INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholorly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 2, No 06, 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-750 PAGE NO: 007-012

STATISTICAL ANALYSIS AND MACHINE LEARNING FOR DISEASE PREDICTION

Dr.S.Shanmuga Priya
Assistant Professor,
Department of Information Technology,
Mannar Thirumalai Naicker College,
Madurai, Tamil Nadu, India.

Ms.J.Geetha
Assistant Professor,
Department of Information Technology,
Mannar Thirumalai Naicker College,
Madurai, Tamil Nadu, India.

Ms.R.Nalini
Assistant Professor,
Department of Mathematics,
Mannar Thirumalai Naicker College,
Madurai, Tamil Nadu, India.

Dr.M.Nagadevi
Assistant Professor,
Department of Mathematics,
Mannar Thirumalai Naicker College,
Madurai, Tamil Nadu, India.

Abstract:

Early disease prediction can help healthcare professionals identify patients who may require further clinical assessment. The increasing availability of healthcare datasets has created opportunities to combine statistical analysis with machine learning for predictive healthcare. Statistical methods help identify relationships among clinical variables, while machine learning algorithms can detect complex patterns and generate predictions. This paper proposes a framework that combines statistical analysis and machine learning for disease prediction. The framework includes data collection, preprocessing, statistical analysis, feature selection, machine learning, and performance evaluation. Algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and XGBoost can be used for prediction.

Evaluation can be performed using accuracy, precision, recall, F1-score, and ROC-AUC. The proposed approach can support disease-risk assessment while maintaining attention to data quality, privacy, interpretability, and clinical validation.

Keywords: Disease Prediction, Statistical Analysis, Machine Learning, Healthcare Analytics, Predictive Analytics, Data Science.

1. Introduction

Healthcare organizations generate large amounts of data through electronic health records, laboratory tests, medical examinations, and wearable devices. These datasets contain useful information about patient characteristics and disease-related risk factors. Analyzing this information can

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholorly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 2, No 06, 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-750 PAGE NO: 007-012

support early disease-risk identification and healthcare decision-making.

Traditional statistical methods such as Logistic Regression are widely used for disease prediction because they are relatively interpretable and can identify relationships between risk factors and outcomes. However, healthcare data can contain nonlinear relationships and complex interactions that may be difficult to model using conventional statistical techniques alone.

Machine learning provides methods for identifying patterns in complex datasets. Algorithms such as Random Forest, Support Vector Machine, and XGBoost can be used to classify patients according to their predicted disease status. Recent research has investigated the integration of machine learning with statistical methods for disease-risk prediction and suggests that combining the two approaches can provide complementary advantages.

The objectives of this study are to:

1. Analyze healthcare data using statistical methods.
2. Identify important disease-related features.
3. Develop machine learning models for disease prediction.
4. Compare model performance using standard evaluation metrics.
5. Identify challenges related to healthcare prediction.

2. Literature Review

Statistical modeling has traditionally been used for clinical risk prediction.

Logistic Regression is particularly useful for binary outcomes such as disease presence or absence. Its probability model can be represented as:

$$P(Y=1 | X) = 1/(1+e^{-z})$$

Where:

$$z = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

Machine learning methods extend predictive analysis by learning patterns directly from data. Decision Trees provide interpretable rules, Random Forest combines multiple trees, SVM identifies separating boundaries, and XGBoost uses sequential decision-tree learning to improve predictions. A systematic review of statistical and machine learning integration in disease-risk prediction reported that combining statistical methods with machine learning can address limitations associated with using either approach independently. (pubmed.ncbi.nlm.nih.gov)

3. Proposed Methodology

The proposed system contains the following stages:

Healthcare Dataset
↓
Data Preprocessing
↓
Statistical Analysis
↓
Feature Selection
↓

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholorly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 2, No 06, 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-750 PAGE NO: 007-012

Machine Learning Models
↓
Performance Evaluation
↓
Disease Prediction

3.1 Data Collection

The framework can be applied to public healthcare datasets for diseases such as:

- Diabetes
- Heart disease
- Kidney disease
- Liver disease
- Hypertension

Possible input features include age, gender, BMI, blood pressure, glucose, cholesterol, family history, smoking status, and physical activity.

3.2 Data Preprocessing

Healthcare datasets may contain missing values, outliers, duplicate records, and categorical variables.

The preprocessing stage includes:

- Handling missing values
- Removing duplicate records
- Detecting outliers
- Encoding categorical variables
- Scaling numerical features

Standardization can be calculated as:

$$Z = (X - \mu) / \sigma$$

Where μ is the mean and σ is the standard deviation.

3.3 Statistical Analysis

Descriptive statistics such as mean, median, variance, and standard deviation are used to understand the dataset.

Correlation analysis can identify relationships between variables:

$$r = \mathrm{Cov}(X,Y)/(\sigma X \sigma Y)$$

Hypothesis tests such as the t-test and chi-square test can be used to examine relationships between patient characteristics and disease outcomes.

3.4 Feature Selection

Feature selection reduces unnecessary variables and can improve computational efficiency. Possible techniques include:

- Correlation-based selection
- LASSO
- Recursive Feature Elimination
- Random Forest feature importance

LASSO uses an L1 penalty that can reduce some coefficients to zero, thereby selecting relevant variables.

4. Machine Learning Models

Logistic Regression

Logistic Regression provides an interpretable statistical baseline for binary disease prediction.

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholorly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 2, No 06, 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-750 PAGE NO: 007-012

Decision Tree

A Decision Tree creates a series of rules to classify patients. It is simple to understand but can overfit without appropriate control.

Random Forest

Random Forest combines multiple decision trees and can model nonlinear relationships while providing feature-importance estimates.

Support Vector Machine

SVM identifies a decision boundary that separates different classes. It can be effective for datasets with many features.

XGBoost

XGBoost is a gradient-boosting algorithm that uses decision trees to achieve strong predictive performance on structured datasets.

5. Model Evaluation

The dataset should be divided into training and testing sets. Cross-validation can be used for model selection and parameter tuning.

The main evaluation metrics are:

Accuracy

Accuracy = (TP + TN)/(TP + TN + FP + FN)

Precision

Precision = TP/(TP + FP)

Recall

Recall = TP/(TP + FN)

F1-Score

F1 = 2 × (Precision × Recall)/(Precision + Recall)

ROC-AUC can also be used to measure classification performance across different decision thresholds. Standard machine learning tools provide these and other evaluation measures. (scikit-learn.org)

Proposed Comparison Table

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	—	—	—	—
Decision Tree	—	—	—	—
Random Forest	—	—	—	—
SVM	—	—	—	—
XGBoost	—	—	—	—

The blank values should be replaced with actual results obtained from the selected dataset.

6. Challenges

Several challenges must be considered when developing disease prediction systems.

- **Data Quality:** Missing, noisy, or inconsistent data can reduce model performance.

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholorly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 2, No 06, 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-750 PAGE NO: 007-012

- **Class Imbalance:** Unequal numbers of disease and non-disease cases can bias predictions.
- **Overfitting:** Complex models may perform well on training data but poorly on new patients.
- **Interpretability:** Some machine learning models are difficult to explain.
- **Privacy:** Healthcare information requires strong security and privacy protection.
- **Bias:** Models trained on unrepresentative datasets may produce unequal performance across populations.
- **Clinical Validation:** Good performance on a dataset does not automatically demonstrate clinical usefulness.

The WHO recommends that AI used in healthcare should consider safety, transparency, accountability, equity, privacy, and human autonomy. (who.int)

7. Future Scope

Future research can improve the proposed system through Explainable AI, deep learning, federated learning, and real-time IoT-based monitoring. Combining electronic health records with wearable-device data could enable continuous disease-risk assessment. External validation using datasets from different hospitals and patient populations is also important. Future studies should evaluate not only predictive accuracy but also model fairness, calibration, interpretability, and clinical usefulness.

8. Conclusion

This paper presented a framework combining statistical analysis and machine learning for disease prediction. Statistical methods help understand clinical variables and identify important relationships, while machine learning algorithms can detect complex patterns and generate predictions.

Logistic Regression, Decision Tree, Random Forest, SVM, and XGBoost can be compared using accuracy, precision, recall, F1-score, and ROC-AUC. The combined approach can support early risk assessment and healthcare decision-making. However, reliable disease prediction requires high-quality data, appropriate validation, privacy protection, interpretability, and clinical testing. Future work can focus on Explainable AI, federated learning, IoT-based monitoring, and personalized disease prediction.

References

1. M. Zhang, Y. Zheng, X. Maidaiti, et al., “Integrating Machine Learning into Statistical Methods in Disease Risk Prediction Modeling: A Systematic Review,” Health Data Science, vol. 4, 2024. doi: 10.34133/hds.0165. (pubmed.ncbi.nlm.nih.gov)
2. L. Breiman, “Random Forests,” Machine Learning, vol. 45, pp. 5–32, 2001.
3. C. Cortes and V. Vapnik, “Support-Vector Networks,” Machine Learning, vol. 20, pp. 273–297, 1995.
4. T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholorly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 2, No 06, 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-750 PAGE NO: 007-012

International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.

5. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
6. World Health Organization, Ethics and Governance of Artificial Intelligence for Health: WHO Guidance, Geneva: WHO, 2021.