International Journal of Science, Technology, Engineering and Mathematics
Volume 6 Issue 2 June 2026
DOI: https://doi.org/10.53378/ijstem.353345

An empirical analysis of Bayesian-optimized boosting ensembles for medical relief demand forecasting

[1]Roman B. Villones, [2]Jonilo C. Mababa, [3]Jovy Jay D. Cabrera & [4]Jaime P. Pulumbarit

Abstract

In humanitarian logistics, the planning of medical relief supply is sensitive to accurate demand forecasting due to its intermittent and volatile demand pattern that constrain the usefulness of conventional statistical techniques. This paper provides an empirical analysis of the increase in ensemble learning models that are optimized through Bayesian searching of hyperparameters to predict demand of medical relief. Using the Team Data Science Process (TDSP) framework, a quantitative methodology is used to evaluate the performance of the model based on predictive accuracy and robustness performance with extreme values, and generalizability on a real-world dataset in the National Capital Region (NCR), Philippines. There are five ensemble models such as AdaBoost, CatBoost, Gradient Boosting, LightGBM, and XGBoost. Findings indicate that Bayesian optimization produces a tangible performance gain, especially on Gradient Boosting and LightGBM. The CatBoost model produces the lowest RMSE, WAPE, and MASE, and the most consistent cross-validation results, which means that it is more accurate and stable in the model tested. On the contrary, XGBoost and AdaBoost exhibits relatively poorer performance and low robustness. Although the results illustrate the efficiency of optimized boosting ensembles to handle complex and irregular demand shapes, the research study is limited by the coverage of datasets, possible temporal leakage and lack of actual deployment. Thus, it is possible to arrive at conclusions only based on the assessed data and experimental conditions.

Keywords: *boosting ensemble learning, Bayesian hyperparameter optimization, medical relief supply chain, humanitarian logistics, machine learning*

Article History:

Received: *January 18, 2026* ***Revised***: *April 3, 2026*
Accepted: *April 8, 2026* ***Published online***: *April 18, 2026*

Suggested Citation:
Villones, R.B., Mababa, J.C., Cabrera, J.J.D. & Pulumbarit, J.P. (2026). An empirical analysis of Bayesian-optimized boosting ensembles for medical relief demand forecasting. *International Journal of Science, Technology, Engineering and Mathematics*, 6(2), 1-21. https://doi.org/10.53378/ijstem.353345

About the author:
[1]Corresponding author. Master in Information Technology. Graduate School Department, La Consolacion University Philippines, Philippines. Email: roman.villones@email.lcup.edu.ph
[2]Doctor of Information Technology & PhD in Educational Leadership and Management. Graduate School Department, La Consolacion University Philippines, Philippines. Email: jonilo.mababa@email.lcup.edu.ph
[3]Doctor of Information Technology. Graduate School Department, La Consolacion University Philippines, Philippines. Email: jovyjay.cabrera@email.lcup.edu.ph
[4]Doctor of Information Technology. Graduate School Department, La Consolacion University Philippines, Philippines. Email: jaime.pulumbarit@email.lcup.edu.ph

1. Introduction

Accurate demand forecasting is increasingly critical for managing supply chains in dynamic and complex environments, and as traditional statistical methods often struggle to capture nonlinear patterns and sudden demand shifts (Dalimunthe et al., 2023). The adoption of machine learning (ML) and ensemble learning methods has grown, as these approaches can effectively extract the data-driven patterns and improved the forecasting accuracy (Özüpak et al., 2025). Ensemble strategies, such as boosting algorithms can combine multiple weak learners to produce stronger predictive performance, correcting errors in iteratively, and capturing complex relationships that individual models may miss (Li et al., 2025).

In humanitarian logistics, particularly for medical relief supply planning, the accurate demand forecasting is vital because underestimating needs can lead to shortages and exacerbate human suffering (Ahatsi & Olanrewaju, 2025). Despite its importance, the empirical studies applying advanced ML and ensemble methods to humanitarian supply chains remain limited (Kumar et al., 2025). The use of Bayesian hyperparameter optimization has demonstrated its effectiveness in improving the model performance by efficiently identifying optimal configurations, and yet its application in medical relief demand forecasting has not been fully explored (Li et al., 2025).

This study conducts an empirical analysis of boosting ensemble learning models optimized through Bayesian hyperparameter search for forecasting medical relief demand. Specifically, it aims to evaluate model performance under real-world conditions characterized by intermittent and zero-heavy demand patterns, which are common in humanitarian logistics. The study is guided by the following research questions:

1. To what extent does Bayesian hyperparameter optimization improve predictive accuracy compared to baseline models?
2. How do optimized boosting ensemble models perform in terms of stability across validation folds, consistency under extreme demand values, and resilience to zero-demand observations?
3. Which ensemble model provides the most reliable balance between accuracy and robustness within the given dataset and evaluation setting?

The contributions of this research are threefold. First, it provides systematic empirical benchmarking of Bayesian-optimized boosting ensembles in a critical humanitarian application, with performance evaluated using quantitative error metrics and cross-validation

stability. Second, it demonstrates the measurable impact of Bayesian hyperparameter optimization, including reductions in prediction error and improvements in generalization, particularly under irregular and sparse demand conditions. Third, it offers practical insights for decision-makers in medical relief supply planning by linking forecasting improvements to potential operational outcomes while acknowledging that these impacts require further real-world validation.

2. Literature Review

Recent research shows that ML models are increasingly central to demand forecasting in supply chain management by offering a performance improvements over traditional statistical approaches and effectively capturing nonlinear and dynamic patterns that classical models like ARIMA that often fail to model (Douaioui et al., 2024; Mohammed & Mandal, 2024). A review of 119 studies highlights the rapid adoption of AI-driven forecasting techniques with ensemble models are recurrent neural networks such as LSTM, and hybrid approaches demonstrating superior predictive accuracy and supply chain responsiveness compared to traditional methods (Douaioui et al., 2024; Ahmed et al., 2025). Despite these advances, the challenges remain, including the data quality issues, high computational costs, difficulties in model interpretability, and the need for context-specific evaluation to ensure robustness. It is defined in this study as consistent performance across validation folds, stability under extreme values, and resilience to zero or sparse demand conditions across different industries and demand scenarios (Mohammed & Mandal, 2024; Ahmed et al., 2025).

2.1. Limited Focus on Humanitarian Forecasting

Although there is extensive work on ML for demand forecasting in commercial supply chains, the research on demand forecasting within humanitarian logistics remains comparatively scarce and underdeveloped relative to the broader supply chain and AI literature with only a modest number of studies directly addressing predictive methods in humanitarian operations (Pantiris et al., 2025; Efe, 2022). Humanitarian demand forecasting involves intermittent, volatile, and highly uncertain demand patterns that pose unique challenges and not fully addressed by many traditional and AI-based models (Pantiris et al., 2025). Recent analyses emphasize that AI and predictive analytics adoption in humanitarian logistics is still emerging, constrained by data quality issues, implementation barriers, and uneven deployment

across regions (Cao, 2023). While artificial intelligence and big data analytics have been shown to enhance predictive capabilities and supply chain responsiveness in crisis contexts. The volume of research is explicitly focused on demand forecasting models for humanitarian relief and it remains limited, indicating significant gaps, and opportunities for future work that develops a robust and context-specific forecasting frameworks tailored to the characteristics of humanitarian operations (Pantiris et al., 2025; Efe, 2022; Altay & Narayanan, 2022).

2.2. Lack of Optimized Ensemble Learning Studies

While ML and ensemble approaches are widely reported to improve forecast accuracy, in a few studies, they systematically apply the advanced hyperparameter optimization methods to these models in demand forecasting research (Douaioui et al., 2024). A variety of ML techniques have been tested its potential performance for improvements and achievable through efficient hyperparameter tuning remain under-examined, particularly in peer-reviewed studies that rigorously compare tuned versus untuned ensemble models (Mohammed & Mandal, 2022). Recent studies suggest that integrating hyperparameter optimization with ensemble and machine learning models could further enhance predictive accuracy and robustness (Douaioui et al., 2024; Mohammed & Mandal, 2024), yet empirical applications in both commercial and humanitarian supply chain forecasting remain limited.

2.3. Insufficient Evaluation of Model Robustness under Volatile Demand

Most forecasting literature focuses on forecast accuracy under relatively stable conditions and its limited empirical evidence on how ML models perform under irregular and volatile demand patterns on typical humanitarian supply chains (Chandran et al., 2024). Ongoing challenges related to time-series complexity and the need for models that adapt to sudden demand shifts and noise (Douaioui et al., 2024). Empirical research on demand forecasting under volatile market conditions also finds that conventional and some machine learning models struggle to maintain accuracy when demand is unstable or influenced by external shocks and underscoring the importance of adaptive or hybrid modeling strategies to cope with nonlinear dynamics (Aldahmani et al., 2024; Chandran & Khan, 2024). These findings suggest that, despite advances in ML forecasting the robust performances interpreted as stability across validation folds, reduced error variance at higher demand levels, and

resilience to zero-demand observations that remains underexplored, particularly when compared with stable forecasting scenarios.

2.4. Gap in Medical Relief Demand Forecasting

Although broad humanitarian logistics research discusses planning, resource allocation, and operational coordination under the specific forecasting studies for medical relief supplies and other critical relief items are notably scarce. Some literature in the humanitarian domain focusing on conceptual frameworks, logistical coordination, and disaster prediction rather than a data-driven and optimized machine-learning demand forecasting (Salamian et al., 2024). Forecasting within humanitarian operations over nearly three decades, it highlights that forecasting research in this context is under-represented compared with commercial supply chain forecasting, and few studies explicitly investigate the advanced predictive analytics for relief needs such as medical supplies where forecast errors can have severe human consequences (Altay & Narayanan, 2022). Beyond foundational reviews, research applying machine learning models to humanitarian demand forecasting is limited. They often rely on conceptual or hybrid methodological proposals rather than systematic and empirically validated demand models (Salamian et al., 2024). This reveals a clear application of a gaps in a domain where accurate prediction directly impacts relief effectiveness and human outcomes.

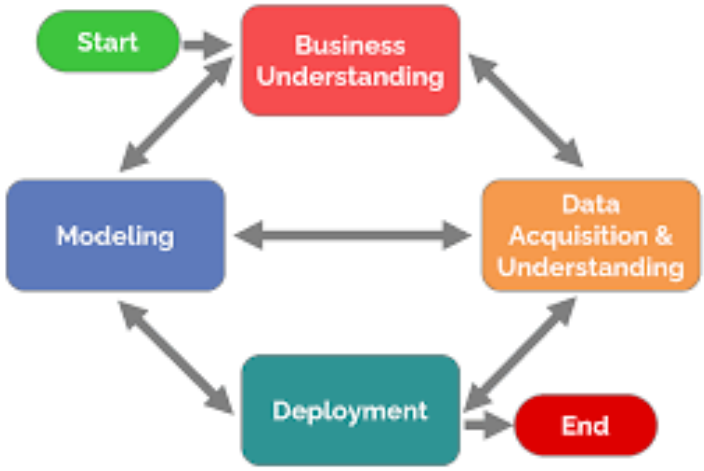
While ML models are increasingly central to demand forecasting research, the key gaps remain including the underexplored use of advanced optimization techniques for ensemble models, limited empirical evaluation under intermittent, zero-inflated demand conditions, and insufficient focus on humanitarian and medical relief forecasting. This study addresses these gaps by conducting an empirical analysis of Bayesian-optimized boosting ensembles.

3. Methodology

This study adopts a quantitative and empirical research design to evaluate the performance of boosting ensemble machine learning models optimized via Bayesian hyperparameter search forecasting demand for medical relief supplies. The Team Data Science Process (TDSP) framework focuses on measuring the model accuracy, robustness, and predictive reliability under realistic humanitarian logistics conditions (Vance, 2021; Stelmaszak & Kline, 2023).

Figure 1

Team Data Science Process (TDSP) framework



***Source**:* Adapted from Saltz and Hotz (2020)

Figure 1 shows the Team Data Science Process (TDSP) framework that guides the data science lifecycle in ensuring the systematically data modeling and evaluation.

Business understanding. In the Team Data Science Process (TDSP) framework, the Business Understanding stage involves defining business objectives, articulating success criteria, and framing the predictive analytics problem to ensure alignment with organizational goals (de Mast & Lokkerbol, 2024). The objective of this stage is to define the problem and establish forecasting goals. For this study, the key goal is to develop accurate and robust demand forecasts for medical relief supplies in humanitarian operations. The stage also identifies performance objectives by its accuracy and robustness and the evaluation metrics that are relevant to decision-making.

Data acquisition and understanding. It involves gathering and preparing relevant data to assess data structure, quality, and suitability for modeling (Shafiq et al., 2021). It is often referred to as data preprocessing or data wrangling and it is the critical phase where raw data is cleaned, integrated, transformed, and structured into a format suitable for analysis and directly impacting the model accuracy and reliability (Fernandes et al., 2023).

The dataset used in this study was collected from relevant humanitarian and medical supply sources within the National Capital Region (NCR), Philippines from the historical records of medical relief demand in a disaster-prone and resource-constrained setting. It consists of more than 76,000 observations and 10 variables both inventory-related and demand-driven attributes. Key variables include inventory level, relief aid (target variable), inventory re-ordered, simple ratio method, weighted demand index, normalized score, and overall demand score, alongside categorical features such as item ID, district, and calamities. The dataset contains no missing values, ensuring completeness; however, its geographic scope is limited to NCR, which may affect generalizability to other regions with different demand dynamics. The target variable, relief aid, represents the quantity of medical supplies required and serves as the basis for demand forecasting. The dataset includes moderate diversity in categorical features, as well as substantial variability in numerical attributes that reflect a heterogeneous demand pattern typical in humanitarian logistics.

For preprocessing, categorical variables such as item ID, district, and calamities were encoded using one-hot encoding to avoid ordinal bias and better represent categorical distinctions. Numerical features were scaled using StandardScaler to normalize feature distributions and improve model convergence, particularly for boosting algorithms. The dataset was then partitioned into training and testing sets using an 80:20 split. However, no explicit time-based splitting was applied, which introduces a potential risk of data leakage, particularly if temporal dependencies exist in demand patterns. This limitation may lead to optimistic performance estimates and should be addressed in future work through time-aware validation strategies. Notably, no feature engineering or feature selection techniques were applied, meaning the models relied solely on the original variables provided. While this ensures transparency and reproducibility, it may limit the models' ability to capture more complex relationships or latent demand drivers.

Several validity threats should be considered. First, the dataset's confinement to NCR limits external validity. Second, the absence of temporal validation may affect internal validity due to potential leakage. Third, the presence of zero-demand values impacts the reliability of certain evaluation metrics, particularly MAPE. Finally, while the dataset is complete, its representativeness of broader humanitarian logistics scenarios remains uncertain, as it may not fully capture variability across different disaster types, regions, or supply chain conditions.

3.3. Modeling

Modeling stage involves selecting appropriate algorithms, which includes training models on prepared data, optimizing their parameters, and evaluating predictive performance against validation criteria (Shafiq et al., 2021). Recent empirical studies continue to demonstrate the strong performance of these methods across diverse domains. For example, a comparison of XGBoost, LightGBM, and CatBoost in construction risk prediction and concrete strength estimation show consistent improvements in accuracy and robustness relative to alternative models when handling nonlinear relationships and feature interactions (Alfath et al., 2025; Mustapha et al., 2024). Similarly, ensemble machine learning frameworks that combine boosted models with deep learning architectures have achieved superior forecast accuracy for environmental time-series problems and it illustrates the flexibility and strength in complex predictive settings (Ahn et al., 2023).

Empirical studies comparing Bayesian hyperparameter search in classification and regression tasks show that Bayesian hyperparameter tuning typically yields better performance with fewer iterations, particularly for complex ensemble and tree-based models by exploiting the structure of the search space more effectively than grid or random search methods (Victoria & Maragatham, 2020). Multiple boosting ensemble models like XGBoost, LightGBM, CatBoost, Gradient Boosting, and AdaBoost are implemented to forecast medical relief demand. Bayesian hyperparameter search is utilized to optimize key hyperparameters, including learning rate, tree depth, and number of estimators, to maximize predictive performance while mitigating overfitting. Model training is conducted iteratively using 5-fold cross-validation to ensure robustness and generalizability across different data partitions.

In the evaluation phase of modeling, the models are developed and assessed to determine their predictive performance, generalizability, and alignment with the business or operational objectives defined during business understanding (Hodson, 2022). Model performance is assessed using quantitative metrics, including Mean Absolute Percentage Error (MAPE), Symmetric MAPE (sMAPE), Median Absolute Error (Median AE), Mean Bias Error (MBE), Explained Variance, and Root Square Mean Error (RSME) which provide complementary perspectives on prediction accuracy, bias, and variance. Comparative analysis across these metrics enables identification of the most accurate and robust model by ensuring reliability and consistency in forecasting results (Makridakis et al., 2022).

3.4. Deployment

In this stage, the models that have been developed, tuned, and evaluated are transitioned into production environments to support decision-making. Deployment involves integrating predictive models into operational systems and monitoring model performance over time (Sarker, 2021; Katya, 2023). This study emphasizes the practical dissemination and usability of the developed forecasting approach rather than full system-level integration. The resulting predictive framework is intended to be made accessible to humanitarian organizations and relevant stakeholders, allowing them to apply it within their own supply chain and decision-making processes. Forecast outputs are presented through visualizations and analytical summaries to demonstrate how predictions can support inventory planning and resource allocation.

Instead of direct deployment within a single operational system, this approach enables flexible adoption where organizations can integrate the forecasting approach into their existing workflows and logistics systems based on their specific needs and constraints. This is particularly relevant in humanitarian contexts where infrastructure, data availability, and technical capacity vary across organizations.

4. Findings and Discussion

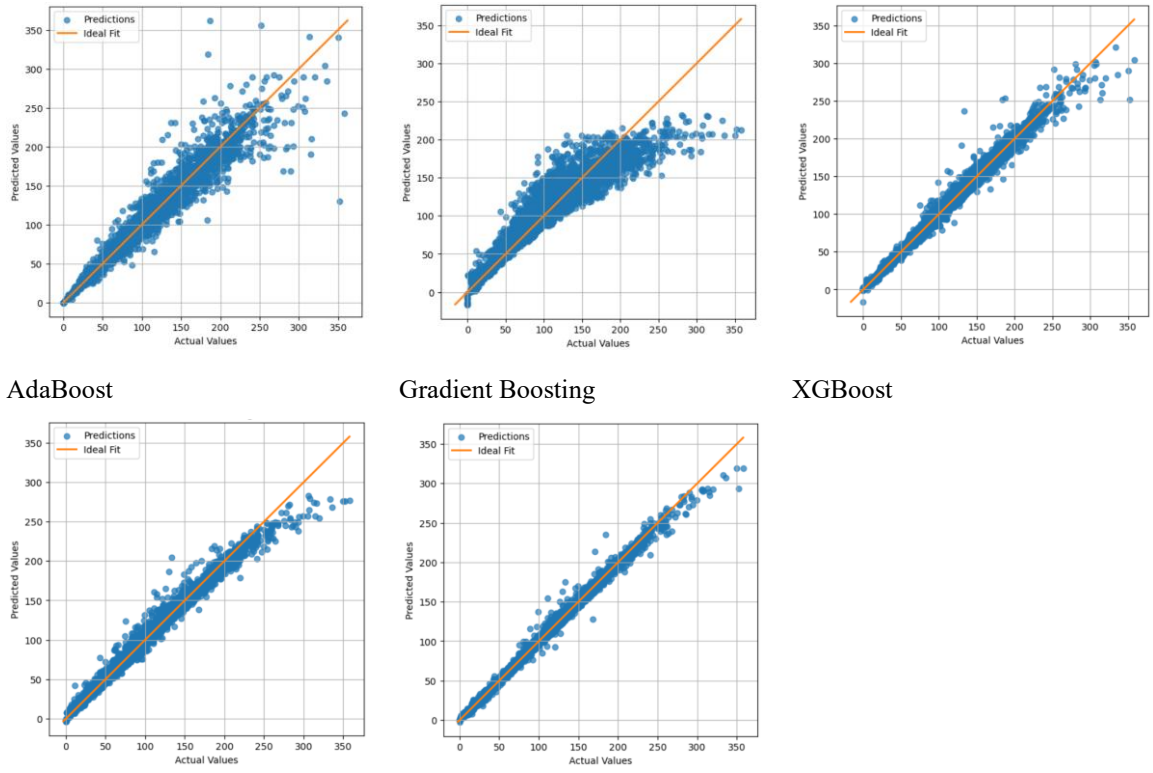
Figure 2 presents the predicted versus actual values for AdaBoost, Gradient Boosting, XGBoost, LightGBM, and CatBoost models. All models demonstrate a strong alignment between predicted and actual values along the ideal diagonal line and indicate their ability to effectively capture the underlying relationships within the dataset and achieve good generalization performance. Across all models, predictions within the lower and mid-range actual values exhibit minimal dispersion. These reflect high predictive accuracy and stable model behavior in these regions; however, as the magnitude of actual values increases, a widening spread of predictions becomes evident. The reduced predictive reliability at higher target values suggests the presence of heteroscedasticity where prediction errors increase with the scale of the response variable.

A common pattern observed among the models is slight to moderate underestimation at higher actual values. This may be attributed to factors such as limited representation of extreme observations in the training data and the regularization mechanisms inherent in ensemble learning techniques. Gradient Boosting shows more noticeable compression and

deviation at higher values, while AdaBoost exhibits relatively higher variance in extreme cases.

Figure 2

Predicted vs actual values across ensemble learnings



AdaBoost | Gradient Boosting | XGBoost

LightGBM | CatBoost

Among the models, XGBoost demonstrates the strongest overall consistency, with tighter predictions around the ideal fit line and only mild heteroscedasticity. LightGBM and CatBoost also show stable and reliable performance, with only minor dispersion and marginal degradation at higher values. Despite the presence of a few outliers across all models, the concentration of data points near the diagonal line confirms that the ensemble approaches maintain robust predictive capability.

Table 1 demonstrates that CatBoost achieved the best overall performance, ranking first across the evaluated metrics. It obtained the lowest RMSE (2.7219), outperforming the next best model, XGBoost (4.211169), by a substantial margin of approximately 1.49 RMSE units. This indicates significantly lower prediction error magnitude. In terms of explained variance,

CatBoost also achieved the highest value (0.9961), exceeding XGBoost (0.9906) by 0.0055, reflecting a stronger ability to capture variability in the data.

Table 1

Comparative performance of boosting ensemble models

Ensemble Learnings	MAPE	sMAPE	Median AE	MBE	Explained Variance	RSME	Rank
CatBoost	inf	3.509547	1.194281	0.000725	0.996125	2.721914	1
XGBoost	inf	4.344884	1.711166	0.031552	0.990602	4.211169	2
LightGBM	inf	4.901453	1.824991	0.017941	0.988793	4.673677	3
AdaBoost	inf	3.11679	1.000000	-0.088849	0.975674	7.761168	4
Gradient Boosting	inf	8.969554	3.476608	0.036789	0.925571	11.998620	5

XGBoost ranked second, with an RMSE of 4.2112, which is 0.46 lower than LightGBM (4.6737), confirming its relatively better predictive accuracy among the remaining models. It also demonstrated competitive performance in Median Absolute Error (1.7112), improving over LightGBM (1.8250) by 0.1138. However, its Mean Bias Error (0.0316) suggests a slight tendency toward overestimation compared to CatBoost (0.0007), which is nearly unbiased.

LightGBM ranked third, showing moderate performance with RMSE (4.6737) and explained variance (0.9888), though it trails XGBoost by 0.46 in RMSE and 0.0018 in explained variance. Its Median Absolute Error (1.8250) is also slightly higher than XGBoost, indicating marginally less precise predictions. AdaBoost presents an interesting case. While it achieved the lowest sMAPE (3.1168) and lowest Median Absolute Error (1.0000), suggesting strong performance in relative and median-based error metrics, its RMSE is significantly higher (7.7612) approximately 5.04 higher than CatBoost.

Gradient Boosting performed the weakest overall, with the highest RMSE (11.9986), which is more than 4 times higher than CatBoost. It also recorded the lowest explained variance (0.9256), trailing CatBoost by 0.0705, indicating reduced capability in explaining data variability. Furthermore, its sMAPE (8.9696) is substantially higher than all other models, confirming poorer predictive accuracy.

Notably, all models reported infinite MAPE values, likely due to the presence of zero or near-zero actual values. This reinforces the importance of relying on alternative metrics such

as Mean Absolute Scaled Error (MASE) and Weighted Absolute Percentage Error (WAPE) for more stable evaluation.

Table 2

Performance analysis using MASE and WAPE

Ensemble Learnings	WAPE (%)	MASE	RSME	Rank
CatBoost	1.901876	0.034852	2.721914	1
XGBoost	2.82795	0.051823	4.211169	2
LightGBM	3.090806	0.05664	4.673677	3
AdaBoost	3.325944	0.060949	7.761168	4
Gradient Boosting	7.610318	0.139461	11.998620	5

Table 2 presents the results with reinforce the superiority of CatBoost, which achieved the lowest WAPE (1.9019%) and lowest MASE (0.0349) among all models. Compared to the next best model, XGBoost, CatBoost improves WAPE by approximately 0.93 percentage points and reduces MASE by 0.0170. This indicates that CatBoost not only minimizes overall percentage error but also performs significantly better than the naïve baseline model.

XGBoost ranked second, with WAPE of 2.8280% and MASE of 0.0518, outperforming LightGBM by 0.26 percentage points in WAPE and 0.0048 in MASE. This confirms that XGBoost maintains stronger predictive accuracy and efficiency relative to LightGBM. Then, LightGBM follows closely in third place, with WAPE (3.0908%) and MASE (0.0566), showing only a modest decline compared to XGBoost. The small differences suggest comparable performance, although LightGBM consistently produces slightly higher errors.

AdaBoost, while previously showing competitive median-based metrics, demonstrates a weaker performance with WAPE of 3.3259% and MASE of 0.0609. This represents a 1.42 percentage point increase in WAPE compared to CatBoost, and a 0.0261 increase in MASE. On the other hand, Gradient Boosting exhibits the poorest performance, with WAPE of 7.6103% and MASE of 0.1395. Its WAPE is 5.71 percentage points higher than CatBoost, while its MASE is approximately 4 times larger. These substantial gaps confirm that AdaBoost and Gradient Boosting has significantly higher forecast errors and weaker performance relative to both the dataset scale and the naïve benchmark.

All models have MASE values below 1, indicating that they outperform the naïve forecasting approach. However, the magnitude of improvement varies considerably, with CatBoost demonstrating the most efficient and reliable forecasting capability.

Table 3

RMSE performance of baseline and tuned boosting ensemble models

Ensemble Learnings	RMSE (Baseline)	RMSE (Tuned)	Improvement	Rank
CatBoost	2.737	2.456605	10.24%	1
LightGBM	4.654318	2.957432	36.46%	2
Gradient Boosting	11.994914	3.128849	73.92%	3
XGBoost	4.262326	3.322875	22.04%	4
AdaBoost	6.826899	7.022274	-2.86%	5

Table 3 shows that the hyperparameter tuning significantly impacted the performance of most ensemble models, as reflected in the RMSE reductions. CatBoost remained the top-performing model, reducing RMSE from 2.737 to 2.4566, representing a 10.24% improvement. Although the percentage improvement is smaller compared to some other models, its absolute RMSE remains the lowest and confirming it as the most accurate and stable model after optimization. LightGBM demonstrated a substantial gain, with RMSE decreasing from 4.6543 to 2.9574, yielding a 36.46% improvement. This improvement elevates LightGBM from a moderate baseline performer to a strong competitor, approaching XGBoost and Gradient Boosting levels of predictive accuracy.

Gradient Boosting exhibited the largest relative improvement of 73.92%, reducing RMSE from 11.9949 to 3.1288. Despite this dramatic reduction, its post-optimization RMSE is still slightly higher than CatBoost (0.672 lower than CatBoost) but reflects a major enhancement in predictive reliability. XGBoost improved its RMSE from 4.2623 to 3.3229, a 22.04% reduction, maintaining its position as a competitive model while still ranking below CatBoost and LightGBM in absolute accuracy. In contrast, AdaBoost showed a negative improvement of −2.86%, with RMSE slightly increasing from 6.8269 to 7.0223, suggesting that the applied tuning did not benefit the model and may have slightly overfitted or destabilized it.

Table 4 presents the 5-fold cross-validation confirms the robustness and stability of the ensemble models after hyperparameter optimization. CatBoost again demonstrates superior

performance, achieving the lowest mean RMSE (2.5821) with a small standard deviation (0.1123). This indicates not only high predictive accuracy but also consistent performance across folds, reinforcing its reliability for generalization on unseen data.

Table 4

5-Fold cross-validation performance of tuned models

Ensemble Learnings	RMSE Mean	RMSE std	Rank
CatBoost	2.582133	0.112313	1
LightGBM	2.989929	0.182611	2
Gradient Boosting	3.008326	0.161016	3
XGBoost	3.27716	0.161452	4
AdaBoost	7.060341	0.163669	5

LightGBM ranks second, with a mean RMSE of 2.9899 and standard deviation of 0.1826, showing good accuracy but slightly higher variability across folds compared to CatBoost. The difference in mean RMSE (~0.4078) highlights CatBoost's clear edge in absolute predictive performance. Gradient Boosting closely follows LightGBM with a mean RMSE of 3.0083 and standard deviation of 0.1610. While the RMSE is slightly higher than LightGBM (~0.0184), the lower standard deviation suggests marginally more stable predictions across folds.

XGBoost has a mean RMSE of 3.2772 with a standard deviation of 0.1615, positioning it fourth. Although competitive, its higher mean RMSE indicates reduced predictive accuracy compared to CatBoost, LightGBM, and Gradient Boosting. AdaBoost remains the weakest model, with a mean RMSE of 7.0603 and standard deviation of 0.1637; its predictive errors are substantially larger and less suitable for reliable forecasting, even with tuning.

Table 5 display the optimized hyperparameters reveal how each ensemble model adapts to the structure and complexity of the dataset. CatBoost, which achieved the best overall performance, utilizes a relatively deep tree structure (depth = 7) and a high number of iterations (900) combined with a moderately high learning rate (0.2). In this configuration, CatBoost benefits from capturing complex nonlinear relationships while maintaining efficient learning through controlled iteration growth. The balance between depth and iterations likely contributes to its low RMSE and stable generalization, as observed in both testing and cross-validation results.

Table 5

Optimized hyperparameters of tuned models

Ensemble Learnings	Best Parameters
CatBoost	'depth': 7, 'iterations': 900, 'learning_rate': 0.2, 'verbose': 0
LightGBM	'learning_rate': 0.1, 'n_estimators': 400, 'num_leaves': 63
Gradient Boosting	'learning_rate': 0.08, 'max_depth': 6, 'n_estimators': 400
XGBoost	'learning_rate': 0.08, 'max_depth': 6, 'n_estimators': 400
AdaBoost	'depth': 4, 'iterations': 1200, 'learning_rate': 0.5

LightGBM employs 63 leaves and 400 estimators with a learning rate of 0.1, which indicates a preference for leaf-wise tree growth with moderate complexity. This configuration enables efficient learning and explains its strong improvement after tuning. However, compared to CatBoost, its slightly higher error metrics suggest that the model may be less effective in capturing extreme variations in demand.

Both Gradient Boosting and XGBoost converge to similar parameter configurations (learning rate = 0.08, max depth = 6, n_estimators = 400), reflecting a shared optimization strategy of moderate depth with controlled learning rates. This setup helps prevent overfitting while improving predictive accuracy. Notably, Gradient Boosting's significant performance improvement that indicates that it is highly sensitive to hyperparameter tuning, transforming it from the weakest baseline model into a competitive alternative. In contrast, AdaBoost adopts a shallower depth (4) but compensates with a very high number of iterations (1200) and a high learning rate (0.5). This configuration suggests reliance on iterative error correction, but the high learning rate may lead to instability and over-adjustment. This is consistent with its observed performance, where tuning did not improve RMSE and resulted in slight degradation, due to limited robustness for this dataset.

Hyperparameter tuning significantly improves most models, especially Gradient Boosting and LightGBM, while CatBoost maintains superior performance and stability. These results highlight the importance of careful model selection, tuning, and validation to support more reliable and context-sensitive forecasting for humanitarian logistics, even if direct operational impacts remain to be measured.

Table 6 shows that Bayesian hyperparameter optimization substantially improves predictive accuracy, particularly for Gradient Boosting (73.9% RMSE reduction) and

LightGBM (36.5%), while CatBoost maintains the lowest overall error with a 10.24% improvement.

Table 6

Summary of results addressing the research questions

Research Question	Key Findings	Evidence
1. To what extent does Bayesian hyperparameter optimization improve predictive accuracy compared to baseline models?	Significant improvement in predictive accuracy across most models	Gradient Boosting: −73.9% RMSE; LightGBM: −36.5%; XGBoost: −22.0%; CatBoost: −10.24%; AdaBoost: +2.86% (decline).
2. How do optimized boosting ensemble models perform in terms of stability across validation folds, consistency under extreme demand values, and resilience to zero-demand observations?	Optimized models show improved stability, but challenges remain at extreme values	CatBoost: RMSE = 2.58, std = 0.112 (most stable); All models: increased dispersion at high values; All models: MASE < 1 (better than naïve); CatBoost: WAPE = 1.90%, MASE = 0.035.
3. Which ensemble model provides the most reliable balance between accuracy and robustness within the given dataset and evaluation setting?	CatBoost provides the most reliable overall performance	Lowest RMSE (2.58), WAPE (1.90%), MASE (0.035), and lowest variability across folds.

In terms of robustness, optimized models demonstrate improved stability across validation folds, with CatBoost achieving the lowest RMSE (2.58) and variability (std = 0.112). However, all models exhibit increased prediction dispersion at higher demand values and sensitivity to zero-demand observations, as reflected by infinite MAPE values. Using alternative metrics, all models outperform naïve baselines (MASE < 1), with CatBoost showing the strongest resilience (WAPE = 1.90%, MASE = 0.035). The CatBoost provides the most reliable balance between accuracy and robustness within the evaluated dataset, while LightGBM and Gradient Boosting serve as competitive alternatives after optimization.

5. Conclusion

Within the scope of this study and the tested dataset, the boosting ensembles demonstrate strong potential for forecasting medical relief demand. CatBoost emerged as the best-performing model, achieving the lowest RMSE, WAPE, and MASE after hyperparameter

tuning and cross-validation, reflecting an improved predictive accuracy and stability in this evaluation setting. Gradient Boosting and LightGBM also showed substantial gains with RMSE reductions respectively that highlights the appropriate tuning can make them viable alternatives depending on operational constraints such as compute resources, interpretability, and data availability during disaster scenarios.

Despite these promising results, limitations must be acknowledged. The dataset was geographically limited to the National Capital Region, Philippines. The time order was not explicitly enforced in all evaluations, and zero-demand observations led to “infinite” MAPE values and requiring a reliance on WAPE and MASE for robust assessment. Additionally, operational impacts such as stockout reduction or safety stock optimization improvements were not measured, meaning the practical applicability of these forecasts remains tentative.

The findings emphasize the importance of careful model selection, hyperparameter tuning, and rigorous validation within the tested context. While CatBoost is recommended as the primary model for its superior performance, Gradient Boosting and LightGBM remain feasible alternatives if appropriately adjusted, supporting data-driven decision-making in humanitarian logistics with awareness of dataset and operational constraints. Future work may extend this effort by developing integrated decision-support platforms or a real-time deployment framework that allows automated forecasting and enabling direct measurement of operational impacts such as stock out reduction and safety stock optimization.

6. Implications

6.1. Practical Implications

The findings of this study provide a context-specific guidance for humanitarian organizations involved in medical relief operations, particularly in disaster-prone and resource-constrained environments like the National Capital Region, Philippines. Within this context, CatBoost consistently achieved the highest predictive accuracy RMSE, WAPE, and MASE and stable cross-validation performance, making it a strong candidate for forecasting medical supply demand. Gradient Boosting and LightGBM also demonstrated substantial gains after hyperparameter tuning, indicating they can serve as practical alternatives when constraints such as computational resources, interpretability, or data availability are considered.

Adopting these predictive models can directly support operational decision-making by improving inventory planning, reducing uncertainty in demand forecasts, and informing the

allocation of critical medical supplies. Specifically, more accurate forecasts could help minimize stockouts and overstock, optimize safety stock levels, reduce operational costs, and improve service levels during emergency response operations. Integrating these models into existing decision-support systems could enhance responsiveness, coordination, and preparedness, though actual operational impact should be empirically validated in practice.

6.2. Theoretical Implications

This research contributes to the literature on machine learning in humanitarian supply chains by providing empirical evidence that Bayesian-optimized boosting ensembles effectively model with high-variance demand patterns common in medical relief logistics. Comparative evaluation of CatBoost, XGBoost, LightGBM, AdaBoost, and Gradient Boosting demonstrates how different frameworks respond to data heterogeneity and extreme values, with optimization significantly improving performance. The study reinforces the importance of systematic hyperparameter tuning and robust validation for ensuring generalizable and stable predictions. By quantifying performance improvements through Bayesian optimization, this work supports the methodological case for combining advanced optimization techniques with ensemble learning in predictive analytics research.

6.3. Research Implications

The study highlights several promising directions for future research. One avenue is extending Bayesian-optimized ensembles to other humanitarian logistics domains while another is developing hybrid “ensemble-of-ensembles” models that integrate complementary algorithms to further enhance predictive accuracy and robustness. There is also potential for real-time adaptive forecasting frameworks by incorporating streaming data, evolving demand patterns, and external disruptions such as natural disasters or epidemics. Integrating these predictive models with broader logistics decision-support and optimization platforms could translate forecast improvements into measurable operational benefits. Future research should also evaluate the trade-offs between predictive performance, computational cost, interpretability, and data requirements to ensure models are both practical and deployable in resource-constrained humanitarian settings.

Disclosure statement

The authors declare no conflict of interest.

Funding

This research did not receive any specific grant from any funding agencies.

AI Declaration

The author declares no Artificial Intelligence–based writing used in the preparation of this manuscript. All analyses, model development, and result generation were conducted using Python programming within the Jupyter Notebook environment. The author takes full responsibility for the integrity and originality of the work.

References

Ahatsi, E., & Olanrewaju, O. A. (2025). Enhancing humanitarian supply chain resilience: Evaluating artificial intelligence and big data analytics in two nations. *Logistics*, 9(2), 64. https://doi.org/10.3390/logistics9020064

Ahmed, K. R., Ansari, M. E., Ahsan, M. N., Rohan, A., Uddin, M. B., & Rivin, M. A. H. (2025). Deep learning framework for interpretable supply chain forecasting using SOM ANN and SHAP. *Scientific Reports, 15*(1), 26355. https://doi.org/10.1038/s41598-025-11510-z

Ahn, J. M., Kim, J., & Kim, K. (2023). Ensemble machine learning of gradient boosting (XGBoost, LightGBM, CatBoost) and attention-based CNN-LSTM for harmful algal blooms forecasting. *Toxins*, 15(10), 608. https://doi.org/10.3390/toxins15100608

Aldahmani, E., Alzubi, A., & Iyiola, K. (2024). Demand forecasting in supply chain using uni-regression deep approximate forecasting model. *Applied Sciences, 14*(18), 8110. https://doi.org/10.3390/app14188110

Alfath, A. S., Wardhana, A. K., & Rumini, R. (2025). Hypertension risk prediction using stacking ensemble of CatBoost, XGBoost, and LightGBM: A machine learning approach. *Journal of Applied Informatics and Computing, 9*(6), 3146–3156. https://doi.org/10.30871/jaic.v9i6.10370

Altay, N., & Narayanan, A. (2022). Forecasting in humanitarian operations: Literature review and research needs. *International Journal of Forecasting, 38*(3), 1234-1244. https://doi.org/10.1016/j.ijforecast.2020.08.001

Cao, L. (2023). AI and data science for smart emergency, crisis and disaster resilience. *International Journal of Data Science and Analytics, 15*(3), 231–246. https://doi.org/10.1007/s41060-023-00393-w

Chandran, J. M., & Khan, M. R. B. (2024). A strategic demand forecasting: Assessing methodologies, market volatility, and operational efficiency. *Malaysian Journal of*

Business, Economics and Management, 150–167. https://doi.org/10.56532/mjbem.v3i2.71

Dalimunthe, S. B., Ginting, R., & Sinulingga, S. (2023). The implementation of machine learning in demand forecasting: A review of method used in demand forecasting with machine learning. *Jurnal Sistem Teknik Industri, 25*(1), 41–49. https://doi.org/10.32734/jsti.v25i1.9290

de Mast, J., & Lokkerbol, J. (2024). DAPS diagrams for defining Data Science projects. Journal of Big Data, 11(1), 50. https://doi.org/10.1186/s40537-024-00916-7

Douaioui, K., Oucheikh, R., Benmoussa, O., & Mabrouki, C. (2024). Machine learning and deep learning models for demand forecasting in supply chain management: A critical review. *Applied System Innovation, 7*(5), 93. https://doi.org/10.3390/asi7050093

Efe, A. (2022). A review on risk reduction potentials of artificial intelligence in humanitarian aid sector. *Journal of Human and Social Sciences, 5*(2), 184–205. https://doi.org/10.53048/johass.1189814

Fernandes, A. A., Koehler, M., Konstantinou, N., Pankin, P., Paton, N. W., & Sakellariou, R. (2023). Data preparation: A technological perspective and review. *SN Computer Science, 4*(4), 425. https://doi.org/10.1007/s42979-023-01828-8

Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions*, 2022, 1-10. https://doi.org/10.5194/gmd-15-5481-2022

Katya, E. (2023). Exploring feature engineering strategies for improving predictive models in data science. *Research Journal of Computer Systems and Engineering, 4*(2), 201-215. https://doi.org/10.52710/rjcse.88

Kumar, V., Goodarzian, F., Ghasemi, P., Chan, F. T., & Gupta, N. (2025). Artificial intelligence applications in healthcare supply chain networks under disaster conditions. *International Journal of Production Research, 63*(2), 395–403. https://doi.org/10.1080/00207543.2024.2444150

Li, T., Wang, S., Nong, T., Liu, B., Hu, F., Chen, Y., & Han, Y. (2025). Bayesian Optimization of LSTM-Driven Cold Chain Warehouse Demand Forecasting Application and Optimization. *Processes, 13*(10), 3085. https://doi.org/10.3390/pr13103085

Makridakis, S., Spiliotis, E., Assimakopoulos, V., Semenoglou, A. A., Mulder, G., & Nikolopoulos, K. (2022). Statistical, machine learning and deep learning forecasting methods: Comparisons and ways forward. *Journal of the Operational Research Society, 74*(3), 840-859. https://doi.org/10.1080/01605682.2022.2118629

Mohammed, I. A., & Mandal, J. (2024). Forecasting accuracy through machine learning in supply chain management. *International Journal of Supply Chain Management, 7*(2), 60-77. https://doi.org/10.47604/ijscm.3074

Mustapha, I. B., Abdulkareem, M., Jassam, T. M., AlAteah, A. H., Al-Sodani, K. A. A., Al-Tholaia, M. M., Nabus, H., Alih, S. C., Aldulkareem, Z., & Ganiyu, A. (2024). Comparative analysis of gradient-boosting ensembles for estimation of compressive strength of quaternary blend concrete. *International Journal of Concrete Structures and Materials, 18*(1), 20. https://doi.org/10.1186/s40069-023-00653-w

Özüpak, Y., Alpsalaz, F., & Aslan, E. (2025). Air quality forecasting using machine learning: Comparative analysis and ensemble strategies for enhanced prediction. *Water, Air, & Soil Pollution, 236*(7), 464. https://doi.org/10.1007/s11270-025-08122-8

Pantiris, P., Pallis, P. L., Chountalas, P. T., & Dasaklis, T. K. (2025). Enhancing coordination and decision making in humanitarian logistics through artificial intelligence: A

grounded theory approach. *Logistics, 9*(3), 113. https://doi.org/10.3390/logistics9030113

Salamian, F., Paksaz, A., Khalil Loo, B., Mousapour Mamoudan, M., Aghsami, M., & Aghsami, A. (2024). Supply chains problem during crises: A data-driven approach. *Modelling, 5*(4), 2001–2039. https://doi.org/10.3390/modelling5040104

Saltz, J. S., & Hotz, N. (2020). Identifying the most common frameworks data science teams use to structure and coordinate their projects. In Proceedings of the 2020 IEEE *International Conference on Big Data (Big Data)* (pp. 2038–2042). IEEE. https://doi.org/10.1109/BigData50022.2020.9377813

Sarker, I. H. (2021). Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science, 2*(5), 377. https://doi.org/10.1007/s42979-021-00765-8

Shafiq, S., Mashkoor, A., Mayr-Dorn, C., & Egyed, A. (2021). A literature review of using machine learning in software development life cycle stages. *IEEE Access, 9*, 140896–140920. https://doi.org/10.1109/ACCESS.2021.3119746

Stelmaszak, M., & Kline, K. (2023). Managing embedded data science teams for success: how managers can navigate the advantages and challenges of distributed data science. *Harvard Data Science Review, 5*(2). https://doi.org/10.1162/99608f92.1f068331

Vance, E. A. (2021). Using team-based learning to teach data science. *Journal of Statistics and Data Science Education, 29*(3), 277-296. https://doi.org/10.1080/26939169.2021.1971587

Victoria, A. H., & Maragatham, G. (2020). Automatic tuning of hyperparameters using Bayesian optimization. *Evolving Systems, 12*(1), 217-223. https://doi.org/10.1007/s12530-020-09345-2