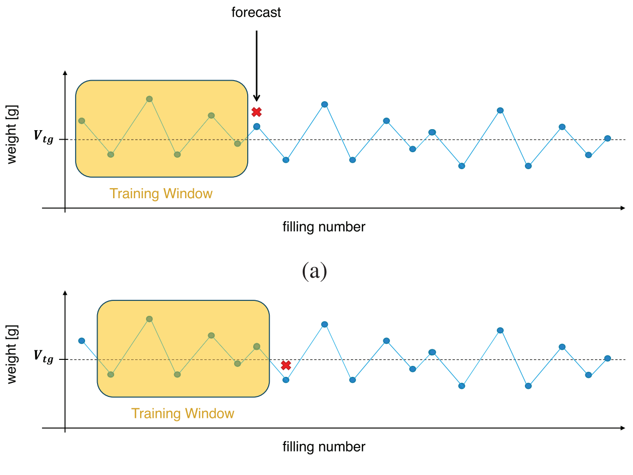
   Considering Figure 9b:
   e) Shift training window one position forward
   f) Update model with new window and forecast next dose (new red cross)
   g) Continue sliding window one step at a time, retraining model until predictions are made for all remaining doses
   h) Repeat process with new subset until entire dataset is considered
2. **Comparison and RMSE calculation:** We compare the predicted values sequence from previous step to actual dispensed volumes, calculating $\text{RMSE}_{\text{Standard}}$ for each run. This provides a standardized measure that allows direct comparison between different model configurations and architectures across identical datasets.
3. **Results aggregation:** We calculate $\overline{\text{RMSE}}_{\text{Standard}}$ for all runs over the considered volume and prediction model m combination. This produces two distinct offline indicators: $\overline{\text{RMSE}}^{\textit{offline}}_{\text{ADCS}(m)_{\text{ON}}}$ calculated exclusively from compensated dataset portions, and $\overline{\text{RMSE}}^{\textit{offline}}_{\text{ADCS}(m)_{\text{OFF}}}$ calculated exclusively from uncompensated dataset portions. This dual-indicator approach enables assessment of model predictive capability under both operational conditions.

All experiments used $TW = 50$ doses to ensure consistent evaluation across both ADCS_{ON} (200 doses) and ADCS_{OFF} (100 doses) segments while providing sufficient training data.

After analyzing both segments, we concluded that $\overline{\text{RMSE}}^{\textit{offline}}_{\text{ADCS}(m)_{\text{ON}}}$ metric effectively predicts real-system performance, showing small deviation from $\overline{\text{RMSE}}^{\textit{online}}_{\text{ADCS}(m)_{\text{ON}}}$ calculated in real-world tests using the same compensation model. Table 5 summarizes findings across volumes and models.

To determine significant relationship between our offline indicator and actual performance, we calculated Pearson correlation coefficient (r) between these values, quantifying strength and direction of linear relationship. The r coefficient measures degree of linear association between two variables:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}}$$



(a)

(b)

FIGURE 9. Sliding window-based volume prediction technique. (a) Initial training window (yellow area) containing historical data points used to forecast next dosing volume (red cross). Blue dots represent actual measured volumes, and horizontal dashed line indicates target volume (V_{tg}). (b) Training window shifted forward by one position, generating new forecast.

TABLE 5. RMSE Comparison between online and offline measurements.

Volume (ml)	Model	$\overline{RMSE}^{online}_{ADCS}$	$\overline{RMSE}^{offline}_{ADCS}$

0.1	AR(10)
	GRU(20,6)
		2.57
		2.54
			2.77
			2.42

0.2	AR(10)
	GRU(20,6)
		3.32
		3.01
			3.63
			3.10

0.3	AR(10)
	GRU(20,6)
		3.51
		3.19
			3.65
			3.23

1.2	AR(10)
	GRU(20,6)
		4.44
		4.24
			4.53
			4.18

2.0	AR(10)
	GRU(20,6)
		5.06
		5.18
			5.53
			4.68

where x_i and y_i are individual data points while $\bar{x}$ and $\bar{y}$ are means of x and y datasets. Coefficient r ranges from -1 to $+1$, where $+1$ indicates perfect positive linear correlation, 0 indicates no linear correlation, and -1 indicates perfect negative linear correlation.

In our study:

- x represents $\overline{\text{RMSE}}^{\textit{offline}}_{\text{ADCS}_{\text{ON}}}$ values (offline indicator predictions)
- y represents $\overline{\text{RMSE}}^{\textit{online}}_{\text{ADCS}_{\text{ON}}}$ values (actual online performance)

Our analysis revealed strong correlation ($r = 0.96$) and Figure 10 visualizes this relationship, showing points clustered around identity line, while Table 6 shows percent deviation of offline predictions from online measurements.

2) QUANTITATIVE ASSESSMENT

To provide rigorous validation, we conducted a statistical analysis of our offline indicator effectiveness using the data from Table 5. The quantitative assessment evaluates

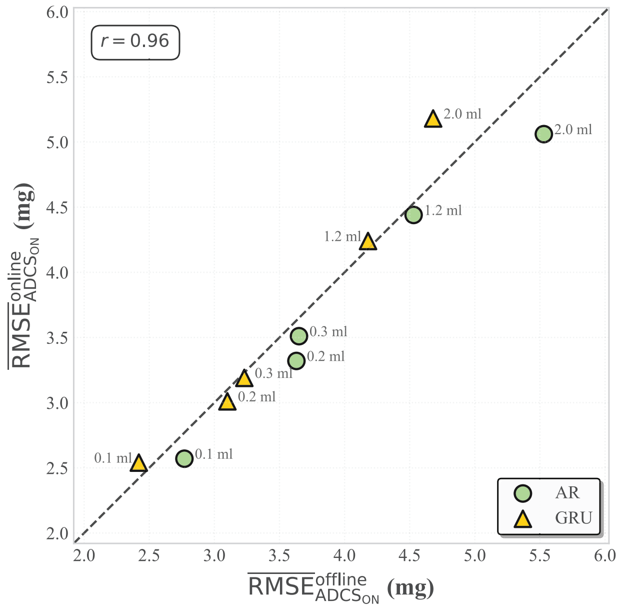


FIGURE 10. Correlation between offline predictions and online measurements for AR and GRU models across different volumes. Dashed line represents identity line (y = x). Strong correlation (r = 0.96, R^2 = 0.92) with MAE = 0.202 and MAPE = 5.25% demonstrates that offline predictions are reliable indicators of actual online performance.

TABLE 6. Percent deviation between online and offline measurements.

Volume (ml)	Model	Deviation (%)

0.1	AR(10)
	GRU(20,6)
		−7.22
		+4.95

0.2	AR(10)
	GRU(20,6)
		−8.67
		−2.76

0.3	AR(10)
	GRU(20,6)
		−3.83
		−1.34

1.2	AR(10)
	GRU(20,6)
		−1.95
		+1.42

2.0	AR(10)
	GRU(20,6)
		−8.48
		+10.74

how accurately our offline methodology predicts real-world compensation performance across different volume-model combinations.

Table 7 presents the complete statistical validation metrics. The coefficient of determination R^2 demonstrates that 92.13% of the variance in online performance is explained by our offline indicator predictions, representing good predictive power for a complex mechanical system. The Mean Absolute Error (MAE) and $RMSE_{Standard}$ indicate that typical prediction errors are well within acceptable tolerances. The Mean Absolute Percentage Error (MAPE) confirms high relative accuracy across the entire range of tested conditions. These metrics collectively validate that our offline indicator seems to provide reliable performance estimates.

TABLE 7. Quantitative validation metrics for offline indicator effectiveness.

Metric	Value
Coefficient of Determination (R^2)	0.9213
Mean Absolute Error (MAE)	0.202
Root Mean Square Error (RMSE)	0.257
Mean Absolute Percentage Error (MAPE)	5.25%
Pearson Correlation Coefficient (r)	0.96

TABLE 8. Error analysis summary by model type and volume regime.

Category	Average Abs. Error	Range
AR Models	0.242	0.090 - 0.470
GRU Models	0.162	0.040 - 0.500
Micro-volumes ($\leq$0.3ml)	0.150	0.040 - 0.310
Standard volumes ($\geq$1.2ml)	0.280	0.060 - 0.500

To analyze error patterns systematically, we calculated absolute errors between offline predictions and online measurements for each volume-model combination from Table 5, then categorized results by model type and volume regime.

Table 8 summarizes the error patterns across different conditions.

Detailed residual analysis reveals specific conditions where the offline indicator shows reduced accuracy, providing valuable insights for practical implementation. The largest prediction errors occur at extreme volumes, particularly 2.0 ml where both AR(10) and GRU(20,6) models show absolute errors of 0.47 and 0.50 respectively (deviations of −8.48% and +10.74%). This represents the upper boundary of reliable offline prediction and suggests that physical validation remains important.

Analysis shows systematic differences in prediction accuracy between model types. GRU-based models demonstrate higher offline prediction reliability (average absolute error: 0.162) compared to AR-based models (average absolute error: 0.242). Notably, AR models exhibit consistent negative residuals (underestimation), while GRU models show both positive and negative deviations, suggesting different error mechanisms between statistical and neural network approaches.

Counter-intuitively, micro-volumes show better offline prediction accuracy, with an average error of 0.150, than standard volumes that achieve an average error of 0.280. This finding indicates that despite the inherent complexity of micro-volume dosing, the offline indicator seems to effectively captures the compensation patterns in this critical regime where precision requirements are most stringent.

The quantitative analysis demonstrates that our offline indicator achieves good potential predictive performance (R^2 = 0.92, MAPE = 5.25%). This established it as a useful tool for preliminary model assessment, significantly reducing development time while maintaining appropriate caution for critical applications requiring physical validation.

C. ENSEMBLE MODELS FOR ADCS OPTIMIZATION

Building upon our validated offline indicator, we focused on three ensemble techniques: XGBoosting, Stacking, and Temporal Multiscale Ensemble (TME). These approaches address unique PP system challenges where traditional methods like AdaBoost and Random Forests face limitations due to the temporal nature and multi-scale dependencies of dosing systems.

XGBoosting [26], discussed in Section III-C1, handles non-linear relationships in PP systems through sequential tree building that targets residual errors. Unlike Random Forests' independent tree averaging, XGBoost optimizes both bias and variance, capturing subtle interactions affecting dosing accuracy.

Stacking, introduced in Section III-C2, integrates diverse regression model architectures through meta-learning, effectively combining XGBoost Regressors, Random Forest Regressors, and Linear Regression models through a meta-learning approach. While AdaBoost improves single-algorithm weak learners sequentially, Stacking's framework learns complex relationships between different modeling paradigms.

Finally in Section III-C3, our TME approach addresses PP systems' temporal characteristics by employing varying time windows and model architectures to capture both short-term fluctuations and long-term trends, surpassing conventional bagging methods that don't explicitly account for temporal dependencies.

1) XGBOOST ENSEMBLE

XGBoost creates sequential weak models (GBDTs - gradient boosted decision trees) optimized for speed and performance. Given dataset $D = \{(x_i, y_i)\}$ with n samples, XGBoost minimizes:

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where $l(y_i, \hat{y}_i)$ is loss function, $\hat{y}_i$ prediction, and $\Omega(f_k)$ regularization term for k-th tree complexity. Each new model focuses on previous residual error reduction:

$$f_m = \arg\min_{f} \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(m-1)} + f(x_i))$$

where $\hat{y}_i^{(m-1)}$ is prediction from previous $m-1$ trees.

XGBoost benefits our ADCS through complex relationship capture via hierarchical structure, L1/L2 regularization preventing overfitting, and natural outlier reduction sensitivity critical for anomalous operational parameters.

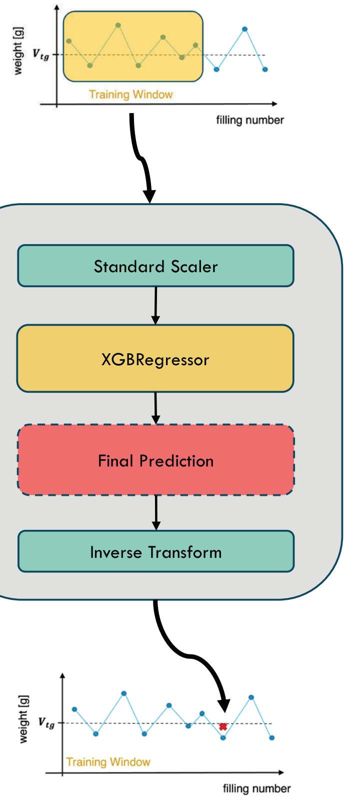


FIGURE 11. XGBoost Ensemble (E-XGB) architecture. Time series data undergoes StandardScaler normalization before XGBRegressor processing with optimal parameters.

Our implementation, depicted in Figure 11, utilizes the XGBoost library [26] with scikit-learn's StandardScaler for feature normalization. This transforms input variables to zero mean and unit standard deviation, improving algorithm stability and convergence speed. Although XGBoost is relatively robust to scale differences due to its tree-based nature, standardization benefits learning processes especially for time series with significant temporal variations. After prediction, values are returned to original scale through inverse transformation.

For the learning objective, our application employs XGBoost's default regression loss function, specifically the squared error:

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$

This loss function aligns with PP control requirements, where minimizing deviations between delivered and target volumes represents the primary optimization goal. The quadratic nature of the loss function appropriately penalizes larger deviations more severely, reflecting the critical importance of accuracy in production planning contexts.

Hyperparameter optimization used grid search maximizing performance gain:

$$\theta^* = \arg\max_{\theta} G(\theta)$$

$$G(\theta) = \left(1 - \frac{\mathrm{Acc}(\theta)^{\mathrm{offline}}_{\mathrm{ADCS}(XGBoost)_{\mathrm{ON}}}}{\mathrm{Acc}^{\mathrm{online}}_{\mathrm{ADCS}_{\mathrm{OFF}}}}\right) \cdot 100$$

where $\mathrm{Acc}(\theta)^{\mathrm{offline}}_{\mathrm{ADCS}(XGBoost)_{\mathrm{ON}}}$ represents estimated XGBoost accuracy using offline indicator $\overline{\mathrm{RMSE}}^{\mathrm{offline}}_{\mathrm{ADCS}_{\mathrm{ON}}}$, and $\mathrm{Acc}^{\mathrm{online}}_{\mathrm{ADCS}_{\mathrm{OFF}}}$ represents uncompensated pump accuracy from real-world tests.

Key optimized parameters with tested ranges:

- Number of estimators: $n_{\mathrm{estimators}} \in \{1, 5, 50, 100, 500\}$
- Learning rate: $\eta \in \{0.01, 0.05, 0.1, 0.2, 0.3\}$
- Maximum depth: $\mathrm{max}_{\mathrm{depth}} \in \{1, 2, 3, 4, 5\}$

Each configuration evaluation used Section III-B's sliding window approach, calculating gains using aggregated RMSE across available volume runs. Remarkably, optimal performance consistently emerged from minimal parameter configurations ($n_{\mathrm{estimators}} \leq 5$, $\eta = 0.01$, $\mathrm{max}_{\mathrm{depth}} = 1$), suggesting that PP dosing patterns favor simple, interpretable models over complex feature interactions.

2) STACKING ENSEMBLE

Stacking combines predictions from M base models through meta-model learning:

$$\hat{y} = g(f_1(x), f_2(x), \ldots, f_M(x))$$

where g is meta-model and f_i base models.

Our configuration uses three base models (XGBoost Regressor for non-linear relationships, Random Forest Regressor for robust predictions, Linear Regression for linear relationships) combined through Linear Regression meta-model:

$$\hat{y}_{\mathrm{final}} = w_0 + w_1 * \hat{y}_{\mathrm{XGBoost}} + w_2 * \hat{y}_{\mathrm{RandomForest}} + w_3 * \hat{y}_{\mathrm{LinearRegression}}$$

where w_i are learned weights.

Before training, data standardization using StandardScaler normalizes features. This step is particularly important for Stacking, where models with different scale sensitivities must operate harmoniously, ensuring each algorithm expresses full predictive potential without scale-induced bias.

Stacking enhances ADCS through diverse model integration capturing varied patterns, hierarchical learning addressing individual model weaknesses, and improved bias-variance trade-off across data distributions.

In this implementation, depicted in Figure 12, we employ scikit-learn's StackingRegressor [27] with custom base model configurations. The optimization follows a two-level hierarchical approach. At the first level, each base model is independently optimized: XGBoost uses the methodology described in Section III-C1, Random Forest optimizes $n_{\mathrm{estimators}}$ and $\mathrm{max}_{\mathrm{depth}}$ parameters, while Linear Regression uses standard implementation without regularization.

At the second level, meta-model training employs cross-validation through StackingRegressor from scikit-learn. The strategy uses TimeSeriesSplit with $n_{\mathrm{splits}} = 5$, specifically designed for time series. This method chronologically subdivides data ensuring each validation fold is temporally subsequent to training data. Unlike standard KFold's random subdivision, TimeSeriesSplit preserves temporal integrity, essential for sequential dosing series where predictions rely on historical data only.

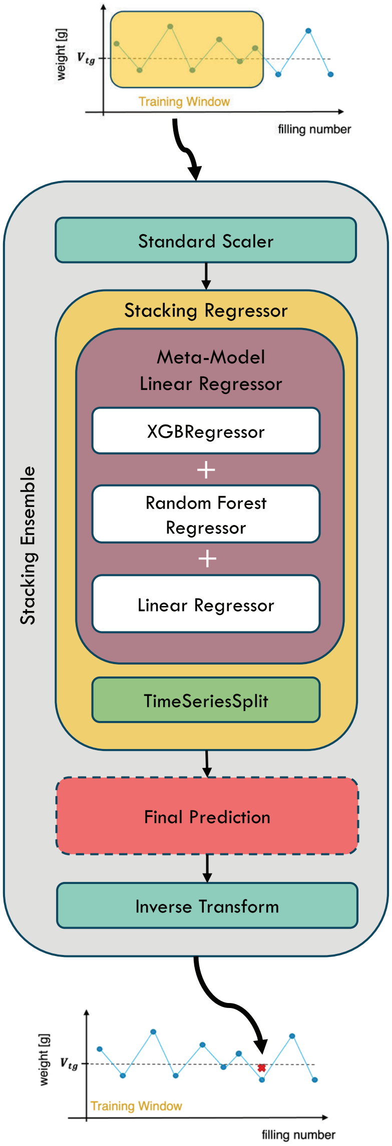


FIGURE 12. **Stacking Ensemble (E-STACK) with TimeSeriesSplit cross-validation. Three base models (XGBRegressor, Random Forest, Linear Regression) feed a Meta-Model that learns optimal combination weights while preserving temporal integrity.**

Meta-model weights are optimized by minimizing Mean Squared Error (MSE) between combined predictions and observed values. This loss function appropriately penalizes larger errors, aligning with dosing accuracy goals.

3) TEMPORAL MULTI SCALE ENSEMBLE

TME captures PP system dynamics through temporal window variation and model parameterization. For M distinct models at different temporal scales, ensemble prediction uses uniform averaging:

$$f_{\mathrm{TME}}(x_t) = \frac{1}{M} * \sum_{i=1}^{M} f_i(x_t)$$

where $f_{\mathrm{TME}}(x_t)$ is prediction at time t, M model number, and $f_i(x_t)$ i-th model prediction.

Uniform averaging is theoretically justified by inherent ensemble diversity. Models operating at different temporal scales with varying configurations naturally induce error decorrelation, capturing distinct system behavior aspects. This decorrelation improves through architectural diversity in temporal pattern processing. Under these conditions, simple averaging effectively combines temporal scales while maintaining robustness [29], [43].

Each model f_i uses Section III-B's sliding window approach with two diversification strategies:

1. Temporal Scale Diversification - Models use varying TW with consistent internal parameters:
    - Short-term models (TW = 10-15): immediate dosing variations
    - Medium-term models (TW = 15-20): intermediate patterns
    - Long-term models (TW = 25-30): gradual system changes
2. Model Configuration Diversification - Fixed temporal window with varying internal parameters:
    - AR models: varying order parameter
    - GRU models: varying neurons and hidden layers

TME fundamentally differs from conventional bagging techniques through temporal-oriented diversification. While bagging uses bootstrap sampling, potentially disrupting temporal structures, TME preserves temporal relationships while achieving diversity through systematic temporal scale and architecture variation.

Our implementation, shown in Figure 13, employs a custom ensemble framework with parallel execution capabilities where each model $f_i(x_t)$ is processed independently. Parameter selection builds on proven AR(10) and GRU(20,6) models' effectiveness in capturing PP dosing patterns [14], [15]. Previous analyses showed excessive historical data can deteriorate performance, as recent data points more strongly influence future behavior. TME therefore uses slight variations in model complexity and temporal scope around proven models.

For AR-based ensembles, we leverage the statsmodels library [24], while GRU-based ensembles utilize Keras [25] with TensorFlow backend.

Table 9 summarizes all configurations, detailing model architectures, training windows, and specific variations.

D. RESULTS AND DISCUSSION

This section presents experimental results and analysis across three main areas: estimated ensemble model performance comparison, real-world validation, and benchmarking against high precision mechanical solutions and standard ADCS.

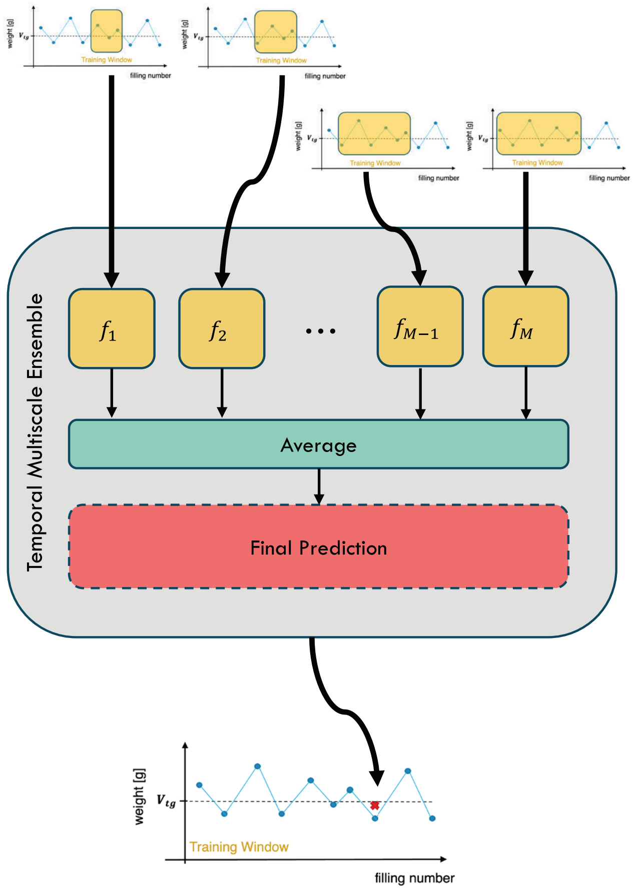


FIGURE 13. Temporal Multi-Scale Ensemble (TME) architecture. Multiple models with different temporal windows and orders process data in parallel. Final prediction uses simple arithmetic averaging.

TABLE 9. Ensemble model configurations.

Model	Configuration	Training Window	Focus/Variation
AR-based Ensembles			
E-AR	AR(5)...AR(15)	50	Short/long-term patterns

E-AR-RW	AR(5),AR(6)
	AR(7),AR(8)
	AR(9),AR(10)
	AR(11),AR(12)
	AR(13)...AR(15)
		10
		15
		20
		25
		30
			Short-term
			Medium-term
			Long-term

GRU-based Ensembles			

E-GRU	GRU(20,4)
	GRU(20,6)
	GRU(20,8)
		20	Neuron count

E-GRU-RW	GRU(10,6)
	GRU(15,6)
	GRU(20,6)
		25
		35
		50
			Window size

E-GRU-P	GRU(15,6)
	GRU(20,6)
	GRU(25,6)
		50	Input sequence

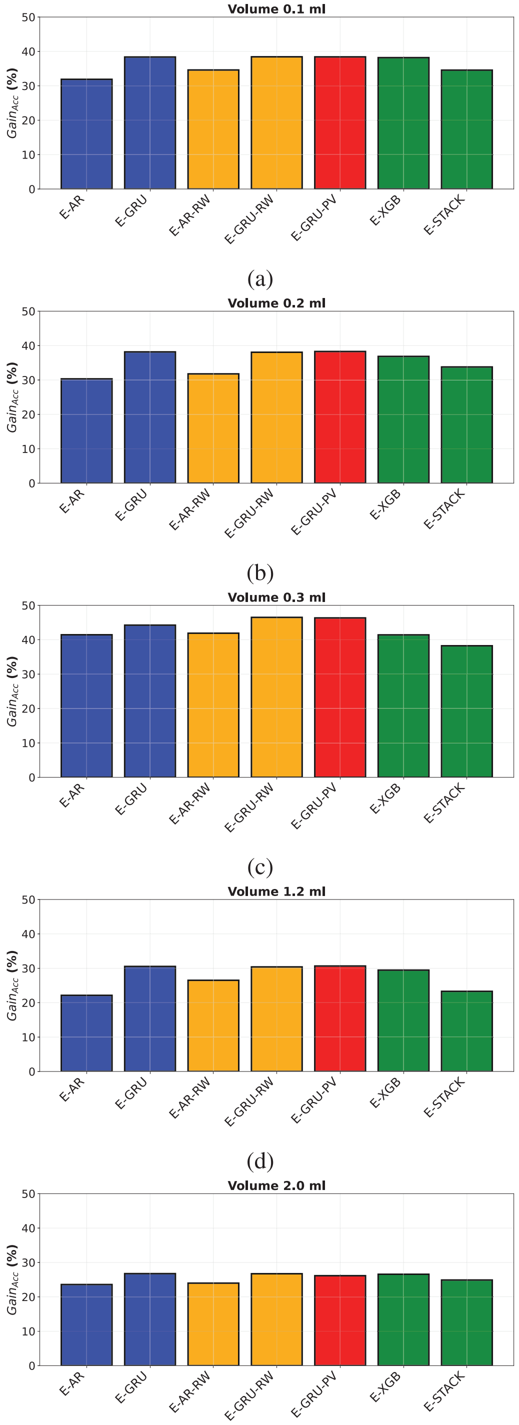
Note: The notation X...Y indicates all models from X to Y inclusive

1) ENSEMBLE MODELS PRELIMINARY PERFORMANCE ANALYSIS

We evaluated proposed ensemble models using our validated offline indicator methodology across volumes from 0.1 ml to 2.0 ml, as shown in Table 10 and Figure 14. Results reveal distinct performance patterns across configurations and

TABLE 10. Performance comparison of ensemble configurations showing Gain% for each volume.

Vol (ml)	Model	Acc_{OFF}^{online}	$Acc_{ON}^{offline}$	Gain (%)
0.1	E-AR	4.0	2.72	31.89
	E-GRU		2.47	38.37
	E-AR-RW		2.62	34.60
	E-GRU-RW		2.46	38.43
	E-GRU-PV		2.46	38.41
	E-XGB(1,0.01,1)		2.47	38.19
	E-STACK(5,0.01,1)		2.62	34.56
0.2	E-AR	2.5	1.74	30.32
	E-GRU		1.55	38.16
	E-AR-RW		1.71	31.77
	E-GRU-RW		1.55	38.06
	E-GRU-PV		1.54	38.28
	E-XGB(5,0.01,1)		1.58	36.85
	E-STACK(5,0.01,1)		1.66	33.78
0.3	E-AR	2.0	1.17	41.44
	E-GRU		1.12	44.24
	E-AR-RW		1.16	41.90
	E-GRU-RW		1.07	46.50
	E-GRU-PV		1.07	46.34
	E-XGB(5,0.01,1)		1.17	41.41
	E-STACK(1,0.01,1)		1.24	38.24
1.2	E-AR	0.5	0.39	22.14
	E-GRU		0.35	30.54
	E-AR-RW		0.37	26.51
	E-GRU-RW		0.35	30.40
	E-GRU-PV		0.35	30.66
	E-XGB(1,0.01,1)		0.35	29.47
	E-STACK(5,0.01,1)		0.38	23.30
2.0	E-AR	0.5	0.26	23.58
	E-GRU		0.23	26.75
	E-AR-RW		0.26	23.99
	E-GRU-RW		0.23	26.72
	E-GRU-PV		0.23	26.15
	E-XGB(5,0.01,1)		0.23	26.57
	E-STACK(5,0.01,1)		0.25	24.89



(a)

(b)

(c)

(d)

(e)

FIGURE 14. Accuracy comparison showing offline ADCS performance versus online ADCS reference across different dosing volumes. GRU-based ensembles demonstrate superior performance particularly at micro-volumes while simpler models maintain more consistent performance across the entire volume range.

volumes, with the optimal ensemble method varying based on dosing volume. Overall, GRU-based ensembles demonstrated superior predicted gains across the entire volume range, with volume-specific variations in which particular GRU configuration performed best. At 0.3 ml, where performance improvements were most pronounced, E-GRU-RW achieved the highest gain of 46.50%, closely followed by E-GRU-PV at 46.34%. For micro-volumes at 0.1 ml, E-GRU-RW again led with 38.43% improvement, with E-GRU-PV showing nearly identical results at 38.41%. At 0.2 ml, E-GRU-PV slightly outperformed other configurations with a 38.28% gain. For larger volumes, E-GRU-PV outperformed at 1.2 ml with 30.66% improvement, while standard E-GRU achieved the best results at 2.0 ml with 26.75%.

This superior performance can be explained by the fact that GRU architectures are particularly effective for micro-volumetric dosing where the pump exhibits complexity not adequately captured by simpler statistical models [15], as evidenced by the empirical results in Figure 6 where GRU models consistently outperform AR models, particularly in the critical micro-volume range. GRU's non-linear activation functions and specialized gating mechanisms allow these networks to model the intricate dynamics that characterize smaller volumes, where mechanical variabilities have proportionally larger effects on dosing precision. The combination of multiple GRU models in an ensemble seems to further leverages these architectural advantages, capturing different aspects of the system's behavior and achieving generally higher improvements compared to other ensemble approaches. Unfortunately this consistent excellence comes at higher computational cost requiring longer training times and specific hardware [15]. It is worth mentioning that other developments in RNN architectures, such as YamRNN [44], suggest that it is possible to design alternative architectures with fewer parameters while maintaining excellent learning capabilities. Such approaches report training time reductions of up to 80% compared to GRU while using only 67% of its parameters, highlighting promising directions for future ADCS optimization. Despite these potential improvements, in our current investigation, AR-based models offered a practical balance between computational efficiency and performance enhancement.

On the other hand, E-AR achieved 41.44% improvement at 0.3 ml, decreasing to 22.14% at 1.2 ml and rising slightly to 23.58% at 2.0 ml. The rolling window variant (E-AR-RW) offered minor improvements over the standard implementation, reaching 41.90% at 0.3 ml and 23.99% at 2.0 ml. While not achieving the peak performance of GRU-based models, AR ensembles provided these improvements with significantly lower computational requirements and higher model interpretability.

XGBoost (E-XGB) and Stacking (E-STACK) methods demonstrated competitive results with conservative configurations. Interestingly, E-XGB performed best with minimal parameters ($n_{\text{estimators}} \leq 5$, learning rate $= 0.01$, $\max_{\text{depth}} = 1$), achieving performance comparable to E-GRU at 2.0 ml (26.57%) and nearly matching E-GRU at 0.1 ml (38.19% vs. 38.37%). More complex configurations consistently degraded performance. E-STACK showed solid improvements ranging from 23.30% at 1.2 ml to 38.24% at 0.3 ml using basic parameters. This preference for simpler configurations suggests PP system interactions follow stable patterns, with prediction relying more on recent temporal patterns than complex feature interactions.

For real-world testing, we selected the E-AR model at 0.3 ml and 1.2 ml volumes, based on a careful balance between performance metrics and practical implementation considerations. While several GRU-based models showed higher predicted gains in our offline analysis, the following considerations guided us toward choosing E-AR. First, the simplicity of AR models and their basic interpretability make them quite suitable for industrial applications where the behavior of the system has to be understood and validated easily; of course, this becomes very important in regulatory compliance of pharmaceutical manufacturing, where model transparency and predictability are strict requirements. Second, while E-XGB showed promising results, our parameter analysis revealed a concerning sensitivity to hyperparameter settings, indicating potential stability issues in real-world applications.

Furthermore, E-AR has considerable practical advantages regarding computational efficiency. Unlike the GRU-based models, which require heavy computation and a long training time, E-AR can be easily deployed and retrained during operation with small system overhead. As shown in Table 4, individual AR(10) models consistently maintain execution times below 1 second across all volumes, while GRU(20,6) models require 10-12 seconds. For our E-AR implementation, we developed a parallel processing architecture, described in Section III-C3, where each constituent AR model within the ensemble runs simultaneously in a separate thread. This technical approach ensures that the computational complexity of E-AR remains comparable to that of a single AR model, with the total execution time primarily determined by the slowest AR variant in the ensemble. This parallel implementation effectively mitigates what would otherwise be a significant computational bottleneck, enabling runtime application in production environments where cycle times are critical. In contrast, a parallel implementation of E-GRU would still be constrained by the inherently longer processing times of individual GRU models, making such ensembles less suitable for time-sensitive industrial applications despite their accuracy advantages. This property is of particular importance for industrial settings, in which processing time hits production efficiency directly.

The selection of the 0.3 ml and 1.2 ml volumes for the validation of our system was tuned to follow previous research [14], [15] with, however, very different predicted gains of 41.44% versus 22.14%. The strong contrast presented an ideal case for the validation of our offline indicator reliability in scenarios where performance does differ. Moreover, these volumes encompass conditions both below and above the critical 1.0 ml threshold, thereby allowing complete validation of our method across a range of dosing regimens. This selection strategy is consistent with industrial best practices, in which reliability, interpretability, and operational efficiency often take precedence over marginal performance gains that may be challenging to maintain in real-world conditions.

2) REAL-WORLD VALIDATION ANALYSIS

Following offline analysis, we perform extensive real-world testing of the E-AR at volumes of 0.3 ml and 1.2 ml using the setup described in Section II-A. Tests follow our established protocol described in Section II-B and the compensation technique analyzed in Section II-D1, collecting a large sample size of 5,000 doses for 0.3 ml and over 6,000 for 1.2 ml to ensure robust statistical validation of our results.

As shown in Figure 15 and Table 11, our real-world experiments demonstrated good accuracy improvements across the various volume ranges. In Figure 15, bars with solid borders represent real measurements obtained from machine testing, while bars with no borders indicate values estimated through our offline indicator. Note that the height of each bar represents the accuracy of the system: lower bars indicate better accuracy, as they represent smaller deviations from the target volume.

To provide a general view of the model landscape we also included the offline indicator estimates for E-GRU since it consistently showed promising estimated performances. At a volume of 0.3 ml, the E-AR model reduced measurement uncertainty from 2.0% (as demonstrated by the Flexicon pump with no compensation) to 0.92%, representing a 53.93% increase in accuracy in real-world testing. This substantial improvement surpassed both baseline ADCS implementations: AR(10) at 1.17% and GRU(20) at 1.06%. Furthermore the offline estimates for E-AR suggested performance levels close to these actual measurements, strengthening the reliability of our estimation approach.

At 1.2 ml, we observed different but equally interesting dynamics. The system showed a 17.02% improvement in real testing, achieving an accuracy of 0.41% compared to the base accuracy of 0.5%. This real-world gain was lower than both the offline indicator's prediction of 22.14% and the performance of simpler models (AR(10) at 26.0% and GRU(20) at 28.0%).

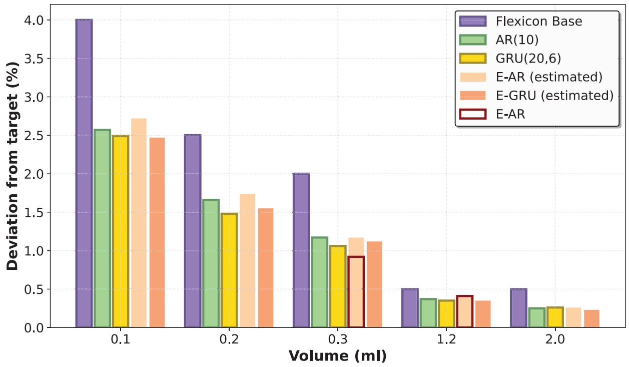


FIGURE 15. Model accuracy comparison across dosing volumes. Lower values indicate better accuracy. The ensemble models demonstrate superior performance, with E-AR achieving up to 53.9% improvement at 0.3 ml compared to baseline. Solid-bordered bars represent actual measurements, while borderless bars show offline indicator estimates. The correlation between estimated and real E-AR measurements validates our offline indicator methodology.

Analysis of our offline indicator's performance reveals notable volume-dependent variations. At 0.3 ml, the indicator reported in Table 10 predicted an accuracy of 1.17% while achieving 0.92% in practice, representing a 21.33% deviation that exceeds the typical ±10% range observed with simpler models. In contrast, at 1.2 ml, the prediction (0.39%) aligned more closely with the actual performance (0.41%), showing only a −5.04% deviation within the expected range. This pattern demonstrates that while our indicator maintains directional accuracy in predicting performance trends, its precision varies significantly with volume and model complexity.

Regarding performance, as shown in Table 11, our E-AR implementation achieves execution times very close to those of individual AR models, confirming the effectiveness of our parallel implementation approach. The execution times represent only a minimal overhead compared to single AR(10) models, while remaining an order of magnitude faster than GRU-based models.

To provide rigorous validation of these results, we implemented our statistical framework across both volume configurations. Table 12 presents the statistical validation metrics for the E-AR implementation.

The statistical analysis reveals distinctive patterns across different volumes. For the 0.3 ml volume, we observe improvements at the mean level ($t = -11.53$, $p < 0.001$), indicating that compensation has systematically shifted the distribution of dispensed volumes closer to the target value. This shift is accompanied by a change in the overall distribution, as evidenced by the Mann-Whitney U test (statistic $= 1520.42$, $p < 0.001$), with 100.0% of subset pairs showing significant differences. The variability of the dosing process is also modified (Levene statistic $= 237.69$, $p < 0.001$), with 52.9% of subset pairs showing differences in variance. The mixed linear model analysis further confirms this systemic effect (statistic $= 1516.84$, $p < 0.001$), demonstrating that compensation is effective even considering variability across different calibration regimes.

In contrast, the 1.2 ml configuration presents a different pattern, with non-significant changes at the mean level ($t = -1.24$, $p > 0.05$), suggesting that compensation has not substantially altered the mean value of the dosages. However, changes are observed at the distributional level (Mann-Whitney U statistic $= 408.63$, $p < 0.001$), with 66.7% of subset pairs showing significant differences, and at the variability level (Levene statistic $= 156.84$, $p < 0.001$), although only 38.1% of subset pairs show relevant differences in variance. The mixed model analysis remains relevant (statistic $= 17.79$, $p < 0.001$), confirming the global effectiveness of compensation despite the absence of an indicative effect on the mean.

Further insights into the error distribution characteristics are provided by the ECF analysis which reveals distinct compensation patterns: at 0.3 ml, the compensated curve (Figure 16) demonstrates a consistently steeper trajectory across the entire error range, indicating systematic improvement in error reduction across all operational conditions.

The 1.2 ml configuration shows more modest improvements (Figure 17), consistent with our accuracy metrics and statistical findings, which indicated lower performance gains compared to simpler models at this volume.

The statistical validation framework, combined with detailed ECF analysis, provides robust evidence for the effectiveness of the E-AR approach, while also highlighting volume-dependent variations in its performance characteristics. It is worth noting that while it might seem appealing to develop highly specialized compensation algorithms for each specific error profile, such an approach would face significant practical limitations in industrial settings where multiple factors (employed pumps, fluid viscosity, temperature fluctuations, tube elasticity changes, calibration drift)

TABLE 11. Performance comparison of different models across volume ranges.

Volume (ml)	PD12 (%)	FSP (%)	AR(10)			GRU(20)			E-AR		
			Acc(%)	Gain(%)	Time(s)	Acc(%)	Gain(%)	Time(s)	Acc(%)	Gain(%)	Time(s)
0.1	4.0	2.5	2.57	35.8	0.89	2.49	37.8	10.85	–	–	–
0.2	2.5	2.0	1.66	33.6	0.90	1.48	40.8	10.92	–	–	–
0.3	2.0	2.0	1.17	41.5	0.88	1.06	47.0	10.87	0.92	53.9	0.96
1.2	0.5	0.8	0.37	26.0	0.91	0.35	28.0	10.94	0.41	17.0	0.97
2.0	0.5	0.5	0.25	49.4	0.90	0.26	48.2	10.89	–	–	–

TABLE 12. Statistical validation metrics for E-AR implementation.

Volume (ml)	Mean-level Analysis	Distribution Analysis	Variance Analysis	Global Effects
	t-stat	MW-stat	Levene-stat	Mixed model
0.3	-11.53^{***}	1520.42^{***} (100%)	237.69^{***} (52.9%)	1516.84^{***}
1.2	-1.24	408.63^{***} (66.7%)	156.84^{***} (38.1%)	17.79^{***}

Note: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$
Percentages in parentheses indicate the proportion of subset pairs with significant differences

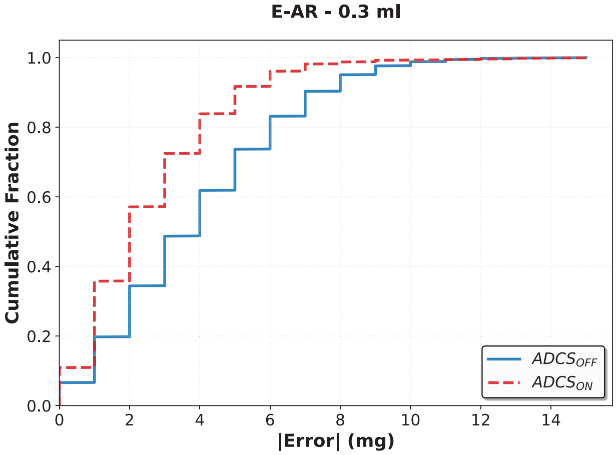


FIGURE 16. Error Cumulative Distribution Function at 0.3 ml.

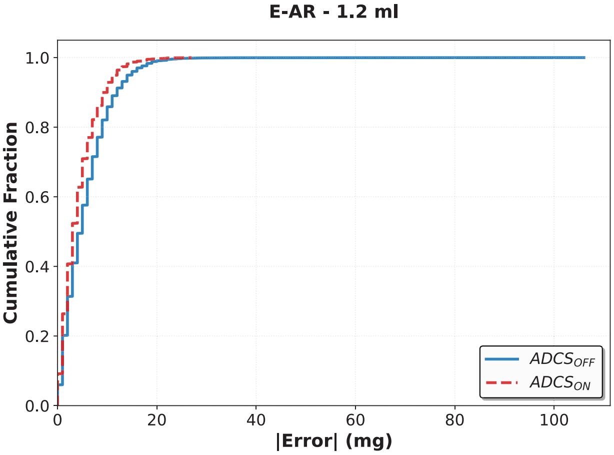


FIGURE 17. Error Cumulative Distribution Function at 1.2 ml.

dynamically influence error characteristics. The adaptive nature of our ADCS methodology represents a more robust approach, allowing the system to respond effectively to diverse error patterns without requiring their explicit a priori characterization.

3) BENCHMARKING AGAINST HIGH-PRECISION MECHANICAL SOLUTIONS

To put our results in the perspective of the dosing technology landscape, we conducted a comparison with the Colanar FSP pump, a state-of-the-art mechanical solution featuring several key innovations. Its 8-roller design (versus industry-standard 6 rollers) creates more frequent but smaller compression points, reducing fluid flow pulsations for smoother delivery. A planetary gear system distributes forces more evenly than conventional direct-drive mechanisms, reducing vibration and ensuring precise roller movement. Advanced tube compression mechanisms with engineered occlusion settings and tube guide elements prevent lateral movement during operation, addressing traditional PP challenges of pulsation, flow inconsistency, and tube wear.

Figure 18 and Table 11 compare FSP performance with our ADCS implementations across the tested volume range.

Our comparison yields excellent results for all volumes and models tested. At the more challenging micro-volume range of 0.1 ml, our ADCS implementation achieves similar performance to the FSP's 2.5% accuracy specification, with AR(10) and GRU(20) reaching 2.57% and 2.49% respectively. The advantages of our software-based approach become more visible at 0.2 ml, where AR(10) attains 1.66% accuracy and GRU(20) reaches 1.48%, outperforming the FSP's 2.0% specification by a substantial margin.

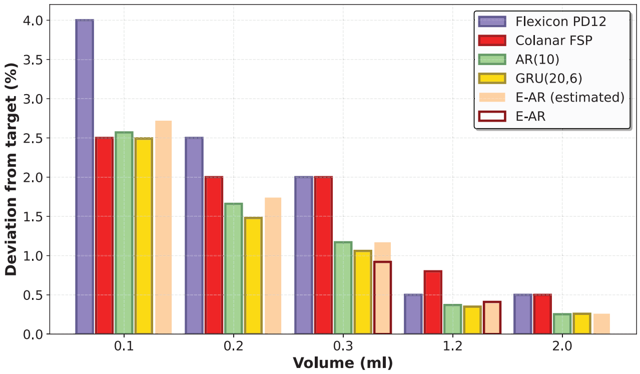


FIGURE 18. Accuracy comparison between our solutions and FSP high-precision pump across volume ranges. E-AR outperforms FSP at 0.3 ml (0.92% vs 2.0%), demonstrating software compensation's effectiveness. Solid-bordered bars show measured values while borderless bars represent estimated E-AR performance. Lower values indicate better accuracy.

At intermediate volumes, E-AR shows excellent results. At 0.3 ml, it achieves 0.92% accuracy, surpassing both FSP's 2.0% specification and base ADCS implementations. This significant improvement demonstrates ensemble methods' enhanced capability in micro-volume dosing conditions.

At larger volumes, all implementations show strong performance outperforming FSP's specification.

Our software-based approach offers substantial advantages beyond performance metrics. While FSP represents significant mechanical optimization investment, our solution achieves superior results through algorithmic enhancement of standard hardware. This provides cost-effective performance improvements without hardware modifications, enables continuous enhancement through software updates, and offers adaptive capability across varied operating conditions and fluid types.

IV. CONCLUSION, LIMITATIONS AND FUTURE WORK

This study has advanced several aspects of PP control. We developed and validated an offline indicator for assessing compensation models without extensive machine testing. Despite limitations with complex models, where deviations can exceed the normal ±10% range especially for ensemble methods, it provides valuable preliminary model assessment and optimization capabilities, though further refinement is needed for more complex model architectures.

Our extended volume range study (0.1–2.0 ml) revealed that while no single ensemble configuration achieves optimal performance across all volumes, GRU-based ensembles consistently demonstrate superior gains, with E-GRU-RW reaching 38.43% improvement at 0.1 ml and 46.50% at 0.3 ml and E-GRU-PV excelling at 0.2 and 1.2 ml (38.28-30.66%). This volume-dependent performance pattern suggests that optimal dosing accuracy may require volume-specific model selection in practical applications. The broad validation across multiple volumes strengthens software-based compensation as a viable precision enhancement approach, particularly for challenging micro-volumes.

For real-world testing, we selected the E-AR model despite its generally lower gains compared to GRU variants, prioritizing its interpretability and computational efficiency. While our real-world validation was limited to E-AR at specific volumes (0.3 ml and 1.2 ml) due to equipment time availability, the results revealed impressive accuracy improvements, achieving 53.93% increase at 0.3 ml and surpassing standard ADCS implementations. Though 1.2 ml showed moderate gains (17.02%), these results confirm ensemble methods' value for challenging volume applications, while highlighting the need for comprehensive evaluation of advanced ensemble methods like E-GRU and E-XGB across all volumes.

Regarding environmental adaptability, our E-AR ensemble demonstrates robust performance through its diverse temporal windows, effectively capturing both short-term fluctuations and long-term trends as evidenced by statistical validation across multiple calibration regimes. However, a significant limitation is our validation using only purified water, leaving fluid type adaptability unvalidated across the diverse pharmaceutical formulations encountered in real manufacturing scenarios. The responsive nature of our Online Training methodology suggests theoretical capability to accommodate different fluid types through continuous retraining, but this hypothesis requires experimental validation across fluids with varying viscosities and rheological properties, representing a critical area for future research. Finally from a computational point of view, our parallel implementation architecture ensures that E-AR maintains execution times comparable to single AR models (<1 second), representing a significant advantage over GRU-based approaches that require approximately 11 seconds per cycle.

Comparative testing against FSP mechanical pump demonstrated software compensation can match or exceed sophisticated hardware solutions, offering cost-effective high-precision dosing without expensive hardware investment.

Concerning the broader applicability of our ADCS, while the validation focused on the Flexicon PD12 system, the methodology was designed for generalizability through its closed box approach that relies on standard API interfaces rather than pump-specific mechanics. The core requirements for successful implementation include pump controllers with API access, hiqh quality weight measurement capability, adequate computational resources for Online Training, and stable operational environments. Our approach should theoretically extend to other systems sharing these characteristics, with the observed volume-dependent performance patterns likely reflecting fundamental physics rather than equipment-specific behaviors. However, the effectiveness of ADCS in various hardware configurations - especially those without control systems - remains to be investigated, and definitive cross-platform validation across diverse pump architectures, control systems, and operational environments remains a critical area for future research to establish comprehensive applicability boundaries.

In conclusion looking forward, several promising research directions emerge:

1. Comprehensive real-world evaluation of advanced ensemble methods across volumes, addressing computational efficiency for runtime applications
2. ADCS extension to different filling systems and pump architectures, including non-advanced controller systems
3. Enhanced evaluation methodologies for complex models, improving offline indicator accuracy for ensemble methods
4. Investigation of hybrid approaches combining mechanical optimization with intelligent software compensation, potentially leveraging volume-specific model selection

ADCS success in improving PP performance indicates new precision fluid handling possibilities across industries. Advancing computational capabilities make sophisticated compensation methods increasingly feasible, suggesting a future where software-enhanced systems become industry standard, enabling more efficient, precise, and cost-effective solutions for pharmaceutical manufacturing and beyond.

ACKNOWLEDGMENT

The authors would like to thank Pharma Integration for providing with the opportunity to use the necessary equipment to conduct the experiments.

REFERENCES

[1] U.S. Food Drug Admin. (2015). *Allowable Excess Volume and Labeled Vial Fill Size in Injectable Drug and Biological Products: Guidance for Industry*. Accessed: Sep. 17, 2024. [Online]. Available: https://tinyurl.com/fda-allowable-excess-volume

[2] R. A. Tariq, R. Vashisht, A. Sinha, and Y. Scherbak, *Medication Dispensing Errors and Prevention*. Treasure Island, FL, USA: StatPearls Publishing, 2023. [Online]. Available: http://europepmc.org/books/NBK519065

[3] R. Patel, A. Vhora, D. Jain, R. Patel, D. Khunt, R. Patel, S. Dyawanapelly, and V. Junnuthula, "A retrospective regulatory analysis of FDA recalls carried out by pharmaceutical companies from 2012 to 2023," *Drug Discovery Today*, vol. 29, no. 6, Jun. 2024, Art. no. 103993.

[4] T. Natof and M. V. Pellegrini, *Food and Drug Administration Recalls*. Treasure Island, FL, USA: StatPearls, Jan. 2025.

[5] S. S. Farid, M. Baron, C. Stamatis, W. Nie, and J. Coffman, "Benchmarking biopharmaceutical process development and manufacturing cost contributions to R&D," *MAbs*, vol. 12, no. 1, Jan. 2020, Art. no. 1754999.

[6] J. Klespitz and L. Kovács, "Peristaltic pumps—A review on working and control possibilities," in *Proc. IEEE 12th Int. Symp. Appl. Mach. Intell. Inform. (SAMI)*, Jan. 2014, pp. 191–194.

[7] W. Shieu, D. Lamar, O. B. Stauch, and Y.-F. Maa, "Filling of high-concentration monoclonal antibody formulations: Investigating underlying mechanisms that affect precision of low-volume fill by peristaltic pump," *PDA J. Pharmaceutical Sci. Technol.*, vol. 70, no. 2, pp. 143–156, Mar. 2016.

[8] S. C. Gasoto, B. Schneider, and J. A. P. Setti, "Study of the pulse of peristaltic pumps for use in 3D extrusion bioprinting," *ACS Omega*, vol. 7, no. 28, pp. 24091–24101, Jul. 2022.

[9] M. Nuijten, "Pricing zolgensma—The world's most expensive drug," *J. Market Access Health Policy*, vol. 10, no. 1, Dec. 2022, Art. no. 2022353.

[10] P. Ferretti, C. Pagliari, A. Montalti, and A. Liverani, "Design and development of a peristaltic pump for constant flow applications," *Frontiers Mech. Eng.*, vol. 9, Jul. 2023, Art. no. 1207464, doi: 10.3389/fmech.2023.1207464.

[11] H. Wang, Z. Liu, and Z. Han, "HO2RL: A novel hybrid offline-and-online reinforcement learning method for active pantograph control," *IEEE Trans. Ind. Electron.*, vol. 72, no. 6, pp. 6286–6296, Jun. 2025.

[12] H. Wang, Z. Liu, G. Hu, X. Wang, and Z. Han, "Offline meta-reinforcement learning for active pantograph control in high-speed railways," *IEEE Trans. Ind. Informat.*, vol. 20, no. 8, pp. 10669–10679, Aug. 2024.

[13] D. Han, H. Qi, S. Wang, D. Hou, and C. Wang, "Adaptive step-size forward–backward pursuit and acoustic emission-based health state assessment of high-speed train bearings," *Structural Health Monitor.*, vol. 2024, Sep. 2024, Art. no. 14759217241271036, doi: 10.1177/14759217241271036.

[14] D. Privitera, S. Bellissima, and S. Bartolini, "Adaptive dosing control system through ARIMA model for peristaltic pumps," *IEEE Access*, vol. 11, pp. 99558–99572, 2023.

[15] D. Privitera, S. Bellissima, and S. Bartolini, "Exploring recurrent neural network approaches for enhancing peristaltic pump accuracy," *SSRN*, Jan. 2024. [Online]. Available: https://ssrn.com/abstract=4978229

[16] M. Tan, M. A. Merrill, V. Gupta, T. Althoff, and T. Hartvigsen, "Are language models actually useful for time series forecasting?" 2024, *arXiv:2406.16964*.

[17] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" 2022, *arXiv:2205.13504*.

[18] P. A. B. Andrade, M. Santos, J. E. Sierra-García, and J. P. Pazmiño-Piedra, "Comparison of LSTM, GRU and transformer neural network architecture for prediction of wind turbine variables," in *Proc. 18th Int. Conf. Soft Comput. Models Ind. Environ. Appl. (SOCO)*, P. García Bringas, H. Pérez García, F. J. M. de Pisón, F. M.Álvarez, A. T. Lora, Á. Herrero, J. L. C. Rolle, H. Quintián, and E. Corchado, Eds., Jan. 2023, pp. 334–343.

[19] X. Wei, G. Wang, B. Schmalz, D. F. T. Hagan, and Z. Duan, "Evaluation of transformer model and self-attention mechanism in the Yangtze river basin runoff prediction," *J. Hydrology: Regional Stud.*, vol. 47, Jun. 2023, Art. no. 101438. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214581823001258

[20] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022.

[21] H. Wu and D. Levinson, "The ensemble approach to forecasting: A review and synthesis," *Transp. Res. C, Emerg. Technol.*, vol. 132, Nov. 2021, Art. no. 103357.

[22] F. Jørgensen. (2008). *PD12 OEM Operators Manual*. [Online]. Available: https://archive.org/details/manualzilla-id-6886200

[23] J. Jeppesen. (2009). *MC100 Pump Control Module User's Manual*. Accessed: Aug. 2, 2024. [Online]. Available: https://www.wmfts.com/globalassets/literature/m-flexicon-mc100-profibus-en.pdf

[24] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with Python," in *Proc. 9th Python Sci. Conf.*, Jan. 2010, pp. 92–96.

[25] F. Chollet. (2015). *Keras*. Accessed: Aug. 2, 2024. [Online]. Available: https://keras.io

[26] 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.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[28] P. Lambert and F. Joergensen, "Accurate dispensing of biopharmaceuticals," *World Pumps*, vol. 2008, no. 498, pp. 22–24, Mar. 2008.

[29] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, vol. 14. Boca Raton, FL, USA: CRC Press, 2012.

[30] L. Chen, "A review of the applications of ensemble forecasting in fields other than meteorology," *Weather*, vol. 79, no. 9, pp. 285–290, Sep. 2024.

[31] U. Pasupulety, A. Abdullah Anees, S. Anmol, and B. R. Mohan, "Predicting stock prices using ensemble learning and sentiment analysis," in *Proc. IEEE 2nd Int. Conf. Artif. Intell. Knowl. Eng. (AIKE)*, Jun. 2019, pp. 215–222.

[32] Y. Li and Y. Pan, "A novel ensemble deep learning model for stock prediction based on stock prices and news," *Int. J. Data Sci. Anal.*, vol. 13, no. 2, pp. 139–149, Mar. 2022.

[33] T. Hashino, A. A. Bradley, and S. S. Schwartz, "Evaluation of bias-correction methods for ensemble streamflow volume forecasts," *Hydrol. Earth Syst. Sci.*, vol. 11, no. 2, pp. 939–950, Feb. 2007.

[34] J. Das, V. Manikanta, K. N. Teja, and N. V. Umamahesh, ''Two decades of ensemble flood forecasting: A state-of-the-art on past developments, present applications and future opportunities,'' *Hydrolog. Sci. J.*, vol. 67, no. 3, pp. 477–493, Feb. 2022.

[35] J. Wu, Z. Wang, J. Dong, X. Cui, S. Tao, and X. Chen, ''Robust runoff prediction with explainable artificial intelligence and meteorological variables from deep learning ensemble model,'' *Water Resour. Res.*, vol. 59, no. 9, p. 2023, Sep. 2023.

[36] M. M. Ghiasi and S. Zendehboudi, ''Application of decision tree-based ensemble learning in the classification of breast cancer,'' *Comput. Biol. Med.*, vol. 128, Jan. 2021, Art. no. 104089.

[37] G. Marques, A. Ferreras, and I. De La Torre-Diez, ''An ensemble-based approach for automated medical diagnosis of malaria using EfficientNet,'' *Multimedia Tools Appl.*, vol. 81, no. 19, pp. 28061–28078, Aug. 2022.

[38] L. Breiman, ''Bagging predictors,'' *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[39] Y. Freund, ''Boosting a weak learning algorithm by majority,'' *Inf. Comput.*, vol. 121, no. 2, pp. 256–285, Sep. 1995.

[40] P. Smyth and D. Wolpert, ''Linearly combining density estimators via stacking,'' *Mach. Learn.*, vol. 36, nos. 1–2, pp. 59–83, Jul. 1999.

[41] L. Breiman, ''Random forests,'' *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[42] X. Wang, H. Jiang, M. Mu, and Y. Dong, ''A dynamic collaborative adversarial domain adaptation network for unsupervised rotating machinery fault diagnosis,'' *Rel. Eng. Syst. Saf.*, vol. 255, Mar. 2025, Art. no. 110662. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0951832024007336

[43] L. K. Hansen and P. Salamon, ''Neural network ensembles,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 10, pp. 993–1001, Jan. 1990.

[44] K. D. Yamada, F. Lin, and T. Nakamura, ''Developing a novel recurrent neural network architecture with fewer parameters and good learning performance,'' *Interdiscipl. Inf. Sci.*, vol. 27, no. 1, pp. 25–40, 2021, doi: 10.4036/iis.2020.R.01.

DAVIDE PRIVITERA received the bachelor's and master's (cum laude) degrees in computer and automation engineering from the University of Siena, Italy. Since 2016, he has been an Automation Engineer at Pharma Integration, where he contributes to the development of innovative robotics solutions in next-generation filling systems for pharmaceutical. In 2021, he embarked on a doctoral program at the University of Siena, focusing on the development of intelligent systems for Industry 4.0. His research interest includes studying methods to improve the accuracy of dosing systems used in the pharmaceutical industry.

ALESSANDRO MECOCCI is currently a Full Professor of artificial vision with the University of Siena, where he heads the Vision and SMART Sensors Laboratory (VISLab). He was a member of the National Security Observatory under the Italian Ministry of Defense, from 2006 to 2010, and managed the approval of highway traffic control systems for the Ministry of Infrastructure and Transport, from 2008 to 2016. He maintained a 28-year collaboration with Autostrade per l'Italia developing artificial vision systems for traffic safety. In 2019, he co-patented the Drone-Box with the Italian Railway Network Company. He holds seven patents in signal and image processing and has co-founded seven startups. His research interests include multisensor monitoring, action recognition, the IoT, 3D object recognition, and real-time anomaly detection for security applications. He received the ''In onore dell'Italia che ci onora'' Prize, in 2024. He was an Italian National Delegate for the Telematics Program Committee, Brussels, from 1996 to 2000, and directed the 3rd Consortium of the Tuscany Region for technology transfer, from 1997 to 2006.

SANDRO BARTOLINI is currently an Associate Professor with the Department of Information Engineering and Mathematical Sciences, University of Siena, Italy. He has led and participated in various research and development projects. His main research interests include high-performance chip multiprocessors (CMPs), new approaches to productive programming of heterogeneous architectures (CPUs and GPUs), integrated photonics for CMPs, feedback-driven compiler optimizations for cache hierarchy performance and low power, and hardware accelerators. He is an active member of the HiPEAC NoE. He is Associate Editor of the *EURASIP Journal of Embedded Computing*. He has been a Co-Guest Editor of *Transactions on High Performance Architectures and Compilation* (Springer) journal, *ACM SigArch Computer Architecture Newsletter*.