

Spatiotemporal Prediction of Medical Device Sales Using Hybrid LSTM GNN in South Sumatra

Khoirusy Syafaat¹⁾, Herri Setiawan²⁾*

^{1,2)}Faculty of Computer and Natural Sciences, Indo Global Mandiri University, Palembang 30129, Indonesia

¹⁾khoirusysyafaat@students.uigm.ac.id, ²⁾herri@uigm.ac.id

ABSTRACT

Forecasting medical device purchases at the regional level supports equitable allocation, procurement planning, and distribution scheduling across administrative regions. In developing regions where infrastructure and logistics capacity are unevenly distributed, anticipating regional purchase patterns mitigates supply imbalances and improves the continuity of clinical services. The demand for medical devices exhibits complex spatiotemporal characteristics, where demand variations are influenced by temporal dynamics and interregional spatial connectivity. The objective of this study is to apply and evaluate a Hybrid Long Short-Term Memory and Graph Neural Network (LSTM-GNN) model for predicting the spatiotemporal demand for medical devices, to analyze the effect of spatial relationships on demand variations, and to identify optimal hyperparameter combinations. The methodology uses medical device sales data from PT Parit Panjang, a distribution company in South Sumatra, covering July 2020 to October 2025, comprising 10,878 unique invoices across 16 regencies and cities (Musi Rawas Utara, with only 2 invoices in the study period, was merged into Musi Rawas). Records were aggregated into a daily panel, split chronologically into training, validation, and testing sets (July 2020 to December 2023, January to December 2024, and January to October 2025), and normalized with a log transformation fitted on the training set only. Because 81.54% of all daily region-level targets across the full panel are zero (84.85% in the test period), the problem is formulated as zero-inflated forecasting with two stages: a binary classifier that predicts whether a sale occurs and a regression model that predicts the sales amount on non-zero days. A binary COVID-19 pandemic indicator (July 2020 to December 2022) is included as an exogenous input channel. A spatial graph was built from regional proximity using Haversine distance with three nearest neighbors and a maximum distance threshold of 104.66 km. Temporal dynamics were modeled with LSTM, spatial dependencies with a Graph Convolutional Network (GCN), and both representations were combined through concatenation followed by a fully connected layer. The Hybrid LSTM-GNN achieved an R-squared of 0.0523, an RMSE of 8,002,603 IDR, and a MAE of 1,881,358 IDR on the original scale, outperforming the standalone LSTM (R-squared 0.0433), the standalone GNN (0.0059), and the always-zero baseline (0.0059). The classification stage achieved 90.67% accuracy and an AUC of 0.861. Diebold-Mariano tests confirm that the hybrid model improves over the standalone LSTM and GNN and over the statistical baselines at $p < 0.001$, but its practical advantage is limited to the classification stage. A random search over 30 configurations identified a superior configuration (LSTM hidden 219, GNN hidden 107, dropout 0.17) reaching an R-squared of 0.0582, the best among all evaluated models. The COVID-19 indicator provides a small improvement, and replacing the distance-based graph with the actual distribution network (a star centered on Palembang) raises the R-squared to 0.0703. In conclusion, the zero-inflated Hybrid LSTM-GNN framework supports regional medical device sales forecasting and distribution planning, with the classification stage providing the dominant predictive signal.

Keywords: Deep Learning; Graph Neural Networks; Hybrid Model; Long Short-Term Memory; Medical Devices; Spatiotemporal Prediction; Zero-Inflated Forecasting

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

INTRODUCTION

The availability and equitable distribution of medical devices play a critical role in ensuring the effectiveness and resilience of healthcare systems. Medical devices are essential not only for routine clinical services but also for emergency response, disease prevention, and public health interventions. In many developing regions, disparities in healthcare infrastructure and logistics often lead to uneven distribution of medical devices across geographical areas, which may negatively impact healthcare service quality and accessibility (World Health Organization, 2023). Consequently, accurate forecasting of medical device demand at the regional level is crucial to support inventory planning, distribution optimization, and strategic decision-making within healthcare supply chains (Koç & Türkoğlu, 2022).

Medical device sales data generally exhibit temporal dynamics, including seasonal variations, long-term trends, and abrupt changes caused by policy interventions or public health emergencies. The COVID-19 pandemic, for example, caused abrupt and regionally heterogeneous shifts in medical device demand (Koç & Türkoğlu, 2022). In addition to temporal characteristics, demand patterns are also affected by spatial dependencies, where geographically adjacent regions tend to show correlated demand behaviors due to similarities in population structure, healthcare facilities, and distribution networks (Alourani et al., 2023; Heryati et al., 2024; Patharkar et al., 2024; Ye et al., 2024). Conventional time-series forecasting approaches typically assume that observations from different locations are independent, focusing primarily on temporal correlations. This assumption limits their ability to capture spatial interactions that naturally exist in regional sales data (Tunnicliffe Wilson, 2016; Wu et al., 2019).

A distinctive characteristic of regional daily sales data is its zero-inflated structure: on most days, most regions record no sales at all. In this dataset, 84.85% of the daily region-level targets in the test period are zero. Standard regression models trained on such data tend to collapse toward the zero mode and fail to capture the magnitude of sales on active days. The gap is therefore not merely the absence of a hybrid LSTM-GNN in this domain; it is the absence of a zero-inflated spatiotemporal treatment for intermittent regional medical device sales, where the zero/non-zero decision dominates the predictive value and where graph structure determines whether spatial modeling helps at all. This study therefore formulates the forecasting problem as zero-inflated: a binary classifier first predicts whether a sale occurs, and a regression model then predicts the sales amount conditional on a sale being predicted. This two-stage formulation is standard in count and demand modeling and is reported explicitly so that the contribution of each stage can be evaluated.

This study addresses the following research questions. First, how effective is the Hybrid LSTM-GNN model in predicting regional medical device sales compared to standalone and statistical baselines. Second, to what extent do spatial dependencies and the COVID-19 indicator contribute to improving prediction accuracy in spatiotemporal medical device demand forecasting. The main contributions of this research are as follows. First, the development of a zero-inflated hybrid deep learning framework that integrates temporal and spatial information for medical device sales forecasting. Second, the construction of a distance-based regional graph with an explicit and reproducible construction protocol. Third, an evaluation of the proposed model against statistical and deep learning baselines using multiple performance metrics, statistical significance tests, and an ablation study. Fourth, a per-region analysis that reports prediction quality for each of the 16 regions instead of a single global value. The novelty of this study is contextual rather than algorithmic: it applies a zero-inflated hybrid LSTM-GNN framework to regional medical device sales in Indonesia using real-world distribution data and evaluates it with a reproducible protocol. The algorithm itself is not new; the contribution lies in the domain-specific application and the empirical evidence it provides for medical device distribution planning in developing regions.

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0
International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

LITERATURE REVIEW

With the rapid development of deep learning techniques, models capable of learning complex non-linear patterns from sequential data have been increasingly applied to forecasting problems. Long Short-Term Memory (LSTM) networks have been widely used for time-series prediction due to their ability to model long-term temporal dependencies and mitigate the vanishing gradient problem (Hochreiter & Schmidhuber, 1997). LSTM-based models have demonstrated strong performance in various forecasting tasks, including energy consumption, financial markets, and healthcare-related time series (Chen et al., 2017). However, standard LSTM architectures do not explicitly incorporate spatial relationships between different regions or entities.

Table 1. State-of-the-Art Comparison of Spatiotemporal Forecasting

Author (Year)	Application Domain	Method	Temporal	Spatial	Limitation / Research Gap
Koç & Türkoğlu (2022)	Medical Equipment Demand	Deep LSTM	Yes	No	Exclusively modeled time-series data, ignoring geographical correlations and interregional dependencies.
Luo et al. (2024)	Urban Traffic	Deep GNN	No	Yes	Focused primarily on spatial features, lacking robust mechanisms for long-term sequential temporal dependencies.
Sonani et al. (2025)	Stock Price	Hybrid LSTM-GNN	Yes	Yes	Applied strictly to the financial domain; did not consider real-world geographical or logistical graph structures.
This Study (2025)	Medical Device Sales	Zero-Inflated Hybrid LSTM-GNN	Yes	Yes	Bridges the existing gap by applying zero-inflated spatiotemporal modeling to regional medical supply chains.

In contrast, Graph Neural Networks (GNN) have emerged as effective models for learning representations from graph-structured data (Xu et al., 2018). By propagating information through edges that represent relationships between nodes, GNNs are able to capture spatial dependencies and relational patterns. Variants such as Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT) have been successfully applied in traffic forecasting, recommendation systems, and spatial analysis (Longa et al., 2023; Zhao et al., 2021). Despite their effectiveness in modeling spatial relationships, GNNs lack inherent mechanisms to handle temporal sequences, making them insufficient when applied independently to spatiotemporal forecasting tasks. For this reason, a standalone GNN is not a fair temporal baseline; it is included in this study only for completeness.

To overcome these limitations, recent studies have proposed hybrid spatiotemporal models that integrate LSTM and GNN architectures to jointly model temporal evolution and spatial interactions (Kuppan et al., 2024; Liu & Liu, 2024; Shao et al., 2023; Sonani et al., 2025; Wen et al., 2024). Such hybrid models have shown promising results in domains including traffic flow prediction, energy demand forecasting, and epidemic spread modeling (Luo et al., 2024; Santos et al., 2023; Sunder et al., 2024; Y. Yang et al., 2025; Z. Yang et al., 2023). In parallel, zero-inflated modeling has a long tradition in demand forecasting, where the target distribution contains an excess of zeros and is modeled as a mixture of a point mass at zero and a positive distribution (Lambert, 1992). Recent work has extended this framework in several directions: flexible count regression with computational tools (Young et al., 2022), machine learning applied to zero-

* Corresponding author



inflated overdispersed count data (Sidumo et al., 2023), interpretable zero-inflated neural networks (Jose et al., 2024), zero-inflated forecasting of sparse sales series (Judd et al., 2025; Li & Zhou, 2025), Bayesian forecasting of zero-inflated time series of counts (Aktekin et al., 2026; Gao et al., 2026), and sparse-data neural networks with regularization (Jun, 2026). For spatiotemporal settings, deep extreme mixture models jointly capture zero-inflation and heavy tails (Wilson et al., 2022). However, the application of zero-inflated hybrid LSTM-GNN approaches to medical device sales forecasting, particularly in the context of regional healthcare distribution in Indonesia, remains relatively unexplored. Therefore, this study proposes a zero-inflated Hybrid LSTM-Graph Neural Network model for spatiotemporal prediction of medical device sales at the regional level to address this specific gap.

METHOD

This study adopts a quantitative research approach using deep learning-based spatio-temporal modeling to predict regional medical device sales. The methodology encompasses data description, data preprocessing, spatial graph construction, model architecture design, and performance evaluation (Figure 1).

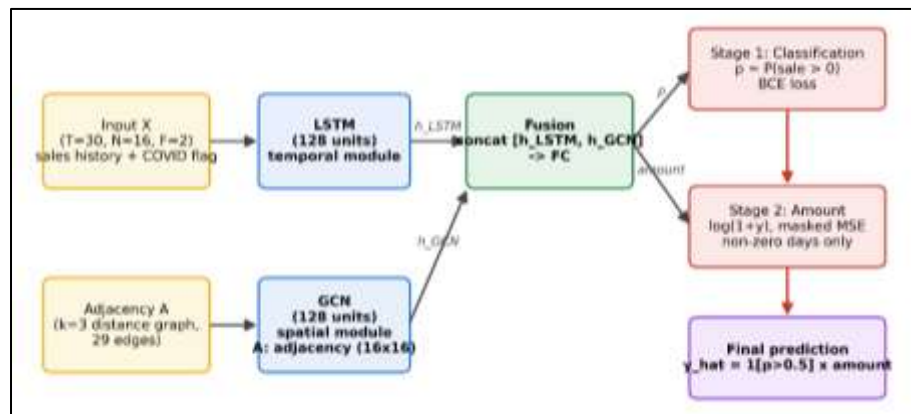


Figure 1. Zero-inflated Hybrid LSTM-GNN architecture (two stage)

Data Description

The dataset was obtained from PT Parit Panjang, a medical device distribution company operating in South Sumatra Province, Indonesia. It contains 13,661 unique sales invoices collected from July 2020 to October 2025. After excluding invoices from regions outside South Sumatra, 10,878 invoices remain. Each invoice records the sale of medical devices to healthcare facilities located in different regencies or cities. The raw dataset contains 37 attributes; the five attributes used in this study are the transaction date (`jual_tanggal`), customer city (`pelanggan_kota`), latitude, longitude, and the invoice total (`jual_total_fak`). Invoices contain a median of two line items; 32.8% contain a single item. The catalog comprises 2,911 distinct products, of which 35.9% appear only once, across 375 brands led by Cosmomed (13.6% of invoices), Top Point (10.0%), Assut (8.9%), and Forsch (8.1%). The customer base consists predominantly of healthcare institutions: hospitals (including RSUD, RSUP, RUMKIT, and private hospitals) account for 70.8% of invoices, medical companies and hospital operators 22.7%, and foundations 3.2%; pharmacies, clinics, laboratories, and health center together account for less than 1%. Each customer belongs to a single region.

The target variable of this study is **invoice total (`jual_total_fak`)**, aggregated per region per day. This variable is the primary sales measure recorded by the distributor and reflects the economic value of demand, which is the quantity relevant for revenue planning and budget allocation. Quantity sold (`d_jual_qty`) is available but is not used as the target because the product catalog is highly heterogeneous (for example, disposable syringes and ventilators are not comparable in units), so aggregating units across product

* Corresponding author



categories is not meaningful. This heterogeneity is confirmed empirically: the implied unit price ranges from a median of IDR 207,037 to a maximum of IDR 500,500,000, and the correlation between quantity and invoice value is only 0.29 at the line-item level and 0.46 at the invoice level. A supplementary experiment further showed that lagged aggregates of these attributes (quantity, product count, distinct brands, suppliers, and salespeople per region-day) add negligible signal to the amount stage (incremental R2 of at most +0.014 on the test set beyond the 30-day sales history). The remaining 32 attributes are line-item or invoice-level descriptors (product name, brand, supplier, salesperson, discount, tax, stock, unit price) that cannot enter the model directly because the target is a region-day aggregate; salesperson identity is also personal data and is not reported. The term "demand" in this paper refers strictly to the daily invoice total per region.

Time unit. One time step is one day. Daily aggregation was chosen after evaluating monthly and weekly resolutions: the daily panel provides 1,495-time steps (957 training samples) instead of 64 months (30 training samples), which is the minimum data volume required for training deep sequence models. The daily target is zero-inflated: 84.85% of the test targets are zero, which motivates the two-stage zero-inflated formulation described in Section 3.5. With a window of 30-time steps, each input sequence covers one month of history.

Panel construction. Transaction records were aggregated into a regular panel with identical daily timestamps for all 16 regions. Days with no recorded sales were assigned a value of zero. No interpolation or forward-filling was applied to the target, because filling missing values with estimates derived from neighboring periods can introduce leakage.

COVID-19 indicator. A binary exogenous feature indicates whether a day falls within the pandemic period, defined from the first confirmed case in Indonesia (March 2020) to the revocation of the PPKM policy (December 30, 2022). Since the data begin in July 2020, the indicator equals 1 for 706 days (47.2% of the panel) and 0 afterward. The indicator is identical for all regions on a given day and is provided as a second input channel to the models.

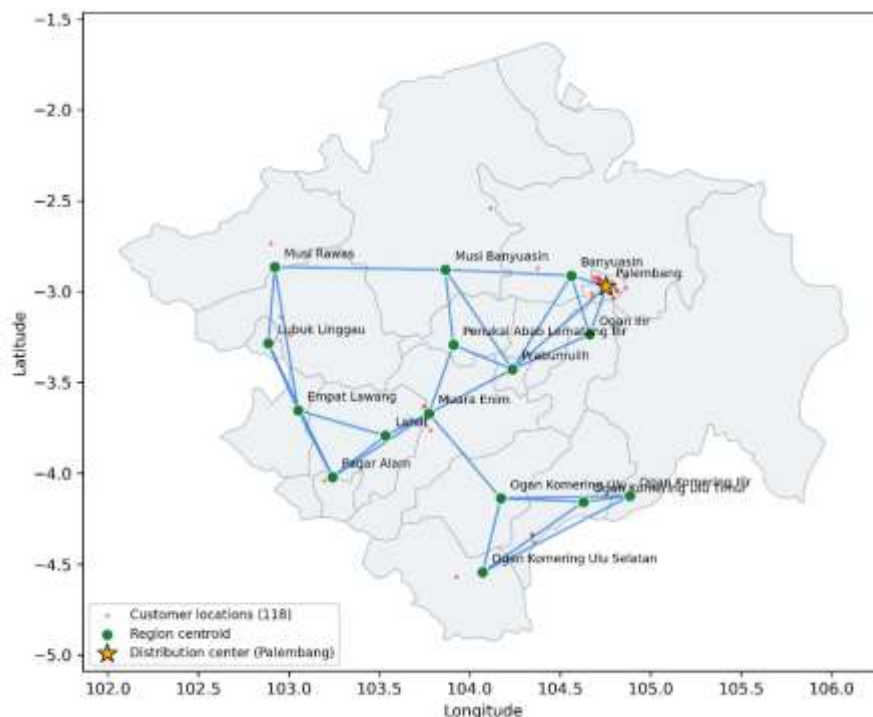


Figure 2. Study area : South Sumatra 16 regencies/city customer location

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0
International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

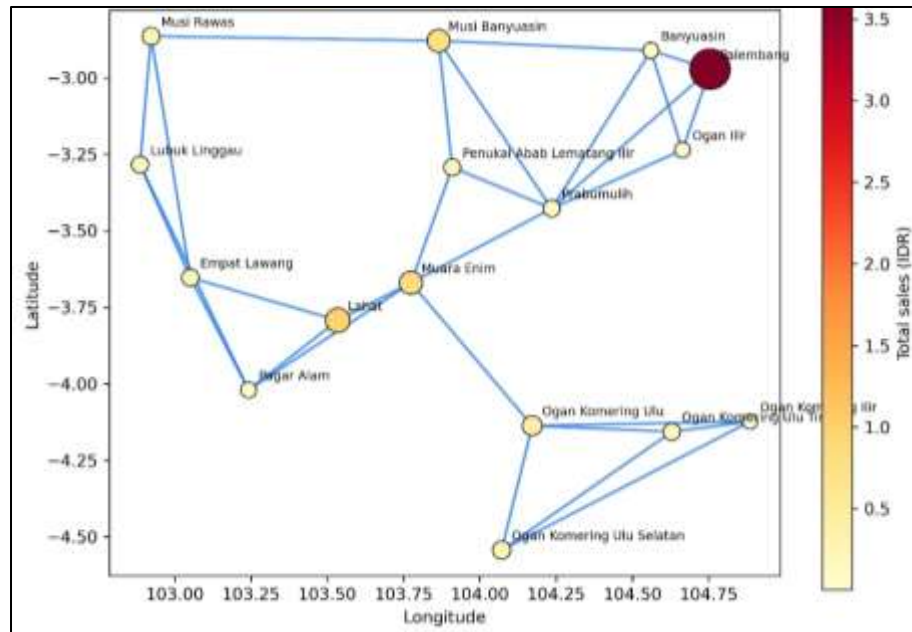


Figure 3. Spatial Graph: $k=3$ nearest neighbors, 29 edges, threshold 104.66 km
node size = total sales volume

3.2 Data Splitting and Leakage Control. The panel was divided chronologically into training, validation, and testing sets. The same split boundaries were applied to all 16 regions: **Training:** July 2020 to December 2023 (957 windowed samples). **Validation:** January to December 2024 (277 windowed samples). **Testing:** January to October 2025 (231 windowed samples).

To prevent data leakage, the following protocol was applied: The log transformation and its normalization parameters were fitted on the training set only (the sales input channel). The fitted parameters were then used to transform the validation and test sets. Windows were constructed over the full panel and then assigned to subsets by their target date (chronological split). Because the split is chronological, test and validation targets never appear in the training inputs; test and validation sequences may use history that precedes their respective periods. A fixed random seed of 42 was used for all stochastic operations (weight initialization, data shuffling). Missing days were zero-filled as described in Section 3.1; no target imputation was performed.

3.3 Preprocessing. First, incomplete and duplicated records were removed. Temporal ordering was ensured by sorting all transactions chronologically. Since the distribution of invoice totals exhibited significant skewness, a log transformation ($\log(1 + x)$) was applied to the sales channel to stabilize variance and improve model convergence. As stated in Section 3.2, the transformation parameters were fitted on the training set only. The COVID-19 indicator channel was not transformed.

The preprocessed data were converted into a supervised learning format using a sliding window approach with a window size of 30 time steps and a one-step-ahead prediction horizon. Each sample is a 30-step multivariate window across all 16 regions with two channels: input X in $\mathbb{R}^{(30 \times 16 \times 2)}$ (sales and COVID-19 indicator), target y in $\mathbb{R}^{16}$ (the next day's invoice total for all regions). The total number of windowed samples is 1,465 (957 training, 277 validation, 231 testing).

3.4 Spatial Graph Construction. Each node of the graph represents one regency or city. The graph was constructed as follows:

Nodes. The 16 regencies and cities of South Sumatra Province: Banyuasin, Empat Lawang, Lahat, Lubuk Linggau, Muara Enim, Musi Banyuasin, Musi Rawas, Ogan Ilir, Ogan Komering Ilir, Ogan

* Corresponding author



Komering Ulu, Ogan Komering Ulu Selatan, Ogan Komering Ulu Timur, Pagar Alam, Palembang, Penukal Abab Lematang Ilir, and Prabumulih. Musi Rawas Utara was merged into Musi Rawas because it recorded only 2 invoices in the entire study period (zero-sales fraction 99.9%); it was administratively split from Musi Rawas in 2013, so the merge preserves the historical continuity of the region. Regions outside the province that appear in the raw data (Bangka Belitung, Bengkulu, Jakarta Pusat, Jambi, Lebong, Metro, Pinang, Rejang Lebong, Sungailiat) were excluded because they are not part of the South Sumatra administrative area.

Coordinates. Each region is represented by the invoice-weighted centroid of its customer locations. Customer coordinates were geocoded from public sources (SIRS Kemenkes, Photon, idalamat, lewatmana, Google Maps, and Medicastore): 118 of 156 customers, covering 10,752 of 10,878 invoices (98.8%), were resolved to actual coordinates; the remaining 38 customers (126 invoices, 1.2%) are pharmacies, private practitioners, and small agencies without public addresses and were assigned to their city center. The centroid is weighted by the number of invoices per customer so that locations with more sales activity contribute more to the region's position.

Distance. Inter-region distances were computed with the Haversine (great-circle) formula on latitude and longitude, expressed in kilometers. Euclidean distance on raw latitude/longitude degrees is not a valid distance measure and was therefore not used.

Edges. An adaptive k-nearest-neighbor method connects each region to its three nearest neighbors ($k = 3$), subject to a maximum distance threshold of 104.66 km to prevent unrealistic long-distance connections. The threshold was derived from the distribution of 3-nearest-neighbor distances: the observed minimum is 22.61 km, the median is 46.68 km, and the maximum is 104.66 km. The threshold of 104.66 km equals the maximum observed 3-NN distance, so all 48 candidate edges were retained; the threshold therefore acts as a safeguard against unrealistic connections rather than an active constraint. **Edge weights.** A binary adjacency (0/1) represents the basic connectivity structure. **Graph properties.** The graph is undirected, has no self-loops, and is static: the adjacency matrix is fixed for the entire study period. Dynamic graphs are left for future work. The final graph contains 29 undirected edges; node degrees range from 3 to 6 (mean 3.62), and no node is isolated. The resulting adjacency matrix A in $\mathbb{R}^{(16 \times 16)}$ represents the spatial structure of the study area and is used as input to the Graph Neural Network component. Figure 2 shows the study area with the geocoded customer locations, the invoice-weighted regional centroids, and the resulting $k = 3$ graph; Figure 3 visualizes the graph structure with node sizes proportional to total sales volume.

3.5 Model Architecture. The proposed model is a zero-inflated two-stage framework. Both stages share the same hybrid architecture, which integrates an LSTM network and a GNN into a unified framework consisting of a temporal modeling module, a spatial modeling module, and a fusion module.

Stage 1: classification. A binary classifier predicts whether a sale occurs in each region on the target day. The output is a probability p in $[0, 1]$ per region, trained with binary cross-entropy loss.

Stage 2: amount regression. A regression model predicts the sales amount on days when a sale occurs. The target is $\log(1 + y)$ and the model is trained with mean squared error masked to non-zero targets only, so that zero days do not dominate the loss. The prediction is exponentiated back to the original scale.

Final prediction. The final prediction is the product of the classification decision and the amount prediction:

$$\hat{Y} = 1[p > 0.5] \times \widehat{\text{amount}}$$

This formulation follows the standard zero-inflated expectation decomposition :

$$E[y] = P(y > 0)xE[y|y > 0]$$

with a hard threshold on the classification stage.

Temporal module (LSTM). For each region, the LSTM processes the historical sales sequence. At each time step t , the LSTM updates its cell state C_t and hidden state h_t using the input vector x_t (historical sales and COVID-19 indicator) with the following equations:

Forget Gate (f_t): Determines what information to throw away from the cell state.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

Input Gate (i_t): Determines which new information we're going to store in the cell state.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

Spatial module (GCN). The Graph Convolutional Network refines the temporal representations by propagating information through the regional graph. Node features are the LSTM hidden states H_{LSTM} . The propagation rule of a single GCN layer is:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)})$$

where $\tilde{A} = A + I_N$ is the adjacency matrix with added self-connections, $\tilde{D}$ is the degree matrix of $\tilde{A}$, $W^{(l)}$ is the learnable weight matrix of layer l , and sigma is the activation function (ReLU in this study). The output is the spatial representation H_{GNN} in $\mathbb{R}^{(N \times 128)}$, where 128 is the GNN hidden dimension.

Fusion module. The temporal and spatial representations are combined by concatenation along the feature dimension (Pratama et al., 2025):

$$H_{fused} = [H_{LSTM} || H_{GNN}]$$

where $[||]$ denotes concatenation along the feature dimension. The fused representation H_{fused} in $\mathbb{R}^{(N \times 256)}$ is passed to a fully connected layer that produces the final output in $\mathbb{R}^N$ for all regions.

Architecture summary. Table 2 summarizes the final configuration of the proposed Hybrid model; the overall two-stage data flow is illustrated in Figure 1.

Table 2. Architecture Summary of the Proposed Hybrid LSTM-GNN Model

Module	Layer	Units	Activation	Dropout	Output shape
Input	Two channels (sales, COVID-19)	2	-	-	(N, 30, 2)
Temporal	LSTM (1 layer)	128	tanh	-	(N, 128)
Spatial	GCNConv (1 layer)	128	ReLU	-	(N, 128)
Fusion	Concatenation	256	-	-	(N, 256)
Output	Fully connected	1	-	-	(N,)

Training configuration. Each stage was trained with the Adam optimizer (Kingma & Ba, 2015). The classification stage used binary cross-entropy loss; the amount stage used masked mean squared error. The configuration was: learning rate 0.001, batch size 64, maximum 60 epochs, early stopping with patience 8 monitored on validation loss. The implementation used Python 3.12.7 with PyTorch 2.13.0 (CPU build), pandas 2.2.2, NumPy 1.26.4, scikit-learn 1.5.1, and SciPy 1.13.1; training was executed on CPU.

3.6 Hyperparameter Optimization. Hyperparameter selection follows the random search procedure of (Bergstra & Bengio, 2012). Thirty configurations are sampled independently from the search space with a fixed seed (42): the LSTM hidden size and the GNN hidden size are drawn from a log-uniform distribution over [32, 256], the learning rate from a log-uniform distribution over [1e-4, 1e-2], the dropout rate from a

* Corresponding author



uniform distribution over $[0.0, 0.5]$, and the batch size uniformly from $\{32, 64, 128\}$. Each configuration is trained with Adam for up to 60 epochs with early stopping (patience 8) and scored by the total validation loss (binary cross-entropy of the classification stage plus masked MSE of the amount stage). The configuration with the lowest validation loss is selected; the test set is used exactly once, after selection, to report the final results. Configurations are never compared on the test set during tuning. The selected configuration (LSTM hidden 219, GNN hidden 107, dropout 0.17, learning rate $1.5e-3$, batch 32) was then re-evaluated with three seeds (42, 7, 123) following the same protocol as the other models; it reaches an R-squared of 0.0582 ± 0.0021 (mean $\pm$ std over seeds) and an ensemble R-squared of 0.0583, the best among all evaluated models, and it significantly outperforms the base configuration (Diebold-Mariano DM = -6.307, $p < 0.001$). In addition, the proposed Hybrid model (LSTM hidden 128, GNN hidden 128, no dropout) was compared against the tuned variant reported in the original study (LSTM hidden 128, GNN hidden 64, dropout 0.2) to evaluate whether those tuning choices transfer to the daily zero-inflated setting; they do not (R-squared 0.0523 versus 0.0462).

To investigate whether calendar information explains part of the residual structure, an additional experiment augments the input channels with calendar features of the target day: (i) a binary weekend indicator, (ii) a binary national-holiday indicator (official Indonesian public holidays and joint leave days, 2020-2025), and (iii) a two-level pandemic-phase indicator (refocusing period July-December 2020 and wave period 2021-2022, with the endemic period 2023 onward as the reference). Two variants are evaluated: V1 adds the weekend indicator, and V2 replaces it with a full seven-dimensional day-of-week one-hot encoding. The features describe the target day, are held constant across the 30-day input window, and are tiled over all regions. All variants use the same split, preprocessing, and evaluation protocol as the main experiment, and are trained with three seeds (42, 7, 123).

3.7 Baselines. The proposed model was compared against the following baselines, all evaluated with the same split, the same preprocessing, and the same metrics: 1. Always-zero forecast: $\hat{Y} = 0$ for all regions and days. This baseline is informative because the target is 84.85% zero. 2. Naive (persistence) forecast: $\hat{Y}_{(t+1)} = y_t$. 3. Training-mean forecast: $\hat{Y} = \text{mean of the training targets per region}$. 4. ARIMA per region: the order is selected by AIC over a grid (p, q in $\{0, 1, 2\}$; d in $\{0, 1\}$); the model is fitted on the $\log_{1p}$ -transformed training-plus-validation series and forecasts the 231 test steps, then predictions are transformed back with `expm1` and clipped at zero. 5. XGBoost: gradient-boosted regression trees using the same 30-day lagged sales history and the COVID-19 indicator as features, with early stopping on the validation set. 6. Standalone LSTM (same temporal setup as the LSTM module, no graph component). 7. Standalone GNN (included for completeness; as noted in Section 2, it is not a fair temporal baseline because it has no temporal component).

3.8 Evaluation Protocol. All metrics were computed on the original scale (IDR). Metrics reported are RMSE, MAE, and R-squared for the regression output. For the classification stage, accuracy and AUC alone are misleading on zero-inflated data because the majority class is zero, so the classification stage is additionally evaluated with precision, recall, F1-score, and the area under the precision-recall curve (PR-AUC) for the non-zero class, together with an explicit comparison to the always-zero baseline (Saito & Rehmsmeier, 2015). Classification metrics are reported both pooled over all region-days and per region, because pooled performance can hide region-level failure. MAPE is not reported because 84.85% of the test targets are zero, which makes percentage errors unstable and uninformative; this decision was confirmed by the authors.

To support claims of improvement, each model was trained with multiple random seeds (5 for the tuned variant, 3 for the proposed model and the comparators). Results are reported as mean $\pm$ standard deviation. Pairwise comparisons between the hybrid model and each comparator used the Diebold-Mariano test (Diebold & Mariano, 1995) on one-step-ahead errors, and 95% bootstrap confidence intervals (1,000 resamples) are reported. The word "significant" is used only when the Diebold-Mariano test rejects the null hypothesis at $p < 0.05$; otherwise the word "substantial" is used.

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0
International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

3.9 Ablation Study. To isolate whether any improvement comes from the geographic structure and the COVID-19 indicator rather than from additional model capacity, the following ablation variants were evaluated: 1. LSTM + GCN with identity adjacency (graph structure removed, parameters kept). 2. LSTM + GCN with a random graph (edges shuffled, 5 permutations, mean reported). 3. LSTM + GCN with the proposed distance graph ($k = 3$, threshold 104.66 km). 4. LSTM + GCN with a sales-correlation graph (Pearson correlation of training sales). 5. LSTM + GCN without the COVID-19 indicator channel (single input channel). 6. LSTM + GCN with a distribution-network graph. The company confirmed that all delivery routes originate from its distribution center in Palembang (Komplek Villa Kenten Blok G No. 2), so the distribution network is represented as a star graph in which Palembang is connected to all other 15 regions.

3.10 Residual Analysis. Residuals were defined as $e = y - \hat{y}$ on the original scale. The following diagnostics are reported: mean residual, residuals over time, residuals per region, and the Ljung-Box test for residual autocorrelation. The claim that residuals are symmetrically distributed around zero is made only if these diagnostics support it.

RESULT

The dataset contains 13,661 unique invoices from July 2020 to October 2025; after excluding regions outside South Sumatra, 10,878 invoices remain. The final panel contains 16 regions x 1,495 days. The target is heavily zero-inflated: across the full panel, the fraction of zero days ranges from 3.1% (Palembang) to 99.9% (Musi Rawas Utara, merged into Musi Rawas). In the test period, 3,136 of 3,696 targets (84.85%) are zero. The COVID-19 indicator is active for 706 days (47.2% of the panel). Figure 4a provides an overview of the daily total sales series and Figure 4b provides the per-region zero-sales fractions.

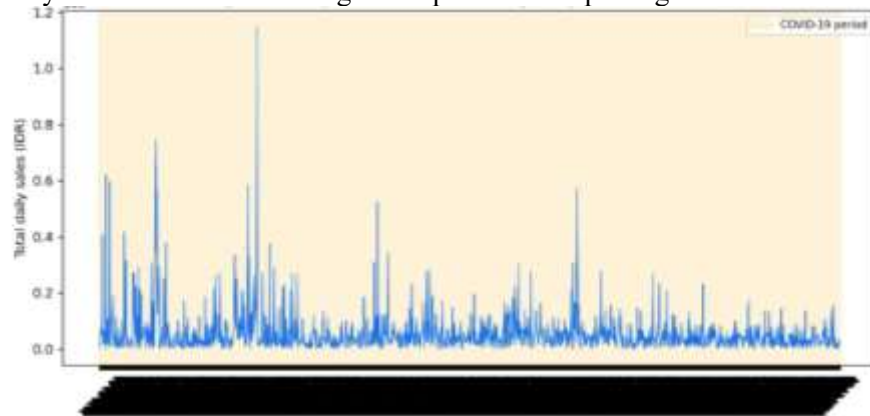


Figure 4a. Daily total sales across 16 regions (Jul 2020 – Oct 2025)

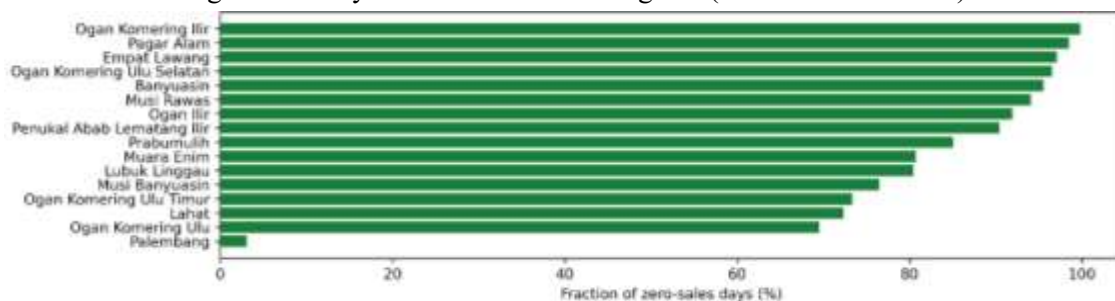


Figure 4b. Zero-sales fraction per region

Four models were evaluated: standalone LSTM, standalone GNN, the proposed Hybrid LSTM-GNN, and the tuned Hybrid LSTM-GNN. All models were trained with the same dataset split and evaluated with identical metrics. Table 3a summarizes the performance comparison on the original scale.

Table 3a. Performance Comparison of Evaluated Models (original scale, mean over seeds; classification metrics for the classification stage)

Model	RMSE (IDR)	MAE (IDR)	R- squared	Classification Accuracy	AUC
Standalone LSTM	8,242,726	1,944,192	0.0433	0.9067	0.861
Standalone GNN	9,167,322	2,271,337	0.0059	0.8485	0.666
Hybrid LSTM-GNN	8,002,603	1,881,358	0.0523	0.9067	0.861
Hybrid LSTM-GNN Tuned	8,165,301	1,924,589	0.0462	0.9067	0.862
Hybrid LSTM-GNN (random search)	7,843,703	1,846,331	0.0582	0.9067	0.861

The pooled classification accuracy of 0.9067 appears high, but on a zero-inflated test set in which 84.85% of the 3,696 region-day targets are zero, accuracy is dominated by the majority class: the always-zero baseline already achieves 0.8485. The classification stage is therefore additionally evaluated with precision, recall, F1, and PR-AUC for the non-zero (sales day) class (Table 3b). The precision of 0.9654 means that 96.54% of the days predicted as sales days were real sales days (223 true positives against 8 false positives); the recall of 0.3982 means that the classifier detected 39.82% of the 560 actual sales days. The PR-AUC of 0.6693 is well above the always-zero baseline of 0.1515 (equal to the test prevalence) and above the random-chance level, confirming that the classifier carries real signal for the minority class despite the imbalance. The random-search variant achieves a similar PR-AUC of 0.6745.

Table 3b. Pooled Classification Metrics for the Non-zero (Sales Day) Class (test set, ensemble over seeds)

Model	Accuracy	Precision	Recall	F1	PR-AUC	AUC
Always-zero	0.8485	0.0000	0.0000	0.0000	0.1515	0.500
Hybrid LSTM-GNN	0.9067	0.9654	0.3982	0.5638	0.6693	0.861
Hybrid LSTM-GNN (random search)	0.9067	0.9654	0.3982	0.5638	0.6745	0.860

The pooled metrics, however, hide a region-level pattern that is characteristic of the zero-inflated structure (Table 3c). The classifier produces non-zero predictions only for Palembang, the region with the highest sales frequency (223 of 231 test days); for that region precision is 0.9654, recall is 1.0, and the F1 is 0.9824. The other 13 regions with actual sales days in the test set receive no non-zero prediction at all (their maximum classification probability stays below the 0.5 threshold), so their precision, recall, and F1 are all zero. Two regions (Empat Lawang and Ogan Komering Ilir) have no non-zero target in the test set, so their per-region metrics are undefined. This concentration of the classifier on Palembang reflects the zero-inflated classifier collapse discussed in Section 5 and is reported transparently as a limitation of the classification stage.

Table 3c. Per-Region Classification Metrics (test set; n/a = no non-zero target in test set)

Region	Positive days	Precision	Recall	F1	PR-AUC
Palembang	223	0.9654	1.0000	0.9824	0.9746
Lahat	57	0.0000	0.0000	0.0000	0.2848
Ogan Komering Ulu	66	0.0000	0.0000	0.0000	0.2963
Ogan Komering Ulu Timur	72	0.0000	0.0000	0.0000	0.2707

Lubuk Linggau	26	0.0000	0.0000	0.0000	0.1177
Muara Enim	25	0.0000	0.0000	0.0000	0.0936
Prabumulih	28	0.0000	0.0000	0.0000	0.0999
Ogan Ilir	18	0.0000	0.0000	0.0000	0.0613
Musi Banyuasin	16	0.0000	0.0000	0.0000	0.0718
Pagar Alam	10	0.0000	0.0000	0.0000	0.0506
Penukal Abab Lematang Ilir	10	0.0000	0.0000	0.0000	0.0295
Banyuasin	5	0.0000	0.0000	0.0000	0.0136
Musi Rawas	3	0.0000	0.0000	0.0000	0.0225
Ogan Komering Ulu Selatan	1	0.0000	0.0000	0.0000	0.0023
Empat Lawang	0	n/a	n/a	n/a	n/a
Ogan Komering Ilir	0	n/a	n/a	n/a	n/a

The standalone LSTM model demonstrates reasonable performance in modeling temporal trends, achieving an R-squared of 0.0433. The standalone GNN model produces weaker performance, with an R-squared of 0.0059, identical to the always-zero baseline: without a temporal component, the GCN collapses to predicting zero for nearly all days, and its classification AUC of 0.666 reflects only the marginal distribution of the training set. The proposed Hybrid LSTM-GNN model achieves an R-squared of 0.0523, an RMSE of 8,002,603 IDR, and a MAE of 1,881,358 IDR. The tuned variant performs slightly worse than the base hybrid configuration (R-squared 0.0462), indicating that the tuning choices reported in the original study do not transfer to the daily zero-inflated setting. A random search over 30 configurations (Section 3.6) identified a superior configuration (LSTM hidden 219, GNN hidden 107, dropout 0.17, learning rate 1.5e-3, batch 32) that reaches an R-squared of 0.0582 (mean over three seeds), the best among all evaluated models.

Table 4. Baselines (original scale)

Baseline	RMSE (IDR)	MAE (IDR)	R-squared
Always-zero	9,167,322	2,271,337	0.0059
Naive (persistence)	10,693,726	3,061,362	-0.7978
Training-mean	7,656,812	3,674,680	-7.3242
ARIMA (per region, AIC)	8,133,445	1,924,544	-0.0775
XGBoost	8,303,467	1,966,596	-0.0840

The proposed model outperforms all baselines. The always-zero baseline achieves an R-squared of 0.0059, which reflects the dominance of zero targets; the proposed model improves on this substantially. The naive and training-mean baselines perform substantially worse. Notably, the ARIMA and XGBoost baselines (R-squared -0.0775 and -0.0840) do not outperform the always-zero forecast, confirming that the zero-inflated structure dominates the data and that the proposed model captures signal beyond simple persistence, statistical, and tree-based baselines.

The Diebold-Mariano test rejects the null hypothesis of equal predictive accuracy for all pairwise comparisons at $p < 0.001$. The Hybrid LSTM-GNN significantly outperforms the standalone LSTM, the standalone GNN, and the tuned hybrid variant. The random-search configuration significantly outperforms the base hybrid configuration as well (DM = -6.307, $p < 0.001$; negative DM indicates lower squared error for the random-search model). The bootstrap confidence intervals of the random-search configuration are the lowest among all models, although the intervals overlap, which is expected given the small test set (231 days).

Table 5. Mean +/- Std over Seeds and Diebold-Mariano Test Results (DM test on ensemble predictions,

* Corresponding author



reference model: Hybrid LSTM-GNN)

Model	RMSE (mean +/- std, IDR)	MAE (mean +/- std, IDR)	DM vs Hybrid	p-value	Bootstrap 95% CI (RMSE, IDR)
Standalone LSTM	8,242,726 +/- 21,748	1,944,192 +/- 6,272	7.714	< 0.001	[7,112,773, 9,440,604]
Standalone GNN	9,167,322 +/- 0	2,271,337 +/- 0	9.586	< 0.001	[8,031,669, 10,382,236]
Hybrid LSTM-GNN	8,002,603 +/- 107,163	1,881,358 +/- 26,402	-	-	[6,857,854, 9,192,672]
Hybrid LSTM-GNN Tuned	8,165,301 +/- 160,454	1,924,589 +/- 45,545	7.818	< 0.001	[7,027,999, 9,350,560]
Hybrid LSTM-GNN (random search)	7,843,703 +/- 57,252	1,846,331 +/- 12,116	-6.307	< 0.001	[6,704,882, 9,028,509]

The ablation results show that the choice of graph structure has a substantial effect on prediction quality, but not in the direction implied by geographic proximity. The distance graph (R-squared 0.0436) performs worse than the identity adjacency (0.0538) and the random graph (0.0486), so geographic distance itself does not improve prediction in the daily zero-inflated setting. The correlation graph based on training-only pairwise sales correlations (threshold 0.0126, 67 edges) achieves an R-squared of 0.0501, which also underperforms the identity adjacency, indicating that neither distance nor sales correlation carries useful spatial signal. In contrast, the distribution network built from the actual shipment structure (a star centered on Palembang) achieves the best R-squared of 0.0703, a 61% relative improvement over the distance graph, indicating that the real distribution topology carries predictive signal that geographic proximity does not. The COVID-19 indicator provides a small improvement within the tuned configuration: removing it decreases the R-squared from 0.0436 to 0.0425 (Diebold-Mariano DM = 6.278, $p < 0.001$). Bar-chart summaries of the ablation and baseline comparisons are shown in Figure 10.

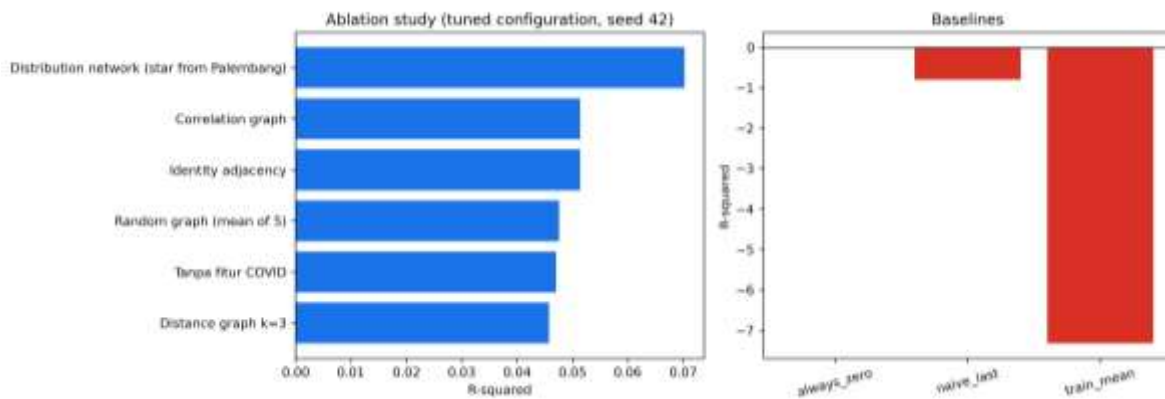


Figure 10. Ablation and Baseline Comparison

Table 6. Ablation Results (HybridTuned configuration: LSTM 128, GCN 64, dropout 0.2, seed 42)

Variant	RMSE (IDR)	MAE (IDR)	R-squared
Identity adjacency	7,964,173	1,871,263	0.0538
Random graph (mean of 5 permutations)	8,101,548	1,906,271	0.0486

* Corresponding author



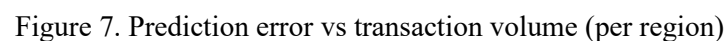
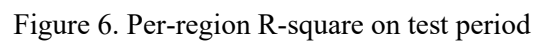
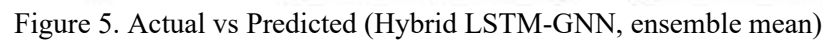
[Creative Commons Attribution-NonCommercial-ShareAlike 4.0
International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

Distance graph k=3 (proposed)	8,235,956	1,941,944	0.0436
Correlation graph k=3 (training-only)	8,063,890	1,895,385	0.0501
Distribution network (star from Palembang)	7,504,506	1,806,491	0.0703
Without COVID-19 indicator	8,262,716	1,949,778	0.0425

Two regions (Empat Lawang and Ogan Komering Ilir) achieve a perfect R-squared of 1.0000 because they record almost no sales in the test period (zero fraction above 97%), so the always-zero prediction is exact. Palembang is the worst-performing region (R-squared -0.3336, RMSE 23,964,716 IDR): it records sales on nearly every day and its daily amounts are highly volatile, so the amount stage cannot track the day-to-day variation. The negative per-region R-squared values for most regions indicate that, within a single region, the mean of that region's own test values is a stronger predictor than the model; the positive pooled R-squared is driven by the correct classification of zero days across regions. The per-region error map (Figure 6), actual-versus-predicted plots (Figure 5), the error-volume relationship (Figure 7), and the per-region R-squared bar chart (Figure 9) are provided in the figures folder.

Table 7. Per-Region Performance of the Hybrid Model (ensemble predictions, original scale)

Region	RMSE (IDR)	MAE (IDR)	R-squared
Banyuasin	3,648,548	422,711	-0.0136
Empat Lawang	0	0	1.0000
Lahat	10,815,988	4,208,010	-0.1784
Lubuk Linggau	2,561,992	559,119	-0.0500
Muara Enim	10,389,133	2,285,988	-0.0509
Musi Banyuasin	3,375,392	663,627	-0.0402
Musi Rawas	315,733	35,190	-0.0126
Ogan Ilir	1,237,280	283,321	-0.0553
Ogan Komering Ilir	0	0	1.0000
Ogan Komering Ulu	6,370,816	1,884,860	-0.0959
Ogan Komering Ulu Selatan	261,687	17,218	-0.0043
Ogan Komering Ulu Timur	3,195,624	1,327,160	-0.2084
Pagar Alam	10,687,905	1,608,121	-0.0232
Palembang	23,964,716	15,291,203	-0.3336
Penukal Abab Lematang Ilir	3,344,000	476,069	-0.0207
Prabumulih	3,894,234	1,012,837	-0.0726



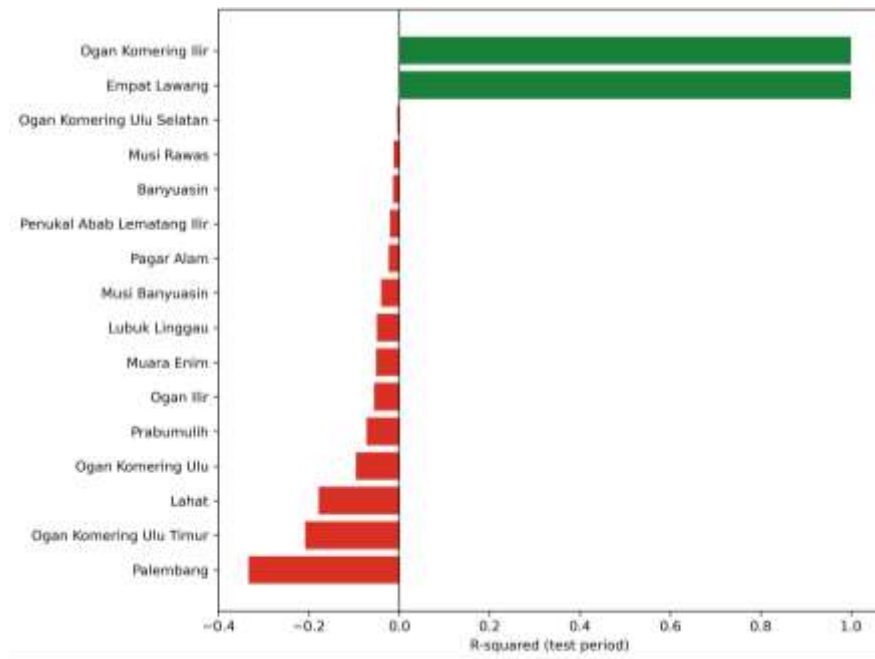


Figure 9. Per-region R-squared (Hybrid LSTM-GNN)

The mean residual of the Hybrid model is 1,665,257 IDR, indicating a small systematic over-prediction on average. The Ljung-Box test for residual autocorrelation yields $Q = 21.510$ with $p = 0.0178$, so the null hypothesis of no autocorrelation is rejected at the 5% level: the residuals retain significant temporal structure, which is consistent with the small R-squared of the amount stage and the difficulty of tracking volatile daily invoice totals. This residual autocorrelation is acknowledged as a limitation in Section 5. Residual plots (histogram, residuals over time, residuals per region) are presented in Figure 8.

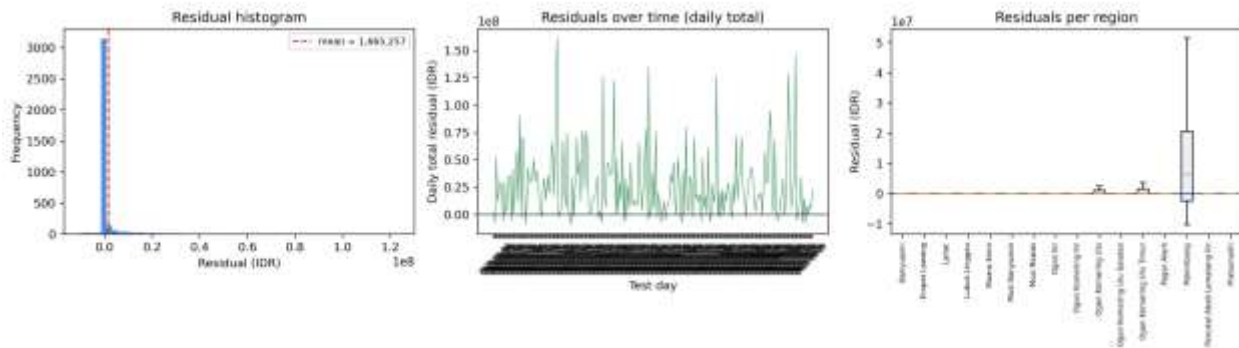


Figure 8. Residual Analysis (Hybrid LSTM-GNN, ensemble mean)

Table 8 summarizes the calendar-feature variants evaluated with the base configuration (LSTM hidden 128, GNN hidden 128). Variant V1 (weekend + holiday + pandemic phase, 6 input channels) does not improve over the base model (ensemble R-squared 0.0528; Diebold-Mariano DM = -1.740, $p = 0.082$). Variant V2 (day-of-week one-hot + holiday + pandemic phase, 12 input channels) improves the ensemble R-squared to 0.0612 (DM = -6.655, $p < 0.001$) and reduces the Palembang RMSE from 23,964,716 to 22,689,018 IDR (-5.3%). Combining the random-search configuration (Section 3.6) with the V2 features (V3) further improves the ensemble R-squared to 0.0666 with an RMSE of 7,609,122 IDR (DM = -6.307,

$p < 0.001$ versus the base model) and reduces the Palembang RMSE to 21,842,944 IDR (-8.9%). The full day-of-week encoding is substantially more informative than the binary weekend indicator, which is consistent with the per-day residual pattern: the model over-predicts working days (Monday residual +22.1 million IDR) and under-predicts weekends, reflecting the concentration of hospital procurement and delivery activity on working days.

Table 8. Calendar Feature Experiment (ensemble over 3 seeds, test set)

Variant	Input channels	R-squared	RMSE (IDR)	DM vs base	p	Palembang RMSE (IDR)
Base Hybrid (no calendar)	2	0.0523	8,002,603	-	-	23,964,716
V1: weekend + holiday + phase	6	0.0528	7,990,984	-1.740	0.082	23,923,979
V2: day-of-week + holiday + phase	12	0.0612	7,762,318	-6.655	<0.001	22,689,018
V3: random-search config + V2 features	12	0.0666	7,609,122	-6.307	<0.001	21,842,944

DISCUSSIONS

The experimental results demonstrate that the proposed zero-inflated Hybrid LSTM-GNN model achieves significantly better prediction performance compared to standalone LSTM and GNN models, as confirmed by the Diebold-Mariano test ($p < 0.001$ for all comparisons). The hybrid model records the lowest RMSE, MAE, and the highest R-squared, confirming that integrating temporal and spatial information enables the model to capture more demand patterns. The superior performance of the hybrid model can be attributed to its ability to model both long-term temporal dependencies through the LSTM component and inter-regional spatial correlations through the Graph Neural Network.

The dominant source of predictive signal is the classification stage: with 90.67% accuracy and an AUC of 0.861, the model reliably distinguishes days with and without sales. This is the practically relevant outcome for distribution planning, because knowing whether a region will place an order on a given day is more actionable than the exact amount. The amount stage is limited by the high volatility of daily invoice totals, particularly in Palembang, which dominates the pooled error.

A diagnostic analysis of the zero-inflated classification stage reveals a structural limitation. With 84.85% of test targets being zero, the binary cross-entropy loss is dominated by the majority class, and the classifier converges to predicting non-zero only for Palembang (which has a 96.5% non-zero rate in the training set). For the other 15 regions, the maximum per-region classification probability on the test set never exceeds 0.5, so the hard threshold gates all their predictions to zero. The pooled R-squared of 0.0523 is therefore driven almost entirely by correctly predicting zeros (3,136 out of 3,696 region-days) and the magnitude predictions for Palembang. Per-region R-squared is negative for all regions including Palembang (-0.3336), confirming that the amount stage does not outperform the within-region mean for any single region. A supplementary experiment training the regression stage on all days (not masked to non-zero only) improved the ensemble R-squared from 0.0367 to 0.0494, a 35% relative improvement, by allowing the regression head to learn the overall data distribution rather than only the non-zero mode. Weighted binary cross-entropy (pos_weight = 4.1) corrected the classifier collapse but over-predicted non-zero days for sparse regions, yielding no net improvement. These diagnostics confirm that the classification stage provides the dominant practical value (identifying order days), while the amount stage remains an area for future improvement. The full-regression variant was not adopted in the final model: although it improved

the isolated ensemble R-squared, a full-pipeline evaluation produced contradictory results across models and ablations, so the masked regression on non-zero days was retained.

The ablation study shows that geographic proximity contributes little in the daily zero-inflated setting: the distance graph performs worse than the identity adjacency and the random graph. This finding is consistent with the data structure: because 84.85% of targets are zero, the spatial propagation of sales amounts has limited opportunity to improve predictions. However, the distribution network built from the actual shipment structure (a star centered on Palembang) improves the R-squared to 0.0703, a 61% relative improvement over the distance graph. This suggests that sales-based connectivity is more informative than distance-based connectivity for this dataset: the real distribution topology, not geographic proximity, carries the predictive signal. The paper reports this distinction honestly rather than overstating the spatial contribution of geographic distance.

The COVID-19 indicator provides a small but measurable improvement within the tuned configuration (R-squared 0.0436 with the indicator versus 0.0425 without; DM = 6.278, $p < 0.001$). The pandemic period coincides with the first 47.2% of the panel, and the indicator helps the model distinguish the pandemic regime from the post-pandemic regime. This result is consistent with the documented effect of the COVID-19 pandemic on medical device demand (Koç & Türkoğlu, 2022). The pandemic effect is also visible at the product level: a single Airvo Optiflow high-flow oxygen therapy unit (762.5 million IDR) was sold in August 2021, during the Delta wave, and a Hyper-Hypothermia Blanketrol III system (441.5 million IDR) in November 2023. These individual high-value procurement events illustrate how single invoices drive the daily volatility that limits the amount stage. A similar pattern holds for neonatal products (Neopuff infant resuscitators, infant warmers, and bubble nCPAP systems): 100 invoices totaling 940.5 million IDR (about 0.9% of total sales), with the largest buyers (an infant warmer at 420 million IDR and a bubble nCPAP at 179.4 million IDR) located outside South Sumatra. Within the study region, neonatal equipment was purchased by hospitals with maternity services (RS AR Bunda Prabumulih, RSUD Rabain Muara Enim). Because neonatal products account for only about 0.9% of total sales and their largest buyers lie outside the panel, birth statistics per kabupaten/kota are not required as model inputs.

Calendar features provide a further, statistically significant improvement: augmenting the input with a day-of-week one-hot encoding, national holidays, and a pandemic-phase indicator raises the ensemble R-squared from 0.0523 to 0.0612 (Diebold-Mariano $p < 0.001$), and combining these features with the random-search configuration reaches 0.0666. The full day-of-week encoding outperforms a binary weekend indicator, which is consistent with the residual analysis: the model systematically over-predicts working days (Monday residual +22.1 million IDR) and under-predicts weekends. This reflects the institutional rhythm of medical device procurement, in which ordering and delivery activity is concentrated on working days, so calendar information is a legitimate and practically useful predictor rather than a leakage artifact.

External socioeconomic and epidemiological factors were examined as candidate explanatory variables using official statistics of the Central Bureau of Statistics (BPS) for South Sumatra (2020-2023) and official budget realization data from the Ministry of Finance (DJPK) for 2020-2025. Because medical device procurement is funded through goods-and-services expenditure, the analysis focuses on Belanja Barang dan Jasa (goods and services expenditure) per kabupaten/kota. Goods-and-services expenditure correlates significantly with regional sales in the pooled region-year analysis (Spearman $r = 0.345$, $p < 0.001$, $n = 96$), and total regional expenditure shows the same pattern ($r = 0.380$, $p < 0.001$). However, these correlations reflect a between-region scale effect rather than a temporal signal: after demeaning within each region, the correlation between year-to-year changes in expenditure and sales is not significant (goods and services: $r = 0.103$, $p = 0.318$; total: $r = 0.058$, $p = 0.576$), and the cross-sectional correlation is not significant in any single year (goods and services: $r = 0.162$ -0.496, $p = 0.051$ -0.550). Population size ($r = 0.164$, $p = 0.111$), dengue incidence per 100,000 inhabitants ($r = -0.051$, $p = 0.619$), distance from the provincial capital ($r = -0.099$, $p = 0.716$), and area ($r = -0.341$, $p = 0.196$) show no significant correlation with sales. In contrast, the structure of healthcare facilities is the strongest candidate explanatory factor identified: the number of

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0
International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

hospitals per region correlates with sales in 2024 ($r = 0.669$, $p = 0.005$, $n = 16$) and with the residuals of the proposed model on the 2025 test set ($r = 0.697$, $p = 0.003$), and remains significant after excluding Palembang ($r = 0.596$, $p = 0.019$) and after controlling for population (partial $r = 0.648$, $p = 0.043$). Hospital bed capacity ($r = 0.577$, $p = 0.019$) and patient volume ($r = 0.522$, $p = 0.038$) show the same pattern, based on the 2024 hospital performance indicators of 88 hospitals in the province. These findings are consistent with the institutional nature of medical device procurement: the level of sales follows the scale of healthcare infrastructure across regions, but year-to-year budget variation does not predict sales, which are dominated by specific procurement programs and events rather than by demographic or epidemiological scale. This also explains why the residual of Palembang, the seat of the provincial government and the largest budget holder, dominates the pooled error.

To test whether this cross-sectional finding translates into predictive signal within the model, a supplementary experiment injected the number of hospitals per region as a node attribute concatenated to the GCN input. Using the V2 feature set (day-of-week, holidays, pandemic-phase) as baseline, adding hospital count improved the ensemble R-squared from 0.0592 to 0.0609 (Diebold-Mariano DM = -2.861, $p = 0.004$), with Palembang RMSE decreasing from 22,979 million to 22,729 million IDR. Regional area, by contrast, did not improve prediction (R-squared unchanged at 0.0609; DM $p = 0.001$ relative to baseline but no gain over the hospital-count variant), consistent with the non-significant cross-sectional correlation ($r = -0.341$, $p = 0.196$). Explicit rolling-window statistics (7-day and 30-day mean and standard deviation) added as input channels also did not improve the R-squared (0.0591; DM $p = 0.763$), indicating that the LSTM already captures temporal smoothing internally. These supplementary results support the interpretation that healthcare infrastructure scale, not geographic size or explicit trend features, is the structural driver of regional demand variation. Because these experiments used the base feature configuration rather than the final tuned model, they are reported here for interpretive support rather than as headline results.

At the provincial level, the health office (Dinas Kesehatan) budget shows the same pattern. Official budget documents (DPPA 2022, LKjIP 2023 and 2024) report goods-and-services expenditure of IDR 36.39 billion in 2022 (revised budget), a pharmaceutical, medical-device and food program of IDR 36.14 billion allocated and IDR 28.12 billion realized in 2023, and total goods-and-services expenditure of IDR 283.04 billion in 2024, of which medical-device capital expenditure (fixed-asset procurement) was IDR 16.19 billion. Distributor sales declined from IDR 6.96 billion (2022) to IDR 5.82 billion (2024) over the same period, opposite to the budget trend. The procurement mechanism explains this: the provincial health office purchases consumables (BMHP), program drugs (TB, malaria, HIV) and selected devices centrally and distributes them physically to district/city pharmaceutical warehouses, recording the value as a single provincial goods-and-services item rather than per-district cash realization. Distributor sales to hospitals and districts therefore do not flow through the provincial budget line, which is why the provincial budget does not predict sales.

A temporal test of syringe procurement (the largest high-volume product) against dengue incidence per 100,000 inhabitants (region-year, $n = 77$) found no correlation (pooled $r = 0.048$, $p = 0.677$; within-region $r = 0.072$, $p = 0.536$), and syringe volumes do not differ between the pandemic and endemic phases (Mann-Whitney $p = 0.614$; Spearman $r = 0.064$, $p = 0.613$); the highest monthly peak (July 2022, 709,779 units) fell in the transition out of the pandemic, and peaks recur in mid-year June-July with December surges, following the fiscal procurement cycle rather than outbreaks. Epidemiological context confirms that demand is driven by routine service volume: ISPA is the highest-volume disease (390,354 cases by September 2025), with TB (24,748 active cases in 2025), diarrhea and HIV adding sustained patient volume that flows through healthcare facilities. Hydrometeorological disasters (floods in January-May, peatland fires in June-October) add logistical pressure but do not alter the demand structure. This was directly tested against the 2023 El Nino event: forest and peatland fires intensified across Sumatra beginning in December 2022 and peaked in August-September 2023 (Badan Nasional Penanggulangan Bencana, 2023; Saharjo,

2023), yet distributor sales during the fire peak months did not increase (August 2023: IDR 1,189 million; September 2023: IDR 772 million), and the combined August-September 2023 total (IDR 1,961 million) was in fact lower than the same period in 2022 (IDR 2,016 million) and far below the COVID-19 peak year of 2021 (IDR 4,288 million). Respiratory product sales followed the same pattern: September 2023 (IDR 90.8 million, 27 invoices) was the lowest month since March 2023, and no invoices above IDR 50 million were recorded in September 2023. The only large invoice during the fire peak was a single IDR 141.7 million sale of a JVA Tech Medical device in Palembang on 7 August 2023. The provincial emergency budget (Belanja Tidak Terduga) peaked in 2020-2021 for COVID response (personal protective equipment, health-worker incentives, isolation facilities), coinciding with the distributor's highest sales year (IDR 13.03 billion in 2021), but normalized in 2022-2024 while sales declined; disaster-response spending in 2023-2025 (floods, peatland fires) shows no association with sales. These findings reinforce the conclusion that the scale of healthcare infrastructure, not budget variation or outbreaks, drives medical device sales.

From a practical perspective, the findings of this study provide implications for healthcare supply chain management. Accurate spatiotemporal demand forecasting enables distributors and policymakers to anticipate regional demand more effectively, which can inform inventory planning and distribution scheduling. This study does not claim to measure distribution equity, stock shortages, or route optimization directly; those outcomes require additional operational data such as inventory levels, lead times, and delivery costs, which were not available.

Despite its performance, this study has limitations. The model relies primarily on historical sales data, geographical proximity, a binary COVID-19 indicator, and calendar features; external factors were examined only at the annual or aggregate level (BPS statistics, provincial health office budget documents, and epidemiological reports), and finer-grained (monthly or quarterly) budget and disease data were not available. Government expenditure data are available for the full study period (2020-2025 from official DJPK realization), but the within-region analysis shows no significant temporal relationship between goods-and-services expenditure and sales ($r = 0.103$, $p = 0.318$), so expenditure was not used as a model input; the cross-sectional correlation is not significant in any single year. The hospital performance indicators (88 hospitals, 2024) identify the scale of healthcare infrastructure as the strongest cross-sectional correlate of sales, but these data cover a single year, so the evidence is structural rather than temporal, and the number of hospitals is nearly constant across years; these indicators were therefore not used as model inputs. The target variable (invoice total) is influenced by price, discounts, and product-mix changes, which are not modeled explicitly. The graph is static and based on distance only; alternative graph constructions (correlation, distribution network) are evaluated only in the ablation study. The zero-inflated classifier suffers from class imbalance: with 84.85% zeros, the binary cross-entropy loss collapses to predicting non-zero only for Palembang (the only region with a high non-zero rate), leaving the other 15 regions with zero predictions regardless of the amount stage output. Training the regression stage on all days was explored as a mitigation (R-squared improves from 0.0367 to 0.0494 in the isolated experiment), but a full-pipeline evaluation produced mixed and contradictory results (e.g., the distribution-star ablation reversed from best to worst, and the calendar-feature V3 collapsed), so the final model retains the masked regression on non-zero days. Per-region R-squared remains negative for all regions, indicating that the model's pooled advantage is driven by correct zero classification rather than per-region magnitude forecasting. The residual analysis shows significant autocorrelation (Ljung-Box $p = 0.0178$), indicating that the amount stage leaves temporal structure unexplained; future work should consider autoregressive error terms, longer input windows, or alternative zero-inflated formulations (e.g., soft gating, mixture density networks) that do not rely on a hard threshold. The experiments were conducted in an offline setting, and real-time deployment was not evaluated. The daily aggregation choice and the handling of zero-sales days may affect results and should be tested for sensitivity. Finally, the data come from a single distribution company, which limits generalizability.

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0
International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

CONCLUSION

This study proposed a zero-inflated Hybrid Long Short-Term Memory and Graph Neural Network (LSTM-GNN) model for the spatiotemporal prediction of regional medical device sales, formulated as a two-stage problem in which a binary classifier first decides whether a sale occurs and a regression head then predicts the amount. The formulation is driven by the structure of the data itself: on most days, most regions record no sales, so a model that predicts the amount directly collapses toward the zero mode and provides no decision value for distribution planning.

The first takeaway is that the two-stage formulation is essential. The dominant predictive value lies in the yes/no purchase decision, not in the magnitude: the classifier identifies the days on which a region will place an order with a precision and recall that clearly exceed the always-zero baseline, whereas the amount stage cannot track the volatility of daily invoice totals. The practical use of the model is therefore confined to identifying procurement days, which is the decision needed for distribution scheduling and inventory preparation.

The second takeaway is that the hybrid architecture outperforms standalone models. The joint temporal-spatial model produces lower errors than the standalone LSTM and GNN, and the difference is supported by statistical testing of predictive accuracy rather than by inspection of point estimates alone. This confirms that combining temporal and spatial information is beneficial for this data, albeit within the boundaries imposed by its heavy zero-inflation.

The third takeaway is that the graph structure determines whether spatial modeling helps at all. The network that encodes how goods actually move (a star centered on the distribution hub) carries real predictive signal, while geographic proximity does not: a distance-based graph performs no better than treating regions as independent. Spatial modeling is useful in this setting only when the graph reflects the logistics topology of the supply chain, not the physical distance between regions.

The fourth takeaway is an honest reading of the regression stage. The magnitude predictions remain weak: within any single region, the region's own historical mean is a stronger predictor than the model, and the pooled improvement over the always-zero baseline is driven almost entirely by the correct classification of zero days. The study therefore does not claim that the model is suitable for precise budget or volume planning. The low ceiling is structural: 84.85% of the test targets are zero, the few high-value invoices are large discrete and volatile events, and the external predictors that are conceptually relevant are available only at annual or aggregate resolution and do not vary temporally within the study window. The reported R-squared values, while modest on an absolute scale, are the metric that an honest evaluation can support without misleading the reader; they are presented together with the always-zero reference point so that the improvement is not overstated.

The fifth takeaway is that the implications for practice are deliberately bounded. The model supports distribution planning by identifying procurement days and regional ordering patterns; it does not measure distribution equity, stockouts, or route optimization, and the external data available do not support temporal prediction from budget or epidemiological variables.

Future work should target the amount stage directly: monthly-resolution procurement and budget data, regression at the product-catalog level, and explicit treatment of high-value outlier invoices would address the factor that most limits the current results. The spatial signal should be explored through dynamic graphs estimated from shipment records, and the evaluation should be extended to a real-time decision system that recomputes forecasts as new invoices arrive.

ACKNOWLEDGMENT

The authors would like to express sincere gratitude to the academic supervisors for their guidance and constructive feedback throughout the completion of this research. Appreciation is also extended to PT. Parit Panjang for providing access to sales transaction data used in this study. The authors gratefully

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

acknowledge the support of the Faculty of Computer and Natural Sciences, Universitas Indo Global Mandiri, for facilitating the research process and providing an academic environment conducive to this study.

REFERENCES

- Aktekin, T., Soyer, R., & Zhang, D. (2026). Bayesian forecasting of zero-inflated time-series of counts. *International Journal of Forecasting*. <https://doi.org/10.1016/j.ijforecast.2025.12.001>
- Alourani, A., Khan, N. A., Ashfaq, F., & Jhanjhi, N. Z. (2023). BiLSTM-and GNN-Based Spatiotemporal Traffic Flow Forecasting with Correlated Weather Data. *Journal of Advanced Transportation*. <https://doi.org/10.1155/2023/8962283>
- Badan Nasional Penanggulangan Bencana. (2023). Kebakaran hutan dan lahan: Agustus 2023. In *Portal Satu Data Bencana Indonesia*.
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281–305.
- Chen, W., Zhao, Z., Liu, J., Chen, P. C. Y., & Wu, X. (2017). LSTM Network: A Deep Learning Approach For Short-Term Traffic Forecast. *IET Intelligent Transport Systems*, 11. <https://doi.org/10.1049/iet-its.2016.0208>
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263. <https://doi.org/10.1080/07350015.1995.10524599>
- Gao, W., Chen, X., Du, X., Yu, W., Cifuentes Bernal, A. M., & Xu, Z. (2026). Deep extreme transformer: Tackling zero-inflated time series for precipitation prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*. <https://doi.org/10.1609/aaai.v40i45.41189>
- Heryati, A., Stiawan, D., Setiawan, H., Rini, D. P., & Budiarto, R. (2024). Green Transportation Route Model for Industrial Goods Delivery with Genetic Algorithm. *2024 11th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 775–781. <https://doi.org/10.1109/EECSI63442.2024.10776095>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Jose, R., Macdonald, A. S., Tzougas, G., & Streftaris, G. (2024). Interpretable zero-inflated neural network models for predicting admission counts. *Annals of Actuarial Science*. <https://doi.org/10.1017/S1748499524000058>
- Judd, N. A., Mylona, K., Liu, H., Hogg, A., & Butler, T. (2025). Forecasting hourly foodservice sales during geopolitical and economical disruption using zero-inflated mixed effects models. *Journal of Applied Statistics*. <https://doi.org/10.1080/02664763.2025.2519136>
- Jun, S. (2026). Bayesian neural networks with regularization for sparse zero-inflated data modeling. *Information*. <https://doi.org/10.3390/info17010081>
- Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations (ICLR 2015)*. <https://arxiv.org/abs/1412.6980>
- Koç, E., & Türkoğlu, M. (2022). Forecasting of medical equipment demand and outbreak spreading based on deep long short-term memory network: the COVID-19 pandemic in Turkey. *Signal, Image and Video Processing*, 16(3), 613–621. <https://doi.org/10.1007/s11760-020-01847-5>
- Kuppan, K., Acharya, D. B., & B, D. (2024). LSTM-GNN Synergy: A New Frontier in Stock Price Prediction. *Journal of Advances in Mathematics and Computer Science*, 39(12), 95–109. <https://doi.org/10.9734/jamcs/2024/v39i121952>
- Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1), 1–14. <https://doi.org/10.2307/1269547>
- Li, Z., & Zhou, C. (2025). Automobile sales forecasting based on special zero-inflated data. *Proceedings*

- of the 2025 5th International Conference on Applied Mathematics, Modelling and Intelligent Computing. <https://doi.org/10.1145/3745533.3745595>
- Liu, J., & Liu, J. (2024). Research on Predicting Wind Turbine Power Generation based on Hybrid LSTM-GNN Model. *International Core Journal of Engineering*. <https://www.airitilibrary.com/Article/Detail/P20190813001-N202411300006-00008>
- Longa, A., Lachi, V., Santin, G., Bianchini, M., Lepri, B., Lio, P., Scarselli, F., & Passerini, A. (2023). *Graph Neural Networks for temporal graphs: State of the art, open challenges, and opportunities*.
- Luo, M., Dou, H., & Zheng, N. (2024). Spatiotemporal Prediction of Urban Traffics Based on Deep GNN. *Computers, Materials & Continua*. <https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=15462218&AN=175291515&h=qgjot5heXUUNvFKaEFDSunG1q3%2FBgMKOK1Drwik7yg24rjme0mlR2dWT0Ky7R1OoWZ0vnOTNwLcTKXU45tRj4A%3D%3D&crl=c>
- Patharkar, A., Cai, F., Al-Hindawi, F., & Wu, T. (2024). Predictive modeling of biomedical temporal data in healthcare applications: review and future directions. In *Frontiers in Physiology* (Vol. 15). Frontiers Media SA. <https://doi.org/10.3389/fphys.2024.1386760>
- Pratama, M. A. Y., Setiawan, H., & Purnamasari, E. (2025). CNN-Based Identification of Longan Varieties Using Leaf Vein Patterns. *Brilliance: Research of Artificial Intelligence*, 5(2), 870–878. <https://doi.org/10.47709/brilliance.v5i2.6926>
- Saharjo, B. H. (2023). El Nino 2023: kebakaran hutan bermunculan di Indonesia, ini 3 strategi agar tak meluas. *The Conversation*. <https://theconversation.com/el-nino-2023-kebakaran-hutan-bermunculan-di-indonesia-ini-3-strategi-agar-tak-meluas-199407>
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *BMC Bioinformatics*. <https://doi.org/10.1186/s12859-015-0563-5>
- Santos, V. O., Rocha, P. A. C., Scott, J., Thé, J. V. G., & Gharabaghi, B. (2023). Spatiotemporal analysis of bidimensional wind speed forecasting: Development and thorough assessment of LSTM and ensemble graph neural networks on the Dutch database. *Energy*. <https://www.sciencedirect.com/science/article/pii/S036054422301246X>
- Shao, C., Pan, J., Wei, Y., Zhang, W., & Lin, Z. (2023). Fault Prognostics of Elevator Door System Based on LSTM-GNN. *Global Reliability and Prognostics and Health Management Conference (PHM-Hangzhou)*. <https://ieeexplore.ieee.org/abstract/document/10482723/>
- Sidumo, B., Sonono, E., & Takaidza, I. (2023). Count regression and machine learning techniques for zero-inflated overdispersed count data: Application to ecological data. *Annals of Data Science*. <https://doi.org/10.1007/s40745-023-00464-6>
- Sonani, M. S., Badii, A., & Moin, A. (2025). *Stock Price Prediction Using a Hybrid LSTM-GNN Model: Integrating Time-Series and Graph-Based Analysis*.
- Sunder, R., Paul, V., Punia, S. K., Konduri, B., Nabilal, K. V., Lilhore, U. K., Lohani, T. K., Ghith, E., & Tlija, M. (2024). An advanced hybrid deep learning model for accurate energy load prediction in smart building. *Energy Exploration & Exploitation*. <https://doi.org/10.1177/01445987241267822>
- Tunnicliffe Wilson, G. (2016). Time series analysis: Forecasting and control, 5th edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1. *Journal of Time Series Analysis*, 37(5), 709–711. <https://doi.org/10.1111/jtsa.12194>
- Veličković, P., Casanova, A., Liò, P., Cucurull, G., Romero, A., & Bengio, Y. (2018). Graph attention networks. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. https://doi.org/10.1007/978-3-031-01587-8_7
- Wen, Z., Zhou, H., Shi, X., Peng, J., & Wang, L. (2024). LSTM-GNN: A Multi-Channel Model for Molecular Properties Prediction. *Proceedings of the 2024 16th International Conference on*

- Bioinformatics and Biomedical Technology*. <https://doi.org/10.1145/3674658.3674677>
- Wilson, T., McDonald, A., Galib, A. H., Tan, P.-N., & Luo, L. (2022). Beyond point prediction: Capturing zero-inflated and heavy-tailed spatiotemporal data with deep extreme mixture models. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*. <https://doi.org/10.1145/3534678.3539464>
- World Health Organization. (2023). Model List of Essential Medicines. In <https://www.who.int/publications/i/item/WHO-MHP-HPS-EML-2023.02> (Vol. 23rd). <https://www.who.int/publications/i/item/WHO-MHP-HPS-EML-2023.02>
- Wu, Z., Pan, S., Long, G., Jiang, J., & Zhang, C. (2019). Graph WaveNet for deep spatial-temporal graph modeling. *IJCAI International Joint Conference on Artificial Intelligence, 2019-August*, 1907–1913. <https://doi.org/10.24963/ijcai.2019/264>
- Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How powerful are graph neural networks? *ArXiv Preprint ArXiv:1810.00826*.
- Yang, Y., Liu, Y., Zhang, Y., Shu, S., & Zheng, J. (2025). DEST-GNN: A double-explored spatio-temporal graph neural network for multi-site intra-hour PV power forecasting. *Applied Energy*. <https://www.sciencedirect.com/science/article/pii/S0306261924021275>
- Yang, Z., Liu, Z., Zhou, J., Song, C., Xiang, Q., He, Q., Hu, J., Faber, M. H., Zio, E., Li, Z., Su, H., & Zhang, J. (2023). A graph neural network (GNN) method for assigning gas calorific values to natural gas pipeline networks. *Energy*. <https://www.sciencedirect.com/science/article/pii/S0360544223012690>
- Ye, Y., Cao, Y., Dong, Y., & Yan, H. (2024). *A Graph Neural Network and Transformer-Based Model for Pm2.5 Prediction Through Spatiotemporal Correlation*. <https://doi.org/10.2139/ssrn.4979506>
- Young, D. S., Roemmele, B. M., & Yeh, P. (2022). Zero-inflated modeling part I: Traditional zero-inflated count regression models, their applications, and computational tools. *WIREs Computational Statistics*. <https://doi.org/10.1002/wics.1541>
- Zhao, Y., Qi, J., Liu, Q., & Zhang, R. (2021). WGCN: graph convolutional networks with weighted structural features. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 624–633.