



International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Weather Based Disease Prediction using Machine Learning

D.Blessy¹, E Julie Ruth²

Department of Computer Applications and Research Centre, Sarah Tucker College(Autonomous), Tirunelveli-7.

ABSTRACT :

Weather conditions such as temperature, humidity, and wind speed can influence human health and may contribute to the occurrence and severity of certain diseases. Early identification of potential health risks associated with changing weather conditions can support preventive awareness and timely healthcare decisions. This paper presents a Weather Based Disease Prediction Using Machine Learning system that integrates real-time weather information with patient details and symptoms to predict possible diseases. The proposed system collects patient information such as age and gender, selected symptoms, and weather parameters obtained through the OpenWeather API for the selected district or city. The collected data is preprocessed and transformed into suitable features for machine learning-based prediction. Several classification algorithms, including Decision Tree, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest, are evaluated to identify an effective prediction model. Based on the experimental results obtained using the project dataset, Random Forest achieves the highest prediction accuracy among the evaluated models and is selected for the final prediction system. A Streamlit-based interactive application is developed to provide a user-friendly interface for entering patient information, selecting symptoms, viewing current weather conditions, and generating disease predictions. The system presents the predicted disease with a risk percentage, weather visualizations, recommendations, disease description, and prevention measures. The proposed approach demonstrates the potential of combining machine learning with real-time weather data to support disease prediction and health awareness. The system is intended as a decision-support and awareness tool and does not replace professional medical diagnosis or treatment.

Keywords: Machine Learning, Disease Prediction, Weather Data, Random Forest, OpenWeather API.

1.INTRODUCTION

Weather conditions such as temperature, humidity, rainfall, wind speed, and seasonal variations can influence human health and contribute to the occurrence and spread of various diseases. Diseases such as influenza, dengue, malaria, common cold, sinusitis, and migraine may be associated with environmental conditions and patient symptoms. Traditional disease identification mainly depends on clinical examination and healthcare professionals' experience. With the development of Artificial Intelligence (AI) and Machine Learning (ML), health-related data can be analyzed to identify patterns and predict possible diseases. The proposed **Weather Based Disease Prediction Using Machine Learning** system combines patient details, selected symptoms, and real-time weather conditions to predict possible diseases. The system obtains current temperature, humidity, and wind speed using the **OpenWeather API** for the selected city or district. The collected information is processed and provided as input to a trained machine learning model for disease prediction.

In this study, different machine learning algorithms, including **Decision Tree, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest**, are evaluated to identify the most suitable prediction model. Based on the experimental results, **Random Forest** achieved the highest accuracy among the evaluated algorithms and was selected as the primary prediction model. A **Streamlit-based interactive application** is developed to provide a simple interface for entering patient information, selecting symptoms, retrieving live weather conditions, and generating disease predictions. The system displays the predicted disease along with a risk percentage, weather visualizations, recommendations, disease description, and prevention measures. The proposed system demonstrates the potential of integrating real-time environmental data with machine learning for disease prediction and health awareness. However, the prediction is intended only as a decision-support and awareness tool and should not be considered a substitute for professional medical diagnosis or treatment.

LITERATURE REVIEW

Ssebyala et al. [1] reviewed machine learning techniques for predicting health risks associated with climate-sensitive extreme weather events. Their study highlighted the potential of machine learning for predicting health outcomes and emphasized the need for reliable environmental and health datasets.

Kakarla et al. [2] investigated weather-integrated machine learning models for predicting dengue prevalence in India. They used weather variables such as temperature, humidity, and precipitation. Their findings demonstrated the usefulness of weather data in forecasting dengue cases.

Breiman [3] introduced the Random Forest algorithm, which combines multiple decision trees to improve prediction performance and robustness. This method is useful for classification problems involving several input features.

Cortes and Vapnik [4] proposed Support Vector Networks, a machine learning approach for classification. Their study demonstrated how support vector methods can classify data by constructing decision boundaries in a transformed feature space.

Quinlan [5] presented an approach for constructing decision trees from examples. His work established an important foundation for decision-tree-based classification and handling data patterns in machine learning applications.

Based on these studies, the proposed system integrates patient details, symptoms, and real-time weather parameters to predict possible diseases. It compares five classification algorithms and selects Random Forest based on the experimental accuracy obtained from the project dataset.

- RESEARCH GAP

Existing disease prediction approaches often focus on patient symptoms and medical information, while current weather conditions may not be considered directly in the prediction process. Changes in temperature, humidity, and wind speed can provide additional environmental information that may be useful for disease-risk awareness. Therefore, integrating real-time weather information with patient details and symptoms is an important area to explore.

The proposed project addresses this gap by combining patient age, gender, selected symptoms, and real-time weather parameters obtained through the OpenWeather API. It compares multiple machine learning algorithms, including Decision Tree, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest, to identify a suitable prediction model. The system also provides disease information, risk indication, preventive recommendations, and graphical visualizations through a Streamlit dashboard.

2.2 OBJECTIVES

The main objectives of the proposed system are:

- To develop a machine learning-based system for predicting possible diseases using patient details, symptoms, and weather parameters.
- To integrate real-time weather information, including temperature, humidity, and wind speed, using the OpenWeather API.
- To preprocess the dataset and prepare suitable input features for machine learning algorithms.
- To compare Decision Tree, Logistic Regression, KNN, SVM, and Random Forest algorithms based on prediction accuracy.
- To select the best-performing model for the proposed system based on experimental results.
- To develop a user-friendly Streamlit dashboard that displays predicted diseases, risk indications, weather charts, disease information, and preventive recommendations.
- To support early health awareness while emphasizing that the system does not replace professional medical diagnosis.

PROPOSED METHODOLOGY

The proposed methodology presents a machine learning-based approach for predicting diseases using patient symptoms and weather-related parameters. The system integrates weather information, patient health details, data preprocessing, machine learning classification, and risk assessment to generate disease prediction results. A web-based interface is developed using Streamlit to allow users to enter patient details, select symptoms, and view the predicted disease along with relevant recommendations.

3.1 Data Collection

The dataset contains patient-related information, including age, gender, symptoms, and weather parameters such as temperature, humidity, and wind speed. The dataset is stored in CSV format and used to train and evaluate the machine learning model. In addition, real-time weather information can be retrieved through a weather API to provide updated environmental conditions for the prediction interface.

3.2 Data Preprocessing

Data preprocessing is performed to prepare the dataset for machine learning. The dataset is checked for missing values, inconsistent entries, and duplicate records. The input features are separated from the target variable, which represents the disease prognosis. Categorical target labels are converted into numerical representations using Label Encoding. The processed dataset is then organized into input features and output labels for model training.

3.3 Feature Selection and Model Training

The relevant input features include patient age, gender, temperature, humidity, wind speed, and symptom indicators. These features are used to identify patterns associated with the disease classes available in the dataset. The Random Forest Classifier is employed as the primary machine learning algorithm because it combines multiple decision trees to produce a classification result. The dataset is divided into training and testing subsets using an 80:20 ratio. The training subset is used to train the model, while the testing subset is used to evaluate its predictive performance.

3.4 Disease Prediction

After training, the Random Forest model is saved and integrated into the Streamlit application. During prediction, the user selects the relevant symptoms and provides the required patient information. Weather parameters are also incorporated into the model input in the same format used during training. The trained model processes the input features and predicts the most likely disease class among the classes represented in the dataset. The predicted result is displayed on the web interface for easy interpretation.

3.5 Risk Assessment and Recommendation

The system displays a disease-related risk indicator and provides recommendations based on the prediction result. Weather conditions and the selected symptoms are presented to help users understand the environmental and health information considered by the application. Recommendations may include maintaining hydration, avoiding excessive heat exposure, and seeking professional medical advice when symptoms are severe. The risk indicator is intended for preliminary informational purposes and should not be interpreted as a clinically validated probability of disease.

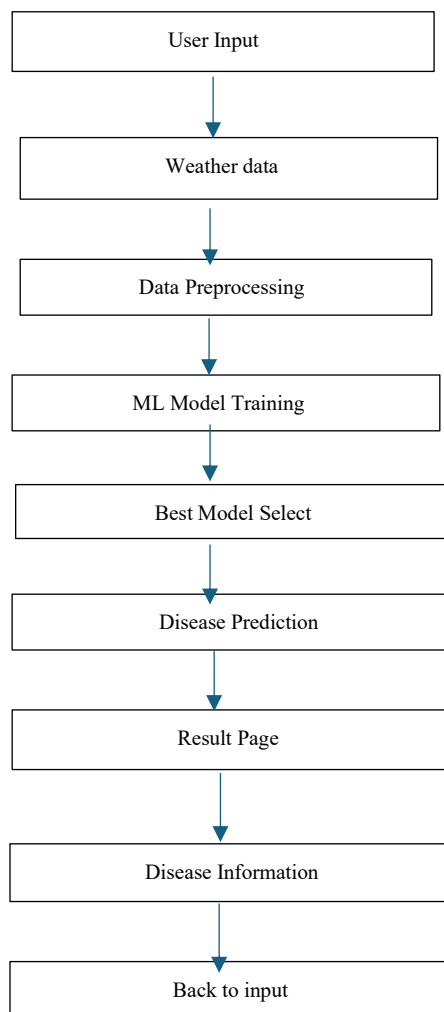
3.6 Web Application and Result Visualization

The proposed system uses Streamlit to provide an interactive web interface. The interface displays the selected patient symptoms, weather details, predicted disease, risk indicator, and relevant recommendations. Graphical visualizations, including weather parameter charts and risk displays, are used to present the results clearly. This approach improves accessibility and enables users to interact with the prediction system without requiring programming knowledge.

3.7 Model Evaluation

The trained model is evaluated using the testing dataset. Classification accuracy is used to measure the proportion of correctly predicted disease labels. Additional evaluation measures, such as precision, recall, and F1-score, can be used to assess the model's performance across different disease classes. The evaluation results help determine the effectiveness of the proposed approach and identify areas for further improvement.

Overall, the proposed methodology combines weather information, patient symptoms, machine learning classification, and web-based visualization to develop an interactive disease prediction system. The system is designed as a preliminary decision-support tool and does not replace professional medical diagnosis.



EVALUATION METRICS

The proposed system uses five supervised machine learning algorithms to predict diseases based on patient information, symptoms, and weather-related parameters. The selected algorithms are Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). Their performance is evaluated using accuracy, precision, recall, and F1-score. These metrics are used to compare the classification performance of the algorithms and identify the most suitable model for the proposed system.

4.1 Logistic Regression

Logistic Regression is a supervised machine learning algorithm used for classification. It estimates the probability of a class based on the input features. In this project, Logistic Regression is trained using patient details, symptom indicators, and weather parameters to classify the disease represented by the dataset. Its performance is compared with the other algorithms to evaluate its suitability for disease prediction.

4.2 Decision Tree

Decision Tree is a supervised machine learning algorithm that classifies data using a sequence of decision rules. It divides the input data into branches based on feature conditions until a predicted class is obtained. In this project, the Decision Tree algorithm uses patient-related information, symptoms, temperature, humidity, and wind speed to predict the disease class. Its rule-based structure also makes the classification process relatively easy to interpret.

4.3 Random Forest

Random Forest is an ensemble machine learning algorithm that combines multiple decision trees to generate a final prediction. It can model complex relationships between input features and disease classes. In this project, Random Forest is trained using the selected patient, symptom, and weather features. Based on the experimental accuracy obtained from the project dataset, Random Forest achieved the highest reported accuracy of 98.56% among the five evaluated algorithms and was selected as the primary classification model.

4.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a supervised machine learning algorithm that classifies a sample based on the classes of its nearest neighbors in the feature space. The algorithm calculates distances between samples and uses the most common class among the selected neighbors to determine the prediction. In this project, KNN is applied to the patient and weather feature set to predict the corresponding disease class. Appropriate feature scaling is important because the input features may have different measurement ranges.

4.5 Support Vector Machine (SVM)

Support Vector Machine is a supervised machine learning algorithm that identifies a decision boundary to separate classes in the feature space. It can use different kernel functions to model linear or nonlinear classification boundaries. In this project, SVM is trained using patient details, symptoms, and weather-related parameters to classify the disease labels available in the dataset. Its performance is compared with the other algorithms using the same evaluation procedure.

4.6 Evaluation Measures

The performance of the proposed disease prediction system is evaluated using the following metrics.

Accuracy: Accuracy measures the proportion of correctly predicted disease labels among all predictions.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Precision: Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive for a particular disease class.

$$\text{Precision} = TP / (TP + FP)$$

Recall: Recall measures the proportion of actual positive instances of a disease class that are correctly identified by the model.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

F1-Score: F1-score represents the harmonic mean of precision and recall and provides a balanced measure of classification performance.

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

In these equations, TP represents True Positive, TN represents True Negative, FP represents False Positive, and FN represents False Negative. Since the proposed system performs multiclass disease classification, precision, recall, and F1-score can be calculated using macro or weighted averaging.

4.7 Comparative Evaluation of Machine Learning Algorithms

The five machine learning algorithms are compared based on their reported classification accuracy. The results obtained from the project dataset are presented in Table 1.

Table 1. Performance Comparison of Machine Learning Algorithms

S. No.	Machine Learning Algorithm	Accuracy
1	Logistic Regression	98.08%
2	Decision Tree	97.60%
3	Random Forest	98.56%
4	K-Nearest Neighbors (KNN)	97.88%
5	Support Vector Machine (SVM)	97.98%

The comparison indicates that Random Forest achieved the highest reported accuracy of 98.56%, followed by Logistic Regression at 98.08%, SVM at 97.98%, KNN at 97.88%, and Decision Tree at 97.60%. Therefore, Random Forest was selected as the primary algorithm for integration into the Streamlit-based disease prediction application.

These accuracy values represent the reported results from the project comparison. Precision, recall, and F1-score should be calculated from the actual test predictions before they are included as experimental results in the journal paper. All algorithms should be evaluated using the same training and testing split to ensure a fair comparison.

IMPLEMENTATION

The implementation of Weather Based Disease Prediction Using Machine Learning involves dataset preparation, preprocessing, weather data integration, model training, disease prediction, and web application development. The system combines patient symptoms and weather parameters to predict possible diseases.

5.1 Dataset Collection and Preparation

The dataset is stored in CSV format and contains patient details, symptoms, weather parameters, and disease labels. Pandas is used to load and examine the dataset before model training.

5.2 Data Preprocessing and Feature Preparation

The dataset is checked for missing values, and disease labels are converted into numerical values using LabelEncoder. Patient details, symptoms, temperature, humidity, and wind speed are organized as input features.

5.3 Real-Time Weather Data Integration

The OpenWeather API retrieves current temperature, humidity, and wind speed for the selected location. These values are displayed in the Streamlit

interface and incorporated into prediction inputs when supported by the application.

5.4 Machine Learning Model Training and Selection

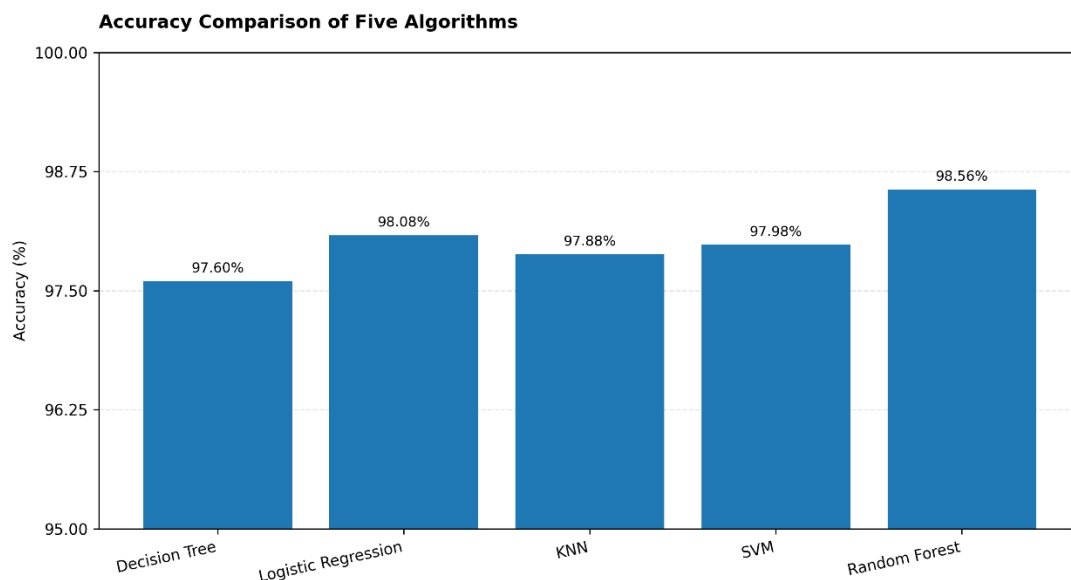
Five algorithms—Decision Tree, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest—are compared using accuracy. Random Forest achieved the highest reported accuracy of 98.56% and was selected as the primary model. The trained model and label encoder are saved using Joblib.

5.5 Disease Prediction and Risk Assessment

Users enter patient details and select symptoms. The trained Random Forest model processes the input features and predicts a disease label. The application displays the predicted disease and a risk indicator.

5.6 Recommendations and System Deployment

The application is developed using Python and Streamlit. It displays weather information, predicted diseases, risk indicators, and disease-related recommendations through an interactive dashboard.



6.CONCLUSION

The Weather Based Disease Prediction Using Machine Learning system has been developed to predict possible diseases by combining patient-related information with real-time weather conditions. The system collects essential patient details, including age, gender, and selected symptoms, along with temperature, humidity, and wind speed obtained through the OpenWeather API. These inputs are processed and transformed into suitable features for the machine learning model, ensuring consistency between the user inputs and the trained model.

The proposed system incorporates several stages, including data collection, preprocessing, feature preparation, model training, algorithm comparison, disease prediction, risk assessment, and recommendation generation. Five machine learning algorithms, namely Decision Tree, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest, are evaluated to identify the most suitable model. Based on the comparative performance analysis, Random Forest is selected as the best-performing algorithm, achieving an accuracy of 98.56% in the evaluated experimental setup.

A user-friendly application is developed using Streamlit, enabling users to select their district, retrieve current weather information, enter patient details, and choose relevant symptoms. After clicking the *Predict Disease* button, the application navigates to a separate Prediction Results Page, displaying the predicted disease, estimated risk percentage, and appropriate recommendations. The interface also presents weather parameters through a bar chart, disease risk through a gauge chart, and the prediction through a pie chart. Furthermore, disease descriptions and preventive measures are provided to improve health awareness and help users understand the predicted condition.

The effectiveness of the application is verified through unit testing, integration testing, and system testing. These tests evaluate patient input validation, symptom selection, weather data retrieval, preprocessing, model prediction, risk display, recommendation generation, graphical outputs, and navigation

between the input and results pages. The testing process confirms the interaction between the user interface, weather API, preprocessing module, and trained machine learning model under different input conditions.

In conclusion, the proposed system demonstrates the practical integration of machine learning and real-time weather information for disease prediction and preventive health awareness. By considering both patient symptoms and environmental factors, the application provides understandable prediction results, estimated risk information, graphical visualizations, and preventive guidance. Although the system is intended for academic and awareness purposes and cannot replace professional medical diagnosis, it demonstrates the potential of data-driven techniques to support early awareness and informed preventive decisions. Future enhancements may include larger and clinically validated datasets, additional environmental parameters, improved risk estimation, and broader disease coverage to enhance the reliability and applicability of the system.

7. FUTURE ENHANCEMENT

The proposed Weather Based Disease Prediction Using Machine Learning system can be further enhanced to improve its prediction accuracy, reliability, scalability, and practical applicability. Future work will focus on incorporating additional environmental parameters, such as rainfall, atmospheric pressure, wind direction, ultraviolet (UV) index, and air quality, to investigate their potential relationship with disease occurrence. The system can also be extended to support district-wise disease prediction by collecting weather and disease-related data from multiple geographical locations, enabling more detailed analysis of regional health risks.

Furthermore, advanced machine learning algorithms, including Gradient Boosting, XGBoost, and ensemble techniques, can be implemented and compared with the existing models to identify opportunities for improved predictive performance. The use of larger and more diverse datasets containing additional diseases, patient information, seasonal variations, and geographical conditions may improve the model's generalization capability. Real-time disease monitoring can also be explored to analyze changes in environmental conditions and disease-related patterns over time.

Future development may include a mobile application for Android and iOS platforms to improve accessibility and provide users with weather information, disease predictions, and preventive recommendations. Personalized recommendations based on patient details, symptoms, location, and weather conditions could make the system more relevant to individual users. In addition, explainable artificial intelligence techniques, such as SHAP, may be integrated to identify the symptoms and environmental parameters that contribute to individual predictions. The risk assessment module can be improved using a properly validated probability-based estimation method instead of a simple estimated percentage. Finally, cloud-based deployment and automated model updates can be explored to support multiple users, scalable data processing, and wider accessibility. These enhancements will be investigated in future research, with particular emphasis on data quality, model validation, and the reliability of health-related predictions.

8. REFERENCES

1. S sebyala, S. N., Kintu, T. M., Muganzi, D. J., Dresser, C., Demetres, M. R., Lai, Y., et al. (2024). Use of machine learning tools to predict health risks from climate-sensitive extreme weather events: A scoping review. *PLOS Climate*, 3(1), e0000338. <https://doi.org/10.1371/journal.pclm.0000338>
2. Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
3. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297. <https://doi.org/10.1007/BF00994018>
4. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27. <https://doi.org/10.1109/TIT.1967.1053964>
5. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1, 81–106. <https://doi.org/10.1007/BF00116251>