Exploring the Molecular Footprint of Dengue Infection in Patients: An Integrative Approach to Analysing Gene Expression in Dengue Fever and Dengue Hemorrhagic Fever

Abstract

This study investigates the intricate impact of the Dengue virus (DENV) on human gene expression, exploring variations across different disease states, including Dengue fever (DF), Dengue hemorrhagic fever (DHF), convalescent stage, and healthy controls. Using the GEO dataset GDS5093, it aims to distinguish between these states and identify significant gene expression variations. Employing methods like Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA), Volcano Plot Analysis, and a Support Vector Machine (SVM) algorithm, this research provides insights into the molecular dynamics of Dengue infection. PCA revealed significant variance in gene expression related to Dengue infection, with clear clustering observed among the disease states. HCA offered insights into relationships among different health conditions, and Volcano plots identified significant genes differentially expressed across these states. However, the SVM model showed limitations in accurately distinguishing between Dengue fever types, highlighting the disease's complexity. The study underscores the applicability of advanced data science techniques in biological research. It identifies significant genes like PSMA4 and CAMK1D, offering insights into the immune mechanisms in Dengue fever. The limitations of the machine learning approach indicate the need for further refinement of predictive models. This research contributes to a deeper understanding of the genetic response to Dengue infection, setting the stage for future studies in this critical area of global health.

Introduction

Dengue fever, a mosquito-borne viral disease, presents a spectrum of clinical manifestations, from mild febrile illness to severe life-threatening conditions. The intricacies of its impact on gene expression remain a significant scientific and public health inquiry due to the widespread prevalence and increasing incidence of the disease globally (Clements et al., 2005). This arthropod-borne pathogen presents a significant global health challenge, affecting an estimated 50 to 100 million individuals annually, as per the World Health Organization (WHO). The complexity of DENV's clinical presentation, which includes four major serotypes and ranges from mild Dengue Fever (DF) to severe Dengue Haemorrhagic Fever (DHF), poses unique challenges in understanding and managing the disease. The scientific community has shown a keen interest in elucidating the determinants that differentiate DF from DHF, with several studies (Coffey et al., 2009; World Health Organisation, 1997) endeavouring to identify these key factors. Concurrently, there has been a surge in research focusing on predictive models for dengue disease outbreaks (Tanner, 2008). Despite these advances, accurately distinguishing DF and DHF patients from uninfected individuals remains an intricate challenge, largely due to the complex nature of the disease and its presentation.

Objectives

In this study, we explore the complexities of gene expression patterns in various populations impacted by Dengue fever, using the GEO dataset GDS5093. This dataset includes four key groups: patients diagnosed with Dengue fever; those suffering from Dengue haemorrhagic fever; individuals in the convalescent stage after Dengue fever; and healthy individuals serving as controls. The objective is to analyse the extent of variation in gene expression profiles between these disease states and healthy controls, employing exploratory data analysis methods. These methods include Principal Component Analysis (PCA), a statistical technique that simplifies the complexity of high-dimensional data by reducing dimensions to access key information (Jolliffe and Cadima, 2016), and Hierarchical Clustering (HCA) (Zhang et al., 2017). We aim to understand differences in gene expression profiles between these groups, identifying significant gene expression variations across disease states using volcano plots. Machine learning techniques, such as the Support Vector Machine (SVM) algorithm and bootstrapping, will be implemented to potentially distinguish between Dengue fever and Dengue haemorrhagic fever, focusing on specific genes that exhibit significant expression differences between the disease states. Finally, a key aspect of our data analysis is to assess the feasibility of using transcriptomics data to differentiate between patients infected with DF and DHF. This approach aims to contribute to the understanding of the molecular mechanisms underlying Dengue fever. Through this analysis, we intend to demonstrate the applicability of advanced data science techniques in biological research, particularly in the context of infectious diseases. The insights gained from this study are expected to provide a deeper understanding of the genetic response to Dengue infection and pave the way for future research in this critical area of global health.

Method

Data Acquisition and Preprocessing

The primary datasets, **dengue_data.csv** for gene expression and **dengue_metadata.csv** for patient metadata, were acquired. The gene expression data were transposed to align samples with metadata. The datasets were merged based on sample identifiers to integrate gene expression levels with clinical information.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was utilized as a technique to reduce the dimensionality of the gene expression data while retaining the most significant patterns and variability inherent in the dataset. In this approach, we focused on extracting the first two principal components, which were then plotted to visually explore and identify any discernible patterns or groupings among different disease states. This visualization aids in understanding how these disease states differ in terms of gene expression. Additionally, by examining the loadings of these principal components, we were able to identify the genes that contributed most significantly to the variance in the data. This step is crucial for gaining insights into the specific gene expression profiles characteristic of each disease state, thus laying a groundwork for more detailed and targeted analyses in future studies.

Hierarchical Clustering Analysis (HCA)

We utilized Hierarchical Cluster Analysis (HCA) to examine transcriptomic data from various stages of dengue infection, including Dengue Fever (DF), Dengue Hemorrhagic Fever (DHF), convalescent patients, and healthy controls. The process began by transposing raw gene

expression data and aligning samples for efficient clustering. We then standardized the data using the StandardScaler to reduce gene variance and highlight significant patterns. Ward's linkage method was chosen for HCA due to its ability to form compact clusters and minimize within-cluster variance, improving data interpretability. The resulting dendrogram was color-coded to indicate the disease state of each sample, offering a clear visual of the data structure. This detailed clustering complemented Principal Component Analysis (PCA), especially in revealing complex hierarchies and nuances within the gene expression data, proving beneficial for our study where PCA's effectiveness is limited in large, complex samples.

Volcano Plot Analysis

Volcano plots were constructed to highlight genes with statistically significant differences in expression between disease states and healthy controls. These plots juxtaposed log2 fold changes against negative log10 p-values to discern genes that were significantly upregulated or downregulated across different disease states, to highlight differentially expressed genes across Convalescent (CONV), Healthy Control (HC), Dengue Fever, and Dengue Hemorrhagic Fever (DHF). A p-value threshold of 0.05 and log2fold changes. This step ensured the analysis was focused on relevant samples. Separate dataframes were created for each disease state and the control group, focusing specifically on gene expression data for each condition. The function, differential_expression_analysis, was created to calculate log fold change and perform a t-test between the disease and control groups. This function was pivotal for identifying genes with significant expression differences. The differential expression analysis was executed for each disease state against the control group and between different DF and DHF. These plots showed significant genes based on p-value and log fold change thresholds, using different colours and edges for better visualization.

Machine Learning

We used Machine learning, specifically a Support Vector Machine (SVM) algorithm, to classify disease states, focusing on distinguishing between Dengue Hemorrhagic Fever and Dengue Fever. The SVM was selected due to its effectiveness in classification tasks and its wide application in bioinformatics (Gold, and Sollich, 2003). To improve the model's reliability and robustness, we implemented bootstrapping. This involved creating a bootstrapping ensemble to test our model multiple times and calculate a weighted accuracy score. This method helped reduce overfitting and enhance model stability by introducing randomness and allowing us to estimate variance and confidence intervals. Our approach aimed to accurately classify disease states, providing a valuable tool for differentiating between Dengue Hemorrhagic Fever and Dengue Fever.

Independent t-Test with Benjamini-Hochberg FDR Correction

The study involved a detailed gene expression analysis comparing 'Dengue Fever' patients and 'Healthy Controls'. Initially, gene expression data was transposed to align with metadata, allowing for efficient merging based on sample identifiers. Groups were defined using metadata, isolating 'Dengue Fever' and 'Healthy Control' data for comparison. The statistical approach included independent t-tests for each gene, using a policy to omit NaN values, assessing the significance of expression differences between groups. To address multiple testing, Benjamini-Hochberg FDR correction was applied, adjusting p-values to control the false discovery rate. Results were compiled in a DataFrame, listing genes with their p-values and adjusted p-values, facilitating interpretation. The most significant genes, indicated by the lowest adjusted p-values, were identified as top markers differentiating between the two groups.

Distribution of samples across disease states

The metadata file is loaded into a panda DataFrame. Within the DataFrame, we focus on the column labelled disease.state, which contains categorical data representing the different disease states in the study. We employ the value_counts() method from pandas, which efficiently calculates the frequency of each unique value in this column. This method returns a Series object, with the index representing the unique disease states and the corresponding values indicating the number of samples for each state. The output is a Series object where the indices represent the disease states, and the values denote the count of samples in each state. This output provides a clear quantitative overview of the dataset's composition in terms of disease state distribution.

Results and discussion:

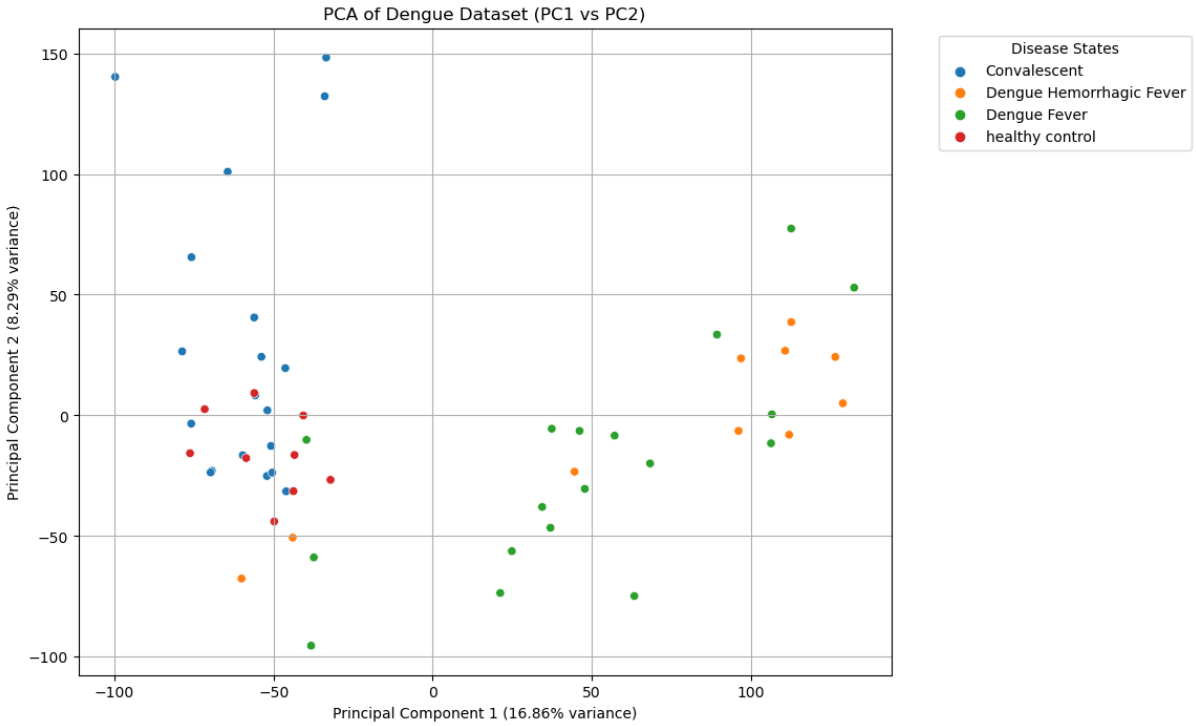
Exploratory Analysis of Gene Expression Across Populations

Table 1. Explained Variance Ratios for the First Ten Principal Components

Principle Components	Variance Ratios
PC1	0.16861576
PC2	0.08293707
PC3	0.06446877
PC4	0.04503
PC5	0.03452972
PC6	0.03197811
PC7	0.02443298
PC8	0.02045925
PC9	0.01964794
PC10	0.01863954

The principal component analysis (PCA) conducted on the dengue gene expression dataset reveals the distribution of explained variance across the first ten principal components. This analysis is a critical step in understanding the underlying structure of high-dimensional gene expression data and in facilitating the identification of patterns that could be biologically significant. The analysis of the explained variance ratios for the first ten principal components (Table 1) provides significant insights into the data structure and the effectiveness of the Principal Component Analysis (PCA) in reducing dimensionality while retaining pertinent information. The first principal component (PC1) accounts for the largest proportion of variance in the dataset (16.86%). This dominant principal component captures the most significant pattern or trend within the data.

Figure 1. Principal Component Analysis (PCA) Plot of Disease States



In Figure 1, the PCA plot provides a visual demarcation of the disease states, emphasizing the stark contrasts in gene expression profiles among them. The first two principal components are cumulatively explained about 25.15% of the total variance in the dataset, reveals discernible clustering among the various disease states. The convalescent state (blue points) are somewhat spread out across the plot, indicating a degree of variability in the gene expression of patients recovering from Dengue. Patients with Dengue Hemorrhagic Fever (orange points) are somewhat clustered together but also spread out, indicating some commonality in gene expression patterns with individual differences. Patients with Dengue Fever (green points) are dispersed across the plot, showing a wide range of gene expression profiles among these individuals. Healthy controls (red points) are scattered throughout but tend to be more on the right side of the plot, suggesting some distinction in gene expression profiles from those with Dengue.

There is overlap between all the disease states, suggesting that while there are differences in gene expression profiles associated with each state, there are also similarities. The healthy controls overlap with the disease states, but there appears to be a trend where the healthy controls are more separated from the convalescent state than from the active disease states, which might reflect the transition of gene expression back to a 'normal' state in recovery. Potential Insights: This plot may suggest that gene expression profiles can somewhat distinguish between healthy individuals and those with Dengue infections, but there is still a significant overlap that could be due to the complexity of the disease, individual variations, or other factors not captured in the first two principal components. These patterns could potentially serve as a basis for further investigation into the molecular underpinnings of disease progression and recovery.

Differential Gene Expression in Disease States

Figure 2. Dendrogram from Hierarchical Cluster Analysis showing the clustering of gene expression profiles among four populations.

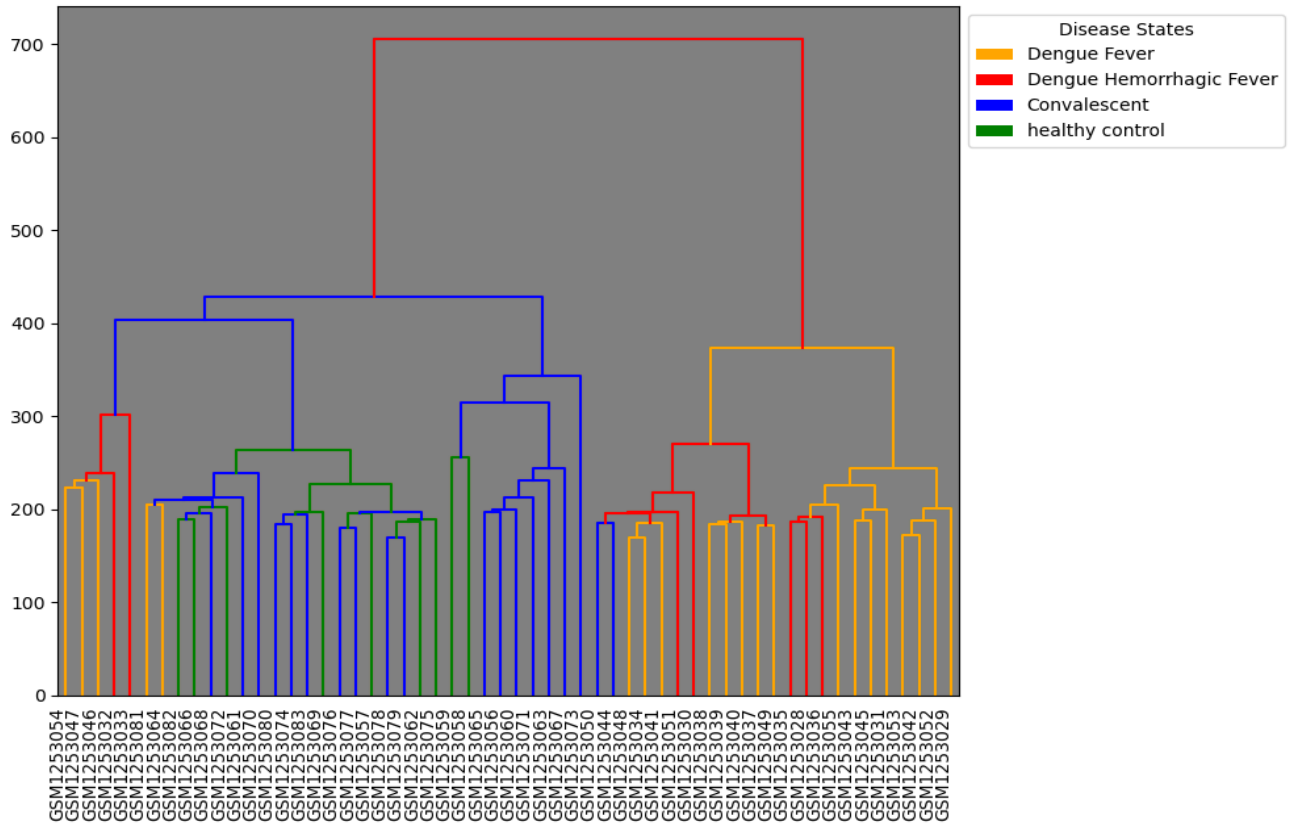
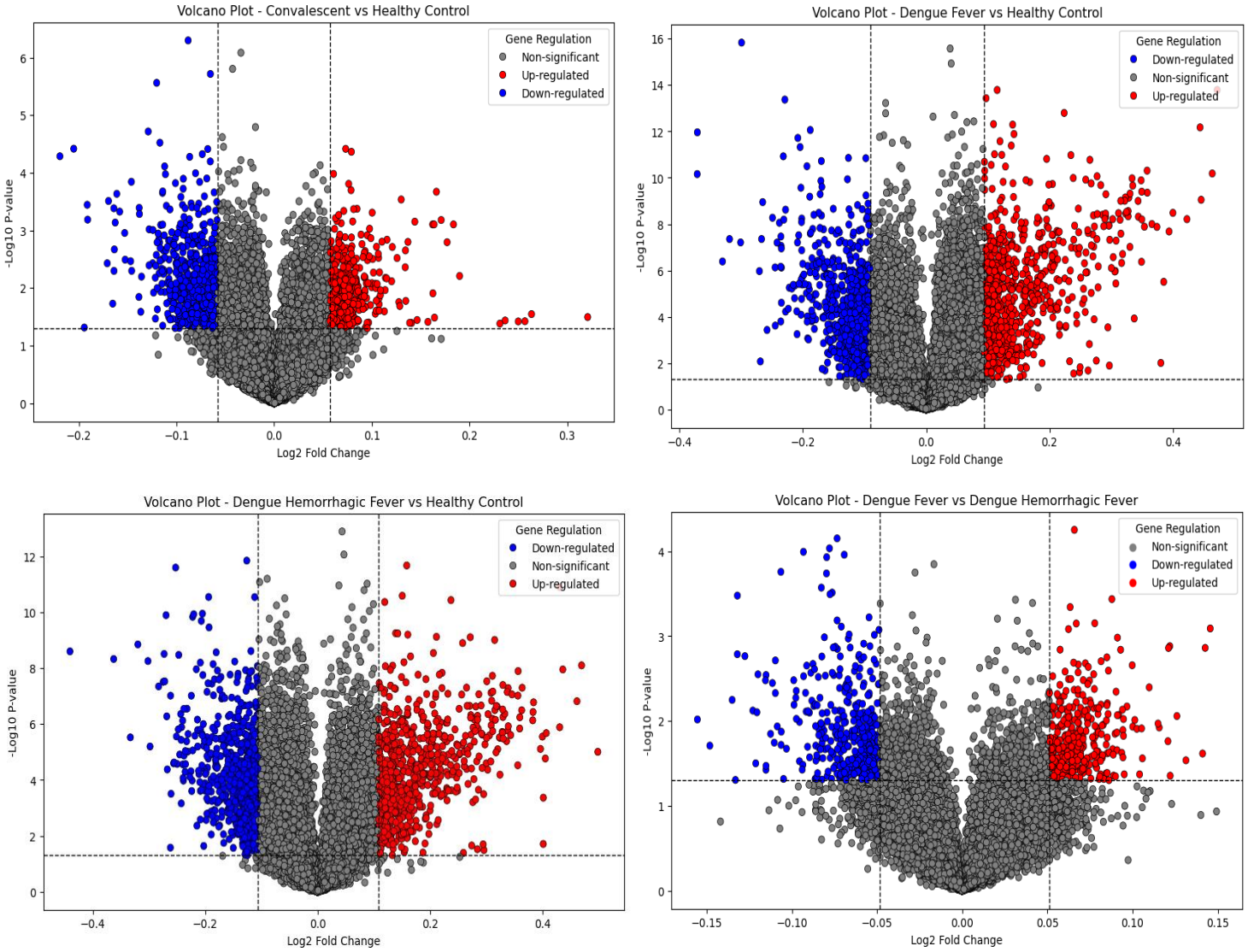


Figure 2 illustrates the results of an agglomerative hierarchical clustering analysis, which organizes samples by their gene expression similarities across several disease states. This method incrementally merges samples or clusters based on their proximity in gene expression space, thus providing insights into the relationships among different health conditions. The dendrogram's vertical axis represents the degree of dissimilarity between clusters. Lower values on this axis suggest more closely related gene expression patterns, whereas higher values denote greater differences. The horizontal axis lists each sample, marked with a distinct identifier. The color-coding—orange for DF, red for DHF, blue for Dengue Fever, and green for Healthy Control—indicates the disease state of each sample, offering an immediate visual distinction of the clustering.

Analysis of the dendrogram shows clear segregation based on disease state. The Healthy Control samples (green) cluster separately, highlighting a distinct gene expression profile from those of diseased individuals. The DF (orange) and Dengue Haemorrhagic Fever (red) samples also cluster according to their respective conditions, demonstrating unique gene expression patterns. Notably, the Convalescent samples (blue) interpose between the diseased and healthy clusters, suggesting a gradual return of gene expression to normal levels as recovery ensues. The pattern revealed by the dendrogram suggests consistent, identifiable differences in gene expression among the various disease states and through the recovery process. Such distinctions are crucial for comprehending disease progression and recovery at a molecular level. Moreover, the dendrogram hints at possible subclusters within the disease states, potentially indicating different disease severities or subtypes, warranting further detailed analysis to substantiate these initial observations.

Figure 3. Comparative Volcano Plots of Gene Expression in Dengue Disease States versus Healthy Control and Dengue Fever vs Dengue Hemorrhagic Fever



Volcano plots are a type of scatter plot that are commonly used to show the results of multiple comparisons of gene expression data. They are useful for identifying genes that are significantly differentially expressed across different conditions or treatments. The Volcano plots display the different gene expressions of DF, DHF, and Convalescent disease states when compared to healthy controls to identify genes with significant changes in expression levels. On the right, the red shows significant upregulated genes, grey is non-significant genes, and blue is downregulated genes that are significantly expressed.

Each volcano plot compares gene expression levels between a disease state and healthy control, showing both the magnitude of expression change (log2 fold change) and the statistical significance (-log10 p-value). In these plots, the x-axis represents the log2 fold change in gene expression, and the y-axis represents the negative log10 of the p-value, which is a measure of statistical significance (P-value <0.05). Typically, genes that are significantly up-regulated are shown in red, significantly down-regulated genes are shown in blue, and non-significant changes are in grey.

Figure 3 suggest distinct gene expression profiles in response to each disease state when compared to healthy controls. In convalescent individuals, the balance between up-regulated

and down-regulated genes suggests a return to homeostasis after an infection or disease. However, both DF and DHF show significant deviations in gene expression from healthy controls, with a larger number of genes being up-regulated than down-regulated. This up-regulation may indicate a robust immune response or activation of pathways involved in inflammation and antiviral response. Notably, the response in DHF is more pronounced than in DF, which may reflect the more severe clinical manifestation of the disease. This could be due to the overactivation of immune pathways leading to the haemorrhagic condition, suggesting that some of these up-regulated genes may be involved in the pathogenesis or progression of the disease.

In Figure 3, Dengue Fever to Dengue Hemorrhagic Fever plot, the upregulated genes, which have a log2 fold change greater than zero and are statistically significant, indicating that these genes are more active in DHF compared to DF. Conversely, the blue dots represent downregulated genes with a log2 fold change less than zero, signifying reduced activity in DHF. In the centre, the distribution of non-significant genes indicates no substantial difference in their expression between the two conditions. The horizontal dashed line likely represents a threshold of significance for the P-value, while the vertical dashed lines may indicate fold change cut-offs. The plot shows a substantial number of both upregulated and downregulated genes, suggesting a complex gene regulation pattern in the progression from DF the more severe DHF, which may be reflective of the underlying biological mechanisms differentiating these disease states.

Figure 4. Confusion Matrix for Dengue Disease Classification

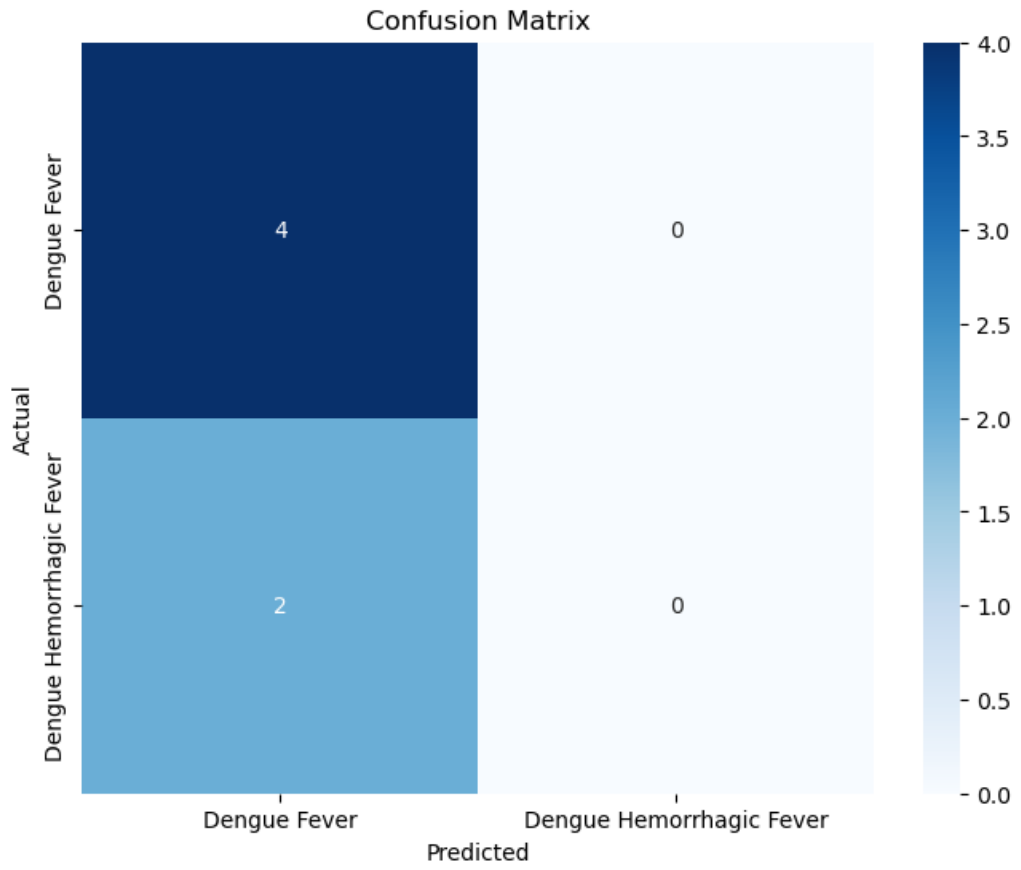


Figure 4 displays the confusion matrix for the comparison of Dengue Fever vs Dengue Hemorrhagic Fever. Confusion matrix showed 4 cases were correctly identified as Dengue Fever. There were no cases of true positives or false negatives for False Negatives (FN): 2 cases were incorrectly identified as Dengue Hemorrhagic Fever when they were actually Dengue Fever. This matrix indicates that the model is more proficient at correctly identifying Dengue Fever than Dengue Hemorrhagic Fever, with a recall of 100% for Dengue Fever (all actual Dengue Fever cases were correctly identified), but a recall of 0% for Dengue Hemorrhagic Fever (no actual Dengue Hemorrhagic Fever cases were correctly identified).

Figure 5. Machine Learning Accuracy

- Bootstrap Mean Accuracy: 0.6316666666666666
- Bootstrap 95% Confidence Interval for Accuracy: (0.6133333333333332, 0.6516666666666666)

Figure 5 presents a machine learning model with a bootstrap mean accuracy of approximately 63.16%. This figure is indeed above the 50% mark, which would suggest random chance, yet it doesn't meet the standard clinical accuracy range of 70-90% that's often cited for medical diagnostics (Fiddler.ai, n.d.). In clinical settings, this range is typically considered the threshold for a test to be reliable and useful. The confidence interval for accuracy ranges from about 61.3% to 65.16%. The accuracy does not exceed 70% This suggests a significant variability in the model's performance, which could be due to the underlying data or the model itself. The upper limit of this interval does not reach the clinical standard; however, the lower limit is well below, indicating that the model's accuracy does not reach the standard for clinical usage.

Table 2. A classification report of Machine Learning (ML) Model

Classification Report				
	Precision	Recall	F1-score	Support
0	0.67	1	0.8	4
1	0	0	0	2
Accuracy			0.67	6
Macro average	0.33	0.5	0.4	6
Weighted average	0.44	0.67	0.53	6

The model shows a significant bias towards identifying dengue fever (Class 0) at the expense of dengue hemorrhagic fever (Class 1). This is evident from the high recall for Class 0 and zero recall for Class 1. The relatively high overall accuracy (67%) is misleading in this context, as it primarily reflects the model's ability to detect only one class of the disease. The F1-score for dengue fever is at 0.8 and at 0 for DHF which displays a strong performance in detecting dengue fever and a poor performance in identifying dengue haemorrhagic fever which raises serious concerns, especially given the more severe nature of this condition. It suggests that the model may not have learned the necessary patterns to differentiate between the two classes effectively, possibly due to limited or imbalanced training data, or a lack of relevant features.

Table 3. Permutation testing with bootstrapping.

Observed accuracy	0.6666666666666666
Permutation	0.6666666666666665
P-Value	1
Observed accuracy = Not statistically significant	

Given that the permutation accuracy is almost identical to the observed accuracy and the p-value is 1 1 in Table 3, suggests that the model's ability to predict or classify based on this dataset is not better than random chance. This outcome could be due to the model not capturing the underlying patterns in the data, or it could indicate that the gene expression levels do not have a strong or consistent relationship with the aspects of dengue infection being modelled (e.g., disease progression, patient response). The results imply that the transcriptomics data, as used in the model, may not be a reliable prediction model for differentiating between dengue fever and the severe state, dengue haemorrhagic fever.

Table 4. Distribution of samples across disease states

Disease State	Samples
Convalescent	**19**
Dengue Fever	**18**
Dengue Hemorrhagic Fever	**10**
Healthy Control	**9**

In Table 4, the dataset is relatively balanced between Convalescent and Dengue Fever states, with 19 and 18 samples, respectively. However, there are fewer samples for Dengue Hemorrhagic Fever (10 samples) and Healthy Control (9 samples). This slight imbalance may be affect the performance of the performance model, as it may show bias towards the classes with more samples such as dengue fever. In a clinical setting, the ability to accurately distinguish between these states can be crucial for patient treatment and management,

especially between the more severe Dengue Hemorrhagic Fever and the other states. The relatively small sample size for each class can limit the generalizability of any predictive model (Schreier, 2018). It might be necessary to collect more data, particularly for the underrepresented classes, to improve model training and predictive accuracy.

Significant genes

The genes PSMA6, GRAMD1C, PSMA4 were the top 3 highly significant genes expressed in all disease states. The identified biomarkers are implicated in various biological processes, such as immune response, cell signalling, and apoptosis.

Table 4. Differential Expression Analysis of Selected Genes

	Gene	P-Value	Adjusted P-Value
13193	PSMA6	2.654478e-16	3.952120e-12
17601	GRAMD1C	1.467777e-16	3.952120e-12
9347	PSMA4	1.174742e-15	1.166009e-11
10532	VRK2	1.628420e-14	9.742794e-11
17772	KCTD14	1.635960e-14	9.742794e-11
20835	TOMM40L	3.700422e-14	1.806820e-10
1491	CASS4	4.247486e-14	1.806820e-10
18078	BANP	5.963128e-14	2.219551e-10
19264	AGAP7P	1.654532e-13	4.926701e-10
20755	FAM72A	1.592182e-13	4.926701e-10

PSMA6

PSMA6 (Proteasome Subunit Alpha 6) is a gene that encodes a component of the proteasome, which is a crucial complex within the cell responsible for degrading unneeded or damaged proteins by proteolysis, a chemical process that breaks down peptides into amino acids. This process is vital for maintaining the cell's health and function (Fagerberg et al., 2014). PSMA6 was identified as a critical gene for the survival of pancreatic cancer cells. It was noted that PSMA6 is part of the proteasome subunit, and its inhibition, particularly with bortezomib (a proteasome inhibitor), was highly effective in reducing cell viability in Pancreatic ductal adenocarcinomas PDAC cells. underlining it's potential of PSMA6 as a therapeutic target, with further research needed to explore this possibility (Bakke et al., 2019). PSMA6, like PSMA4, is one of the top 3 highly significant genes expressed in all disease states of DF and DHF. The extremely low p-value of 2.654478e-16 and adjusted p-value of 3.952120e-12 for PSMA6 suggest a highly significant statistical difference in its expression between the disease states and possibly controls. In biomedical research, a p-value below 0.05 is generally considered significant. The given p-values for PSMA6 far exceed this threshold, indicating a strong statistical significance. The adjusted p-value accounts for multiple comparisons, ensuring that the significance is not a result of random chance but is genuinely associated with the disease conditions.

GRAMD1C

GRAMD1C (GRAM Domain Containing 1C) is a gene that encodes a protein belonging to the GRAMD family. The GRAMD family of proteins is characterized by the presence of a GRAM domain, a protein domain believed to be involved in binding to phosphoinositide's, which are a class of lipids involved in cellular signalling (Ng et al., 2022). GRAMD1C shows a p-value of 1.467777e-16 and an adjusted p-value of 3.952120e-12. These values are extremely low, indicating a highly significant statistical difference in the expression of GRAMD1C in the studied conditions (presumably Dengue fever and Dengue haemorrhagic fever). A p-value below 0.05 is generally considered significant in biomedical research. GRAMD1C's p-value is much lower than this threshold, suggesting that the difference in its expression is not due to random chance. The adjusted p-value accounts for multiple comparisons in the study. The fact that this value remains extremely low strengthens the argument that the differential expression of GRAMD1C is statistically significant and relevant to dengue disease.

GRAMD1C is involved in the regulation of lipid composition and distribution within cell membranes (Giordano, F. and Kiger, A., 2020). Since dengue virus enters cells via receptor-mediated endocytosis, alterations in membrane lipid composition can influence viral entry and replication. Changes in the lipid environment might affect the ability of the virus to fuse with the host cell membrane or to form replication complexes (Mazzon, and Mercer, 2014). GRAMD1C ability to change the composition of lipids can influence the severe inflammatory response would be involved in the pathophysiology of DHF. This gene's role in lipid metabolism, might influence the stability and permeability of endothelial cells, contributing to the plasma leakage characteristic of severe dengue infection as lipids maintain endothelial cell integrity and function (Srikiatkhachorn, A., 2009).

PSMA4

PSMA4 (Proteasome Subunit Alpha 4) plays a critical role as part of the proteasome, a complex responsible for degrading unneeded or damaged proteins through proteolysis. This degradation process is vital for maintaining cellular homeostasis and is implicated in numerous cellular functions, including the regulation of the cell cycle, signalling pathways, transcriptional regulation, and immune responses (Chiao et al., 2021). In the context of DF and DHF, a heightened expression of PSMA4 suggests an enhanced proteasomal activity, which could be part of the host's immune response to the dengue virus infection. An upregulated proteasome activity may contribute to the generation of antigenic peptides presented on the cell surface to T cells, thereby influencing the immune response against the dengue virus (Goldberg et al., 2002).

However, this increased activity could also lead to altered degradation of cellular proteins that are crucial for maintaining normal cell functions. In diseases such as non-small cell lung cancer (NSCLC), PSMA4's role in modulating antigen presentation has been studied with respect to its impact on anti-tumour immunity. An upregulated expression of PSMA4 is associated with the immune system's capacity to recognise and destroy tumour cells (Bai et al., 2002). However, in the setting of DF and DHF, the implications of elevated PSMA4 levels are more complex and may not be entirely beneficial. The clinical impact of PSMA4 upregulation in DF and DHF patients can be multifaceted. While it may aid in the immune response to clear the virus, Liu et al. (2009) state that the expression of PSMA4, when downregulated, can induce apoptosis and cell proliferation due to its ability to act as a mediator for lung cancer cell growth, and may contribute to the pathogenesis of the disease by affecting cell proliferation and apoptosis. In severe cases of DHF, where the immune response is overly

activated, PSMA4 upregulation may exacerbate the condition by influencing the cytokine storm or by altering the balance between cell survival and death. The increased risk associated with lung cancer and the dysregulation observed in cell proliferation and apoptosis raises concerns about the long-term effects of sustained high levels of PSMA4. From the chart we can infer that, unlike in convalescent individuals (CONV) and healthy controls (HC), where PSMA4 expression appears to be within a lower range, patients with active DF and DHF exhibit a pronounced expression, indicating an ongoing response to infection.

Exclusion of Heatmap for Top 50 genes

In this study, we employed a focused approach to analyse gene expression patterns in DF and DHF, utilizing Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA), Volcano Plot Analysis, and a Support Vector Machine (SVM) algorithm. While a heatmap for the top 50 genes by variance could offer visual insights into gene expression variability, its omission is justifiable given our study's specific analytical framework and objectives. Our methodology prioritized quantitative analyses and specific gene identification over broad pattern visualization. The PCA and HCA already provided substantial insights into gene expression patterns and relationships among different disease states, while the volcano plots effectively highlighted genes with significant differential expression. Additionally, the limited interpretability and potential complexity of a heatmap in this context, where the focus was on precise gene identification and classification accuracy, made its inclusion less pertinent to our study's goals.

Limitations

This analysis based on the GEO dataset GDS5093, which may have inherent limitations such as sample size, sample diversity, and data quality. The representation of the four populations (dengue fever, dengue hemorrhagic fever, convalescent patients, and healthy controls) in the dataset could affect the generalizability of the findings. PCA reduces dimensionality but might oversimplify complex gene expression patterns. The hierarchical clustering and volcano plots provide insights but are limited by the statistical thresholds and criteria set for significance. Using the 2.5th and 97.5th percentiles for log fold change thresholds might not always capture biologically meaningful changes. These thresholds are arbitrary and data-dependent, which means they can vary significantly between datasets. The p-value threshold of 0.05 is a conventional choice but does not account for multiple testing corrections. In high-throughput experiments like gene expression studies, adjusting for multiple comparisons (e.g., using Bonferroni or False Discovery Rate methods) is crucial to reduce the risk of false positives.

Samples

Convalescent: 19 samples

Dengue Fever: 18 samples

Dengue Hemorrhagic Fever: 10 samples

Healthy Control: 9 samples

The machine learning approach may have limitations related to the feature selection, model overfitting, or under-representation of certain disease states in the training data. The accuracy and recall rates in your analysis indicate possible model biases or limitations in differentiating between dengue fever and dengue hemorrhagic fever. The dataset does show some

imbalance, but it is not extremely imbalanced. The classes "Dengue Hemorrhagic Fever" and "Healthy Control" have fewer samples compared to "Convalescent" and "Dengue Fever". This could potentially impact the performance of machine learning models, as they might be biased towards the classes with more samples. In the context of disease prediction, even a slight imbalance can be significant, especially if the minority class is of greater interest (DHF).

While significant genes and pathways were identified, the biological interpretation of these findings might be limited by the current understanding of dengue fever's pathogenesis. While the significant genes have established roles in immune response and cell signalling, their specific functions in the context of dengue infection are not fully understood. The extrapolation of their roles from other diseases or general biological processes to dengue fever might not capture the unique pathophysiological mechanisms of DF and DHF. Their pleiotropic nature (having multiple effects) and possible redundancy with other genes or pathways in the immune response might complicate the interpretation of their roles in DF and DHF (Ma et al., 2017). Translating these findings to clinical outcomes, such as disease severity or treatment responses in DF and DHF, is challenging. The relationship between gene expression levels and the actual functional impact in the disease context is not straightforward. The findings from this GEO dataset and analysis methods might not be generalizable to all dengue fever cases or other datasets. Reproducibility of the results with different datasets or analytical approaches is an important consideration.

Conclusion

In conclusion, the study aimed to discern the molecular distinctions between different stages of Dengue infection with the application of explanatory analysis and machine learning on transcriptomics data to analyse and understand the molecular distinctions between patients with Dengue Fever (DF) and Dengue Hemorrhagic Fever (DHF). Volcano plots further highlighted specific genes significantly dysregulated in DF and DHF, suggesting their potential roles in the disease's pathogenesis. Particularly noteworthy were genes like PSMA4 and CAMK1D, which are implicated in immune response and cell signalling, offering insights into the immune mechanisms in Dengue fever. However, the machine learning approach, using a Support Vector Machine (SVM) algorithm, exhibited limitations in distinguishing between DF and DHF accurately, indicating the need for further refinement of predictive models. The model’s performance, while above random chance, did not meet the desired accuracy standards for clinical diagnostics, underscoring the complexity of the disease and the challenges in developing reliable diagnostic tools based solely on gene expression data. The convalescent stage is ambiguous and is unclear if it accounts for the recovery stage of both DF and DHF. To address the limitation of not accounting for patients recovering from DHF, future work should include gene expression analysis during the convalescent phase of DHF. Investigating how environmental factors (such as geographic location, climate, or vector density) interact with gene expression in dengue patients could provide insights into the variable epidemiology and clinical presentations of the disease.

References

Anwar, A., Hosoya, T., Leong, K.M., Onogi, H., Okuno, Y., Hiramatsu, T., Koyama, H., Suzuki, M., Hagiwara, M. and Garcia-Blanco, M.A., (2011). The kinase inhibitor SFV785 dislocates dengue virus envelope protein from the replication complex and blocks virus assembly. *PLoS one*, *6*(8), p.e23246.

Bai, Y., Zheng, J., Cheng, L., Liu, Q., Zhao, G., Li, J., Gu, Y., Xu, W., Wang, M., Wei, Q. and Zhang, R., 2022. Potentially functional genetic variants of VAV2 and PSMA4 in the immune-activation pathway and non-small cell lung cancer survival. *The Journal of Gene Medicine*, *24*(10), p.e3447.

Bakke, J., Wright, W.C., Zamora, A.E., Oladimeji, P., Crawford, J.C., Brewer, C.T., Autry, R.J., Evans, W.E., Thomas, P.G. and Chen, T., (2019). Genome-wide CRISPR screen reveals PSMA6 to be an essential gene in pancreatic cancer cells. *BMC cancer*, *19*(1), pp.1-12.

Chiao, C.C., Liu, Y.H., Phan, N.N., An Ton, N.T., Ta, H.D.K., Anuraga, G., Minh Xuan, D.T., Fitriani, F., Putri Hermanto, E.M., Athoillah, M. and Andriani, V., (2021). Prognostic and genomic analysis of proteasome 20S Subunit Alpha (PSMA) family members in breast cancer. *Diagnostics*, *11*(12), p.2220.

Clements, A., Gray, D., Adhikary, R., Furuya-Kanamori, L. and Wangdi, K., (2021). Clinical predictors of severe dengue: a systematic review and meta-analysis.

Coffey, L.L., Mertens, E., Brehin, A.C., Fernandez-Garcia, M.D., Amara, A., Després, P. and Sakuntabhai, A., (2009). Human genetic determinants of dengue virus susceptibility. Microbes and infection, 11(2), pp.143-156.

Fagerberg, L., Hallström, B.M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K. and Asplund, A., (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & cellular proteomics*, *13*(2), pp.397-406.

Fiddler.ai, (n.d.). Which is more important: model performance or model accuracy. 'https://www.fiddler.ai/model-accuracy-vs-model-performance/which-is-more-important-model-performance-or-model-accuracy#:~:text=Good%20accuracy%20in%20machine%20learning,demand%2099%25%20accuracy%20and%20up.'

Giordano, F. and Kiger, A., (2020). Lipid regulation and transport in membrane remodeling. *Molecular Biology of the Cell*, *31*(6), pp.403-404.

Goldberg, A.L., Cascio, P., Saric, T. and Rock, K.L., (2002). The importance of the proteasome and subsequent proteolytic steps in the generation of antigenic peptides. *Molecular immunology*, *39*(3-4), pp.147-164.

Gold, C. and Sollich, P., (2003). Model selection for support vector machine classification. *Neurocomputing*, *55*(1-2), pp.221-249.

Jolliffe, I.T. and Cadima, J., (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), p.20150202.

Liu, Y., Liu, P., Wen, W., James, M.A., Wang, Y., Bailey-Wilson, J.E., Amos, C.I., Pinney, S.M., Yang, P., De Andrade, M. and Petersen, G.M., (2009). Haplotype and cell proliferation

analyses of candidate lung cancer susceptibility genes on chromosome 15q24-25.1. *Cancer research*, *69*(19), pp.7844-7850.

Ma, M.C.J., Pettus, J.M., Jakoubek, J.A., Traxler, M.G., Clark, K.C., Mennie, A.K. and Kwitek, A.E., (2017). Contribution of independent and pleiotropic genetic effects in the metabolic syndrome in a hypertensive rat. *PloS one*, *12*(8), p.e0182650.

Mazzon, M. and Mercer, J., (2014). Lipid interactions during virus entry and infection. *Cellular microbiology*, *16*(10), pp.1493-1502.

Ng, M.Y.W., Charsou, C., Lapao, A., Singh, S., Trachsel-Moncho, L., Schultz, S.W., Nakken, S., Munson, M.J. and Simonsen, A., (2022). The cholesterol transport protein GRAMD1C regulates autophagy initiation and mitochondrial bioenergetics. *Nature Communications*, *13*(1), p.6283.

Schreier, M., (2018). Sampling and generalization. *The SAGE handbook of qualitative data collection*, pp.84-97.

Srikiatkhachorn, A., (2009). Plasma leakage in dengue haemorrhagic fever. *Thrombosis and haemostasis*, *102*(12), pp.1042-1049.

Tanner, L., Schreiber, M., Low, J.G., Ong, A., Tolfvenstam, T., Lai, Y.L., Ng, L.C., Leo, Y.S., Thi Puong, L., Vasudevan, S.G. and Simmons, C.P., (2008). Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. PLoS neglected tropical diseases, 2(3), p.e196.

World Health Organization, (1997). *Dengue haemorrhagic fever: diagnosis, treatment, prevention and control*. World Health Organization.

Zhang, Z., Murtagh, F., Van Poucke, S., Lin, S. and Lan, P., (2017). Hierarchical cluster analysis in clinical research with heterogeneous study population: highlighting its visualization with R. *Annals of translational medicine*, *5*(4).