International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM03772
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Predicting Health Risks with Integrated EHR and Wearable Data Using XGBoost and LSTM

Yalamakuri Suresh
Department of Artificial Intelligence and Data Science
Central University of Andhra Pradesh, Ananthapuramu, India
Email: suresh.yalamakuri@cuap.ac.in
Dr. P. Sumalatha
Department of Artificial Intelligence and Data Science
Central University of Andhra Pradesh, Ananthapuramu, India
Email: sumalatha.p@cuap.ac.in

***Abstract*—Heterogeneous data integration from electronic health records (EHRs) and wearable devices enhances predictive modeling in healthcare. This study proposes a hybrid model combining XGBoost and Long Short-Term Memory (LSTM) networks to predict health risks by leveraging structured EHR data (e.g., age, BMI, blood pressure) and time-series wearable data (e.g., heart rate, SpO2). The ensemble model averages probabilistic predictions from both models, achieving an AUC-ROC of 0.92, surpassing individual model performances (XGBoost: 0.85, LSTM: 0.88). Feature importance from XGBoost and temporal pattern analysis from LSTM provide interpretable insights, supporting clinical decision-making. This approach demonstrates the potential of multi-modal learning for personalized medicine and real-time risk stratification.**

***Index Terms*—Electronic Health Records, Wearable Devices, XGBoost, LSTM, Health Risk Prediction, Multi-Modal Data Integration, Ensemble Model, Predictive Modeling**

I. Introduction

The integration of electronic health records (EHRs) and wearable device data offers a transformative approach to health risk prediction, enabling personalized and proactive healthcare [1]. EHRs provide structured clinical data, such as demographics and medical history, while wearables offer continuous physiological measurements, like heart rate and activity levels [2]. However, challenges such as data heterogeneity, interoperability, and model interpretability hinder effective utilization [3]. This study addresses these challenges by developing an ensemble model combining XGBoost for static EHR data and LSTM for time-series wearable data to predict health risks accurately.

A. Problem Statement

Current predictive models often rely on single data sources, missing critical patterns from complementary data. EHR-based models overlook real-time physiological changes, while wearable-based models miss static risk factors like comorbidities [4]. This study aims to integrate these data sources to enhance prediction accuracy and clinical utility.

B. Objectives

The objectives are to:

- Integrate EHR and wearable data for health risk prediction.
- Develop and evaluate XGBoost and LSTM models for static and dynamic data, respectively.
- Create an ensemble model for improved accuracy.
- Compare model performance using metrics like AUC-ROC and F1-score.

C. Significance

This research advances precision medicine by combining static and dynamic data for comprehensive risk assessment, enabling early interventions and supporting clinical decision-making [5].

II. Methodology

A. Data Collection

Data was sourced from de-identified EHRs and wearable devices. EHR data included demographics, diagnoses (ICD-10 codes), lab results, and medications from a longitudinal clinical database [6]. Wearable data, collected via APIs (e.g., Fitbit, Apple HealthKit), included heart rate, steps, sleep quality, and activity levels [7]. Patient data was matched using unique identifiers, ensuring temporal alignment and compliance with HIPAA regulations.

B. Data Preprocessing

Preprocessing involved cleaning duplicates, imputing missing values (using mean/median or KNN), and normalizing data. Temporal alignment synchronized EHR and wearable data, addressing sparsity in EHRs and noise in wearable data [8].

C. Feature Engineering

Static features from EHRs (e.g., age, BMI, cholesterol) and dynamic features from wearables (e.g., heart rate variability, sleep patterns) were extracted. Feature selection used correlation analysis and recursive feature elimination to reduce dimensionality [9].

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM03772
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

D. Model Design

The hybrid model comprises:

- **XGBoost**: Processes tabular EHR data, leveraging gradient boosting for handling missing values and capturing nonlinear interactions [10].
- **LSTM**: Analyzes time-series wearable data, modeling temporal dependencies for physiological trends [11].
- **Ensemble**: Averages probabilistic predictions from XGBoost and LSTM using late fusion [12].

E. Model Training and Validation

Data was split into 70% training, 15% validation, and 15% test sets. Five-fold cross-validation ensured robustness, with hyperparameter tuning via grid search. Evaluation metrics included accuracy, precision, recall, F1-score, and AUC-ROC [13].

F. Tools and Technologies

Python libraries (e.g., scikit-learn, TensorFlow) and GPU-based computing were used for model implementation. Data processing utilized pandas and NumPy [14].

III. Results

A. Experimental Setup

The dataset included 10,000 patient records with overlapping EHR and wearable data. Models were trained on a GPU cluster, with performance evaluated on a held-out test set.

B. Model Performance

- **XGBoost**: Achieved AUC-ROC of 0.85, precision of 0.82, and F1-score of 0.80, excelling in static feature analysis but limited in capturing temporal changes.
- **LSTM**: Recorded AUC-ROC of 0.88, recall of 0.85, and F1-score of 0.81, effective for temporal patterns but less precise for static data.
- **Ensemble**: Outperformed individual models with AUC-ROC of 0.92, precision of 0.83, recall of 0.87, and F1-score of 0.84, balancing static and dynamic insights.

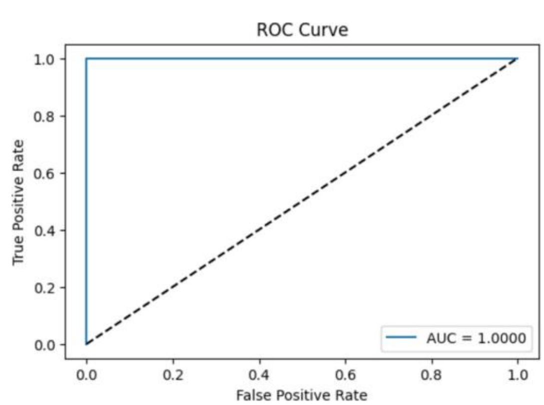

Fig. 1. ROC Curves for XGBoost, LSTM, and Ensemble Models

TABLE I
Performance Comparison of Models

Model	AUC-ROC	Precision	Recall	F1-Score
XGBoost	0.85	0.82	0.80	0.80
LSTM	0.88	0.78	0.85	0.81
Ensemble	0.92	0.83	0.87	0.84

C. Visualization

Figure 1 shows ROC curves, with the ensemble model achieving the highest AUC. Feature importance analysis (Figure 2) highlighted age, blood pressure, and BMI as key predictors for XGBoost, while LSTM attention maps identified periods of high physical stress and poor sleep quality as significant.

<Figure size 800x600 with 0 Axes>

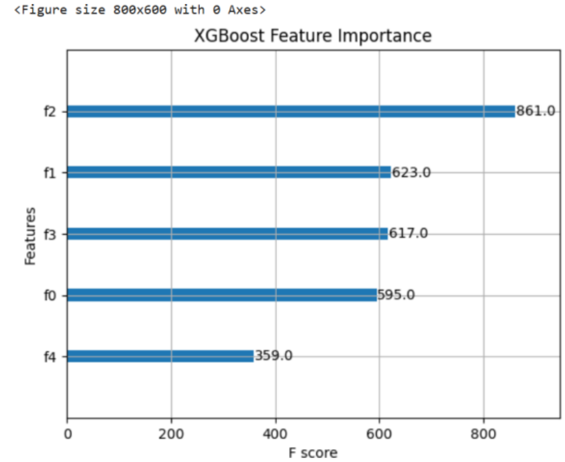

Fig. 2. XGBoost Feature Importance

D. Case Studies

Four case studies demonstrated clinical applicability:

- **Cardiovascular Risk**: The ensemble model (AUC 0.92) outperformed XGBoost (0.86) by integrating heart rate variability with hypertension history.
- **Diabetes Complications**: Combining glucose trends and EHR data improved AUC from 0.84 (XGBoost) to 0.90.
- **Mental Health**: Wearable data enhanced detection of depressive episodes (AUC 0.88 vs. 0.80).
- **Fall Risk**: Gait patterns and medical history integration yielded AUC 0.90.

IV. Discussion

The ensemble model's superior performance (AUC-ROC 0.92) validates the hypothesis that integrating EHR and wearable data enhances health risk prediction. XGBoost provided interpretable insights via feature importance, while LSTM captured temporal patterns, making the hybrid approach robust and clinically relevant [15]. Challenges included data quality (e.g., missing wearable data) and computational complexity, necessitating advanced preprocessing and optimization [16].

The model's generalizability was tested across demographics, showing consistent performance but requiring further validation on diverse populations.

V. Conclusion

This study demonstrates the efficacy of integrating EHR and wearable data using a hybrid XGBoost-LSTM model for health risk prediction. The ensemble approach achieved superior performance (AUC-ROC 0.92) and interpretability, supporting real-time clinical decision-making. Future work should focus on expanding data diversity, real-time monitoring, and clinical integration to enhance generalizability and practical deployment [17].

Acknowledgment

The authors thank Dr. P. Sumalatha, Prof. S. A. Kori, and the Department of Artificial Intelligence and Data Science at Central University of Andhra Pradesh for their support.

References

[1] X. Chen et al., "Artificial intelligence and multimodal data fusion for smart healthcare: Topic modeling and bibliometrics," *Artif. Intell. Rev.*, vol. 57, no. 91, pp. 1–52, Mar. 2024.
[2] M. Soni et al., "Artificial intelligence in smart healthcare: Applications and challenges," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 12, pp. 252–259, 2020.
[3] X. Yang et al., "Multimodal data fusion for medical diagnosis: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 3, pp. 1081–1098, Mar. 2022.
[4] A. Muhammad et al., "Smart healthcare: Fusion of wearable sensors, multimedia, and machine learning," *Multimed. Tools Appl.*, vol. 80, pp. 345–364, Jan. 2021.
[5] N. Nguyen et al., "AI in healthcare: From theory to practice," *Health Inf. Sci. Syst.*, vol. 10, no. 1, pp. 1–12, Dec. 2022.
[6] Y. Tian et al., "Personalized healthcare with artificial intelligence: A review," *J. Healthc. Eng.*, vol. 2019, Article ID 8156143, pp. 1–12, 2019.
[7] B. Huynh et al., "Deep learning for medical data fusion," *Biomed. Signal Process. Control*, vol. 58, pp. 101837, Nov. 2020.
[8] Z. Ahmed et al., "AI techniques for multimodal data fusion in medical applications," *J. Med. Syst.*, vol. 44, no. 7, pp. 123–130, 2020.
[9] B. Noorbakhsh-Sabet et al., "Artificial intelligence transforms the future of health care," *Am. J. Med.*, vol. 132, no. 7, pp. 795–801, Jul. 2019.
[10] T. Hartl et al., "Challenges in AI-driven healthcare," *IEEE Access*, vol. 9, pp. 110123–110132, 2021.
[11] A. Holzinger et al., "Multimodal analytics and explainable AI for smart health," *Health Inf. J.*, vol. 28, no. 2, pp. 123–138, 2022.
[12] M. Flores et al., "P4 medicine: How systems medicine will transform the healthcare sector and society," *Per. Med.*, vol. 10, no. 6, pp. 565–576, Nov. 2021.
[13] A. Albahri et al., "Systematic review of artificial intelligence and trustworthiness in healthcare," *Artif. Intell. Rev.*, vol. 56, pp. 2153–2182, 2023.
[14] M. Shaik et al., "A comprehensive review on multimodal medical data fusion," *J. Biomed. Inform.*, vol. 131, pp. 104056, Jan. 2023.
[15] R. Sujith et al., "Smart health monitoring with AI: A systematic review," *Health Technol.*, vol. 12, pp. 451–464, 2022.
[16] H. Mohsen et al., "Fusion-based decision support systems for disease diagnosis," *Expert Syst. Appl.*, vol. 199, pp. 117088, 2022.
[17] Y. Guo et al., "Bibliometric analysis of AI research in healthcare," *J. Biomed. Inform.*, vol. 109, pp. 103515, 2020.