REPORT

Disease Risk Prediction Using Machine Learning and Deep Learning

MSc Computer Science and Data Science

Advanced Deep Learning

(MCSCIN5A1625)

Submitted by:

Gurpreetkaur Jaykumar MODI

1. Project Overview

- Early identification of individuals at high risk of developing lifestyle-related diseases is essential for preventive healthcare and informed decision-making. Advances in machine learning and deep learning have enabled the analysis of large-scale health and lifestyle datasets to uncover complex patterns that are difficult to capture using traditional statistical approaches.
- This project focuses on developing a **Disease Risk Prediction System** using structured demographic, lifestyle, and physiological health data. The system applies a complete data science pipeline, including data preprocessing, exploratory data analysis, feature engineering, feature selection, model training, hyperparameter tuning, and explainability analysis.
- Multiple classical machine learning models—Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Support Vector Machine (SVM)—are evaluated and compared alongside a simple Artificial Neural Network (ANN). To ensure transparency and trustworthiness, Explainable AI (XAI) techniques such as SHAP are incorporated to interpret model predictions.
- The project emphasizes not only predictive performance but also **model interpretability, robustness, and responsible use**, making it suitable for healthcare-related risk assessment applications.

2. Problem Statement

- The increasing prevalence of lifestyle-related diseases highlights the need for early risk assessment tools that can assist in preventive healthcare. Traditional diagnostic methods often require clinical intervention and may not capture the combined influence of demographic, lifestyle, and physiological factors.

- This project aims to develop a **Disease Risk Prediction System** that classifies individuals into **low-risk or high-risk categories** using machine learning and deep learning techniques. The problem is formulated as a **binary classification task**, where the objective is to predict disease risk based on structured health and lifestyle data. The system is designed to support decision-making and risk awareness rather than provide a medical diagnosis.

3. Dataset Description

- The dataset consists of 100,000 records representing individuals with demographic, lifestyle, and physiological health attributes.

- **Key Characteristics:**
    - Total features: 16 (before feature engineering)
    - Target variable: disease_risk (binary)
    - No missing values
    - Well-structured tabular dataset

- **Feature Categories:**
    - Demographic: Age, Gender
    - Lifestyle: Daily steps, Sleep hours, Water intake, Calories consumed, Smoking, Alcohol consumption
    - Physiological: BMI, Blood pressure, Cholesterol, Resting heart rate
    - Medical History: Family history of disease

- The dataset is suitable for supervised learning and comparative evaluation of multiple ML and DL models.

4. Exploratory Data Analysis (EDA)

- Exploratory Data Analysis was conducted to understand the structure, distribution, and relationships within the data.

- **Key Observations:**
    - The target variable is moderately imbalanced (≈ 75% low risk, 25% high risk).
    - Numerical features show varying distributions with visible outliers, especially in BMI, blood pressure, and cholesterol.
    - Gender distribution is balanced.
    - Correlation analysis revealed no severe multicollinearity, making the dataset suitable for both linear and non-linear models.

- EDA helped validate data quality and guided subsequent preprocessing and feature engineering decisions.

5. Project Pipeline

5.1 Preprocessing

- To ensure consistent and fair evaluation across models, the following preprocessing steps were applied:
    - Train–test split with stratification to preserve class balance
    - Feature standardization to improve convergence for distance-based and neural models
    - Use of reproducible preprocessing pipelines

- This ensured that all models were trained and evaluated under identical conditions**.**

5.2 Feature Engineering

- Feature engineering was performed to capture meaningful health relationships and improve predictive performance. Engineered features include:
    - Derived cardiovascular indicators from blood pressure values
    - Combined lifestyle risk indicators incorporating physical activity and sleep patterns
    - Aggregated health metrics reflecting overall physiological

condition

- These features enhanced both predictive power and interpretability.

5.3 Feature Selection

- Feature relevance was assessed using:
    - Correlation analysis to detect redundant features
    - Random Forest feature importance to rank predictors based on contribution

- Only informative features were retained, reducing noise and improving generalization.

5.4 Machine Learning Models Evaluated

- Multiple machine learning models were implemented to compare performance across different learning paradigms:
    - Logistic Regression
    - K-Nearest Neighbors (KNN)
    - Decision Tree
    - Random Forest
    - Support Vector Machine (SVM)

- These models represent linear, distance-based, tree-based, and margin-based approaches

5.5 Hyperparameter Tuning

- To improve performance and reduce overfitting:
    - RandomizedSearchCV was applied to tune Random Forest hyperparameters
    - Parameters such as number of estimators, maximum depth, and split criteria were optimized
    - Model performance before and after tuning was compared

- Tuning resulted in improved stability and generalization for tree-based models.

5.6 Explainability (XAI)

- Explainable AI techniques were applied to ensure transparency:
    - Random Forest feature importance provided a global understanding of key predictors
    - SHAP (SHapley Additive exPlanations) was used to analyze feature impact at both global and local levels

- SHAP analysis highlighted that lifestyle and physiological factors such as BMI, physical activity, and blood pressure play a significant role in disease risk prediction.

5.7 Deep Learning Model

- A simple **Artificial Neural Network (ANN)** was developed to evaluate deep learning performance on tabular data:
    - Fully connected dense layers with ReLU activation
    - Dropout regularization to mitigate overfitting
    - Sigmoid output layer for binary classification
    - Binary cross-entropy loss and Adam optimizer

- Loss curve analysis showed stable training with minimal overfitting. While the ANN achieved competitive performance, it did not outperform the best classical machine learning models.

6. Model Performance Comparison

- The following table summarizes the final performance of all evaluated models, as obtained from the notebook:

Model	Accuracy	ROC-AUC
Logistic Regression	0.7518	0.5011
KNN	0.6978	0.4931
Decision Tree	0.6152	0.5001
Random Forest	0.7517	0.5044
Linear SVM	0.7518	—
ANN	0.7518	—

7. Key Insights

- Tree-based models, particularly Random Forest, perform well on structured health data
- Feature engineering significantly improves model performance
- ANN achieved comparable accuracy but did not surpass classical ML models
- SHAP explainability revealed lifestyle and physiological indicators as dominant risk factors
- Interpretability is crucial in healthcare-related applications

8. Limitations

- The model estimates disease risk and does not provide medical diagnosis
- External factors such as genetics, stress, and environmental conditions are not included
- The dataset is static and does not capture temporal health changes
- Deep learning models remain less interpretable compared to tree-based models

9. Future Work

- Incorporate longitudinal and time-series health data
- Integrate genetic and environmental features
- Explore ensemble and hybrid deep learning models
- Deploy the system as a clinical decision-support tool

10. Conclusion

- This project successfully developed a Disease Risk Prediction System using machine learning and deep learning techniques. A complete and well-structured pipeline was implemented, covering preprocessing, feature engineering, feature selection, model training, hyperparameter tuning, and explainability analysis.

- Experimental results demonstrate that Random Forest provides the best balance between accuracy and interpretability, while the ANN highlights the limitations of deep learning on structured tabular data. The inclusion of SHAP-based explainability enhances transparency and trust, making the system suitable for healthcare-related risk assessment and decision support.

11. Project Structure

- The project is organized in a clear and minimal structure to ensure ease of understanding, reproducibility, and accessibility of all components.

```
├── Data/
│   └── health_lifestyle_dataset.csv
├── Disease_Risk_Prediction_ML_DL.ipynb
├── README.md
├── Disease_Risk_Prediction_Report.pdf
```

- **Description:**
  - **Data/**: Contains the dataset used for training and evaluating the machine learning and deep learning models.
  - **health_lifestyle_dataset.csv**: Raw dataset consisting of demographic, lifestyle, and physiological health features.
  - **Deep_Learning_Project(preeti).ipynb**: Jupyter Notebook containing the complete implementation of the project pipeline, including EDA, preprocessing, feature engineering, model training, hyperparameter tuning, explainability, and ANN development.
  - **README.md**: Provides an overview of the project,

dataset description, and instructions to run the notebook.

- **Disease_Risk_Prediction_Report.pdf**: Final project report documenting methodology, results, analysis, limitations, and conclusions.