Exploring Differential Gene Expression and Molecular Signatures in Dengue Virus Infections: A Comprehensive Analysis using Exploratory data analysis and Machine Learning"

Abstract

Dengue fever, caused by the Dengue virus, poses a global health threat, affecting a substantial portion of the world's population. This study uses gene expression data obtained through microarray analysis methods to explore the intricate dynamics of Dengue virus infections across different disease states. The methodology involves comprehensive exploratory data analysis, including Principal Component Analysis (PCA) and Hierarchical Clustering Analysis (HCA), alongside statistical analysis and machine learning using the Random Forest classifier.

The results reveal distinct clustering patterns among disease states, suggesting robust molecular differentiation. Unique gene expression profiles associated with Dengue Hemorrhagic Fever, Dengue Fever, healthy controls, and convalescent individuals are unveiled through HCA. Volcano plots highlight differential expression, identifying potential biomarkers such as KCTD14, implicated in substrate recruitment for ubiquitination. Despite challenges in machine learning discernment between Dengue Fever and Dengue Hemorrhagic Fever, this study provides valuable insights into the molecular landscape of Dengue virus infections. The identified genes present promising targets for therapeutic interventions and diagnostic advancements, contributing to a deeper understanding of dengue infection dynamics.

Introduction

Dengue fever is a rapidly disseminating acute systemic viral illness caused by four different serotypes of the dengue virus—DENVs 1 through 4. The primary mode of transmission occurs through the bites of female Aedes mosquitoes, particularly Aedes aegypti and Aedes albopictus. These mosquitoes are common in tropical and subtropical regions, creating an environment conducive to the spread of the disease. The global prevalence of dengue is substantial, with approximately one-third of the world's population residing in areas at risk of infection affecting more than 400 million people per year (Parveen et al., 2023). The impact of dengue extends beyond geographical boundaries, making it a major public health concern that demands comprehensive research and strategic interventions. Understanding the dynamics of dengue transmission, its diverse clinical manifestations and the factors influencing its prevalence is crucial for developing effective preventive measures and therapeutic interventions (Khetarpal & Khanna, 2016).

Although infections with the Dengue virus in humans often go unnoticed or present with mild symptoms, the clinical spectrum extends to severe manifestations, including potentially fatal conditions like dengue hemorrhagic fever (Bhatt S. et al., 2013). This study aims to delve into the intricate dynamics of Dengue virus infections, ranging from asymptomatic cases to severe manifestations such as dengue hemorrhagic fever to identify significant genes associated with the dengue fever disease states to facilitate potential therapies.

The acquisition of gene expression data for Dengue Fever analysis employed microarray technology and related methodologies. Microarrays have significantly advanced biological research, allowing for detailed exploration of gene expression patterns that were previously unattainable (Page et al., 1970). The gene expression data for Dengue Fever analysis was obtained from the GDS5093 dataset, stored in the "dengue_data.csv" file. This dataset contains information on various populations, including patients with Dengue Fever, Dengue Hemorrhagic Fever, convalescent-stage patients, and healthy controls. Additional metadata, crucial for sample categorization, was extracted from the "dengue_metadata.csv" file.

Methods

Exploratory Data Analysis (EDA):

The initial phase of our analysis involved the exploration of gene expression variations across distinct disease states. Principal Component Analysis (PCA) is a technique for dimensionality reduction, facilitating the visualization of sample relationships based on gene expression profiles. In this context, python libraries were leveraged, including pandas for data manipulation, matplotlib for plotting, and scikit-learn for implementing PCA.

The gene expression data, extracted from the "dengue_data.csv" file, was paired with metadata from "dengue_metadata.csv." The metadata, including a column denoting disease states ('disease.state'), was instrumental in guiding the analysis. A mapping was established, assigning numerical labels to each disease state for computational purposes. Subsequently, the gene expression data was transposed to facilitate a sample-centric approach, aligning it with the metadata.

The execution of PCA with scikit-learn involved specifying ten principal components to capture the dataset's major sources of variance. The resulting explained variance ratios for each principal component were computed, offering insights into the proportion of variability retained in the reduced-dimensional space. The visualization of PCA results was accomplished through a scatter plot, where each sample was colour-coded based on its disease state. The colour mapping was achieved using a custom legend, providing clarity regarding the association between numerical labels and disease states.

The resulting PCA plot, exhibited at a size of 10 by 8 inches, showcased the distribution of samples along the first two principal components. The legend, created manually for all disease states, enhanced the interpretability of the plot. Furthermore, the colour bar provided a reference for mapping numerical labels to specific disease states.

Hierarchical Clustering Analysis (HCA):

Following the initial exploration with Principal Component Analysis (PCA), the investigation extended to Hierarchical Clustering Analysis (HCA), providing a more granular understanding of the inherent structures within the gene expression data. This step involved transposing the gene expression data and merging it with metadata to ensure a comprehensive analysis. Subsequently, only numeric data relevant for HCA was selected and standardized to facilitate meaningful comparisons.

Utilizing the 'ward' linkage method, HCA was performed, generating a dendrogram that visually encapsulates the relationships between samples and genes across different disease states. To enhance interpretability, disease states were incorporated as colour annotations in the dendrogram. This colour-coded representation allowed for precise identification of potential associations between gene expression patterns and specific disease conditions. The resulting dendrogram, with branches coloured according to disease states (blue for Convalescent, red for Dengue Haemorrhagic Fever, magenta for Dengue Fever, and green for healthy control), effectively showcased how samples clustered based on their expression profiles.

Moreover, the dendrogram incorporated sample names as labels at the tips, providing a direct link between the hierarchical clustering structure and individual samples. This approach aids in the identification of patterns and relationships within the gene expression data. The combination of PCA and HCA thus equips us with a multifaceted understanding of the dataset, laying the foundation for further biological interpretations and insights into disease-related patterns.

Statistical Analysis:

A statistical analysis employing t-tests was conducted to identify genes with significant expression differences across various disease states. Leveraging the versatility of Python and utilizing essential libraries such as Pandas, NumPy, SciPy, and Matplotlib, the gene expression data was meticulously aligned and subjected to t-test comparisons between pairs of disease states. The outcomes were then visually depicted through the generation of volcano plots, providing an insightful representation of the nuanced relationship between fold change and statistical significance across different disease states.

The custom volcano plots effectively distinguish between upregulated and downregulated genes, employing colour-coded markers to delineate significance levels and log2 fold change thresholds. An arbitrary threshold of log(1.5) or 0.58 was used to establish log fold change significance. This decision was made based on previous studies suggesting that this is an appropriate threshold to establish significance (McCarthy & Smyth, 2009). This strategic visualization allows for the rapid identification of genes exhibiting substantial expression disparities and their corresponding statistical significance in the context of distinct disease states. The inclusion of significance level annotations and the labelling of individual significant genes further enhance the interpretability of the plots, offering a comprehensive overview of the molecular landscape associated with each disease state.

Machine Learning Analysis:

The Random Forest classifier machine learning approach was used due to its ability to navigate complex datasets.

The initial step involved the integration of gene expression data with relevant metadata, creating a merged dataset. The subsequent preparation included defining features (gene expressions) and the target variable, emphasizing clinical relevance. To ensure robust model training, the dataset was split into training and testing sets.

The scikit-learn RandomForestClassifier played a pivotal role in navigating the multidimensional landscape of gene expression patterns. This method was intentionally chosen for its proven capacity to handle intricate datasets and discern subtle patterns within biological data.

Post-training, predictions were made on the testing set, and the model's performance underwent a comprehensive evaluation using key metrics, including accuracy and a detailed classification report.

Results and discussion

Exploring gene expression between different dengue disease states:

The PCA biplot derived from gene expression data reveals distinct clustering patterns among the various disease states, suggesting a robust biological differentiation at the molecular level. Samples corresponding to 'Dengue Hemorrhagic Fever' exhibit a tight cluster, indicating a high degree of similarity in their gene expression profiles, which could be reflective of a specific host response to this severe disease phenotype. In contrast, the 'Dengue Fever' and 'Convalescent' states are more dispersed, implying a broader spectrum of host responses or a transitional state of recovery. Notably, the 'healthy control' samples are distinctly separated from the diseased states, affirming the utility of PCA in distinguishing between normal and pathological conditions. The overlap between the 'Dengue Fever' and 'Convalescent' samples may suggest shared pathways or a continuum of the host response during infection and recovery. The clear separation of 'healthy control' indicates that the molecular signatures of health are markedly different from those during active disease or convalescence, which could be exploited for diagnostic purposes. This biplot thus provides a visual hypothesis-generating tool that suggests specific gene expression profiles are associated with each clinical manifestation of dengue, warranting further investigation into the biological processes and pathways involved.

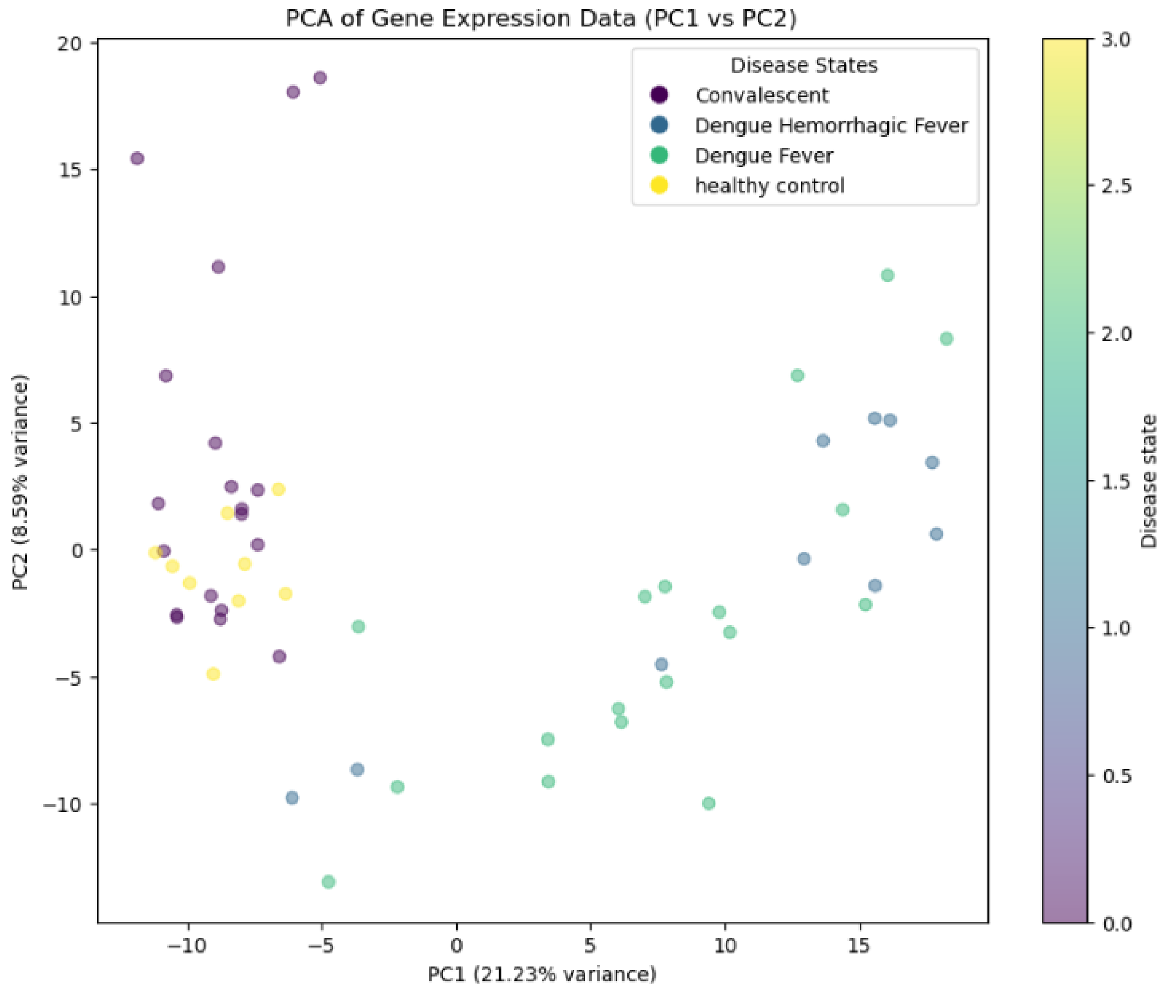


Figure 1 PCA plot illustrating gene profiles of the four disease states; convalescent, DHF, DF and healthy control between the first and second principle components

The principal component analysis (PCA) plot provided significant insights into the intricate nature of the dengue infection-associated gene expression dataset. The revelation that only 29.82% of the dataset's variance is explained by the first two principal components (PC1 and PC2) underscores the complexity of the biological systems influencing gene expression. This limited capture of variance suggests that the biological processes governing the dataset are multidimensional, extending beyond the scope of just two dimensions. The substantial unexplained variance (over 70%) emphasizes the need for further analysis, indicating that additional principal components are essential to unveil more layers of the dataset's underlying structure. Importantly, this unexplained variance holds the potential for harboring overlooked biological signals, including subtle yet biologically relevant patterns related to less common genes or intricate gene interactions. While capturing nearly 30% of the variance may be sufficient for certain analyses, the exploration beyond the initial two dimensions becomes crucial for more detailed investigations. The genes with the highest loadings on PC1 and PC2, highlighted through the biplot, serve as focal points for experimental inquiries, guiding attention to genes central to the primary variation in disease expression. In summary, the PCA analysis, though informative, highlights the complexity of the dataset and underscores the necessity for a comprehensive exploration beyond the first two principal components to unravel the full scope of underlying biological phenomena associated with dengue infection.

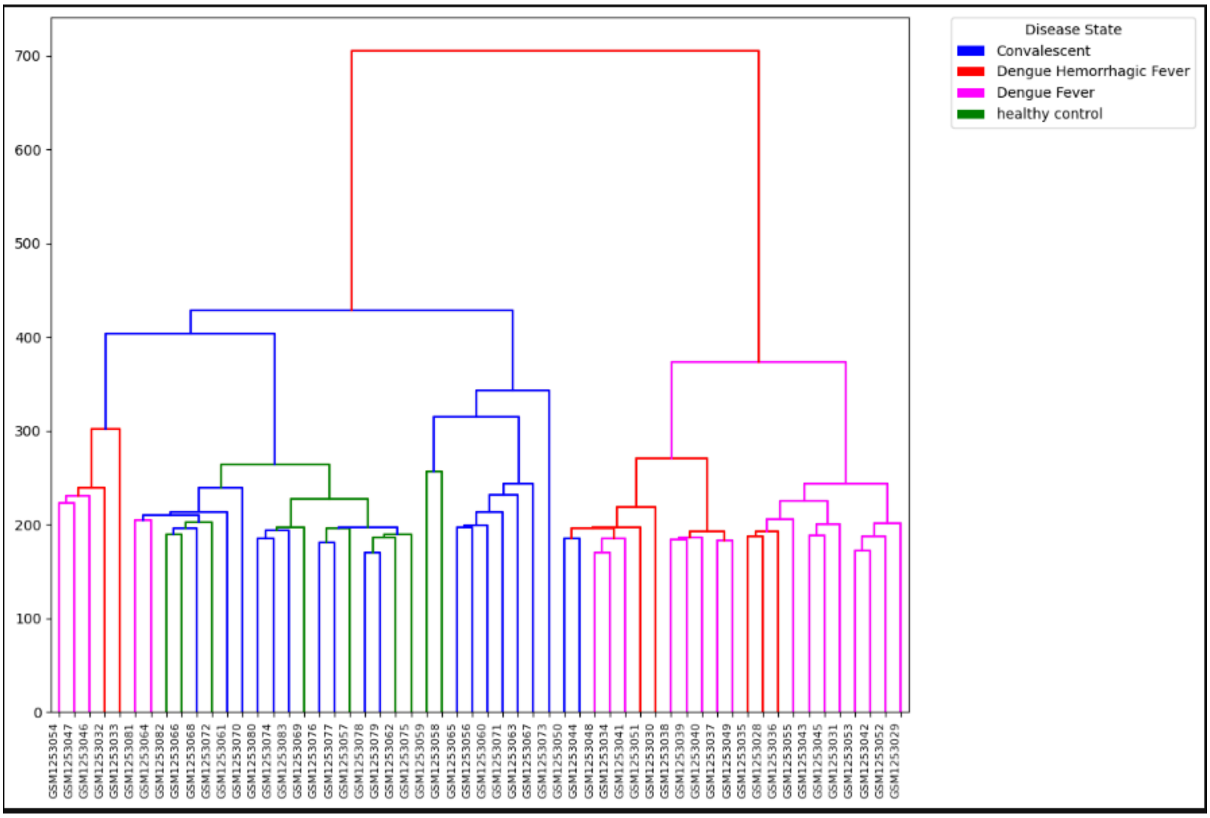


Figure 2 illustrates a dendrogram depicting the clustering of sample patients based on their gene expression profiles. The color-coded representation provides a visual differentiation of disease states.

Exploration of gene expression dynamics through hierarchical clustering analysis, as depicted in figure 2 annotated by disease states, has yielded insightful findings regarding the distinct molecular profiles associated with various stages of dengue infection. The prominent separation of samples into discrete clusters based on disease states unveils unique gene expression patterns characterizing Dengue Hemorrhagic Fever, Dengue Fever, healthy controls, and Convalescent individuals. This distinctiveness underscores the complexity and heterogeneity of the underlying molecular mechanisms at play during different phases of dengue infection. The spatial arrangement of branches, particularly between Dengue Fever and Dengue Hemorrhagic Fever, hints at potential shared molecular pathways or a continuum in disease progression. The clear delineation of healthy controls and the juxtaposition of Convalescent branches suggest identifiable gene expression shifts during recovery, reinforcing the diagnostic potential of transcriptomic data. Moreover, the clustering patterns offer a promising avenue for biomarker identification, with disease-specific clusters holding potential markers for accurate disease discrimination. The observed outliers within clusters raise intriguing questions about individual variability, co-infections, or distinct disease stages. The varying heights of branches in the dendrogram indicate dissimilarities and robustness in clustering, shedding light on the degree of molecular divergence between disease states. These findings collectively contribute to our understanding of the intricate gene expression landscape associated with dengue infection, providing a foundation for further investigations into the underlying biological mechanisms and potential diagnostic applications.

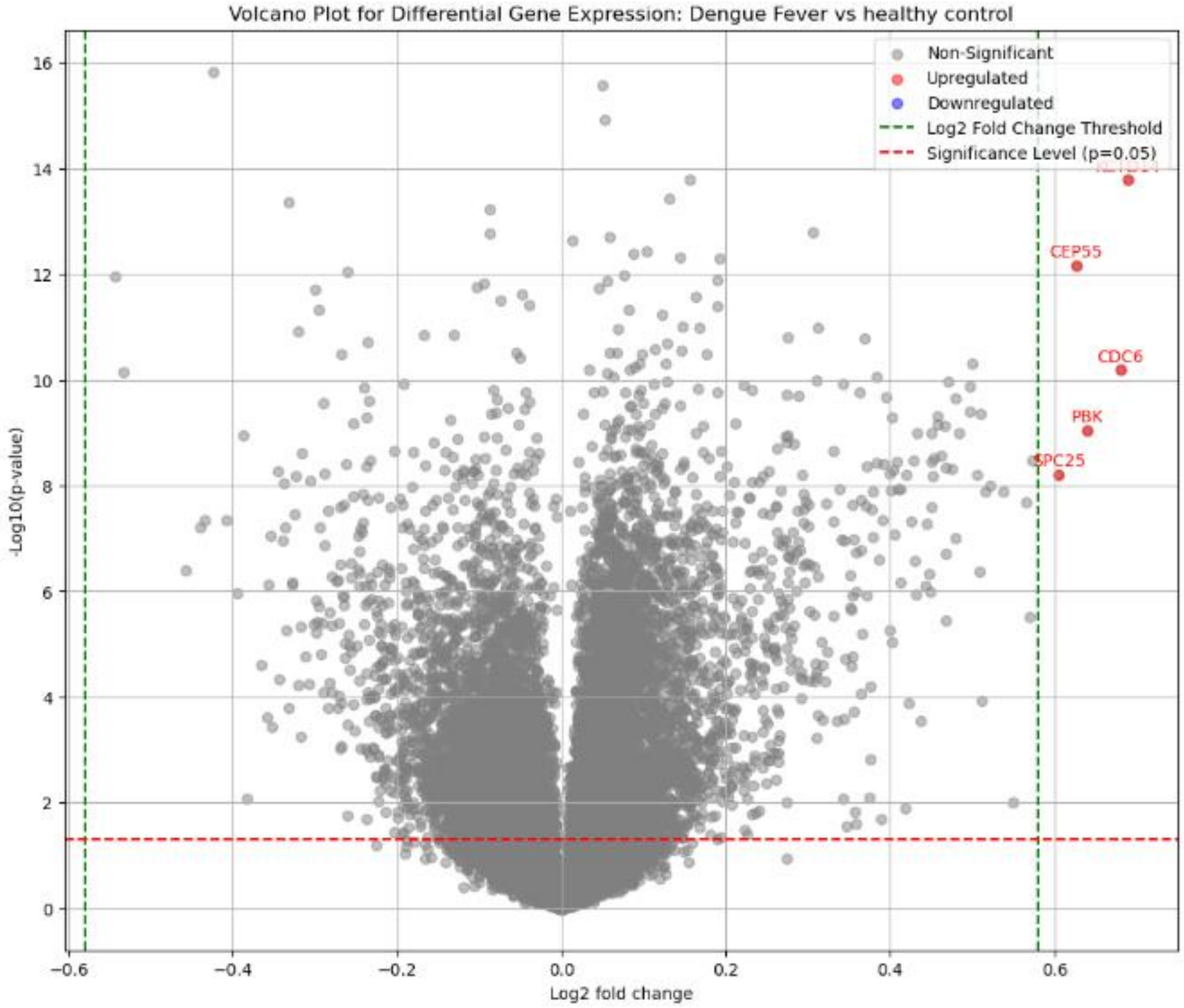


Figure 3 Volcano plot depicting differential expression results between the dengue fever and healthy control disease states where the the log2 fold change threshold is +-0.58

Figure 3 provides an exploration of the differential gene expression landscape between Dengue Fever and healthy controls. Along the log2 fold change axis, genes are positioned, revealing the magnitude and direction of their expression alterations. The negative logarithm of the p-value along the y-axis serves as a metric for the statistical significance of these expression changes. Notably, the significance threshold, set at p=0.05, demarcates genes with substantial alterations, and fold change thresholds help identify biologically relevant changes. Quadrant categorization aids systematic classification, with upregulated genes (top-right quadrant) such as CEP55, CDC6, PBK, and PCP4 emerging as potential biomarkers or therapeutic targets due to their statistical significance. This visualization offers insights into the molecular underpinnings of Dengue Fever, guiding future research directions. Importantly, most data points in the non-significant region emphasizes the discerning nature of the analysis, focusing attention on genes exhibiting significant changes. Overall, the volcano plot serves as a tool for unravelling the molecular intricacies associated with Dengue Fever, laying the groundwork for further investigations into potential diagnostic and therapeutic avenues.

For example, the KCTD14 gene belongs to the human family of Potassium Channel Tetramerization domain proteins (Angrisani et al., 2021). This is a gene which can be seen in Figure 3 to be significantly overexpressed in patients with dengue fever. Further studies into the KCTD14 gene indicate that it is a gene that play a pivotal role in the recruitment of substrates for ubiquitination.

Ubiquitination, a post-translational modification, involves the attachment of ubiquitin molecules to target proteins, marking them for various cellular processes such as degradation, regulation, or signalling (Gu & Jan Fada, 2020). In the context of KCTD proteins, their role as adapters implies that they facilitate the interaction between substrates and the ubiquitination machinery, ensuring the selective modification of specific proteins within the cell (Angrisani et al., 2021). In theory, this would allow the dengue fever to use the host cellular machinery in order to propagate. Novel gene therapeutics could be applied to stop the upregulation of this gene which could result in the disease being unable to invade the host cell.

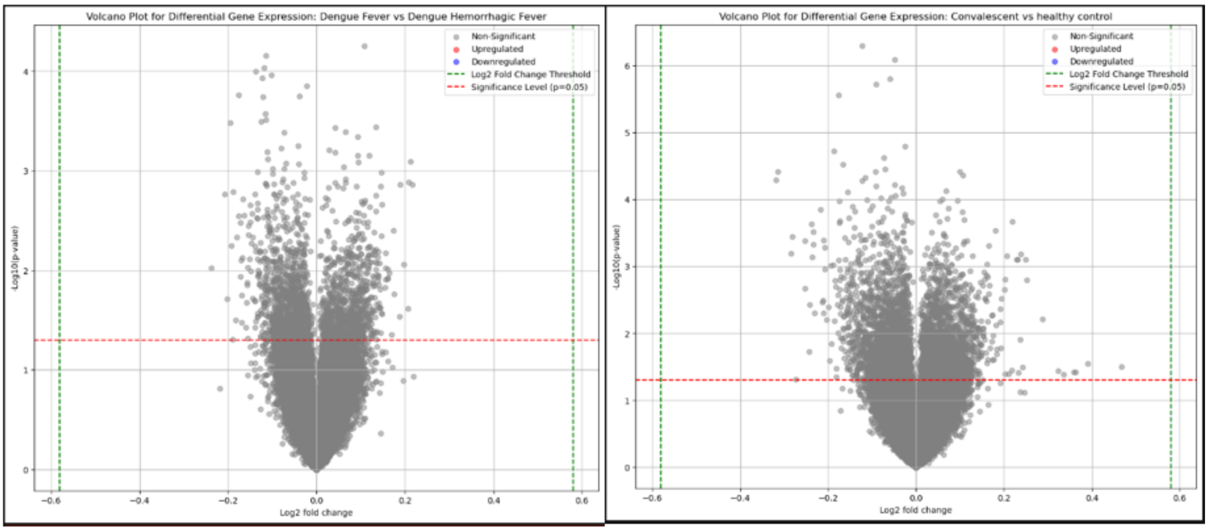


Figure 4 Volcano plot depicting differential expression results between the convalescent vs healthy control as well as DHF vs DF disease states where the log2 fold change threshold is +-0.58

Another observation made was that there were seemingly no significantly over-expressed genes between dengue fever and dengue hemorrhagic fever or convalescent vs health control disease states. This finding is also corroborated by the PCA plot in Figure 1 where there was a significant overlap between the dengue fever and dengue hemorrhagic fever as well as the convalescent vs health control conditions.

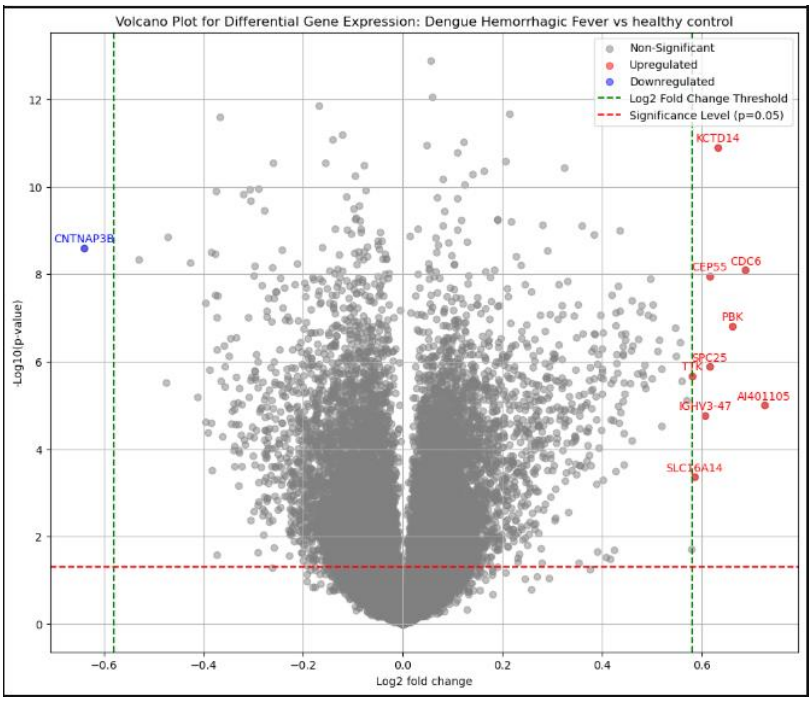
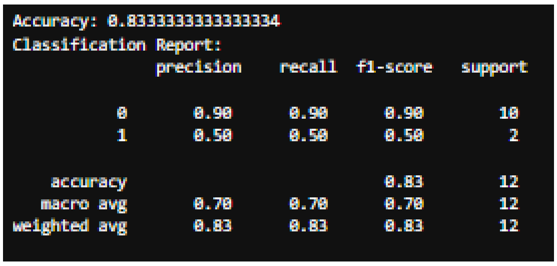


Figure 5 Volcano plot depicting differential expression results between the dengue hemorrhagic fever and healthy control disease states where the the log2 fold change threshold is +-0.58

Figure 5 displays the volcano plot for the Dengue Hemorrhagic fever against the healthy control disease state. All of the significant genes found in dengue fever vs control (figure 3) are also shown to be significant in this plot however there are additional genes which are significant that are not present in figure 3. For example, while CEP55 is shown to be over-expressed in both dengue fever and dengue hemorrhagic fever when compared to the healthy control, the SLCI6A14 gene is also shown to be over expressed in Dengue hemorrhagic fever where it is not overexpressed in figure 3.



```
Accuracy: 0.8333333333333334
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.90      0.90        10
           1       0.50      0.50      0.50         2

    accuracy                           0.83        12
   macro avg       0.70      0.70      0.70        12
weighted avg       0.83      0.83      0.83        12
```

Figure 6 Implementing a Random Forest classifier to distinguish between Dengue Fever (DF) and Dengue Hemorrhagic Fever (DHF) based on gene expression data

The random forest machine learning technique was implemented to determine whether machine learning could be harnessed to determine whether a patient is likely to suffer from dengue fever or dengue hemorrhagic fever, as seen in Figure 6. However, the results obtained in this study indicated that the machine learning technique was unsuccessful in accurately discerning between the two states as the accuracy is not near the standard needed for this method to be viable. This could be for several reasons one of which could be that, as seen in Figure 4, there are no significantly over-expressed genes between the dengue fever and dengue hemorrhagic fever disease states therefore machine learning methods could not be properly implemented. Furthermore the number of samples in this dataset could be another reason why this method didn’t work. A larger sample size will allow the algorithm to assess the variability of the genetic expression in each disease state more effectively. Furthermore, the any machine learning technique could display bias in this sample as the data was collected from just one hospital in Thailand which may not be the most representative dataset. Perhaps further study into other machine learning methods such as support vector machine learning could yield more promising results.

Conclusion

This study provides insights into the molecular landscape of Dengue virus infections, emphasizing significant gene expression differences among disease states. The identified genes offer potential targets for therapeutic discovery and diagnostic advancements. Despite challenges in machine learning discernment between Dengue Fever and Dengue Hemorrhagic Fever, the comprehensive analysis lays the groundwork for further research into specific pathways, contributing to our understanding of dengue infection dynamics.

Bibliography

Bhatt, S. et al. (2013) The Global Distribution and burden of Dengue, Nature News. Available at: https://www.nature.com/articles/nature12060 (Accessed: 15 December 2023).

Khetarpal, N. and Khanna, I. (2016) Dengue fever: Causes, Complications, and vaccine strategies, Journal of immunology research. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4971387/ (Accessed: 15 December 2023).

Page, G.P. et al. (1970) Microarray analysis, SpringerLink. Available at: https://link.springer.com/protocol/10.1007/978-1-59745-530-5_20 (Accessed: 15 December 2023).

McCarthy, D.J. and Smyth, G.K. (2009) Testing significance relative to a fold-change threshold is a treat, Bioinformatics (Oxford, England). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2654802/ (Accessed: 15 December 2023).

Angrisani, A. et al. (2021) The emerging role of the KCTD proteins in cancer - cell communication and signaling, BioMed Central. Available at: https://biosignaling.biomedcentral.com/articles/10.1186/s12964-021-00737-8#Sec1 (Accessed: 15 December 2023).

Gu, H. and Jan Fada, B. (2020) Specificity in ubiquitination triggered by virus infection, International journal of molecular sciences. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7313089/ (Accessed: 15 December 2023).

Parveen, S. Riaz, Z. Saeed, S. Ishaque, U. Sultana, M. Faiz, Z. Shafqat, Z. Shabbir, S. Ashraf, S. Marium, A. (2023). ‘Dengue hemorrhagic fever: a growing global menace.’ Journal of Water and Health, 21(11), pp. 1632–1650. < https://iwaponline.com/jwh/article/21/11/1632/98218/Dengue-hemorrhagic-fever-a-growing-global-menace>