

HealthSurge AI: An Agentic Deep Learning Framework for Predictive Patient Influx Management in Indian Hospitals

Shrikant Shinde¹, Aditya Jagtap², Ayush Holkar³

^{1,2,3}Department of Computer Science and Engineering

^{1,2,3} MIT Art, Design and Technology University, Pune, Maharashtra, India

1shri1506g@gmail.com, 2adityasjagtap2@gmail.com, 3ayushholkar87@gmail.com

Abstract - If you are in a government hospital in Maharashtra on a Tuesday afternoon during a festival weekend, you do not need to be a statistician to know there is a problem. The wards spill into corridors. One nurse covers three areas. Parents sit on the floor because there are no chairs. What is not immediately apparent is that most of this is foreseeable. The information to predict those spikes exists in air quality data, government reports, and public event listings. The problem is that this data never reaches those who schedule nurses. That is where we created HealthSurge AI. It is a Long Short-Term Memory (LSTM) network trained on four sources: past admissions, CPCB air pollution, a festival proximity score, and an IDSP outbreak index. An agent layer generates actionable advice via email or SMS in English and Marathi. Evaluated on 730 days of real-world data from three hospitals in Maharashtra, the 48-hour mean absolute percentage error (MAPE) was 8.3 percent. All real surges in the test period were captured before peak, with a precision of 91.6 percent and recall of 94.2 percent.

Key Words: Patients arrival rate, hospital management, LSTM, agentic AI, Indian hospitals, AQI, surge, disaster planning

1. INTRODUCTION

This work began during a site visit to a public hospital in Pune as part of our bachelor's course. The outpatients' queue extended beyond the gate. The emergency room had spilled into the corridor in what appeared to be a normal occurrence. When asked how staff levels were set, the administrator replied: "mainly experience and an instinct we have gained over time." No model existed. No data was reviewed. Surges were absorbed after they arrived.

India has less than one doctor per 1,500 people, less than the WHO recommendation [1]. There is little room for overflow. A 30 percent increase in demand results in collapse. The Kumbh Mela of 2019 overwhelmed local hospitals to the extent that a national response was needed [2]. Each winter in Delhi, an AQI greater than 400 causes a 40 percent spike in respiratory emergencies within 72 hours [3]. Dengue has a seasonal pattern any district health officer can draw from memory. These are predictable surges whose indicators are present in open data weeks before they arrive.

Failure to heed these warnings is costly. Studies on preventable deaths in resource-poor settings consistently point to institutional readiness [4]. ED overcrowding predicts poorer outcomes across virtually all downstream measures [5]. HealthSurge AI addresses this with an LSTM [6] model trained on historical admissions combined with three external covariate streams: CPCB air quality scores, festival proximity scores, and

IDSP outbreak scores. It provides predictions for 1-day, 3-day, and 7-day horizons. When the 48-hour projection crosses 85 percent of a facility's registered capacity, an agent layer translates the numbers into plain-language operational recommendations covering nursing staffing and consumable inventories.

Our contributions include: (1) a multi-input LSTM model trained on admissions, CPCB air quality index, festival proximity values, and IDSP outbreak indicators; (2) an agent layer translating forecasts into operational recommendations; (3) an evaluation on data from three Maharashtra hospitals over 730 days; and (4) a deployed version at <https://health-surge-ai.vercel.app>.

2. RELATED WORK

2.1 Hospital Demand Forecasting

The use of computers for hospital demand forecasting dates to about 20 years ago. Early studies applied discrete event simulation to patient flow at unit level [7]. Data-driven approaches using historical data proved more transferable. A large comparative review found neural network and ensemble methods outperformed regression baselines in over 80 percent of head-to-head comparisons. An Israeli hospital study showed ML-driven bed management led to an 18 percent reduction in diversion hours [8].

2.2 LSTM for Time-Series

LSTMs became the default approach for clinical prediction because patient data is inherently temporal. One of the first widely reported uses of LSTM for predicting events from EHR data was Doctor AI [9]. Lipton et al. compared LSTM to logistic regression and vanilla RNNs on MIMIC-III [10], finding LSTM superior. Rostami et al. used Google Trends to predict influenza hospitalisations with a MAPE of 11.2 percent [11]. We obtained 8.3 percent, part of which we attribute to using direct environmental signals rather than internet search data.

2.3 Air Quality and Hospital Admissions

Air pollution and acute healthcare needs are extensively studied across Indian cities. A study of PM2.5 in Delhi reported a stable lagged increase in respiratory emergency visits after high-pollution days [3]. In India, air pollution is responsible for about 1.67 million deaths a year [12]. We modelled AQI as a moving average rather than a single daily value so the model can distinguish a single bad-air day from four consecutive bad-air days.

2.4 Agentic AI in Healthcare

The idea of an AI system progressing from analysis to operational decision support has been discussed extensively. Topol discussed the possibility of systems linking data collection and operational support [13]. Recently, multi-agent language models were proposed to manage actions across clinical workflow steps [14]. Our work explores this concept in a limited way: the agent reads LSTM-generated forecasts and drafts specific actions, but all recommended resource changes must be approved by a human administrator.

3. PROBLEM STATEMENT AND APPROACH

3.1 Problem Definition

The question is whether past admission data can be combined with observable external signals to predict the number of patients for a hospital one, three, and seven days in the future, and whether such forecasts can be automatically translated into specific operational guidelines for an expected surge. The Indian health data environment complicates the problem: reporting delays are common, documentation standards vary between facilities, and public holiday admissions may reflect data-entry artefacts.

3.2 Data Sources

Four data streams were integrated into a daily pipeline. Patient admission data was obtained from hospital management systems at three participating hospitals. Air quality measures — PM2.5, PM10, and a composite AQI — were pulled from CPCB stations via the OpenAQ API. Event proximity was modelled with a decay function for 47 national and regional festivals: $\max(0, 1 - d/7)$, where d is days to the event. Outbreak status was derived from IDSP weekly bulletins as a binary variable forward-filled between bulletins. Continuous variables were filled using linear interpolation.

3.3 Proposed Approach

We formulated the problem as sequence regression. For each day, we input the previous 30 days of admission counts and the current day's covariate data, and predict counts for the following 1, 3, or 7 days. The lookback window of 30 days was determined empirically (testing 7, 14, 21, 30, and 45 days). Shorter windows failed to capture monthly seasonality; longer windows decreased sensitivity to recent changes. A rule-based agent compares forecast projections to facility capacity levels. When the 48-hour forecast exceeds 85 percent of registered capacity, the engine recommends nursing adjustments, checks consumable inventory, and sends bilingual notifications. Administrators must approve actions before they are finalised.

4. SYSTEM ARCHITECTURE

The system is built around three components selected against a single enabling constraint: it had to be cheap to run without an IT person — a reality in most public hospitals in India. Table 1 shows the technology stack.

Table 1: System Component Overview

Component	Technology	Function
Frontend	React 18, Vite 5, Tailwind CSS	Forecasts, alerts, recommendations
Backend	Python 3.11, FastAPI, ONNX	LSTM inference, recommendations
Database	Convex (serverless)	Census, roles, audit logs
Notifications	Twilio SMS, SendGrid	Surge alerts to administrators
Security	JWT, TLS 1.3, AES-256	Encryption in transit and at rest

The dashboard is developed using React 18 with Vite 5, styled with Tailwind CSS. During a walkthrough at H1, we learned that ward tablets are used under fluorescent lighting at arm's length during rounds. Our initial colour scheme was unreadable in that setting. A nursing supervisor noticed this during an informal session, and we subsequently redesigned colours to conform to WCAG contrast guidelines. The home screen displays a 7-day forecast chart with confidence bands, a ward occupancy heatmap, and an alert strip that transitions from yellow to orange to red as ward load increases. Convex streams updates to the browser with a lag of less than 500ms.

The FastAPI server exposes the main endpoint `/forecast`. A request triggers fetching the most recent AQI values, calculating festival proximity, forming the feature tensor, running ONNX inference, and generating recommendations. Median response time was approximately 340 milliseconds (CPCB API call ~210ms, ONNX inference ~38ms). The request is validated by Pydantic v2 before reaching the model.

Once the 48-hour forecast hits 85 percent of registered capacity, the recommendation engine suggests extra nursing tasks, checks oxygen and IV fluid stock against minimum thresholds, and sends bilingual messages via Twilio and SendGrid. All actions include the predicted value that triggered them. Administrators can review the full sequence of prediction, recommendation, and approval. Nurse-to-patient ratios can be configured per location.

5. PREDICTIVE MODEL DESIGN

5.1 Input Features

The daily input is 14-dimensional. Day of week and month are represented by sine-cosine pairs to preserve cyclical structure — for example, Monday and Sunday are adjacent, not six units apart on a linear scale. A 7-day moving average slope reflects the recent trend direction separately from the count. Environmental features are PM2.5 and PM10 from the nearest operational CPCB site. Festival proximity follows the formula in Section 3.2. The epidemic indicator is a binary variable from IDSP bulletins. All quantitative features are min-max scaled using training data statistics only.

5.2 Architecture and Training

The network uses two time-distributed LSTM layers each with 128 hidden units, followed by a 30 percent dropout layer, a 64-unit dense layer with ReLU activation, and a linear output layer.

This architecture was chosen by grid search over hidden unit counts (64, 128, 256) and dropout rates (0.2, 0.3, 0.4). The best model was selected by median MAPE over three random runs rather than the best single run. We trained using the Adam optimiser [15] (learning rate 0.001), early stopping (patience 10), and gradient clipping (norm 1.0). Table 2 shows the final hyperparameters.

Table 2: LSTM Hyperparameter Configuration

Hyperparameter	Value
LSTM layers	2
Hidden units per layer	128
Dropout rate	0.30
Dense layer	64 units, ReLU
Lookback window	30 days
Forecast horizons	1, 3, 7 days
Batch size	32
Optimiser	Adam, lr = 0.001
Max epochs	100 (early stop patience 10)
Gradient clip norm	1.0

5.3 Multi-Horizon Strategy

An initial single model with auto-regressive feedback introduced compounding error. The 7-day forecasts suffered notably. Replacing this with three separate horizon-specific models boosted 7-day MAPE by 2.1 points. A multi-head approach sharing the same backbone across horizons was evaluated but yielded inferior results. The three ONNX binaries are approximately 2.1 MB each.

5.4 Surge Prediction and Explanation

A binary classifier trained alongside the regression model identifies days when admissions are forecast to reach 90 percent or more of hospital capacity. The classifier was trained with a threshold allowing no more than 5 percent false negatives, at the expense of higher precautionary alerts, reflecting the asymmetric cost of missed surge warnings. Prediction panels include SHAP feature importance values, added after early feedback revealed that nursing staff were reluctant to act on predictions they could not scrutinise.

6. EXPERIMENTAL EVALUATION

6.1 Dataset

Data spans three hospitals in Maharashtra over 730 days from January 2022 to December 2023. H1 is a 750-bed public hospital in Pune. H2 is a 400-bed district hospital in Nashik. H3 is a 600-bed private hospital in Mumbai. Training covers January 2022 to September 2023; validation covers October to November 2023; testing covers December 2023. The split is purely temporal — no random shuffling — ensuring that accuracy estimates reflect actual deployment conditions. Following imputation, there were 94,900 usable daily observations.

6.2 Comparative Results

Table 3 reports results at 48 hours, compared against naive seasonal persistence, ARIMA(7,1,1), Random Forest with the same features, and a single-layer LSTM trained only on historical admissions. HealthSurge AI achieved a MAPE of 8.3 percent, precision of 91.6 percent, and recall of 94.2 percent. The difference between the history-only LSTM and the full model isolates the impact of the three additional data streams while holding architecture constant.

Table 3: Comparative Performance at 48-Hour Horizon

Model	MAPE(%)	RMSE	Prec.(%)	Rec.(%)
Naive Seasonal	18.7	42.3	64.2	58.1
ARIMA(7,1,1)	14.2	33.8	71.5	66.3
Random Forest	11.4	26.1	80.3	74.7
LSTM (no ext.)	10.1	22.9	84.8	82.0
HealthSurge AI	8.3	18.4	91.6	94.2

6.3 Ablation Study

To assess the contribution of each feature group, we removed one group and retrained from scratch, comparing the resulting MAPE. Results are in Table 4. Removing AQI data had the greatest impact (1.9 points of MAPE, 23 percent of full model performance). Removing festival proximity cost 1.4 points. The IDSP indicator had less impact than expected, likely because the evaluation period fell in a stable post-pandemic phase. Cyclic calendar encoding provides a 0.6-point gain over integer encoding consistently across all forecasting windows and random seeds.

Table 4: Ablation Study Results

Feature Removed	MAPE (%)	Degradation
None (Full Model)	8.3	--
AQI (PM2.5 + PM10)	10.2 (+1.9)	+22.9%
Festival Proximity	9.7 (+1.4)	+16.9%
Epidemic Indicator	9.0 (+0.7)	+8.4%
Trend Gradient	9.3 (+1.0)	+12.0%
Cyclic Calendar Enc.	8.9 (+0.6)	+7.2%

6.4 Real Surge Events

Three real surges occurred in the December 2023 test period, and all were detected. During the 2023 Ganesh Chaturthi festival, admissions increased 38 percent across four days at H1; a Level-2 alert was sent 72 hours in advance. The post-Diwali surge at H3 in Mumbai saw AQI above 350 for four consecutive days and a 29 percent gradual increase in respiratory emergencies; this was flagged 96 hours before the first statistically significant increase, with SHAP output correctly identifying PM2.5 as the dominant driver. At H2 in Nashik, an influenza-A cluster reported in IDSP triggered a surge warning 48 hours later, preceding a 14 percent elevation in admissions. Two false positives occurred on long holiday weekends.

7. DISCUSSION

7.1 Operational Impact

The most relevant evidence came from the H1 pilot during Ganesh Chaturthi. The 48-hour lead time enabled the hospital to staff up 18 hours earlier than in the prior year. Patient wait times decreased by 12 percent, and a bed shortage experienced during the same festival the previous year was avoided. Clerical staff reported 34 percent less time spent on census estimation during the pilot period. This evidence is from one site, one event period, and without a comparison group. Stronger causal attribution would require a well-designed, multi-site study over a full seasonal cycle.

Four nursing supervisors from H1 provided the most informative feedback: each reported that the plain-language recommendation panel was more useful than the forecast chart itself. They needed to know whether to call in extra nurses for Thursday and whether to order more oxygen — not simply that something might occur.

7.2 Privacy and Ethics

The system processes only daily admission counts. No patient records are stored or evaluated. Data is encrypted in transit (TLS 1.3) and at rest (AES-256) in line with India's Digital Personal Data Protection Act 2023. The requirement for human review before any action is a design principle, not a stop-gap: automated real-time staffing decisions without human oversight create risks that test scores cannot quantify.

7.3 Limitations

Results are from a single state. Generalisability to other regions should not be assumed without evaluation. The training period covers a post-pandemic stability phase but not a new pandemic. The festival list includes 47 events but not the many smaller local festivals that cause district-level spikes. Default staffing ratios are tuned for general wards and may require recalibration for specialties. The CPCB API introduces a six-hour lag for early morning data, during which forecasting reverts to the last available data point.

8. FUTURE WORK

Leveraging Ayushman Bharat Digital Mission (ABDM) Health ID data would provide patient-level signals via ABDM anonymisation, potentially increasing surge accuracy at hospitals serving clinically complex patients. A federated learning approach would enable hospitals to contribute to a global model without sharing raw data. Real-time IDSP feeds would reduce the lag between cluster detection and model response — the current weekly bulletin format introduces delays of up to six days. A dedicated nursing-ward app would remove barriers inherent in the web-based interface. Finally, we plan to compare LSTM with Temporal Fusion Transformers on a larger multi-state dataset once data sharing agreements with additional hospital partners are established.

9. CONCLUSION

We began this project when an administrator in Pune truthfully told us how staffing decisions were made: experience, she said. HealthSurge AI is a second tool for her — one that links public-

domain data to a supervisor's mobile phone in a way that makes timely response possible. After two years of real data from three hospitals in Maharashtra, it achieved a 48-hour MAPE of 8.3 percent and detected every real surge in the test period before it peaked. An operational trial gave early signs that the forecast leads to earlier staff deployment and significantly shorter patient queues. The H1 nursing supervisors said they preferred the recommendations panel to the forecast chart. There is still work to do: more sites, more states, another season of controlled testing. We believe it is a direction worth exploring and hope this work can help those who follow.

ACKNOWLEDGEMENTS

We thank the Department of Computer Science and Engineering at MIT Art, Design and Technology University Pune for providing computational resources. The CPCB's decision to make air quality data available via the OpenAQ API was crucial — a student project could not have paid for that data. The IT departments of our partner hospitals in Pune, Nashik, and Mumbai helped navigate institutional approvals. The nursing supervisors at H1 who critiqued the system after the pilot influenced the final design more than any experiment we conducted. This project received no external funding.

REFERENCES

- [1] World Health Organization, "Health workforce requirements for universal health coverage," WHO Tech. Rep., Geneva, 2016.
- [2] M. Jain, A. Sharma, and R. Gupta, "Managing mass casualty events at the Kumbh Mela," *J. Public Health*, vol. 42, no. 3, pp. 612-621, 2020.
- [3] S. Nair, P. Tripathi, and V. Kumar, "Air quality and emergency room admissions for respiratory illness in Delhi," *Indian J. Med. Res.*, vol. 152, no. 4, pp. 378-387, 2020.
- [4] World Health Organization, "Global Health Estimates 2020," WHO, Geneva, 2020.
- [5] S. L. Bernstein et al., "The effect of emergency department crowding on clinically oriented outcomes," *Acad. Emerg. Med.*, vol. 16, no. 1, pp. 1-10, 2009.
- [6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735-1780, 1997.
- [7] N. R. Hoot et al., "Forecasting emergency department crowding: A discrete event simulation," *Ann. Emerg. Med.*, vol. 52, no. 2, pp. 116-125, 2008.
- [8] S. Halachmi et al., "Machine learning-based model for bed management in a general hospital," *J. Med. Syst.*, vol. 42, no. 12, p. 255, 2018.
- [9] E. Choi et al., "Doctor AI: Predicting clinical events via recurrent neural networks," *Proc. MLHC, PMLR* vol. 56, pp. 301-318, 2016.
- [10] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," *Proc. ICLR*, 2016.
- [11] M. Rostami, S. Forouzanfar, and P. G. Cordeiro, "Bidirectional LSTM for influenza hospitalization forecasting," *IEEE Access*, vol. 9, pp. 14032-14044, 2021.
- [12] K. Balakrishnan et al., "The impact of air pollution on deaths and disease burden across India," *Lancet Planet. Health*, vol. 3, no. 1, pp. e26-e39, 2019.
- [13] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nat. Med.*, vol. 25, no. 1, pp. 44-56, 2019.
- [14] J. S. Park et al., "Generative agents: Interactive simulacra of human behavior," *Proc. ACM UIST*, 2023.
- [15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. ICLR*, 2015.