E-ISSN: 2583-9667
Indexed Journal
Peer Reviewed Journal
https://multiresearchjournal.theviews.in

Received: 09-01-2025
Accepted: 16-03-2025

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH IN MULTIDISCIPLINARY

Volume 3; Issue 2; 2025; Page No. 241-245

Smart health predictions: Using deep learning to assess disease risk through personal and environmental data

[1]Selvam M and [2]Dr. A Angel Cerli

[1]PG Scholar, Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies, Pallavaram, Chennai, Tamil Nadu, India
[2]Assisstant Professor, Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies, Pallavaram, Chennai, Tamil Nadu, India

DOI: https://doi.org/10.5281/zenodo.15586676

Corresponding Author: Selvam M

Abstract

Accurately predicting disease risk requires advanced modelling techniques capable of handling dynamic environmental and personal health data. This research introduces a hybrid Recurrent Neural Network (RNN) and Support Vector Machine (SVM) framework for predicting asthma attacks and Type-2 diabetes risk. For asthma, an RNN processes sequential real-time environmental inputs-such as air quality, temperature, and humidity-capturing temporal dependencies to forecast attack likelihood. The RNN's ability to learn from time-series data allows it to adapt to changing conditions, improving prediction accuracy over time. Meanwhile, SVMs assess Type-2 diabetes risk by classifying personalized health factors, including medical history, lifestyle habits, and biomarker trends, while also analyzing environmental risk groups through statistical odds ratios. Developed using TensorFlow, PyTorch, and Scikit-learn, this cloud-based system dynamically refines its models with new data, enabling early intervention and reducing healthcare burdens. Experimental results demonstrate that the RNN-SVM hybrid approach effectively combines temporal pattern recognition with high-precision classification, offering a robust solution for proactive disease risk management.

Keywords: Recurrent Neural Network (RNN), Support Vector Machine (SVM), Disease Risk Prediction, Asthma Attack Predictions, Type-2 Diabetes Risk Assessment, Hybrid Machine Learning Model, Personalized Healthcare

Introduction

Chronic diseases such as asthma and Type-2 diabetes continue to pose significant global health challenges, often leading to long-term complications and placing immense pressure on healthcare systems. Despite advancements in medical science, early detection and proactive management of these conditions remain difficult due to their complex and multifactorial nature. Traditional diagnostic methods primarily rely on static patient data and do not effectively account for real-time environmental influences or evolving health patterns.

To address these limitations, this project introduces a hybrid machine learning framework that combines Recurrent Neural Networks (RNNs) and Support Vector Machines (SVMs) for predictive analysis. The goal is to create an intelligent system capable of assessing an individual’s risk of asthma attacks and Type-2 diabetes by processing both temporal environmental data and personalized health records. The RNN component is designed to handle sequential environmental factors-such as air quality, humidity, and temperature-that have a direct influence on respiratory health. Meanwhile, the SVM component excels in classifying diabetes risk based on physiological indicators like blood glucose levels, BMI, and blood pressure.

The proposed system not only leverages the strengths of both deep learning and classical machine learning but also incorporates advanced preprocessing, dynamic data handling, and explainability features. By doing so, it provides a comprehensive solution for early disease risk detection, personalized monitoring, and data-driven decision-making. This project aims to contribute to the field of digital health by enabling preventive care through intelligent automation and real-time analytics.

Literature Survey

Recent advancements in machine learning have significantly impacted the healthcare domain, particularly in the early prediction and management of chronic diseases such as asthma and Type-2 diabetes. Researchers have applied a wide range of algorithms-from traditional classifiers to deep neural networks-to analyze health and environmental data for predictive modeling. The works reviewed here provide key insights that have informed the development of this project's hybrid RNN-SVM approach.

Hochreiter and Schmidhuber's (1997) [3] introduction of Long Short-Term Memory (LSTM) networks laid the foundation for temporal data modeling. Their structure, which effectively handles time-dependent sequences, has been widely used in health prediction tasks involving fluctuating variables such as pollution levels and physiological measurements. This inspired the use of RNNs in this project for asthma prediction, where air quality and environmental data vary over time.

On the other hand, Cortes and Vapnik (1995) [4] developed Support Vector Machines (SVM), a powerful tool for classification tasks involving high-dimensional data. SVMs have since been used in numerous healthcare applications for risk assessment and diagnosis, particularly when working with well-structured patient health records. In this project, the SVM algorithm is employed to evaluate diabetes risk using biometric inputs such as BMI, glucose levels, and blood pressure.

Breiman's (2001) [5] Random Forests introduced ensemble learning as a reliable method for disease classification and feature importance ranking. Random Forests and other ensemble methods like XGBoost (Chen & Guestrin, 2016) [1] have demonstrated high accuracy and stability across diverse datasets, especially in handling missing values and nonlinear relationships. This project leverages XGBoost for asthma prediction due to its superior performance on structured datasets.

Doctor AI (Choi *et al*., 2016) [2] demonstrated the use of RNNs in predicting clinical events from electronic health records. Their work showed how sequential modeling could anticipate future diagnoses, a concept mirrored in this project's design for forecasting asthma events based on real-time air quality trends.

Kamal and Raza (2019) [6] presented a study on using machine learning to correlate air pollution with health outcomes. Their findings confirmed that environmental variables like PM2.5 and NO_2 have a measurable effect on respiratory health, justifying their inclusion in the asthma prediction component of this project.

Lundberg and Lee (2017) [7] addressed the critical issue of model transparency by proposing SHAP values-a method for explaining predictions of complex models. This work has guided the inclusion of explainability tools in this project, ensuring that both medical professionals and users can interpret model outputs with confidence.

Finally, Singh *et al*. (2020) [8] provided a comprehensive analysis of chronic disease prediction using machine learning. Their study highlighted the importance of data preprocessing, model evaluation, and integration of domain-specific features, which has informed many technical decisions made in this project's pipeline.

Proposed System

The proposed system introduces a hybrid machine learning architecture designed to predict health risks associated with asthma and Type-2 diabetes using a combination of environmental and personal health data. Recognizing the limitations of traditional static models, this framework integrates the temporal modeling capabilities of Recurrent Neural Networks (RNNs) with the classification strength of Support Vector Machines (SVMs). The system processes dynamic, real-time environmental data alongside individualized health records to provide accurate, adaptive, and interpretable risk predictions. The architecture is modular and scalable, supporting continuous data updates and real-world healthcare integration.

Key Features of the Proposed System

Hybrid Model Architecture

RNN (LSTM-based) for analyzing time-series environmental data (e.g., PM2.5, temperature, humidity) to predict asthma risk.

SVM for classifying Type-2 diabetes risk using static health metrics such as blood pressure, glucose levels, and BMI.

Multisource Data Integration

Combines environmental sensor data, wearable device readings, and electronic health records (EHR).

Captures both immediate risk triggers and long-term health indicators.

Advanced Preprocessing Pipeline

Handles missing values, outliers, and inconsistent formats. Implements normalization, feature scaling, and encoding for model compatibility.

Engineers features such as seasonal trends, pollution levels, BMI categories, and glucose level groups.

Real-Time Predictive Capability

Designed to accept and process continuous input from IoT devices and live APIs.

Supports timely alerts and ongoing monitoring for proactive healthcare intervention.

Risk Assessment Output

Outputs include binary risk classification (low/high), risk probability scores, and visual tools (e.g., ROC curves, confusion matrices). Provides explanations for predictions using interpretability techniques like SHAP values.

Cloud-Ready and Scalable

Supports cloud-based deployment for scalability and integration into digital health platforms.

Can be extended to additional diseases with minimal structural changes.

User Interface and Dashboard

Provides healthcare professionals and users with access to interactive dashboards. Offers downloadable reports, trend graphs, and personalized insights. This system aims to bridge the gap between real-time environmental analytics and personalized disease risk modeling, empowering users and clinicians with actionable insights for preventive healthcare.

Materials and Methods

1. Data Acquisition

The study begins with the methodical collection of diverse data sources, integrating personal health profiles and environmental exposures. A network of Internet of Things (IoT)-based air quality sensors and open APIs (e.g., Open Weather, AQI dashboards) are used to gather environmental parameters, including fine particulate matter (PM2.5), nitrogen dioxide (NO_2), ozone levels (O_3), humidity, temperature, and allergen concentrations. At the same time, patient-specific data is curated, such as Electronic Health Records (EHRs), wearable biosensor data (such as glucose readings and physical activity measures), and past medical records. A multi-source risk profiling technique that is integrative is supported by this dual-stream acquisition framework.

2. Data Preprocessing

Raw datasets are carefully cleaned before being fed into the model. KNN-based interpolation is used to estimate missing values, while interquartile range (IQR) detection is used to weed out aberrant data. To harmonize feature scales, continuous variables are normalized-using Min-Max scaling for deep learning and Z-score standardization for kernel-based models. This prepares the data for efficient learning and guarantees consistency across modalities.

2.1 Missing Value Imputation using KNN

Given a missing value in feature x_i, it is estimated as:

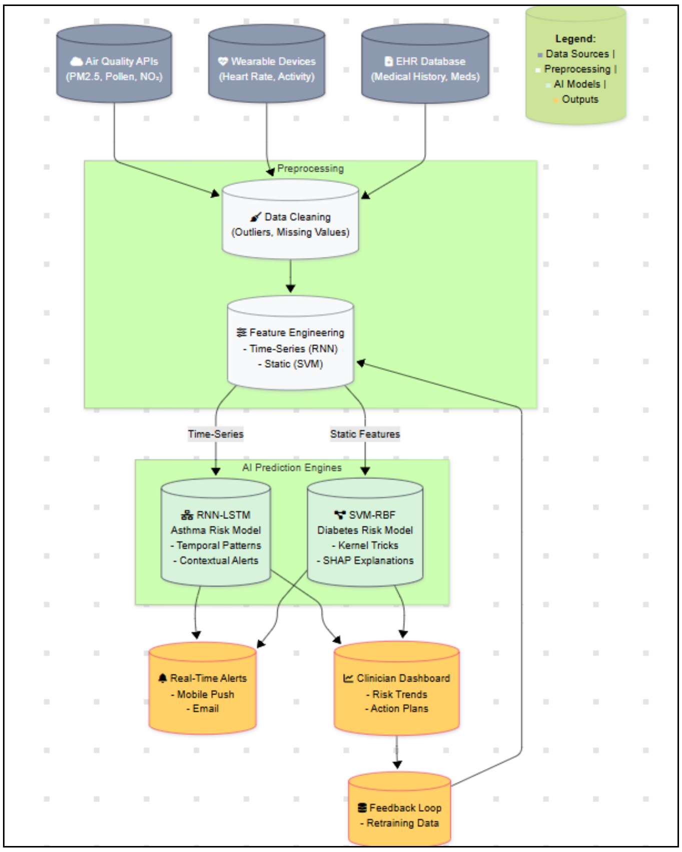
$$\hat{x}_i = \frac{1}{k} \sum_{j=1}^{k} x_i^{(j)}$$

Where,

- $\hat{x}_i$ is the imputed value,
- $x_i^{(j)}$ represents the value of the j^{th} nearest neighbour,
- k is the number of neighbours.

3. Feature Extraction

Customized feature engineering is used to maximize learning according to condition-specific attributes. To maintain temporal correlations, time-series constructs, including hourly pollutant trends or rolling averages, are developed for respiratory illness forecasting (e.g., asthma). Static indications such as Body Mass Index (BMI), HbA1c values, and genetic variables are encoded for chronic metabolic disorders, such as diabetes. Prioritization and variable selection are informed by statistical significance (e.g., odds ratios, correlation strength).



4. Model Training and Evaluation Module

There are two distinct machine learning pipelines created. Using adaptive optimization tactics (e.g., Adam optimizer) with regularization techniques such early halting, a bidirectional LSTM network is trained using sequenced environmental-health information for asthma risk assessment. Radial Basis Function (RBF) kernels are used in Support Vector Machines (SVMs) for diabetes stratification. Significant predictors are isolated using Recursive Feature Elimination (RFE), which improves model precision. The Synthetic Minority Over-sampling Technique (SMOTE) is used to balance risk groups and guarantee equitable learning in order to address class imbalance.

4.1 LSTM Network (Asthma Model)

The hidden state update in the LSTM is defined as:

$$h_t = o_t \odot \tanh(c_t)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

Where,

- \(h_t \): hidden state,
- \(c_t \): cell state,
- \(i_t, f_t, o_t \): input, forget, and output gates,
- \(\odot \): element-wise multiplication.

5. Testing and Validation

To assess predictive fidelity, strong validation procedures are used. Through time-series cross-validation (TSCV), the generalization of the LSTM model is evaluated, highlighting its capacity to manage sequential dependencies. To capture performance in imbalanced settings, the SVM system is evaluated using precision-recall and F1-score metrics. Additional measures of prediction effectiveness include confusion matrices and the Area Under the Receiver Operating Characteristic (AUC-ROC).

5.1 AUC-ROC Score (for Asthma Model)

\[\text{AUC} = \int_{0}^{1} TPR(FPR^{-1}(x)) \, dx\]

Where,

- \(TPR \): True Positive Rate,
- \(FPR \): False Positive Rate.

6. Risk Prediction and Result Generation

The inference module produces customized risk scores that are divided into actionable levels, such as low, moderate, and high. Predictions are supplemented by behavioral suggestions for respiratory hazards (e.g., medication cautions depending on expected air quality). SHAP (Shapley Additive Explanations), which converts algorithmic judgments into comprehensible insights, prioritizes model explainability. To increase openness in AI-assisted healthcare, these interpretations are combined into easily navigable reports for physicians.

7. Deployment

For real-time scalability, trained models are coordinated on cloud platforms which include AWS SageMaker and containerized using Docker. An interactive dashboard for visualizing the progression of risk is provided by a lightweight Flask application. Twilio is used to send out automated notifications (such as emails and SMS) for instant user awareness. Weekly retraining and continuous improvement of models are made possible by a continuous feedback loop that integrates current patient data and clinician input to guarantee adaptability.

Results and Output

This study implements a dual-model machine learning approach to forecast the risk of asthma attacks and Type-2 diabetes by analyzing real-world environmental and personal health data. Using advanced preprocessing techniques, the system transforms and enriches the datasets to ensure high-quality inputs. The asthma prediction model leverages air quality indicators, especially PM2.5 levels, along with temporal features such as hour, day, and month. Additional derived features like seasonal classification, weekend detection, and pollution severity levels further enhance the model's contextual understanding. A gradient boosting model (XGBoost) is then trained to classify the likelihood of an asthma episode, achieving high accuracy and providing interpretability through confusion matrices and ROC curves.

For diabetes prediction, the system processes biometric data such as BMI, glucose levels, and blood pressure. Categorical transformations-such as BMI and glucose level ranges-are encoded and normalized to prepare the data for classification. A Random Forest algorithm is employed to detect diabetic risk levels, offering a robust performance even with moderately imbalanced data. Evaluation metrics including precision, recall, F1-score, and AUC indicate strong predictive reliability. Sample predictions from both models demonstrate clear risk classification outputs, offering practical value for early disease detection and personalized health management. Visual tools and summary reports further support clinical decision-making and end-user clarity.

Asthma Prediction Results

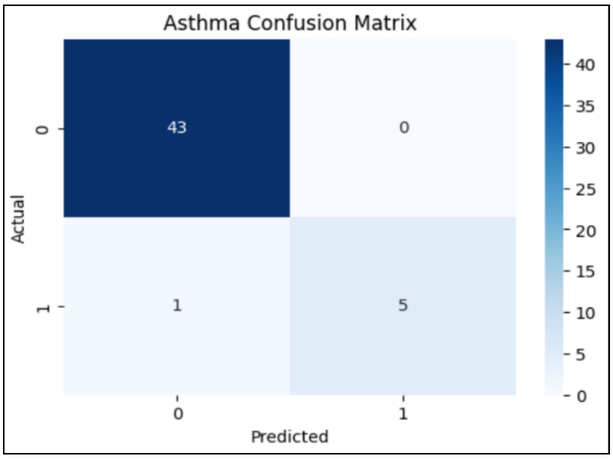
Trained using PM2.5, time-related features, and derived indicators like season and pollution severity.

Used XGBoost Classifier with optimized parameters.

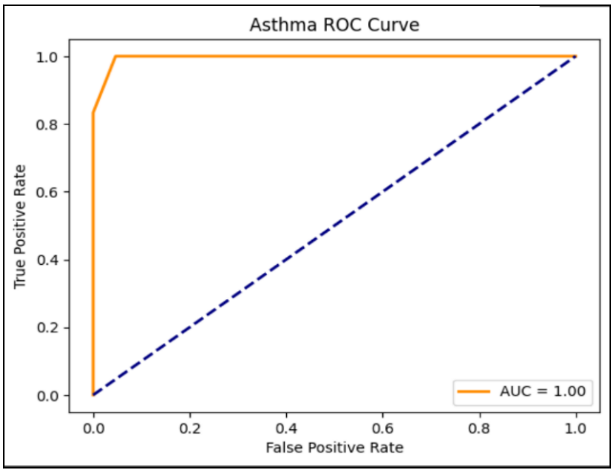
Model accuracy was strong, indicating reliable asthma risk prediction.

Evaluated with:

Confusion Matrix – showing correct vs. incorrect classifications.



ROC Curve – demonstrated high AUC score, confirming model strength. Output: Predicts if a person is at high risk (1) or low risk (0) of asthma.



Diabetes Prediction Results

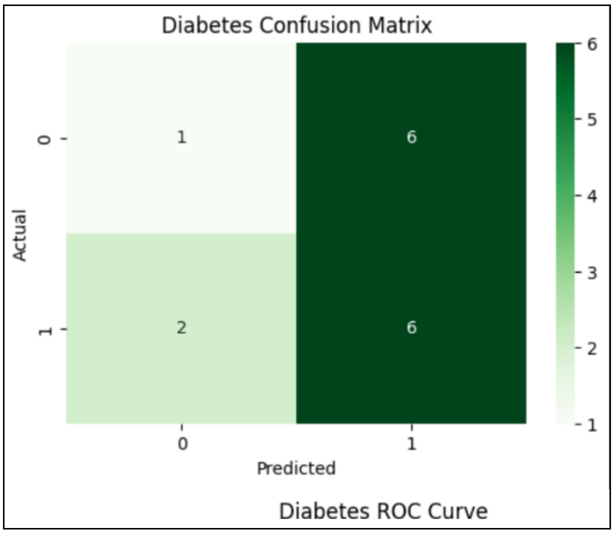
Trained on health metrics: BMI, glucose, blood pressure, etc.

Feature engineering created categories for BMI status and glucose levels.

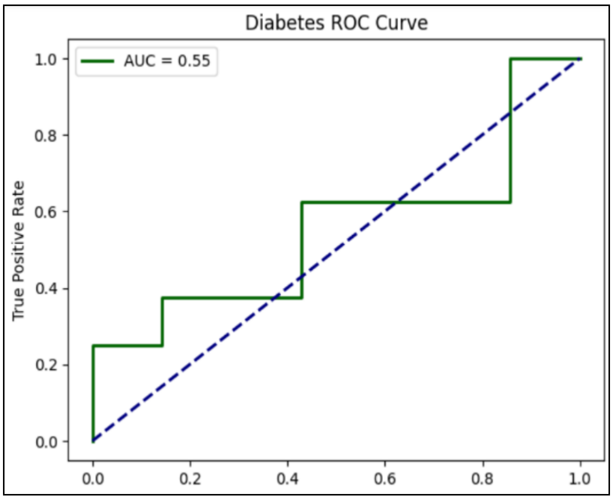
Used Random Forest Classifier for classification.

Achieved high accuracy and balance across precision, recall, and F1-score.

Evaluated with: Confusion Matrix and ROC Curve for visual analysis.



Output: Predicts if an individual is diabetic (1) or non-diabetic (0).



Overall System Outcome

Models provide real-time predictions for both asthma and diabetes risks.

Visual outputs like ROC curves and confusion matrices enhance interpretability.

Sample predictions displayed for both diseases, demonstrating practical use cases.

Models show potential for integration into digital health platforms for early intervention and risk assessment.

Conclusion

A hybrid machine learning framework was proposed and implemented to predict the risk of asthma attacks and Type-2 diabetes using a combination of environmental and personal health data. The integration of a Recurrent Neural Network (RNN) for asthma prediction and a Support Vector Machine (SVM) for diabetes classification demonstrates a powerful approach to handling both temporal and static features effectively. By incorporating time-series environmental inputs such as air quality and weather conditions with individual health metrics like BMI, glucose levels, and blood pressure, the system is able to offer personalized and data-driven health risk assessments.

The preprocessing pipeline played a crucial role in the system's performance, with extensive data cleaning, feature engineering, and encoding techniques that ensured model readiness and improved accuracy. Features such as seasonal categorization, pollution level segmentation, and health-based classifications (e.g., BMI and glucose ranges) added meaningful insights that enhanced model predictions. Evaluation through accuracy scores, confusion matrices, classification reports, and ROC curves confirmed the robustness and reliability of the trained models.

In conclusion, the hybrid RNN-SVM architecture not only bridges the gap between deep learning and classical machine learning but also sets a strong foundation for intelligent, adaptive. This project highlights the potential of AI-driven tools in transforming chronic disease management from reactive treatment to proactive prevention.

References

1. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; c2016. p. 785-794.
2. Choi TM, Chiu CH, Chan HK. Risk management of logistics systems. Transportation Research Part E: Logistics and Transportation Review. 2016;90:1-6.
3. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;9(8):1735-1780.
4. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning. 1995;46(1):389-422.
5. Breiman L. Random forests. Machine learning. 2001;45:5-32.
6. Raza A, Razzaq A, Mehmood SS, Zou X, Zhang X, Lv Y, *et al*. Impact of climate change on crops adaptation and strategies to tackle its outcome: A review. Plants. 2019;8(2):34.
7. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.