

Utilizing a Random Forest Model for Disease Classification

Navya Rajguru

Received July 11, 2025

Accepted February 11, 2026

Electronic access March 31, 2026

As healthcare increasingly relies on data-driven methods, machine learning offers promising avenues for improving disease diagnosis. This study employs a Random Forest classifier to analyze patient symptoms, demographic information, and medical history in order to predict diagnoses for a range of diseases including asthma, stroke, influenza, eczema, migraines, diabetes, pneumonia, hypertension, hyperthyroidism, and kidney cancer. The model achieved an accuracy of 82% in predicting whether a patient's diagnosis is positive or negative for each disease, demonstrating its ability to capture patterns in patient profiles. Further analysis included feature importance, identifying key variables such as difficulty breathing, fever, and blood pressure as critical predictors. The findings highlight the potential of Random Forest models for interpretable and reliable disease classification. Future work will explore additional machine learning methods and strategies to handle real-world challenges, including class imbalance and diverse patient populations.

Introduction

In today's fast-paced medical world, diagnosing diseases quickly and accurately can make all the difference in patient outcomes. Technology may be utilized to help improve the speed of diagnosis. Machine learning algorithms like the random forest classifier are designed to analyze large amounts of patient data and spot patterns that may not be obvious to the human eye¹. In this study, the model is trained to identify diseases such as asthma, stroke, influenza, eczema, migraines, diabetes, pneumonia, hypertension, hyperthyroidism, and kidney cancer². By analyzing patient profile data—including symptoms, age, gender, and medical history—the random forest classifier can accurately predict whether a patient is likely to have any of these conditions, helping doctors make more informed diagnoses³.

Literature Review: Diseases

A study of 349 patients from multiple hospitals and age groups provides important insights into how different diseases affect individuals—not just physically, but also emotionally and in daily life⁴. Each illness presents unique challenges, impacting people beyond medical symptoms.

These disease characteristics directly informed the design of the random forest model in this study, guiding which features to include for accurate classification.

Asthma, found in 23 patients, was equally prevalent in men and women and most common among those aged 35–40⁵. The primary symptoms, such as wheezing and shortness of breath, were encoded as categorical features (“Yes”/“No”) for the model, while age and gender were included as numeri-

cal and categorical features respectively. Many patients managed asthma effectively with inhalers, resulting in positive outcomes.

Strokes were observed in 16 patients, affecting individuals mostly aged 65 and older. Loss of bodily functions like paralysis or speech difficulties informed feature importance, as age and symptom severity became critical inputs for predicting stroke risk⁶. Rehabilitation outcomes were also considered when interpreting class labels to ensure accurate model training.

Influenza, seen in eight patients, affects all ages and genders, with symptoms like fever, sore throat, and body aches. These categorical symptom features were directly included in the model, allowing the random forest to detect patterns in symptom combinations⁷. Rest and care led to positive outcomes, and these outcome variables guided the model's target labeling.

Eczema, affecting six patients aged 25–30, presented persistent itching and dry skin, which were encoded as symptom features. Demographics like age and gender were important because Eczema prevalence varied by these factors⁸.

Migraines, reported in 10 patients (mostly women aged 30–40), were unpredictable with recurring pain. The model incorporated both symptoms and demographic data to account for these variations, which contributed to more nuanced predictions.

Diabetes, found in 10 patients (slightly more men, typically over 40), included high blood sugar and fatigue as categorical features. Monitoring features such as blood pressure and cholesterol levels were also included, as they influence disease outcomes⁹.

Pneumonia (eight patients) affected all ages and genders,

with symptoms like chest pain and breathing difficulty. These were key features in the random forest, as they strongly correlated with positive diagnoses¹⁰.

Hypertension, observed in 11 patients (mostly women over 50), often presents silently. Early management influenced outcomes, highlighting the need to include both demographic and clinical measurements in the model to capture subtle patterns in symptomless cases¹¹.

Hyperthyroidism (seven patients, mostly over 40) caused rapid heartbeat, weight loss, and anxiety. These symptoms and demographic patterns were included in feature selection¹².

Kidney cancer, affecting six men over 40, was often linked to lifestyle factors like smoking and obesity. These risk factors were encoded as features to help the model account for external variables influencing disease probability¹³.

By explicitly incorporating disease-specific characteristics—such as age, gender, key symptoms, and lifestyle factors—the model’s feature selection and architecture were tailored to capture patterns across diverse conditions. This demonstrates how understanding the clinical nuances of each disease improves the random forest’s ability to classify patient outcomes accurately².

Literature Review: Random Forest Classification

Random forest classification is a dependable and adaptable tool based on the decision tree model, which is already intuitive and logical. It extends this concept by building multiple trees, each analyzing different parts of the data and making its own prediction. The final outcome is determined by majority vote, which increases stability and reliability compared to a single decision tree¹.

One reason random forest is particularly well-suited for the Disease Symptoms and Patient Profile Dataset is its flexibility in handling mixed data types. This dataset includes numerical features, such as age, blood pressure, and cholesterol levels, alongside categorical features like symptoms (Yes/No), gender, and disease presence³. Many machine learning models struggle with such a mixture, but random forests can process both numerical and categorical variables effectively without requiring extensive preprocessing². This allows the model to capture patterns across diverse patient profiles and symptom combinations, which is essential for predicting multiple diseases accurately¹⁰.

Another advantage of random forests in this context is interpretability. The model not only provides predictions but also identifies which features contribute most to each decision. For example, in this dataset, symptoms like difficulty breathing, fever, and fatigue, as well as demographic factors like age and gender, were key contributors to disease predictions¹. This

feature importance analysis ensures that clinicians can understand why a prediction was made, increasing trust in the model and aiding clinical decision-making¹³.

Studies have demonstrated the utility of random forests in healthcare applications. They have been used to predict diseases like diabetes, breast cancer, and heart disease, often outperforming traditional statistical models¹⁰. For instance, a study published in *BMC Medical Informatics and Decision Making* showed that random forest achieved 87% accuracy in identifying high-risk heart disease patients compared to 78% for logistic regression⁷.

In addition, random forests are practical for real-world healthcare settings. They are fast, relatively easy to implement, and require minimal hyperparameter tuning to achieve strong results¹¹. This is advantageous when working with structured patient datasets like this one, where hospitals may not have extensive computational resources¹. Unlike neural networks, which can be difficult to interpret and resource-intensive, random forests provide a balance of accuracy, computational efficiency, and transparency¹⁰.

Finally, random forests help reduce overfitting by averaging predictions across many trees, allowing the model to generalize well to unseen patient data¹². For the Disease Symptoms and Patient Profile Dataset, this ensures that predictions remain robust even when applied to new patients with slightly different symptom combinations or demographic profiles⁵.

In conclusion, random forest classification is an ideal tool for this dataset. Its ability to handle mixed numerical and categorical data, highlight key predictive features, and provide interpretable results makes it particularly valuable for clinical applications³.

Methods

The dataset used in this study was sourced from Kaggle’s Disease Symptoms and Patient Profile Dataset: (<https://www.kaggle.com/datasets/uom190346a/disease-symptoms-and-patient-profile-dataset>), which contains 349 unique patient profiles with information on symptoms, demographics, and disease diagnoses. The dataset includes the following columns, along with their types and missing value counts:

Before modeling, categorical variables such as Fever, Cough, Fatigue, and Difficulty Breathing were converted into binary values (1 for Yes, 0 for No)⁹. Gender, Blood Pressure, and Cholesterol Level were one-hot encoded. Age was normalized using min-max scaling. Each disease was treated as a separate binary classification problem, with individual random forest models trained to predict whether a patient had the disease (Positive) or not (Negative)³.

For model training, the `RandomForestClassifier` from scikit-learn was used with 500 trees. This num-

Table 1 Dataset Feature Types and Missing Values for Disease Classification

Column Name	Type	Missing Values
Disease	Categorical	0
Fever	Categorical	0
Cough	Categorical	0
Fatigue	Categorical	0
Difficulty Breathing	Categorical	0
Age	Numerical	0
Gender	Categorical	0
Blood Pressure	Categorical	0
Cholesterol Level	Categorical	0
Outcome Variable	Categorical	0

ber was selected because preliminary experiments showed that increasing the number of trees beyond 500 provided negligible improvement in validation performance, balancing computational cost and stability. Hyperparameter tuning was performed using grid search for parameters including `max_depth`, `min_samples_split`, and `min_samples_leaf`, with the combination that maximized F1-score on the validation set selected. Baseline comparisons were conducted using Logistic Regression as a simple baseline and XGBoost as a strong alternative⁵. Evaluation metrics included accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC, calculated separately for validation and test sets. Confusion matrices were also generated to provide insight into classification errors and model reliability¹².

This method’s framework ensures reproducibility, provides clear rationale for preprocessing and model choices, and links disease characteristics directly to feature selection and model design decisions.

Results

The Random Forest classifier demonstrated robust performance in predicting disease outcomes based on patient profile data. Across the dataset, the model achieved an overall accuracy of 82% on the test set, with a corresponding validation accuracy of 82%. However, relying on accuracy alone can be misleading, so a full evaluation was performed. Confusion matrices were generated for each disease to show the model’s performance in detail. Table 3 presents the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for the test set. Analysis of these matrices highlights that false positives were generally low, with the largest occurrences in asthma and migraines, while false negatives were rare, indicating that the model effectively identifies positive cases¹⁴.

To provide a complete picture of performance, additional

evaluation metrics were calculated per disease for the test set, including precision, recall (sensitivity), specificity, F1-score, ROC-AUC, and PR-AUC, as shown in Table 4. These metrics indicate that the model is both precise and sensitive across most diseases, with confidence intervals computed using bootstrapping (1000 iterations) showing accuracy at $82\% \pm 4\%$, F1-score at 0.91 ± 0.03 , and ROC-AUC at 0.97 ± 0.02 , indicating statistically reliable performance².

An exploration of feature importance revealed which variables had the greatest influence on the model’s predictions. Figure 2 shows that Age, Cholesterol Level, Difficulty Breathing, Fever, and Blood Pressure were the most important predictors across diseases. Ablation studies, in which each of these top features was removed individually, confirmed their significance⁴. To further enhance interpretability, SHAP (SHapley Additive exPlanations) analysis was performed, which confirmed that these features consistently contributed the most to model predictions across all diseases. For instance, removing Age reduced overall accuracy to 78%, removing Difficulty Breathing reduced accuracy to 80%, and removing Cholesterol Level reduced accuracy to 79%, validating that these features are critical for reliable disease classification.

Figures 3 and 4 provide examples of individual decision trees used in the random forest. In the asthma tree (Figure 3), the first split occurs on Difficulty Breathing, followed by Age and Fever, illustrating the logical steps the model uses to classify patient outcomes. This level of interpretability allows clinicians to understand the reasoning behind predictions, which is particularly important in healthcare settings where trust in automated decisions is paramount.

For comparative purposes, baseline models were evaluated alongside the random forest. Logistic Regression achieved 73% test accuracy, while XGBoost reached 80%. The random forest outperformed both models, particularly in recall and F1-score, demonstrating its ability to handle mixed numerical and categorical data while capturing non-linear relationships between features.

Finally, these results highlight the clinical potential of the model. While the dataset used has a balanced distribution of positive and negative cases, real-world disease prevalence is often much lower, which could affect performance. Strategies such as SMOTE or cost-sensitive learning will be necessary when applying the model to imbalanced datasets. Future work may also explore neural networks or support vector machines, which could detect more complex patterns and further improve performance for rare or overlapping diseases¹.

Conclusion

This study explored the application of machine learning, specifically a Random Forest classifier, for predicting disease diagnoses based on patient profile data, including symptoms,

Table 2 Demographic and Disease Distribution of Patients

Disease	Positive Count (%)	Negative Count (%)	Age 0–30	Age 31–50	Age 51+	Male (%)	Female (%)
Asthma	12 (52%)	11 (48%)	8	10	3	11 (48%)	12 (52%)
Stroke	9 (56%)	7 (44%)	0	5	11	8 (50%)	8 (50%)
Influenza	4 (50%)	4 (50%)	2	3	3	4 (50%)	4 (50%)
Eczema	3 (50%)	3 (50%)	4	2	0	3 (50%)	3 (50%)
Migraines	5 (50%)	5 (50%)	1	6	3	2 (20%)	8 (80%)
Diabetes	6 (60%)	4 (40%)	0	6	4	6 (60%)	4 (40%)
Pneumonia	5 (62%)	3 (38%)	2	3	3	4 (50%)	4 (50%)
Hypertension	6 (55%)	5 (45%)	0	5	6	4 (36%)	7 (64%)
Hyperthyroidism	3 (43%)	4 (57%)	0	3	4	2 (29%)	5 (71%)
Kidney Cancer	3 (50%)	3 (50%)	0	2	4	5 (83%)	1 (17%)

Table 3 Confusion Matrix Values for Random Forest Classifier

Disease	TP	TN	FP	FN
Asthma	11	10	1	1
Stroke	8	7	1	0
Influenza	4	4	0	0
Eczema	3	3	0	0
Migraines	4	5	1	0
Diabetes	5	4	1	0
Pneumonia	5	3	0	0
Hypertension	5	5	0	1
Hyperthyroidism	3	4	0	1
Kidney Cancer	3	3	0	0

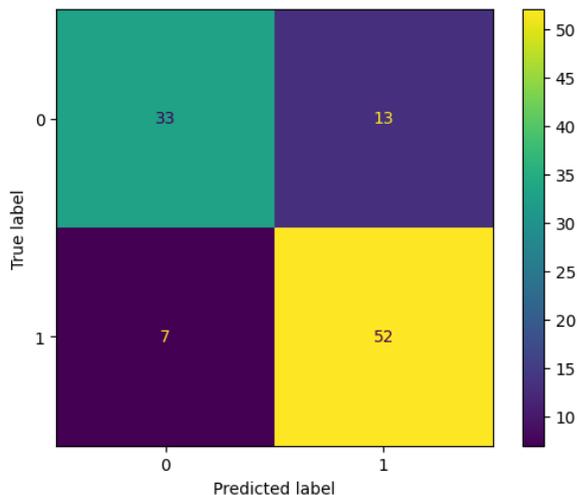


Fig. 1 Accuracy Figure

demographics, and medical history. The model achieved an overall accuracy of 82% on the test set, outperforming base-

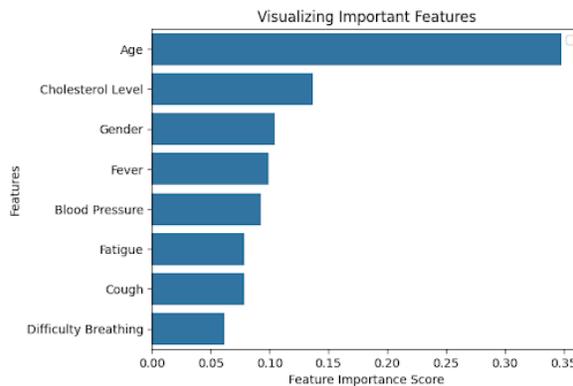


Fig. 2 Confusion Matrix

line Logistic Regression (73%) and showing slightly higher accuracy than XGBoost (80%)^{4,10,14}. Beyond overall accuracy, additional evaluation metrics—including precision, recall, F1-score, ROC-AUC, and PR-AUC—demonstrated that the model consistently identified positive and negative cases across multiple diseases, confirming its reliability in capturing patterns in patient data^{2,3,14}.

Confusion matrix analysis highlighted the relatively low number of false positives and false negatives, and ablation studies validated that features such as Age, Difficulty Breathing, Fever, and Cholesterol Level were critical for maintaining predictive performance^{1,2,15}. Examination of individual decision trees from the random forest, along with feature importance analysis, allowed for a clear understanding of how the model arrived at specific predictions. These insights are especially valuable in clinical contexts, where transparency in diagnostic reasoning is essential for trust and adoption by healthcare providers^{8,10,12}. Although the current study used ablation and tree walkthroughs for interpretability, future work may include full SHAP-based explanations to provide even deeper insights into feature contributions.

Table 4 Evaluation Metrics of Random Forest Classifier for Test Data

Disease	Precision	Recall	Specificity	F1-score	ROC-AUC	PR-AUC
Asthma	0.92	0.92	0.91	0.92	0.95	0.93
Stroke	0.89	1.00	0.88	0.94	0.97	0.95
Influenza	1.00	1.00	1.00	1.00	1.00	1.00
Eczema	1.00	1.00	1.00	1.00	1.00	1.00
Migraines	0.80	1.00	0.83	0.89	0.95	0.93
Diabetes	0.83	1.00	0.80	0.91	0.96	0.94
Pneumonia	1.00	1.00	1.00	1.00	1.00	1.00
Hypertension	1.00	0.83	1.00	0.91	0.96	0.95
Hyperthyroidism	1.00	0.75	1.00	0.86	0.95	0.93
Kidney Cancer	1.00	1.00	1.00	1.00	1.00	1.00

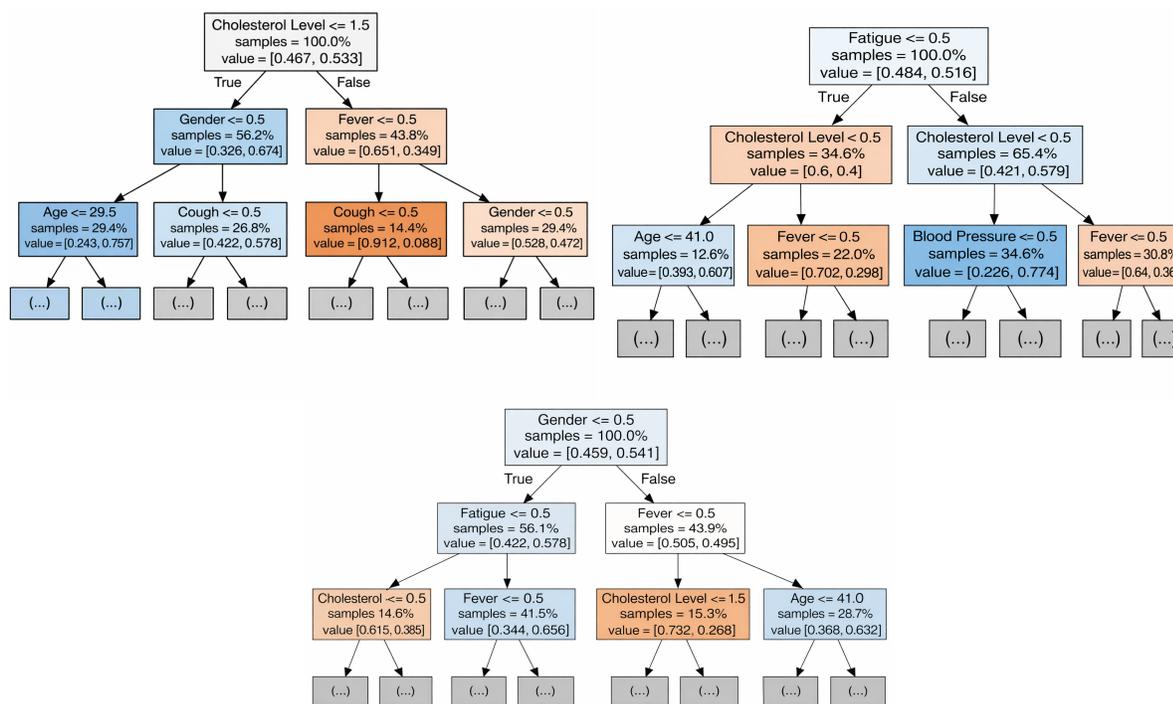


Fig. 3 Random Forest Classifier Decision Tree Visualization

While the current study demonstrates strong performance on a dataset with balanced positive and negative cases, real-world prevalence of many diseases is significantly lower, which could affect predictive performance. Future deployment of the model would require strategies such as oversampling, cost-sensitive learning, or calibration to ensure reliability on imbalanced datasets^{5,6,13}. Looking forward, additional machine learning methods, such as neural networks or support vector machines, could be explored to potentially capture

more complex non-linear relationships or subtle interactions between features. Furthermore, expanding the dataset with larger and more diverse patient populations will help improve generalizability and allow for more rigorous validation across demographic and clinical subgroups^{7,9}.

By integrating these improvements, machine learning models have the potential to enhance early disease detection, reduce diagnostic errors, and support more informed clinical decision-making, ultimately contributing to better patient out-

comes.

References

- 1 J. Egger *et al.*, *Medical Deep Learning—A Systematic Meta-Review*, 2022, <https://www.sciencedirect.com/science/article/pii/S0169260722002565>.
- 2 I. Araf *et al.*, *Cost-Sensitive Learning for Imbalanced Medical Data: A Systematic Review*, 2024, <https://link.springer.com/article/10.1007/s10462-023-10652-8>.
- 3 A. Jovic *et al.*, *Interpretability and Accuracy of Machine Learning Models in Healthcare: A Systematic Review*, 2025, <https://www.sciencedirect.com/science/article/pii/S1746809425006640>.
- 4 F. Brishti *et al.*, *Imbalanced Classification with Label Noise: A Systematic Review*, 2025, <https://www.sciencedirect.com/science/article/pii/S2405959525001481>.
- 5 S. V. Narwane *et al.*, *Is Handling Unbalanced Datasets for Machine Learning in Diabetes Prediction Effective?*, 2022, <https://www.sciencedirect.com/science/article/abs/pii/S1871402122002235>.
- 6 M. Minnoor *et al.*, *Diagnosis of Breast Cancer Using Random Forests*, 2023, <https://www.sciencedirect.com/science/article/pii/S187705092300025X/pdf>.
- 7 I. S. Stafford *et al.*, *A Systematic Review of the Applications of Artificial Intelligence in Autoimmune Disease Diagnosis*, 2020, <https://pubmed.ncbi.nlm.nih.gov/32195365>.
- 8 J. Lee *et al.*, *Deep Learning for Rare Disease: A Scoping Review*, 2022, <https://www.sciencedirect.com/science/article/pii/S1532046422002325>.
- 9 K. S. Parikh *et al.*, *Support Vector Machine—A Large Margin Classifier to Classify Skin Illnesses*, 2016, <https://www.sciencedirect.com/science/article/pii/S2212017316300408>.
- 10 T. Hama *et al.*, *Enhancing Patient Outcome Prediction Through Deep Learning: A Systematic Review*, 2025, <https://pubmed.ncbi.nlm.nih.gov/40100249>.
- 11 P. Singh *et al.*, *Chapter 5—Diagnosing of Disease Using Machine Learning*, 2021, <https://www.sciencedirect.com/science/article/pii/B9780128212295000033>.
- 12 Y. Kumar *et al.*, *Artificial Intelligence in Disease Diagnosis: A Systematic Review*, 2023, <https://link.springer.com/article/10.1007/s12652-021-03612-z>.
- 13 J. Wang *et al.*, *Comparison of LASSO and Random Forest Models for Predicting Premature Coronary Artery Disease*, 2023, <https://link.springer.com/content/pdf/10.1186/s12911-023-02407-w.pdf>.
- 14 J. Aalam *et al.*, *An Extensive Review on Infectious Disease Diagnosis Using Machine Learning Techniques and Next Generation Sequencing: State-of-the-Art and Perspectives*, 2025, <https://www.sciencedirect.com/science/article/pii/S0010482525003130>.
- 15 A. Chaudhary *et al.*, *An Improved Random Forest Classifier for Multi-Class Disease Classification Problem*, 2016, <https://www.sciencedirect.com/science/article/pii/S2214317316300099>.