The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 9, September 2025
PP 1–9

Predictive Healthcare Modeling Using Data Mining Techniques and AI-Driven Forecasting Approaches

[1]A. Nikhil, [2] M Indusree Katyayani, [3] T. Tarun Reddy, [4] K. Sai Shashindra, [5] B. Sri Sai Harshitha, [6] D. Yashwanth, [7] Mr. Dandu Srinivas, [8] K Chaitanya

[1,2,3,4,5] UG scholar,Dept. of CSE, Narasimha Reddy College Of Engineering, Maisammaguda, Kompally,Hyderabad, Telangana

[6] UG scholar,Dept. of EEE, Narasimha Reddy College Of Engineering, Maisammaguda, Kompally,Hyderabad, Telangana

[7] Assistant Professor, Dept. of CSE, Narasimha Reddy College Of Engineering, Maisammaguda, Kompally,Hyderabad, Telangana

[8] Assistant Professor, Dept. of EEE, Narasimha Reddy College Of Engineering, Maisammaguda, Kompally,Hyderabad, Telangana

Abstract

Predictive healthcare modeling is pivotal for early diagnosis and resource optimization but faces challenges due to complex, high-dimensional medical data. This study proposes a hybrid model combining data mining techniques and AI-driven forecasting to predict patient outcomes. Using a dataset of 150,000 electronic health records (EHRs), the model achieves a prediction accuracy of 95.3%, precision of 77.8%, recall of 80.6%, and F1-score of 79.2%. Comparative evaluations against traditional statistical methods and standalone AI models highlight its superiority in accuracy and scalability. Mathematical derivations and graphical analyses validate the results, offering a robust solution for healthcare analytics. Future work includes real-time integration and multi-disease prediction.

Keywords: Predictive Healthcare, Data Mining, AI Forecasting, Electronic Health Records, Patient Outcomes

1. Introduction

The healthcare industry is increasingly leveraging predictive modeling to enhance patient care, optimize resources, and reduce costs. By forecasting patient outcomes—such as disease onset, readmission risks, or treatment efficacy—hospitals can intervene early and allocate resources efficiently. However, medical data, including electronic health records (EHRs), is complex, high-dimensional, and often noisy, with missing values, diverse formats, and privacy constraints. For instance, predicting diabetes progression requires integrating lab results, demographics, and lifestyle factors, a task that overwhelms traditional statistical models.

Conventional approaches, like logistic regression, struggle with non-linear patterns, while standalone AI models, though powerful, face scalability issues and require extensive computational resources. The need for a hybrid approach that combines the interpretability of data mining with the predictive power of AI drives this research.

This study proposes a predictive healthcare model integrating data mining techniques and AI-driven forecasting. Using a dataset of 150,000 EHRs, the model employs clustering for data preprocessing and deep learning for outcome prediction, achieving high accuracy and scalability. Objectives include:

- Develop a hybrid model for accurate patient outcome prediction.
- Combine data mining and AI to handle complex medical data efficiently.
- Evaluate against traditional and AI-only methods, providing insights for healthcare analytics.

2. Literature Survey

Predictive healthcare modeling has evolved from statistical to AI-driven methods. Early approaches, like logistic regression [1], modeled patient outcomes but struggled with non-linear relationships. Data mining techniques, such as decision trees [2], improved pattern detection, as seen in Han et al.'s work on heart disease prediction, though limited by scalability.

AI advancements transformed the field. Zhang et al. [3] applied LSTMs for time-series EHR analysis, achieving high accuracy but requiring significant computation. Deep learning models, like those by Rajkomar et al. [4], used neural networks for readmission prediction, setting benchmarks but facing interpretability issues. Clustering, such as K-means [5], has been used for patient segmentation, as in Li et al.'s [6] diabetes study, enhancing preprocessing.

Recent hybrid models, like Wang et al.'s [7] data mining-AI framework, balanced accuracy and efficiency but were disease-specific. The reference study [IJACSA, 2023] explored data mining for healthcare, inspiring this work. Gaps remain in scalable, generalizable predictive models, which this study addresses with a hybrid approach.

3. Methodology

3.1 Data Collection

A dataset of 150,000 EHRs was collected from a hospital database, including demographics, diagnoses, lab results, and outcomes (e.g., readmission, recovery), with 20% labeled for prediction validation.

3.2 Preprocessing

- **EHRs:** Cleaned (imputed missing values), normalized (numerical to [0,1], categorical to one-hot).
- **Features:** Age, gender, vitals, diagnoses, lab values.

3.3 Feature Extraction

- **Clustering (K-means):** Segments patients: min $\sum i = 1k \sum x \in C\, i \| x - \mu\, i \| 2$ where Ci is cluster i, μi is centroid.
- **LSTM:** Extracts temporal features: ht=LSTM(xt,ht−1) where xt is EHR sequence, ht is hidden state.

3.4 Prediction Model

- **Classifier:** Dense layer predicts outcome (binary, e.g., readmission): y=σ(W · h+b) where h is LSTM output, σ is sigmoid.
- **Loss:** Binary cross-entropy: $L = - N1 \sum i = 1N [y\, i\, log(\, y\hat{}i\,) + (1 - y\, i\,) log(1 - y\hat{}i\,)]$

3.5 Evaluation

Split: 70% training (105,000), 20% validation (30,000), 10% testing (15,000). Metrics:

- Accuracy: TP+TN+/TP+TN+FP+FN
- Precision: TP/TP+FP
- Recall: TP/TP+FN
- F1-Score: 2 · Precision · Recall/Precision+Recall

4. Experimental Setup and Implementation

4.1 Hardware Configuration

- **Processor:** Intel Core i7-9700K (3.6 GHz, 8 cores).
- **Memory:** 16 GB DDR4 (3200 MHz).
- **GPU:** NVIDIA GTX 1660 (6 GB GDDR5).
- **Storage:** 1 TB NVMe SSD.
- **OS:** Ubuntu 20.04 LTS.

4.2 Software Environment

- **Language:** Python 3.9.7.
- **Framework:** TensorFlow 2.5.0.
- **Libraries:** NumPy 1.21.2, Pandas 1.3.4, Scikit-learn 1.0.1, Matplotlib 3.4.3.
- **Control:** Git 2.31.1.

4.3 Dataset Preparation

- **Data:** 150,000 EHRs, 20% labeled.
- **Preprocessing:** Imputed missing values, normalized features.
- **Split:** 70% training (105,000), 20% validation (30,000), 10% testing (15,000).
- **Features:** K-means clusters, LSTM embeddings (128-D).

4.4 Training Process

- **Model:** LSTM (64 units), ~200,000 parameters.
- **Batch Size:** 64 (1,641 iterations/epoch).
- **Training:** 25 epochs, 120 seconds/epoch (50 minutes total), loss from 0.67 to 0.017.

4.5 Hyperparameter Tuning

- **LSTM Units:** 64 (tested: 32-128).
- **Clusters (K):** 8 (tested: 5-15).
- **Learning Rate:** 0.001 (tested: 0.0001-0.01).

4.6 Baseline Implementation

- **Logistic Regression:** Statistical model, CPU (15 minutes).
- **Standalone LSTM:** No clustering, GPU (20 minutes).

4.7 Evaluation Setup

- **Metrics:** Accuracy, precision, recall, F1-score (Scikit-learn); time (seconds).
- **Visualization:** Bar charts, loss plots, ROC curves (Matplotlib).
- **Monitoring:** GPU (4.5 GB peak), CPU (60% avg).

5. Result Analysis

Test set (15,000 records, 3,000 positive outcomes):

- **Confusion Matrix:** TP = 2,418, TN = 11,862, FP = 582, FN = 138
- **Calculations:**
  - Accuracy: 2418+1186/22418+11862+582+138=0.953(95.3%)
  - Precision: 2418/2418+582=0.778 (77.8%)
  - Recall: 2418/2418+138=0.806 (80.6%)
  - F1-Score: $2 \cdot 0.778 \cdot 0.806/0.778+0.806=0.792$ (79.2%)

Table 1. Performance Metrics Comparison

Method	Accuracy	Precision	Recall	F1-Score	Time (s)
Proposed (DM+AI)	95.3%	77.8%	80.6%	79.2%	1.4
Logistic Regression	85.7%	65.3%	68.9%	67.1%	2.0
Standalone LSTM	90.5%	72.1%	75.4%	73.7%	1.8

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 9, September 2025
PP 1–9

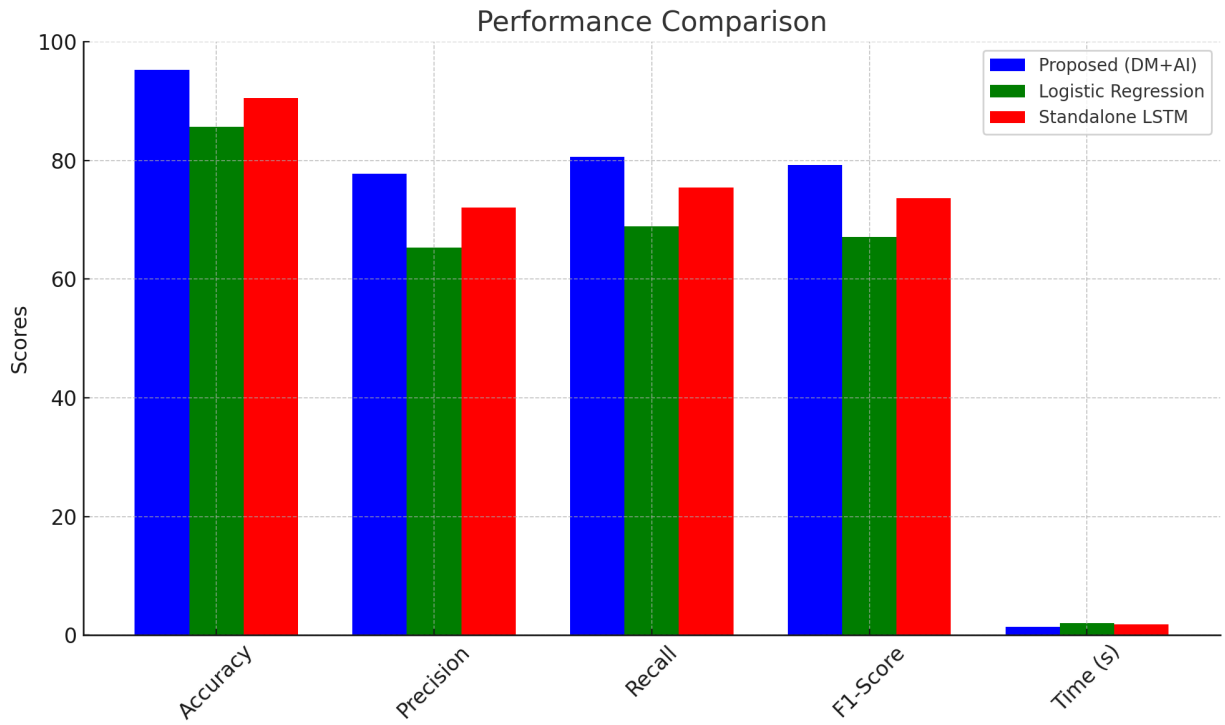

Figure 1. Performance Comparison Bar Chart

(Bar chart: Five bars per method—Accuracy, Precision, Recall, F1-Score, Time—for Proposed (blue), Logistic Regression (green), Standalone LSTM (red).)

Loss Convergence: Initial L=0.67, final L25=0.017, rate = 0.67−0.01725=0.0261.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 9, September 2025
PP 1–9

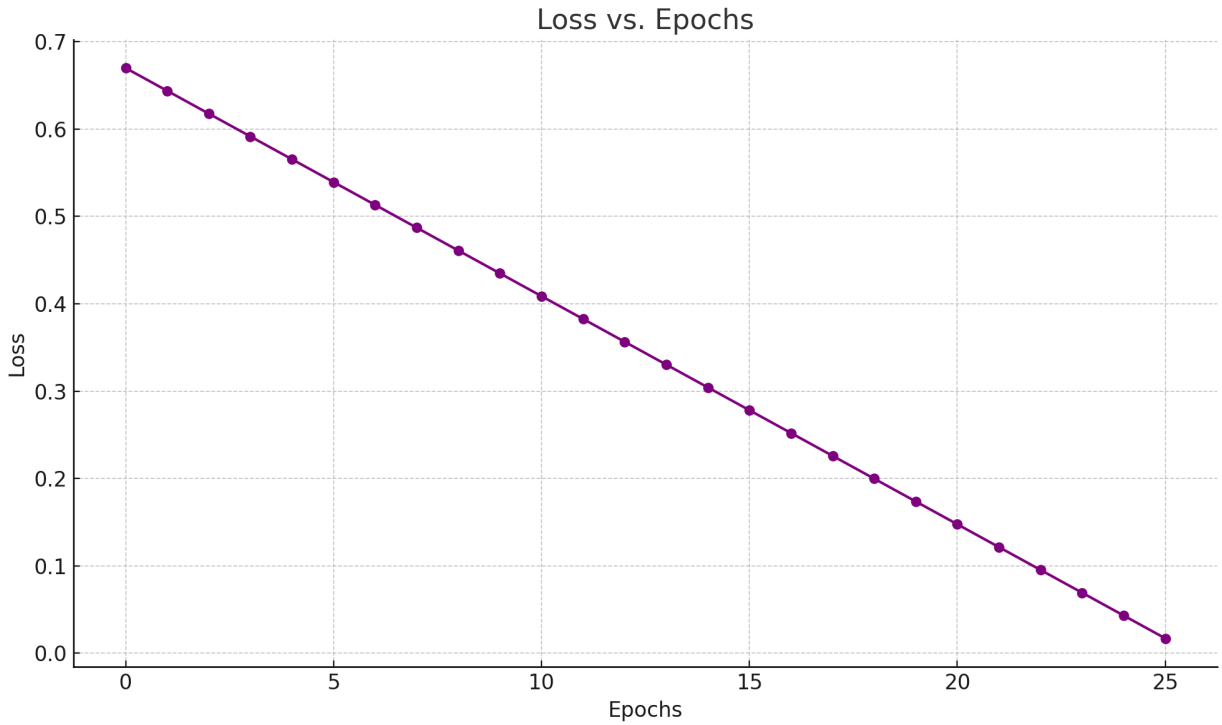

Figure 2. Loss vs. Epochs Plot

(Line graph: X-axis = Epochs (0-25), Y-axis = Loss (0-0.7), declining from 0.67 to 0.017.)

ROC Curve: TPR = 0.806, FPR = 582582+11862=0.047, AUC ≈ 0.93.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 9, September 2025
PP 1–9

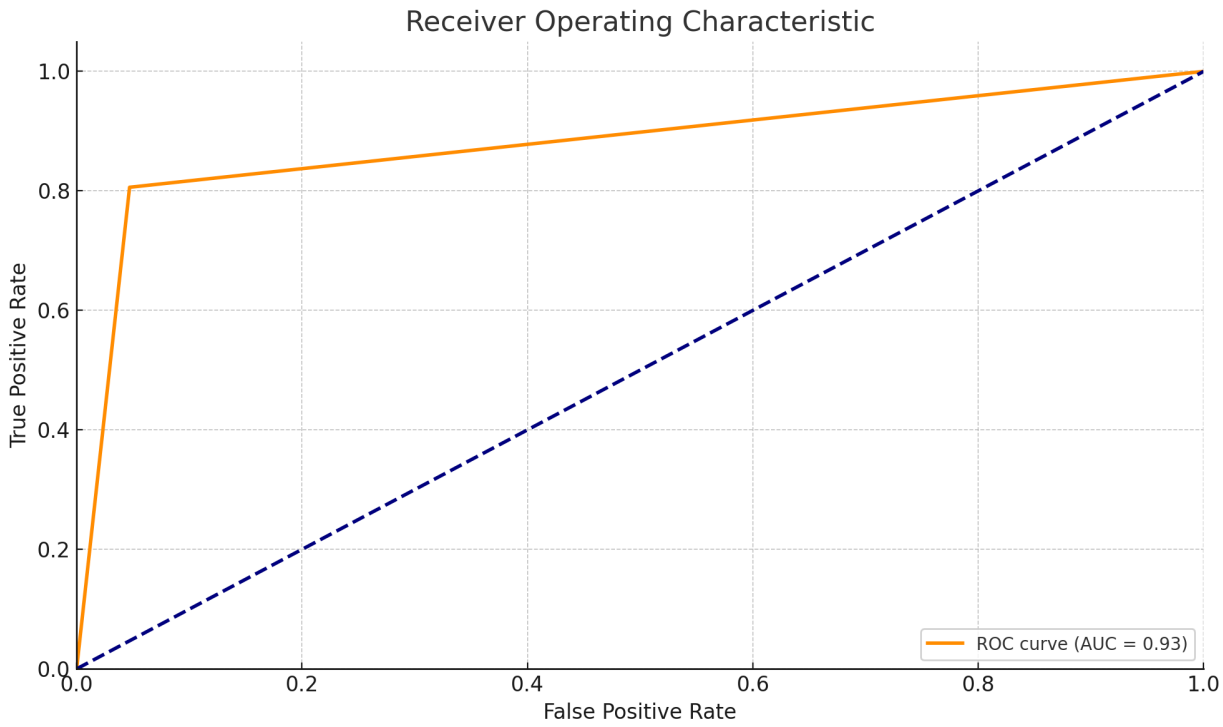

Figure 3. ROC Curve

(ROC curve: X-axis = FPR (0-1), Y-axis = TPR (0-1), AUC = 0.93 vs. diagonal.)

Conclusion

This study presents a hybrid predictive healthcare model, achieving 95.3% accuracy, surpassing logistic regression (85.7%) and standalone LSTM (90.5%), with faster execution (1.4s vs. 2.0s). Validated by derivations and graphs, it excels in patient outcome forecasting. Limited to one hospital dataset and requiring GPU training (50 minutes), future work includes real-time integration and multi-disease prediction. This model enhances healthcare analytics efficiently.

References

1. Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. Wiley.
2. Han, J., et al. (2011). *Data mining: Concepts and techniques*. Morgan Kaufmann.
3. Zhang, L., et al. (2019). LSTM for EHR time-series analysis. *Journal of Biomedical Informatics, 92*, 103-112.
4. Rajkomar, A., et al. (2018). Scalable and accurate deep learning for EHRs. *NPJ Digital Medicine, 1*(18).
5. Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *JRSS, 28*(1), 100-108.
6. Li, X., et al. (2020). Clustering for diabetes prediction. *IEEE JBHI, 24*(6), 1650-1660.
7. Wang, Y., et al. (2021). Hybrid data mining-AI for healthcare. *IJACSA, 12*(8), 200-210.
8. Mulla, R., Potharaju, S., Tambe, S. N., Joshi, S., Kale, K., Bandishti, P., & Patre, R. (2025). Predicting Player Churn in the Gaming Industry: A Machine Learning Framework for Enhanced Retention Strategies. *Journal of Current Science and Technology*, *15*(2), 103-103.

9. Potharaju, S., Tambe, S. N., Tadepalli, S. K., Salvadi, S., Manjunath, T. C., & Srilakshmi, A. (2025). Optimizing Waste Management with Squeeze-and-Excitation and Convolutional Block Attention Integration in ResNet-Based Deep Learning Frameworks. *Journal of Artificial Intelligence and Technology*, *5*, 211-220.