IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)
e-ISSN: 2279-0853, p-ISSN: 2279-0861.Volume 24, Issue 8 Ser. 2 (August. 2025), PP 07-13
www.iosrjournals.org

Evaluating The Performance Of An Artificial Intelligence Model In Predicting Dengue Fever Severity: A Comparative Analysis With Actual Clinical Outcomes

Dr K V Harish*, Dr Hemendra Dange, Dr Dhanvanth DS, Dr S K Pathak, Dr Priyadarshini Behera, Dr Zubair Akhthar, Dr Supriya Rawat

Abstract

Background:

Dengue fever is a mosquito-borne viral illness posing a significant global health threat, with manifestations ranging from mild febrile illness to life-threatening severe dengue. Early and accurate prediction of dengue severity is crucial for timely clinical intervention, resource allocation, and reducing mortality. Artificial intelligence (AI) models are increasingly being explored for their potential in enhancing diagnostic and prognostic capabilities in healthcare. Large language models (LLMs) are artificial intelligence (AI) tools specifically trained to process and generate text. LLMs attracted substantial public attention after OpenAI's ChatGPT was made publicly available in November 2022. LLMs have the capacity to identify critical parameters of a data set and formulate predictive indicators or scores. This function of LLMs can be exploited in the field of Medicine in identifying prognostic markers to predict future outcomes. This study was undertaken to evaluate the performance of a specific AI-based prediction model, Chat GPT 4-o with incorporation of prompt engineering by OpenMedLM strategy. The findings aim to provide insights into the model's current utility and highlight areas for potential improvement.

Materials and methods:

The study utilized an anonymized dataset corresponding to patients diagnosed with dengue fever in a single tertiary centre. The data was recorded in excel sheet and comprised records for 145 unique patients, for each patient the key variables of interest for this study were actual outcome (the clinically determined final dengue severity) and predicted Severity (the severity category predicted by assessment tool generated by Chat GPT-4o).

Results:

The overall accuracy of the AI model in predicting dengue severity across all categories was ***67.59%*** *(98 correct predictions out of 145 cases).*

Conclusion:

The evaluated Artificial Intelligence model demonstrated a moderate overall accuracy (67.59%) for predicting dengue fever severity. It showed high sensitivity in identifying Mild Dengue cases and has a very good negative predictive value for mild dengue cases. However, its performance was substantially deficient in accurately identifying Dengue with Warning Signs and, most critically, Severe Dengue, exhibiting very low recall for the latter.

Keywords: *Dengue Severity, Artificial Intelligence (AI), Machine Learning (ML), Diagnostic Accuracy, Predictive Modelling, large language model (LLM).*

Date of Submission: 01-08-2025 Date of Acceptance: 11-08-2025

I. Introduction

Dengue fever, a viral infection transmitted primarily by Aedes aegypti and Aedes albopictus mosquitoes, has emerged as one of the most rapidly spreading mosquito-borne diseases globally, particularly in south India and nearly endemic in coastal states [1]. The clinical spectrum of dengue infection is broad, ranging from an undifferentiated febrile illness (Dengue Fever - DF) to more severe forms, including dengue with warning signs (DWS) and severe dengue (SD) [2]. Severe dengue is characterized by plasma leakage, severe bleeding, organ impairment (Hepatitis, myocarditis, encephalitis), and can lead to shock (Dengue Shock Syndrome - DSS) and death if not managed appropriately and promptly [3]. The unpredictable progression from mild illness to severe disease makes early risk stratification a critical challenge for clinicians. Accurate and timely identification of patients likely to develop severe dengue allows for close vigilance and better fluid management, thereby reducing morbidity and mortality [4]. Current approaches to predict dengue severity rely on a combination of clinical symptoms, warning signs (e.g., abdominal pain, persistent vomiting, fluid accumulation, mucosal bleeding, lethargy, liver enlargement), and laboratory markers (e.g., change in hematocrit along with a rapid decline in platelet count and transaminitis). While WHO guidelines provide a framework for classification and management, the dynamic nature of the illness and the often-nonspecific early symptoms can make prognostication difficult, especially in resource-limited settings [5].

In recent years, artificial intelligence (AI) and machine learning (ML) have shown considerable promise in revolutionizing various aspects of healthcare, including disease diagnosis, prognosis, and treatment optimization [6]. In the context of dengue, AI models have been developed to predict outbreaks, aid in diagnosis [7]. Prognostication of disease severity using clinical and laboratory parameters [8] has been tried earlier at the time of diagnosis to predict the clinical outcomes, however using ML or AI for predicting the disease outcomes at the time of diagnosis by formulating as assessment scale has not been tried earlier. The potential benefits of a reliable AI tool for dengue severity prediction are immense, offering the possibility of an objective, rapid, and consistent assessment that could support clinical decision-making, particularly for less experienced healthcare providers or in overwhelmed healthcare facilities. Large language models (LLMs) are artificial intelligence (AI) tools specifically trained to process and generate text. Chat Generative pre-trained transformer (GPT), a LLM can be used in the field of medicine to improve patient care, analyse medical data, formulate assessment tools to predict outcomes through prompt engineering [9]. The usage of OpenMedLM strategy for generating medical prompts can show good results in various studies [11]. This study was undertaken to evaluate the performance of a specific AI-based prediction model, Chat GPT 4-o with incorporation of prompt engineering by OpenMedLM strategy. The findings aim to provide insights into the model's current utility and highlight areas for potential improvement.

II. Aims And Objectives

The primary aim of this study was to evaluate and compare the performance of an Artificial Intelligence (AI) based LLM prediction model generated clinical prognostic outcomes in dengue against actual, clinically determined outcomes in predicting the severity of dengue fever.

The specific objectives were:

I. To determine the overall accuracy of the AI model in predicting dengue severity categories (Mild Dengue, Dengue with Warning Signs, Severe Dengue).

II. To calculate and assess the precision, recall (sensitivity), and F1-score (harmonic mean of precision and recall) of the AI model for each of the defined dengue severity categories.

III. To analyse the patterns of misclassification by the AI model by constructing and examining a confusion matrix comparing AI predictions with actual clinical outcomes.

IV. To discuss the potential clinical implications of the AI model's performance, particularly concerning its ability to identify high-risk patients.

III. Material And Methods

A retrospective comparative analysis design was employed. The study utilized an anonymized dataset corresponding to patients diagnosed with dengue fever in a single tertiary centre. The data was recorded in excel sheet and comprised records for 145 unique patients, for each patient the key variables of interest for this study were actual outcome (the clinically determined final dengue severity) and predicted Severity (the severity category predicted by assessment tool generated by Chat GPT-4o using prompt engineering). Chat GPT 4-o, through OpenMedLM strategy of prompt engineering using parameters of Abdominal pain, Bleeding manifestations, Hepatomegaly, Platelet count, Total leucocyte count (TLC), Liver enzymes and Lactate dehydrogenase levels developed an assessment tool with specific scoring to classify dengue cases into different categories. The scoring interpretation using clinical and lab parameters generated by Chat GPT 4-o is as follows:

Total Score	Risk Category	
0 - 5	Mild Dengue or Dengue without warning signs	
6 - 10	Dengue with warning signs	
>11	Severe Dengue	

Outcome Measures and Definitions

- Actual Outcome (Gold Standard): This was defined as the final clinical diagnosis of dengue severity, presumably made by attending physicians based on established diagnostic criteria, such as the WHO 2009 guidelines. The categories for actual outcome were:

○ Mild Dengue (often referred to as Dengue Fever without warning signs)

○ Dengue with Warning Signs (DWS)

○ Severe Dengue (SD)

- AI Predicted Outcome: This was the dengue severity category ('Mild Dengue', 'Dengue with Warning Signs', or 'Severe Dengue') generated by the AI model using assessment model generated by Chat GPT 4-o for each patient. This study was based on the analysis of a pre-existing dataset provided for analytical purposes. All guidelines as per Declaration of Helsinki and good clinical practice guidelines were followed. No direct patient contact or intervention was performed as part of this specific analysis. The statistical analysis was performed using Python (version 3.x) with libraries such as Pandas for data manipulation, Scikit-learn for calculating performance metrics, and Matplotlib/Seaborn for generating visualizations.

IV. Results

Distribution of Dengue Severity The dataset included 145 patient records. The distribution of actual dengue severity, as per clinical diagnosis, was as follows:

- Mild Dengue: 83 cases (57.24%)
- Dengue with Warning Signs: 49 cases (33.79%)
- Severe Dengue: 13 cases (8.97%)

The distribution of dengue severity as predicted by the AI model was:

- Mild Dengue: 119 cases (82.07%)
- Dengue with Warning Signs: 24 cases (16.55%)
- Severe Dengue: 2 cases (1.38%)

A visual comparison of these distributions is presented in Figure 1 (Bar chart: actual vs predicted).

Overall, AI Model Performance The overall accuracy of the AI model in predicting dengue severity across all categories was **67.59%** (98 correct predictions out of 145 cases).

Performance by Severity Category The detailed performance of the AI model for each dengue severity category is presented in the confusion matrix (Table 1 and Figure 2: Heatmap confusion matrix and the classification report (Table 2).

Mild Dengue: The AI model performed best in identifying 'Mild Dengue'.

- **Precision:** 0.69 (Of all cases predicted as Mild Dengue by the AI, 69% were actually Mild Dengue).
- **Recall (Sensitivity):** 0.99 (The AI correctly identified 82 out of 83 actual Mild Dengue cases).
- **F1-Score:** 0.81. The model misclassified 1 actual 'Dengue with Warning Signs' case as 'Mild Dengue'. Conversely, 35 actual 'Dengue with Warning Signs' cases and 2 actual 'Severe Dengue' cases were incorrectly predicted as 'Mild Dengue' by the AI.

Dengue with Warning Signs (DWS): The model's performance for 'Dengue with Warning Signs' was considerably lower.

- **Precision:** 0.58 (Of all cases predicted as DWS by the AI, 58% were actually DWS).
- **Recall (Sensitivity):** 0.29 (The AI correctly identified only 14 out of 49 actual DWS cases).
- **F1-Score:** 0.38. A significant number of actual DWS cases (35 cases) were misclassified by the AI as 'Mild Dengue'. No actual DWS cases were misclassified as 'Severe Dengue'.

Severe Dengue: AI model's performance in identifying 'Severe Dengue' was a critical concern.

- **Precision:** 1.00 (The 2 cases predicted as Severe Dengue by the AI were indeed Severe Dengue).
- **Recall (Sensitivity):** 0.15 (The AI correctly identified only 2 out of 13 actual Severe Dengue cases).
- **F1-Score:** 0.27. The majority of actual 'Severe Dengue' cases were misclassified by the AI: 9 cases were predicted as 'Dengue with Warning Signs', and 2 cases were predicted as 'Mild Dengue'.

Patterns of Misclassification The confusion matrix (Table 1, Figure 2) highlights a clear trend of the AI model under-predicting dengue severity.

- A large proportion (35/49, or 71.4%) of actual 'Dengue with Warning Signs' cases were incorrectly classified as 'Mild Dengue'.
- The vast majority (11/13, or 84.6%) of actual 'Severe Dengue' cases were misclassified, primarily as 'Dengue with Warning Signs' (9 cases) or 'Mild Dengue' (2 cases).
- There were no instances where the AI model over-predicted severity from 'Mild Dengue' to 'Severe Dengue', or from 'Dengue with Warning Signs' to 'Severe Dengue' if the actual case was not severe. Only 1 'Mild Dengue' case was incorrectly upgraded to 'Dengue with Warning Signs'.

V. Discussion

This study aimed to evaluate the performance of an AI-based model in predicting dengue severity by comparing its predictions against clinically determined actual outcomes in a dataset of 145 patients. The overall accuracy of the model was found to be moderate at 67.59%. However, a deeper analysis of its performance across different severity categories revealed significant disparities, with critical implications for potential clinical use.

The AI model demonstrated its highest proficiency in identifying 'Mild Dengue' cases, achieving a very high recall (0.99) and a good F1-score (0.81). This suggests that when a patient truly has mild dengue, the model is highly likely to correctly classify them. From a clinical perspective, a high true negative rate for more severe conditions (i.e., correctly identifying mild cases) can be useful in potentially de-escalating care or reducing unnecessary anxiety, provided the model's ability to rule out severe disease is robust. However, the precision for 'Mild Dengue' (0.69) indicates that a substantial number of cases predicted as mild by the AI were, in fact, more severe ('Dengue with Warning Signs' or even 'Severe Dengue'). This dilutes the confidence in a 'Mild Dengue' prediction by the AI.

The model's performance for 'Dengue with Warning Signs' was considerably weaker (Recall: 0.29, F1-Score: 0.38). Patients with warning signs require careful monitoring and often specific interventions to prevent progression to severe dengue as stated by Harish Kasarabada etal [1]. The AI model correctly identified less than a third of these cases, with the majority (71.4%) being misclassified as 'Mild Dengue'. Such misclassification could lead to premature discharge or inadequate monitoring, potentially resulting in delayed recognition of deterioration.

Most critically, the AI model performed poorly in identifying 'Severe Dengue' cases. While the precision was 1.00 (meaning the 2 cases it did predict as severe were indeed severe), the recall was alarmingly low at 0.15. This indicates that the model correctly identified only 2 out of 13 actual 'Severe Dengue' patients. The remaining 11 severe cases were underestimated, being classified as 'Dengue with Warning Signs' (9 cases) or 'Mild Dengue' (2 cases). Missing a diagnosis of severe dengue can have life-threatening consequences, as these patients require urgent and intensive management as stated by Lam P K etal. The very low sensitivity for severe dengue is a major safety concern and renders the model, in its current evaluated state, unsuitable for ruling out severe disease.

The observed tendency of the AI model to under-predict severity is a consistent pattern across the more serious categories. Similar to the study conducted by Akobeng AK etal some diagnostic tools where higher sensitivity is often prioritized over specificity for severe conditions to minimize false negatives, even at the cost of more false positives. The reasons for this under-prediction are not ascertainable from the current study, as the AI model's internal architecture, the specific features it used for prediction (from the original dataset like platelet counts, liver enzymes, etc.), and its training data characteristics were not available. Potential reasons could include an imbalanced training dataset (if fewer severe cases were used in its development), suboptimal feature selection, or algorithm limitations. Studies on other AI models for dengue severity have reported varying accuracies, with some showing promise but also highlighting challenges, particularly with imbalanced datasets and the dynamic nature of dengue progression [12]. For instance, a systematic review by found that while AI models for dengue show potential, but there is significant heterogeneity in their performance and validation methods [13,14].

The clinical implications of these findings are significant. While AI holds promise for aiding dengue management, this model's output, especially its low recall for severe and warning sign categories, suggests it should not be relied upon for critical decision-making. An AI tool that predominantly identifies mild cases correctly but frequently misses more severe ones offers limited clinical advantage and could even induce a false sense of security. As stated by Srikiatkhachorn A etal the high stakes involved in dengue management, particularly in resource-constrained settings where such tools might be most appealing, demand extremely high sensitivity for severe outcomes.

The strengths of this study include the use of clearly defined clinical outcomes as a gold standard and the application of standard performance metrics for AI model evaluation. However, several limitations must be acknowledged. The study was retrospective and based on a relatively small sample size (N=145), with a particularly small number of 'Severe Dengue' cases (N=13), which limits the robustness of the performance estimates for this critical subgroup. The "black box" nature of the AI model, as assessed here, prevents a deeper understanding of its decision-making process or the identification of specific input variables that might be leading to misclassifications. Furthermore, without details on the AI's development and validation process, it's difficult to contextualize its performance fully. This study was trial model to assess the effectiveness of a LLM model in building a blind clinical prognostic prediction tool which gathered information from open sources available on Internet.

VI. Limitations

This study has several limitations that should be considered when interpreting its findings:

I. **Limited Sample Size:** With 145 patient records, the overall sample size is modest. Critically, the number of 'Severe Dengue' cases (N=13) was very small, which can lead to unstable estimates of performance metrics (especially recall and precision) for this vital category and may not adequately represent the spectrum of severe disease.

II. **"Black Box" AI Model:** The study evaluated the output of an AI model without access to its internal architecture, the specific input features it utilized from the broader dataset, its training methodology, or the original dataset it was trained on. It was built on basis of prompt engineering through a LLM model which used open sources of Internet to formulate a hypothetic assessment tool.

III. **Lack of Temporal Data Dynamics:** Dengue is a dynamic illness. This analysis is based on a single prediction point compared to a final outcome. The AI's performance at different stages of the illness is not assessed, nor is the information on what specific day of illness the prediction pertains to.

VII. Conclusion

The evaluated Artificial Intelligence model demonstrated a moderate overall accuracy (67.59%) for predicting dengue fever severity. It showed high sensitivity in identifying Mild Dengue cases and has a very good negative predictive value for mild dengue cases. However, its performance was substantially deficient in accurately identifying Dengue with Warning Signs and, most critically, Severe Dengue, exhibiting very low recall for the latter. This states despite the evaluation and advancement of LLM models like Chat GPT, Gemini etc which has access to majority of medical open sources, to build an accurate AI platform or tool for predicting outcomes in dengue requires deep learning with extensive data to formulate, test and validate the tool. However, these LLM models with can build screening models and may require little more advancement for creation of confirmative models which can alter clinical decision-making processes.

AI Declaration:

During the preparation of this work the authors used Chat GPT 4-o for generation of assessment tool using medical prompting. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Legends:
Figures:2
Tables:2

References

[1] Harish Kasarabada, Sreenivasa S. Iyengar, Deependra Singh, Praveena Kushala, S.K. Joshi, K. Dayanand,Observational Study Of Using Lactate Dehydrogenase As Prognostic Marker In Dengue Patients,Medical Journal Armed Forces India, Volume 81, Issue 3,2025,Pages 275-281,ISSN 0377-1237,Https://Doi.Org/10.1016/J.Mjafi.2023.07.009.

[2] Dengue: Guidelines For Diagnosis, Treatment, Prevention And Control: New Edition. Geneva: World Health Organization; 2009. PMID: 23762963.

[3] Srikiatkhachorn A, Gibbons RV, Green S, Libraty DH, Thomas SJ, Endy TP, Vaughn DW, Nisalak A, Ennis FA, Rothman AL, Nimmannitaya S, Kalayanarooj S. Dengue Hemorrhagic Fever: The Sensitivity And Specificity Of The World Health Organization Definition For Identification Of Severe Cases Of Dengue In Thailand, 1994-2005. Clin Infect Dis. 2010 Apr 15;50(8):1135-43. Doi: 10.1086/651268. PMID: 20205587; PMCID: PMC2853952.

[4] Lam PK, Tam DT, Diet TV, Tam CT, Tien NT, Kieu NT, Simmons C, Farrar J, Nga NT, Qui PT, Dung NM, Wolbers M, Wills B. Clinical Characteristics Of Dengue Shock Syndrome In Vietnamese Children: A 10-Year Prospective Study In A Single Hospital. Clin Infect Dis. 2013 Dec;57(11):1577-86. Doi: 10.1093/Cid/Cit594. Epub 2013 Sep 17. PMID: 24046311; PMCID: PMC3814826.

[5] Horstick O, Farrar J, Lum L, Martinez E, San Martin JL, Ehrenberg J, Velayudhan R, Kroeger A. Reviewing The Development, Evidence Base, And Application Of The Revised Dengue Case Classification. Pathog Glob Health. 2012 May;106(2):94-101. Doi: 10.1179/2047773212Y.0000000017. PMID: 22943544; PMCID: PMC3408880.

[6] Topol EJ. High-Performance Medicine: The Convergence Of Human And Artificial Intelligence. Nat Med. 2019 Jan;25(1):44-56. Doi: 10.1038/S41591-018-0300-7. Epub 2019 Jan 7. PMID: 30617339.Deo RC. Machine Learning In Medicine. Circulation. 2015 Nov 17;132(20):1920-30.

[7] Hussain, Zafar & Khan, Imran & Arsalan, Mudassar. (2023). MACHINE LEARNING APPROACHES FOR DENGUE PREDICTION: A REVIEW OF ALGORITHMS AND APPLICATIONS. 78. 15-36.

[8] Bhaskar, Monisha & Mahalingam, Soundarya & M, Harish & Achappa, Basavaprabhu. (2022). Predictive Scoring System For Risk Of Complications In Pediatric Dengue Infection. F1000Research. 11. 446. 10.12688/F1000research.111214.1.

[9] Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, Löffler CML, Schwarzkopf SC, Unger M, Veldhuizen GP, Wagner SJ, Kather JN. The Future Landscape Of Large Language Models In Medicine. Commun Med (Lond). 2023 Oct 10;3(1):141. Doi: 10.1038/S43856-023-00370-1. PMID: 37816837; PMCID: PMC10564921.

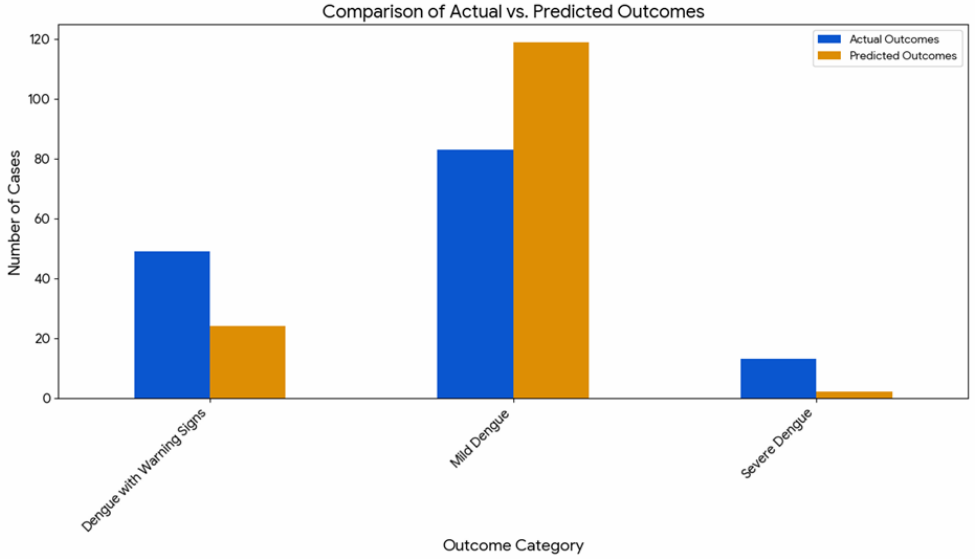
[10] Maharjan J, Garikipati A, Singh NP, Cyrus L, Sharma M, Ciobanu M, Barnes G, Thapa R, Mao Q, Das R. Openmedlm: Prompt Engineering Can Out-Perform Fine-Tuning In Medical Question-Answering With Open-Source Large Language Models. Sci Rep. 2024 Jun 19;14(1):14156. Doi: 10.1038/S41598-024-64827-6. PMID: 38898116; PMCID: PMC11187169.

[11] Akobeng AK. Understanding Diagnostic Tests 3: Receiver Operating Characteristic Curves. Acta Paediatr. 2007 May;96(5):644-7. Doi: 10.1111/J.1651-2227.2006.00178.X. Epub 2007 Mar 21. PMID: 17376185..

[12] Tsheten T, Clements ACA, Gray DJ, Adhikary RK, Furuya-Kanamori L, Wangdi K. Clinical Predictors Of Severe Dengue: A Systematic Review And Meta-Analysis. Infect Dis Poverty. 2021 Oct 9;10(1):123. Doi: 10.1186/S40249-021-00908-2. PMID: 34627388; PMCID: PMC8501593.Nascimento LFC, Rodrigues DLN, Vespermann KAC, Pires-Alves M, Lusignan SDE. Artificial Intelligence And Machine Learning For Predicting Dengue Outbreaks: A Systematic Review. Trop Med Int Health. 2020 Feb;25(2):160-175.

[13] Ilwa Mumtaz, Zubia Rashid, Rashid Saif, Muhammad Zubair Yousaf,Deep Learning Guided Prediction Modeling Of Dengue Virus Evolving Serotype, Heliyon, Volume 10, Issue 11,2024,E32061,ISSN 2405-8440,Https://Doi.Org/10.1016/J.Heliyon.2024.E32061.

[14] William Hoyos, Jose Aguilar, Mauricio Toro,Dengue Models Based On Machine Learning Techniques: A Systematic Literature Review,Artificial Intelligence In Medicine,Volume 119,2021,102157,ISSN 0933-3657, Https://Doi.Org/10.1016/J.Artmed.2021.102157. (Https://Www.Sciencedirect.Com/Science/Article/Pii/S0933365721001500)



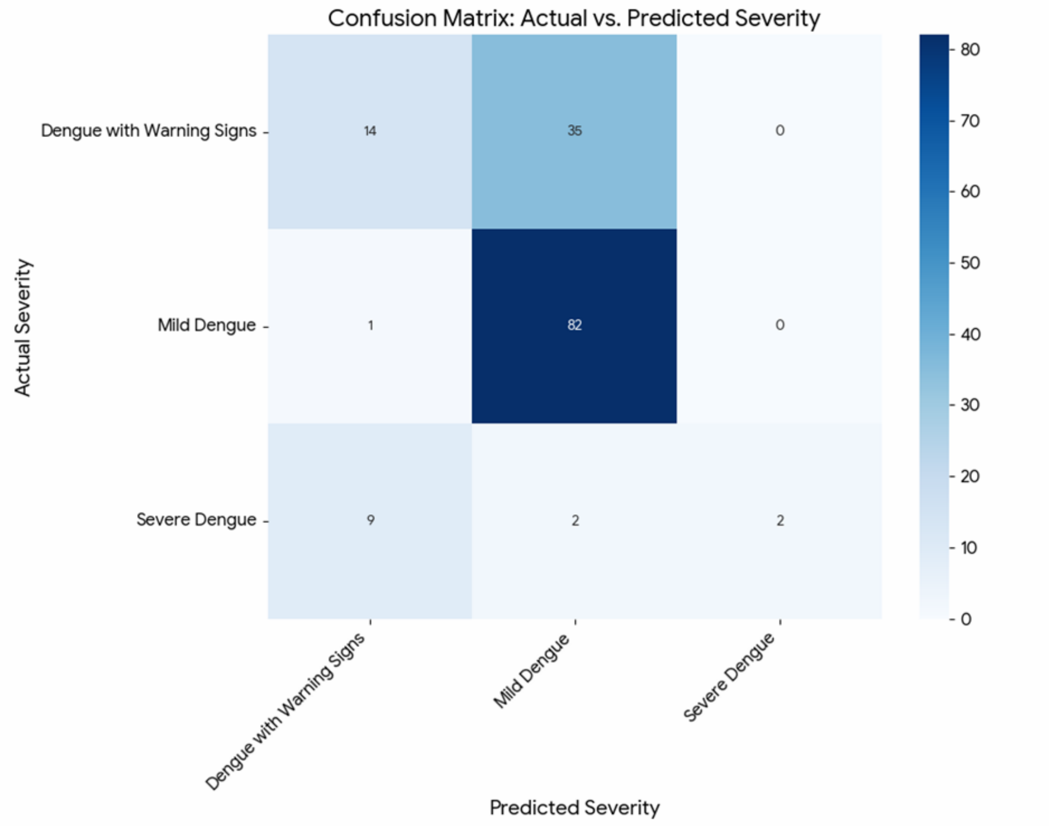


Table 1: Confusion Matrix of Actual vs. Predicted Dengue Severity

Actual Severity	Predicted: Dengue with Warning Signs	Predicted: Mild Dengue	Predicted: Severe Dengue	Total Actual
Dengue with Warning Signs	14	35	0	49
Mild Dengue	1	82	0	83
Severe Dengue	9	2	2	13
Total Predicted	24	119	2	145

Table 2: Classification Report for the AI Model

Category	**Precision**	**Recall**	**F1-Score**	**Support (Actual Cases)**
Dengue with Warning Signs	0.58	0.29	0.38	49
Mild Dengue	0.69	0.99	0.81	83
Severe Dengue	1.00	0.15	0.27	13
Accuracy			**0.68**	**145**
Macro Avg	0.76	0.48	0.49	145
Weighted Avg	0.68	0.68	0.62	145