Disease Prediction System

PWSkills Mini-Hackathon: Healthcare Applications

Submitted by:
Prabal Roy
prabalroy242@gmail.com
Mobile: 9679097884

September 10, 2025

Contents

Chapter 1

Project Overview

The Disease Prediction System is an end-to-end machine learning solution for disease prediction based on patient symptoms. It analyzes 132 symptoms and predicts 42 possible diseases with high accuracy.

Key Features

- Interactive Streamlit web application
- Advanced ML pipeline with hyperparameter optimization
- Multiple algorithm comparison (Random Forest, XGBoost, CatBoost, etc.)
- Feature selection using statistical methods
- Cross-validation for robust evaluation
- Real-time prediction with confidence scoring
- Comprehensive analytics and model insights
- Medical disclaimer and safety considerations

Chapter 2

Dataset Information

- Training Samples: 4,920 records
- Testing Samples: 42 records (1 per disease)
- Features: 132 symptoms (binary encoded)
- Classes: 42 diseases
- Data Quality: No missing values, balanced dataset

Diseases Covered

Examples include fungal infections, allergies, diabetes, hypertension, migraine, pneumonia, tuberculosis, hepatitis, malaria, dengue, and more.

Chapter 3

Project Structure

```
disease-prediction-system/
 app.py                  # Main application
 requirements.txt        # Dependencies
 predictions.csv         # Model predictions
 Notebooks/              # Development notebooks
 src/                    # Source code modules
 data/                   # Raw and processed data
 models/                 # Trained models
 static/                 # Static files
 templates/              # HTML templates
```

Chapter 4

Development Workflow

1. Data preparation and preprocessing
2. Exploratory Data Analysis (EDA)
3. Model training and evaluation
4. Web application deployment

Chapter 5

Machine Learning Pipeline

Model Selection Process

1. Baseline models: Random Forest, XGBoost, CatBoost, SVM, Naive Bayes
2. Feature selection using Chi-square test
3. Hyperparameter optimization with Optuna
4. Ensemble methods: Voting classifier
5. Cross-validation with 5-fold stratification

Performance Metrics

- Test Accuracy: >95%
- Cross-validation: Consistent across folds
- Precision/Recall: Balanced across diseases
- F1-score: High weighted average

Model Comparison Results

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.952	0.951	0.952	0.951
XGBoost	0.948	0.947	0.948	0.947
CatBoost	0.945	0.944	0.945	0.944
Ensemble	0.957	0.956	0.957	0.956

Chapter 6

Web Application Features

Main Interface

- Symptom selection with multi-select dropdown
- Real-time prediction with confidence scoring
- Alternative diagnoses (Top 5 possibilities)

Analytics Dashboard

- Prediction history
- Confidence distribution analysis
- Disease frequency statistics
- Interactive visualizations

Chapter 7

Usage Instructions

Prediction Workflow

1. Select symptoms
2. Configure confidence threshold
3. Generate prediction and review results

Confidence Interpretation

- High (>80%): Strong prediction, consult doctor
- Medium (50–80%): Possible condition, consultation recommended
- Low (<50%): Uncertain prediction, professional diagnosis required

Chapter 8

Deployment Options

- Local development using Streamlit
- Cloud deployment (Render, Heroku, etc.)
- Docker deployment with lightweight image

Chapter 9

Future Enhancements

- Multi-language support
- Voice input for symptoms
- Integration with medical databases
- Real-time model updating
- Mobile-responsive design improvements

Chapter 10

Medical Disclaimer

This system is for educational and research purposes only.
It is not intended for medical use, diagnosis, or treatment. Always consult healthcare professionals for medical advice and emergencies.

Chapter 11

Hackathon Compliance

- Machine learning model with 132 symptoms and 42 diseases
- Multiple algorithm comparison
- Hyperparameter optimization
- Cross-validation evaluation
- Streamlit web application
- Complete documentation

Chapter 12

Acknowledgments

- PWSkills for organizing the hackathon
- Healthcare community for inspiration
- Open-source libraries and ML community