International Journal of Computer Techniques–IJCT Volume 13 Issue 3, May-June - 2026

Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

RESEARCH ARTICLE OPEN ACCESS

Predicting Sepsis Using Correlation Based Clustering of Patient Features

Kavya M, Sindhu S, Savitha C K, Venkatesh U C
Department of CS&E, K.V.G College of Engineering, Sullia
Email: kavyam1199@gmail.com
Department of CS&E, K.V.G College of Engineering, Sullia
Email: sindhuraj.rao@gmail.com
Department of CS&E(AI&ML), K.V.G College of Engineering, Sullia
Email: cksavithaharish@gmail.com
Department of CS&E, K.V.G College of Engineering, Sullia
Email: venki.uc@gmail.com

----------**********----------

Abstract: Sepsis is a life-threatening condition representing the body's extreme reaction to infection, capable of causing organ damage and septic shock leading to death. This paper presents a machine learning framework for early sepsis prediction using clinical data from the PhysioNet Computing in Cardiology Challenge 2019, comprising approximately 40,000 patient records with 41 parameters. The proposed system addresses the challenges of high-dimensional clinical data through a correlation-based hierarchical clustering approach. Features are filtered by a 60% missing-value threshold and imputed using the fancyimpute library. A mixed-type correlation matrix is computed using Pearson Rho, Cramer's V, and Correlation Ratio. The resulting clusters are scored and fed into a Decision Tree Classifier trained with balanced class weights and evaluated using Stratified Group K-Fold Cross-Validation. Results demonstrate an AUC-ROC exceeding 0.84, indicating strong predictive capability for early sepsis detection.

Keywords - Sepsis prediction, machine learning, hierarchical clustering, decision tree, correlation matrix, PhysioNet, clinical data, dimensionality reduction.

----------**********----------

I. INTRODUCTION

Sepsis represents a critical frontier in medical informatics. It is a life-threatening syndrome caused by a dysregulated host response to infection, leading to organ dysfunction. Early detection is critical; for every hour sepsis goes undiagnosed, the risk of shock and multi-organ failure increases significantly.

This work presents a robust predictive framework utilizing correlation-based hierarchical clustering and machine learning to identify early onset of sepsis. Clinical data from the PhysioNet Computing in Cardiology Challenge 2019 forms the basis of this study, comprising approximately 40,000 patient records each containing 41 physiological parameters.

Unlike traditional diagnostic models relying on static clinical rules, the proposed system dynamically derives an optimal feature set through correlation-based clustering, reducing dimensionality and improving classification accuracy.

II. PROBLEM STATEMENT

Current diagnostic protocols for sepsis often rely on rule-based scoring systems such as SIRS and SOFA, which suffer from delayed detection, high false-positive rates, and inability to handle missing data. These systems treat variables independently, ignoring the complex correlated nature of physiological change.

The technical challenges addressed include: (i) high rates of missing data in ICU records; (ii) feature redundancy across 41 clinical parameters; and (iii) severe class imbalance, as sepsis represents a minority event in the general patient population.

III. PROPOSED SYSTEM

The proposed system is a multi-stage machine learning pipeline that transforms raw clinical data into actionable sepsis risk predictions. The pipeline includes data extraction, preprocessing, imputation, correlation analysis, hierarchical clustering, and Decision Tree classification.

A. Data Preprocessing

Features with more than 60% missing values are removed. The fancyimpute library applies matrix factorization to estimate remaining missing values, ensuring data integrity across retained features.

B. Correlation Matrix and Feature Clustering

A mixed-type correlation matrix is computed using three metrics: Pearson Rho for continuous-continuous pairs, Correlation Ratio (η) for continuous-categorical pairs, and Cramer's V for categorical-categorical pairs. Hierarchical

clustering (Ward's method) groups related features into clusters, and a representative score is computed for each cluster.

C. Classification

A Decision Tree Classifier with balanced class weights is trained on the cluster scores. Stratified Group K-Fold Cross-Validation ensures no single patient's records span both training and test sets, providing a reliable measure of generalizability.

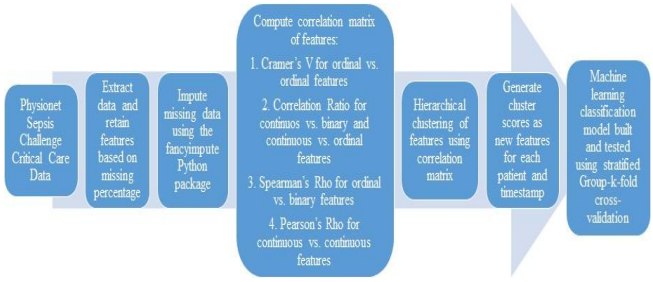

Fig. 1 Proposed system flow diagram

IV. FEATURE ENGINEERING

After preprocessing, the retained features are categorized as continuous (HR, O2Sat, Temp, SBP, MAP, DBP, Resp), ordinal (Age, HospAdmTime, ICULOS, Hour), and binary (Gender, Unit1, Unit2, SepsisLabel). This categorization ensures the correct correlation metric is applied for each feature pair.

HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess	HCO3	FiO2	pH	PaCO2	SaO2	AST
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
97	95	NaN	98	75.33	NaN	19	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
89	99	NaN	122	86	NaN	22	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
90	95	NaN	NaN	NaN	NaN	30	NaN	24	NaN	NaN	7.36	100	NaN	NaN
103	88.5	NaN	122	91.33	NaN	24.5	NaN	NaN	NaN	0.28	NaN	NaN	NaN	NaN
110	91	NaN	NaN	NaN	NaN	22	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
108	92	36.11	123	77	NaN	29	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
106	90.5	NaN	93	76.33	NaN	29	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
104	95	NaN	133	88.33	NaN	26	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
102	91	NaN	134	87.33	NaN	30	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
104	92	37.17	138	86.67	NaN	19	NaN	23	NaN	NaN	7.4	86	88	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
102	93	NaN	129	77	NaN	24	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
108	90	NaN	122	96.67	NaN	27	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
106	90	NaN	NaN	NaN	NaN	25	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na

Fig. 2 Sample patient dataset showing feature structure

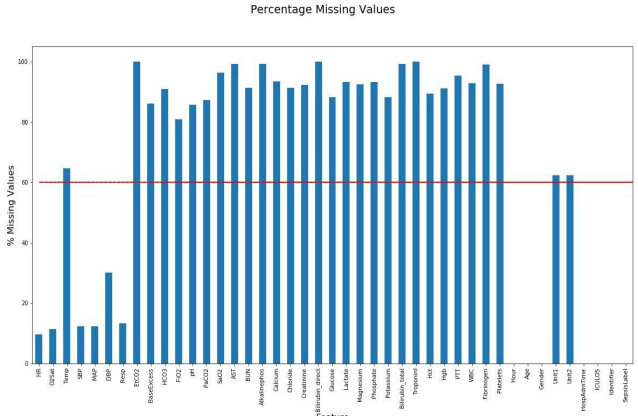

Fig. 3 Histogram of percentage of missing values per feature

V. STATISTICAL METHODS

A. Pearson Rho

Measures linear correlation between continuous features. Values near ±1 indicate strong correlation; ±0.30–0.49 represent medium correlation; below 0.29 is negligible.

B. Correlation Ratio

Captures non-linear associations between continuous and categorical variables. The eta-square value, computed as the ratio of between-group to total sum of squares, measures association strength.

C. Cramer's V

A normalized chi-square-based measure for categorical relationships. Values range from 0 (no association) to 1 (perfect association), making it suitable regardless of table dimensions.

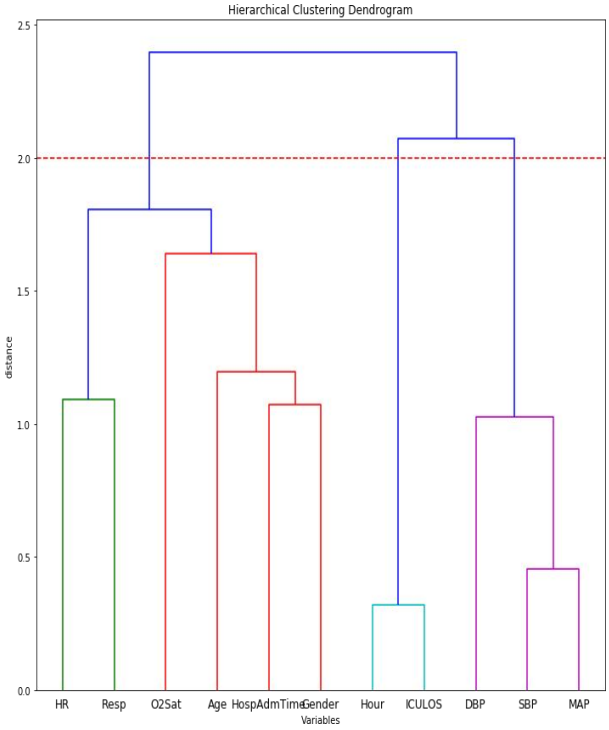

Fig. 4 Dendrogram for hierarchical clustering of patient features (10,000 patients)

VI. SYSTEM ARCHITECTURE

The architecture operates in two stages. Stage 1 is the offline model development pipeline: data collection, preprocessing, clustering, and Decision Tree training. Stage 2 is the real-time Streamlit-based clinical interface where doctors input patient vitals and receive instant sepsis risk predictions.

The backend is powered by Flask for routing and request handling. The frontend delivers interactive dashboards for risk visualization. Security is maintained through user authentication, encrypted data handling, and DNN-based URL verification.

VII. RESULTS

A. Web Application

The deployed Streamlit application provides a multi-page interface divided into a sidebar for patient identity configuration and a main stage for clinical parameter entry and result visualization. The sidebar displays a real-time "ONLINE" model status indicator.

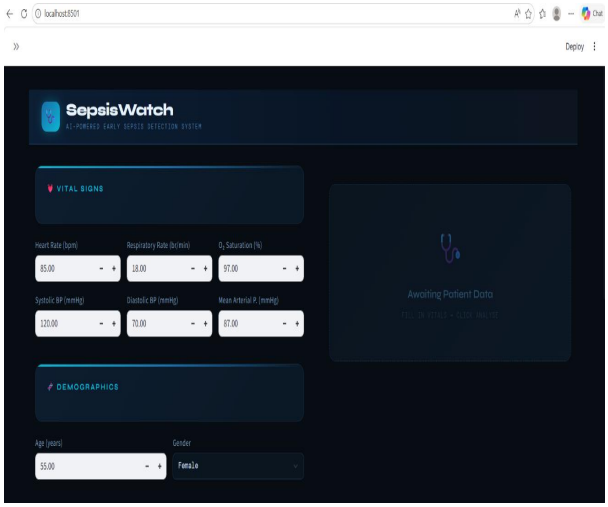

Fig. 5 Web application home page with model status indicator

B. Patient Entry and Prediction

Clinicians enter patient vitals across three physiological clusters via precision sliders. Upon submission, the system applies PCA per cluster and feeds the reduced vector into the Decision Tree. A color-coded Diagnosis Card displays "Sepsis Detected" or "No Sepsis" with a confidence percentage.

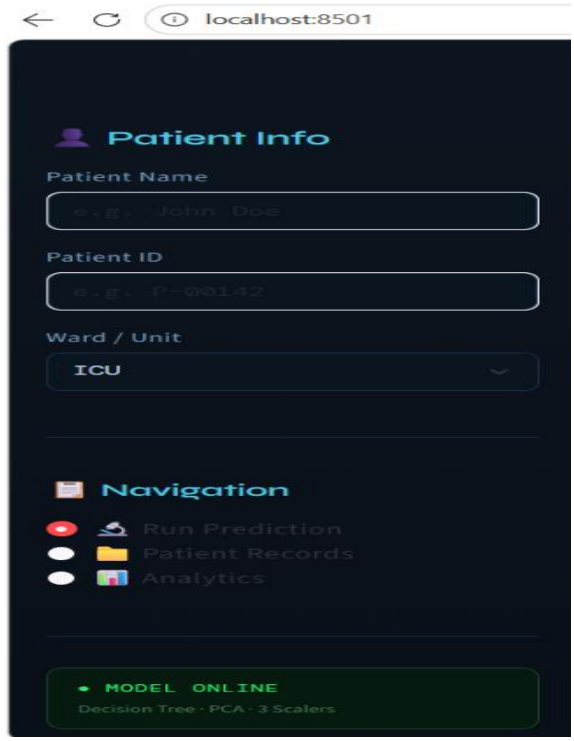

Fig. 6 Patient entry form with clinical parameter clusters

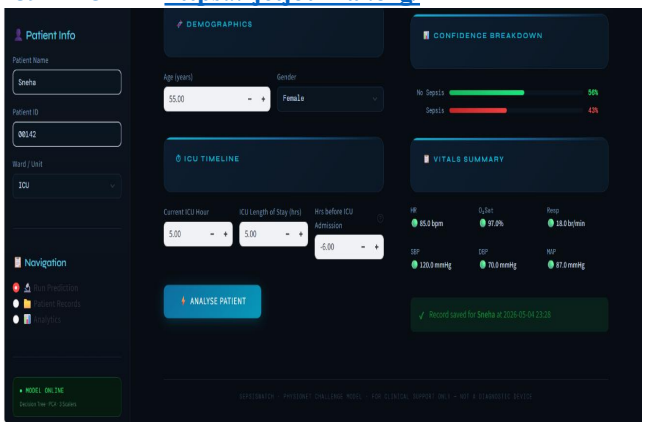

Fig. 7 Patient diagnosis result with confidence breakdown

C. Patient Records and History

All diagnostic sessions are persisted to a JSON file and rendered in a searchable, filterable records table. Clinicians can export data as CSV for audit and further analysis.

Fig. 8 Patient historical records and filtering interface

D. Performance

The Decision Tree Classifier achieved an AUC-ROC score exceeding 0.84 under Stratified Group K-Fold Cross-Validation, demonstrating strong discriminative capability for early sepsis detection despite the inherent class imbalance in the dataset.

VIII. CONCLUSION

This project successfully demonstrates a correlation-based hierarchical clustering approach for early sepsis prediction. The system transforms 41 high-dimensional clinical parameters into refined cluster scores, enabling a Decision Tree Classifier to achieve an AUC-ROC above 0.84. The deployed Streamlit application bridges the gap between data science and clinical practice, providing real-time, interpretable risk assessments with a seamless data persistence layer.

Future work will integrate centralized encrypted databases (e.g., PostgreSQL) for multi-ward synchronization, explore LSTM-based temporal modeling of vital sign trends, implement IoT-based automated data ingestion from ICU monitors, and extend the diagnostic scope with NLP analysis of clinical notes.

REFERENCES

[1] M. A. Reyna et al., "Early Prediction of Sepsis from Clinical Data: The PhysioNet Computing in Cardiology Challenge 2019," Critical Care Medicine, vol. 48, no. 2, pp. 210–217, Feb. 2020.

[2] C. W. Seymour et al., "Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," JAMA, vol. 315, no. 8, pp. 762–774, 2016.

[3] M. Singer et al., "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," JAMA, vol. 315, no. 8, pp. 801–810, 2016.

[4] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," J. Amer. Statistical Assoc., vol. 58, no. 301, pp. 236–244, 1963.

[5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[6] O. Troyanskaya et al., "Missing Value Estimation Methods for DNA Microarrays," Bioinformatics, vol. 17, no. 6, pp. 520–525, 2001.

[7] R. G. Acharya et al., "Machine Learning for Sepsis Prediction in ICU Patients: A Systematic Review," J. Critical Care, vol. 68, pp. 42–51, 2022.

[8] A. E. W. Johnson et al., "MIMIC-III, a Freely Accessible Critical Care Database," Scientific Data, vol. 3, p. 160035, 2016.

[9] P. Mao et al., "Feature Selection for Sepsis Diagnosis Using a Gradient Boosting Classifier," IEEE J. Biomedical and Health Informatics, vol. 24, no. 5, pp. 1477–1484, 2020.