Predicting Dengue Using Blood Test Results

Atkiya Maisha
2022-3-60-110
Dept. of Computer Science and Engineering
East West University
Dhaka, Bangladesh

Mohua Akter
2022-3-60-112
Dept. of Computer Science and Engineering
East West University
Dhaka, Bangladesh

Asiful Alam Sami
2022-3-60-230
Dept. of Computer Science and Engineering
East West University
Dhaka, Bangladesh

Abstract

Dengue fever is a mosquito-borne illness common in tropical and subtropical regions. Early detection helps improve treatment and reduce complications. This study uses routine blood test results and machine learning models—including Random Forest, Decision Tree, SVM, Logistic Regression, and XGBoost to predict dengue infection. The data was cleaned, scaled, and analyzed, and models were evaluated using accuracy, precision, recall, F1-score, and ROC AUC. XGBoost, and Decision performed especially well. These results show that combining simple blood tests with machine learning can offer a fast and cost-effective way to detect dengue early, especially in resource-limited settings.

1. Introduction

Dengue is a fast-spreading viral infection carried by mosquitoes which causes serious health and economic problems in many tropical and subtropical areas and is the most common mosquito-borne virus in the world {1]. It is caused by the dengue virus (DENV), which is transmitted primarily by the Aedes aegypti mosquito. Nearly 50% of the global population is currently at risk of contracting dengue, with an estimated 100 to 400 million cases reported annually [2]. The number of dengue infections has grown a lot over the last 50 years, causing a big impact on people's health worldwide[3]. Dengue research is growing due to disease awareness and vaccine hopes, but better treatment, virus control, and understanding its spread are still needed to fight it effectively[4]. Deepening our

knowledge of dengue can enhance the treatment of individual patients and strengthen efforts to control the disease [5].

Recent advancements in machine learning (ML) offer potential pathways for disease prediction by analyzing complex patterns in clinical data. ML algorithms can be trained to detect subtle variations in hematological parameters, enabling effective dengue detection. This research aims to explore and compare various ML models—including Decision Tree, Random Forest, Support Vector Machine (SVM), Logistic Regression, XGBoost, and Neural Networks—to predict dengue from standard blood test results and evaluate their diagnostic performance.

2. Data Description

The dataset used in this study consists of blood test results collected from multiple hospitals which is available on Kaggle (https://www.kaggle.com/datasets/owmexiaolin/dengue-prediction/data). It contains 499 samples (rows) with 21 features (columns), including clinical and hematological parameters of patients suspected of dengue infection, along with their confirmed dengue status (Positive or Negative). The dataset includes the following features:

- **Age:** Patient's age in years.
- **HB (gm/dl):** Hemoglobin concentration in grams per deciliter.
- **ESR (mm):** Erythrocyte Sedimentation Rate in millimeters.
- **WBC (TC) (/cumm):** White Blood Cell Total Count per cubic millimeter.
- **Neutrophils (%):** Percentage of neutrophils in the white blood cell count.
- **Lymphocytes (%):** Percentage of lymphocytes.
- **Monocytes (%):** Percentage of monocytes.
- **Eosinophils (%):** Percentage of eosinophils.
- **Circulating Eosinophils (/cumm):** Absolute count of circulating eosinophils.
- **RBC (m/ul):** Red Blood Cell count in millions per microliter.
- **HTC/PCV (%):** Hematocrit or Packed Cell Volume percentage.
- **MCV (fl):** Mean Corpuscular Volume in femtoliters.
- **MCH (pg):** Mean Corpuscular Hemoglobin in picograms.
- **MCHC (g/dl):** Mean Corpuscular Hemoglobin Concentration in grams per deciliter.
- **RDW (%):** Red Cell Distribution Width percentage.
- **PDW (fl):** Platelet Distribution Width in femtoliters.
- **Platelet (PC) (/cumm):** Platelet count per cubic millimeter.

	0	1	2	3	4	5	6	7	8	9
Age	14.0	26.0	42.0	35.0	8.0	17.0	32.0	66.0	19.0	70.0
HB (gm/dl)	11.4	12.2	14.9	13.9	10.0	13.2	14.3	14.0	12.6	13.2
ESR(mm)	28.0	43.0	103.0	12.0	28.0	22.0	14.0	32.0	47.0	16.0
WBC(TC) (/cumm)	4200.0	4700.0	19800.0	9700.0	10600.0	8700.0	4900.0	5000.0	6800.0	12500.0
Neutrophils (%)	57	77	82	88	46	54	82	68	68	51
lymphocytes (%)	36	19	15	10	42	27	12	26	26	38
Monocytes (%)	5	3	2	1	2	3	4	3	4	2
Eosinophils (%)	2	1	1	1	10	16	2	3	2	9
Cir Eosinophils (/cumm)	84	47	198	97	1060	1392	98	150	136	1125
RBC (m/ul)	4.64	4.24	5.31	5.49	3.99	4.82	6.19	5.74	5.25	4.74
HTC/PCV (%)	37.0	39.7	47.4	46.3	33.0	43.5	47.8	46.2	41.7	43.2
MCV (fl)	79.7	93.6	89.3	84.3	82.7	90.2	77.2	80.5	79.4	91.1
MCH (pg)	24.6	28.8	28.1	25.3	25.1	27.4	23.1	24.4	24.0	27.8
MCHC (g/dl)	30.8	30.8	30.8	30.8	30.8	30.8	30.8	30.8	30.8	30.8
RDW (%)	15.5	14.3	13.7	14.1	13.7	13.4	14.6	14.1	13.7	14.0
PDW (fl)	20.1	12.9	12.7	12.7	11.5	13.1	11.0	14.7	12.4	10.1
Platelete(PC)(/cumm)	167000	277000	215000	282000	337000	185000	82000	137000	207000	380000
MPV (fl)	10.3	8.5	9.1	8.7	7.6	10.4	10.2	9.1	7.9	7.5
PCT (%)	0.07	0.24	0.2	0.25	0.26	0.19	0.08	0.13	0.16	0.28
Class Identification	Negative	Positive	Negative	Negative	Negative	Negative	Positive	Negative	Negative	Negative

Figure: Partial Preview of the Dataset

- **MPV (fl):** Mean Platelet Volume in femtoliters.
- **PCT (%):** Plateletcrit percentage.
- **Class Identification:** Target variable indicating dengue infection status (Positive or Negative).

The dataset provides a comprehensive set of hematological markers that can be leveraged to predict dengue infection. It contains both continuous numeric variables (e.g., blood counts, percentages) and a categorical target variable indicating the presence or absence of dengue.

3. Methodology

The dataset used in this study was obtained from Kaggle (https://www.kaggle.com/datasets/owmexiaolin/dengue-prediction/data) and comprises blood test results collected from multiple hospitals. This dataset includes various hematological parameters such as white blood cell count, platelet count, hematocrit, and

others, along with labels indicating dengue infection status.

3.1. Data Cleaning

Any missing data in the dataset was removed to ensure only complete records were used. Outliers were handled by replacing values below the lower bound with the lower bound, and values above the upper bound with the upper bound.

3.2. Data Preprocessing

The numerical features were standardized using a technique called StandardScaler, which adjusts the data so that each feature has a mean of zero and a standard deviation of one. This helps the models learn better.

The target variable, **Class Identification** (which tells if the case is dengue positive or negative), was converted from text labels into numbers using LabelEncoder. This makes it easier for the models to process the data.

3.3. Feature Engineering

To identify the most relevant features for predicting dengue infection, we calculated the correlation of each numerical feature with the target variable, **Class Identification**. We selected features that showed a strong relationship with the target, defined as having an absolute correlation value greater than 0.1 (either positive or negative). The 16 selected features include important blood test indicators such as:
MCHC (g/dl), Monocytes (%), MPV (fl), HB (gm/dl), PDW (fl), MCH (pg), Lymphocytes (%), RBC (m/ul), HTC/PCV (%), RDW (%), Neutrophils (%), ESR (mm), Cir Eosinophils (/cumm), WBC (TC) (/cumm). PCT (%), Platelete (PC) (/cumm). These features were chosen because their correlation with the dengue classification indicates they have significant predictive power for distinguishing between positive and negative cases.

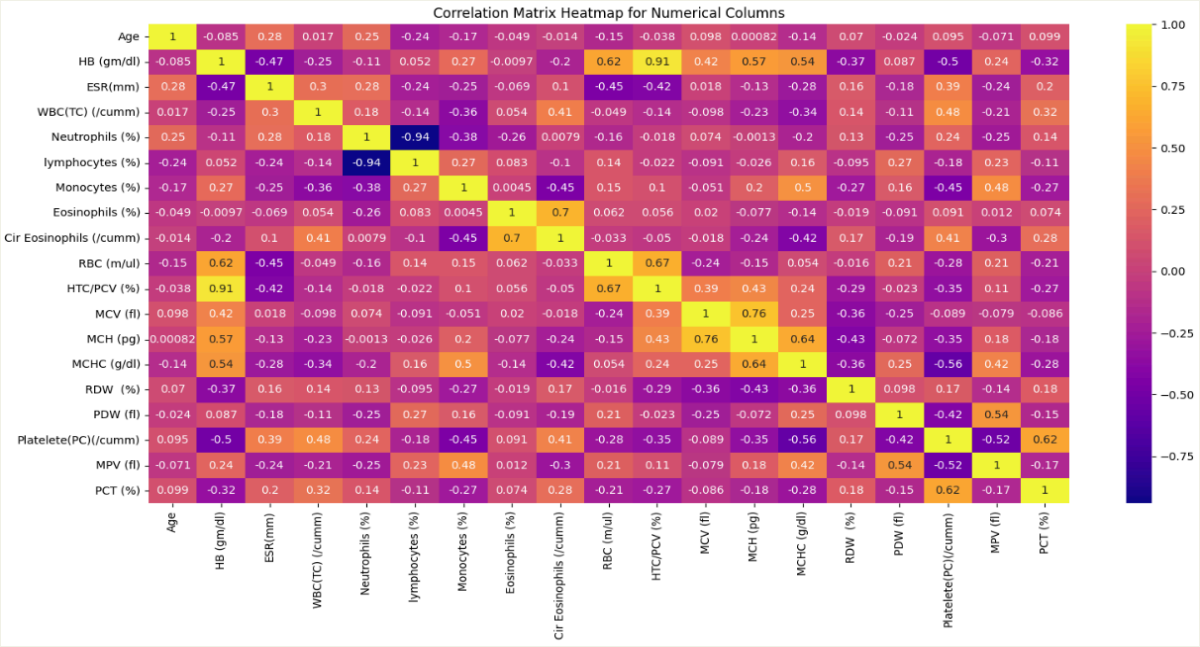


Figure: Correlation Matrix of the dataset

3.4. Train-Test Split

We used 80% of the data to train the model so it can learn patterns, and the remaining 20% to test how well the model works on new, unseen data.

3.5. Model Selection

We chose classification models to predict dengue status, including Logistic Regression, Random Forest, Support Vector Machine (SVM), Neural Network, and XGBoost. These models are good for distinguishing between two classes (positive or negative) and offer simplicity, accuracy, and advanced learning capabilities.

3.6. Model Training

We trained six different models — Random Forest, Logistic Regression, SVM, Decision Tree, XGBoost, and Neural Network — using the training data. Each model was combined with a scaler to standardize the features for better performance. After training, we tested the models on unseen data to evaluate their predictions. We also used cross-validation on the training set to check how well each model might perform on new data by measuring accuracy and F1 score. This process helped us compare which models work best for predicting dengue cases.

4. Result and Discussion

The evaluation results show that the **XGBoost** model performed the best overall, achieving the highest accuracy (0.99), F1-score (0.96), and ROC AUC (0.998), indicating excellent prediction capability. The **Decision Tree** model also showed strong performance with an

accuracy of 0.98 and a perfect recall score of 1.0, meaning it successfully identified all dengue-positive cases. Both the **Neural Network** and **Random Forest** models achieved high accuracy (0.97) and F1-scores above 0.90, suggesting a good balance between precision and recall. Meanwhile, **Logistic Regression** and **Support Vector Machine (SVM)** demonstrated slightly lower accuracy (0.95) and F1-scores around 0.83, but still maintained high ROC AUC values, reflecting reliable performance. Overall, all models showed promising results, with **XGBoost, Decision Tree**, and **Neural Network** emerging as the most effective for dengue prediction using blood test data.

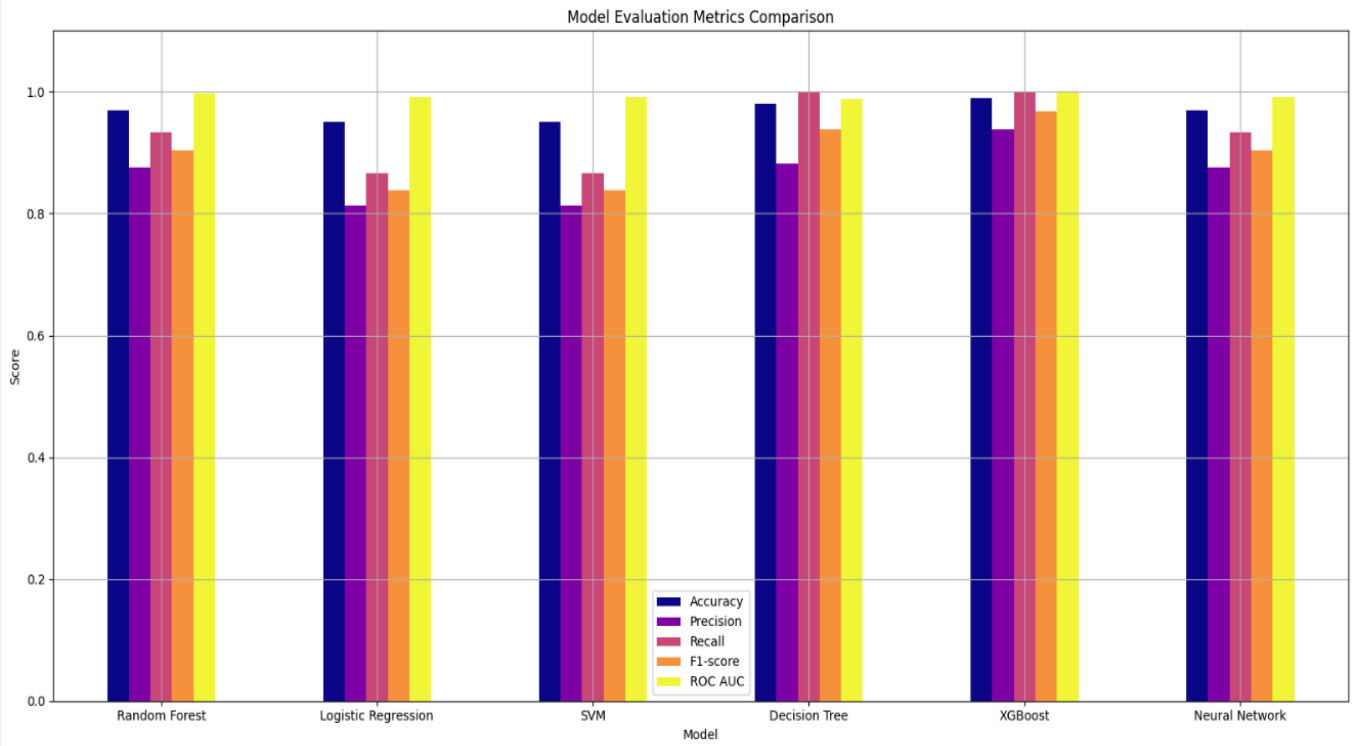


Figure: Comparison the Models

```
=== Model Evaluation Results on Test Set ===
                Model  Accuracy  Precision    Recall  F1-score   ROC AUC
0             XGBoost      0.99   0.937500  1.000000  0.967742  0.998431
1       Decision Tree      0.98   0.882353  1.000000  0.937500  0.988235
2      Neural Network      0.97   0.875000  0.933333  0.903226  0.990588
3       Random Forest      0.97   0.875000  0.933333  0.903226  0.997647
4  Logistic Regression      0.95   0.812500  0.866667  0.838710  0.990588
5                  SVM      0.95   0.812500  0.866667  0.838710  0.991373
```

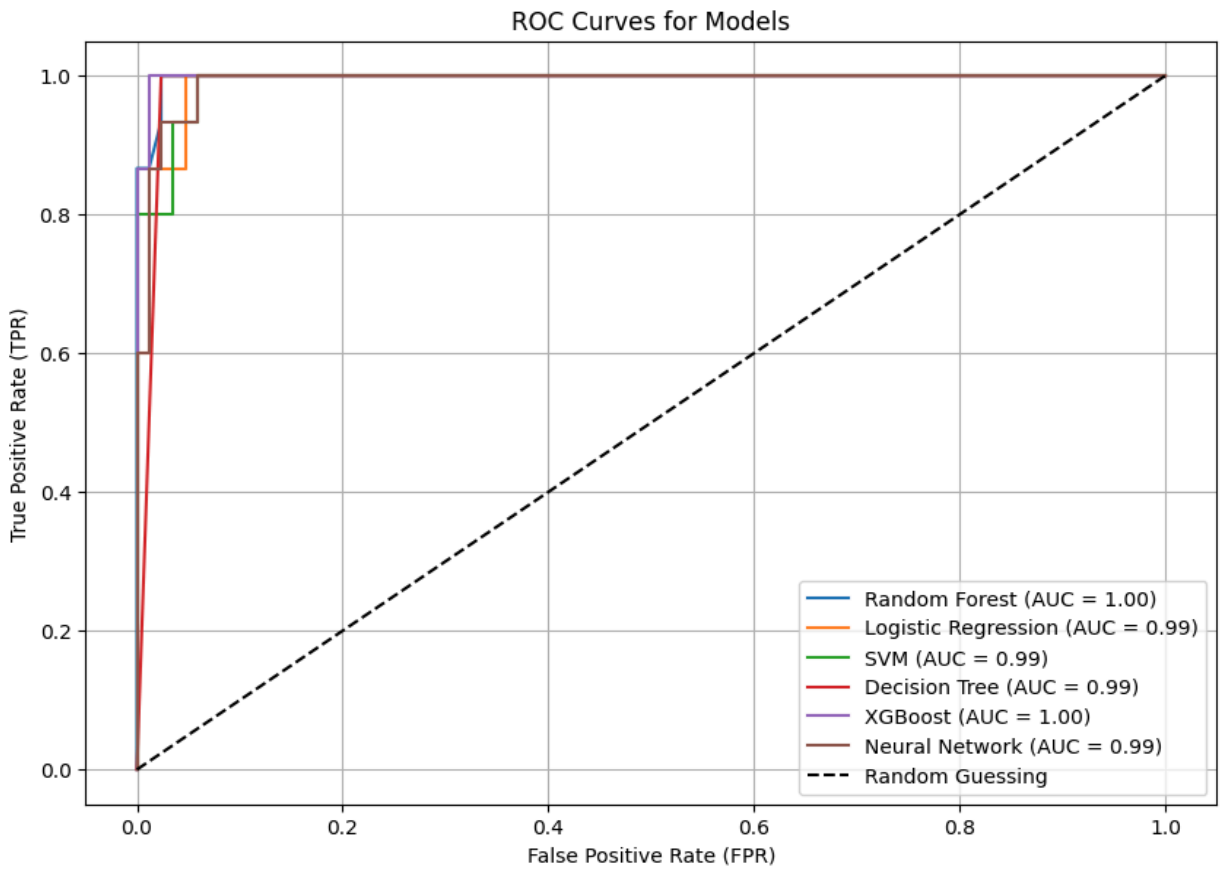
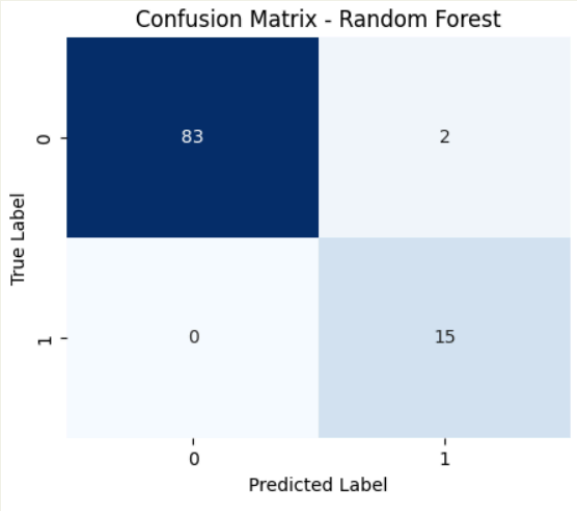
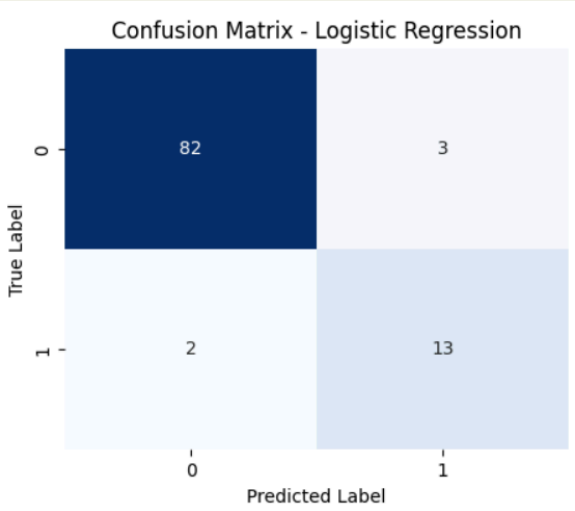
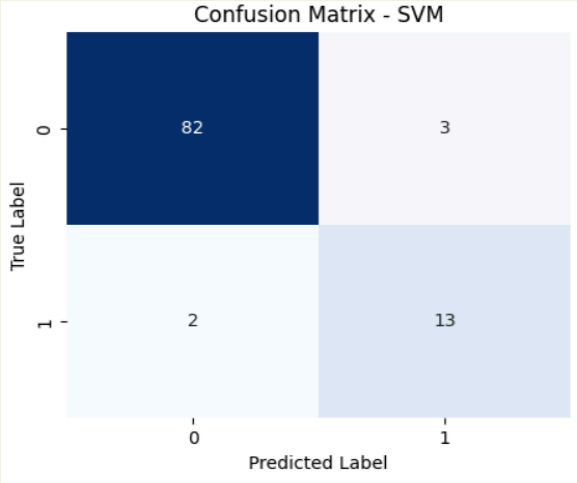


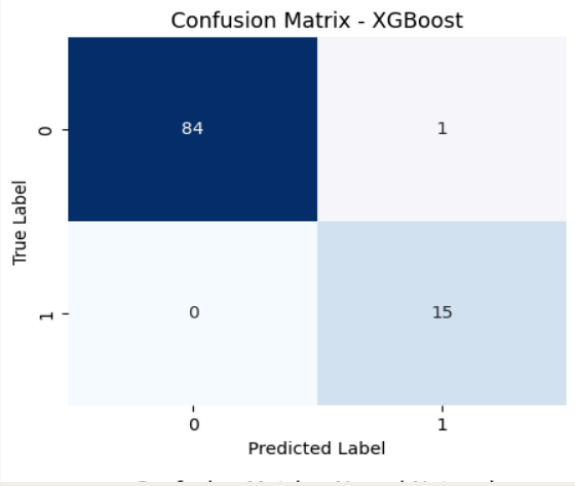
Figure: ROC Curve for the Models

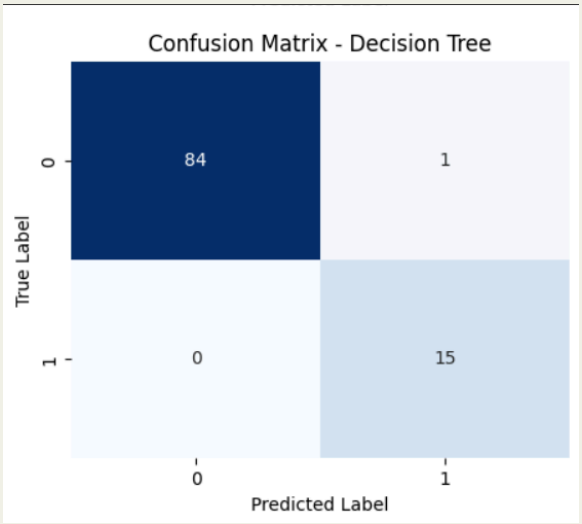
Confusion Matrix for Each Model

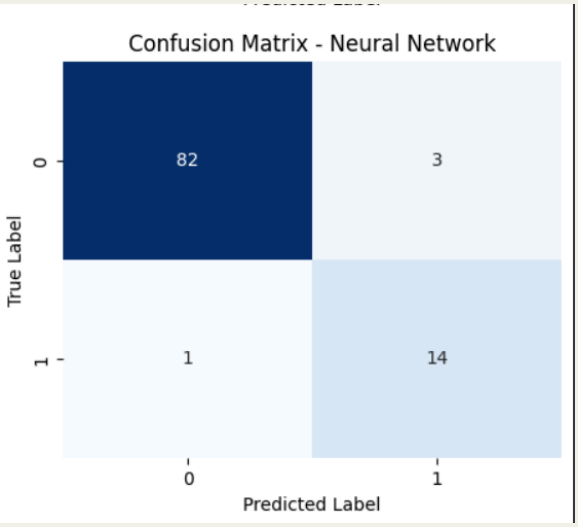












5. Limitations

This study has several limitations. Firstly, the dataset used was relatively small, containing only 499 records, which can limit the generalizability of the machine learning models. A small dataset increases the risk of overfitting, where models perform very well on training data but may not generalize effectively to new, unseen data. Additionally, the data was sourced from a limited number of hospitals, potentially introducing bias and reducing the diversity of the patient population.

The features used were restricted to routine blood test results, which, while useful, may not capture the full clinical picture of dengue. Other factors such as patient symptoms, travel history, or environmental conditions were not included but could enhance prediction accuracy. Moreover, some models may require further hyperparameter tuning and evaluation on larger, more varied datasets to confirm their real-world applicability.

6. Conclusion

This study shows that machine learning can help predict dengue using simple blood test results. Among all the models, XGBoost and Decision Tree gave the best results. Before training, we cleaned the data, chose useful features, and scaled the numbers to improve accuracy. Even though the dataset was small, the models still performed well. With more data in the future, this method could become a helpful and low-cost tool for early dengue detection, especially in hospitals with limited resources.

7. References

1. Wilder-Smith, A., Ooi, E. E., Horstick, O., & Wills, B. (2019). Dengue. The Lancet, 393(10169), 350-363. https://www.jvsmedicscorner.com/Medicine_files/Dengue%20Review%202019.pdf
2. https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue
3. Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q., ... & Ma, W. (2017). Developing a dengue forecast model using machine learning: A case study in China. PLoS neglected tropical diseases, 11(10), e0005973. https://journals.plos.org/plosntds/article/file?id=10.1371/journal.pntd.0005973&type=printable
4. Simmons, C. P., Farrar, J. J., van Vinh Chau, N., & Wills, B. (2012). Dengue. New England Journal of Medicine, 366(15), 1423-1432. https://www.nejm.org/doi/pdf/10.1056/NEJMra1110265
5. Whitehorn, J., & Farrar, J. (2010). Dengue. British medical bulletin, 95(1), 161-173. https://researchonline.lshtm.ac.uk/id/eprint/19251/1/Dengue_BritMedBull2010.pdf


THE
END