Data Analysis and Machine Learning Insights into Dengue Fever

Abstract

This study delves into the gene expression patterns in patients affected by dengue fever and its severe counterpart, dengue haemorrhagic fever, with a dataset sourced from 56 individuals, including 9 healthy controls, the study aims to decipher patterns or genes that may elucidate biological distinctions among the populations or symptoms associated with dengue fever. Employing a combination of exploratory, statistical, and machine learning methods, the analysis encompasses Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA), Volcano Plots, and Support Vector Machines (SVMs).

The exploration begins with an overview of the dataset, categorized into patients with dengue fever, those with dengue haemorrhagic fever, individuals in recovery, and healthy controls. The study's objective is to uncover patterns through exploratory data analysis and machine learning techniques. Initial steps involve data preparation and PCA to visualise complex patterns efficiently. Subsequently, HCA is employed for clustering, providing insights into sample relationships. Volcano plots are utilized to pinpoint significant genes, one such gene showing signs of possibly being the cause of the haemorrhagic symptoms, while SVMs serve as a machine learning model to classify and distinguish between different populations.

In conclusion, the study provides a nuanced exploration of gene expression profiles in dengue patients, shedding light on potential biomarkers and pathways associated with disease progression. The combination of exploratory data analysis and machine learning techniques proves valuable in unravelling complex patterns within the dataset. While individual analyses offer insights, the integration of methods contributes to a comprehensive understanding of the underlying biology. The study underscores the significance of multi-faceted approaches in interpreting high-dimensional gene expression data and lays the groundwork for further investigations into the molecular intricacies of dengue infections.

Introduction

This study embarks on a comprehensive exploration of gene expression patterns of patients with dengue fever and its more severe form, dengue haemorrhagic fever, within a dataset published by the Gene Expression Omnibus (GEO). The whole blood sample of 28 dengue patients, of which 10 had dengue haemorrhagic fever, and as well as 9 healthy non-infected persons were analysed. 19 of the dengue patients had again provided samples 4 weeks or more into their recovery. The samples were grouped as such: (1) patients afflicted with dengue fever, (2) patients experiencing dengue haemorrhagic fever, (3) patients in the recovery phase from dengue fever, and (4) healthy controls[1]. The objective of this study is to use exploratory, statistical, and machine learning methods to understand the data and to find patterns or genes that may indicate any biological explanations for some of the differences in the populations or some of the symptoms of dengue fever.

Method

2.1 **Data Preparation and PCA**

The dataset was two files, one a dataframe which has sample names for the 56 samples as the columns and each sample had an expression level for 29777 genes. The other is a smaller metadata file, which has sample names in the rows and columns for other information such as whether the sample was from an infected person and if so, at what stage of infection as well as a patient number, disease abbreviation code and a short description. To begin we needed to visualise the data to find any obvious patterns, however, the raw data was too large for any simple plots and needed to be refined before moving forward. Principle Components Analysis (PCA) allows us to reduce the complexity of the data while maintaining as much of the information as possible, this was done by reducing its dimensionality. PCA groups several correlated variables and gives them a principal component which maximises variance while reducing data size and keeping as much variability as possible[2]. We set the number of principal components to 10 and then transposed the data to make sure the PCA is sample-centric, this way we can plot and visualise the relationship among them.

2.2 **HCA**

Next, we performed hierarchical clustering analysis (HCA), an unsupervised machine learning method that groups clusters of data points according to the set parameters. To calculate which points should be grouped we use a proximity matrix which is a table showing the shortest distance between any two points and the closest two form a cluster, each cluster forms a cluster with the nearest cluster and so on until all the points are under the same cluster. We then plot a dendrogram as it is the easiest way to visualise the clustering and allows you to see the patterns very quickly.

2.3 **Volcano plots**

A volcano plot is a type of scatter plot that, unlike the PCA or HCA, can plot a data centric graph while still being able to see patterns clearly and identify important genes. We use a volcano plot to compare the dengue haemorrhagic fever patients to the healthy control. We take the mean of both groups and find the ratio of them, the fold change. We also need the p-values for each for the y-axis. We identified some significant genes and then plotted more volcano plots between other populations to look for some common genes. The p-value for significant genes is <0.05 and the fold change threshold was >0.5 to make sure we had only the most significant genes highlighted.

2.4 **SVMs**

To create a machine learning model for our dataset we split the data and used 80% to train the model and then 20% to test the model on. Support vector machines (SVM) are supervised machine learning algorithms that seek to make sure that the maximum gap is formed when making boundaries to perform classification. It can even go scale up to another dimension if necessary. We built a linear SVM and trained it and then tested it on our test sample. To improve our model, we used bootstrapping to improve the testing and create up 100 models, storing the accuracy scores of each to plot.

Results & Discussion

3.1 **PCA plot**

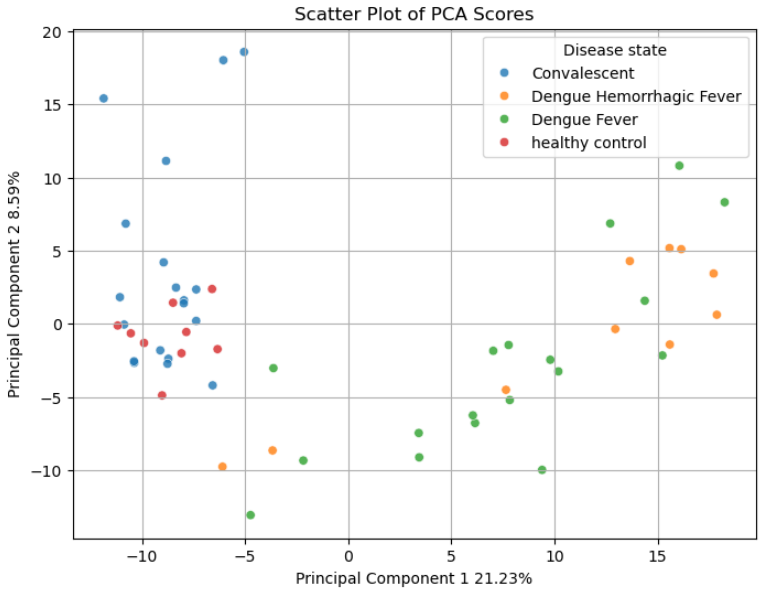


Fig. 1
PCA score scatter plot with the 4 sample groups colour coordinated, two clear groups have formed.

The figure above shows the PCA scores plot, we can see a general pattern of convalescent individuals and the control group having similar gene expression levels, and the dengue fever and severe dengue fever groups following a trend. Issues with this plot is, that only ~30% of the data is being captured in the first 2 PCAs, so too much is lost to draw significant conclusions. That's why we looked at the HCA dendrogram.

3.2 **Dendrogram**

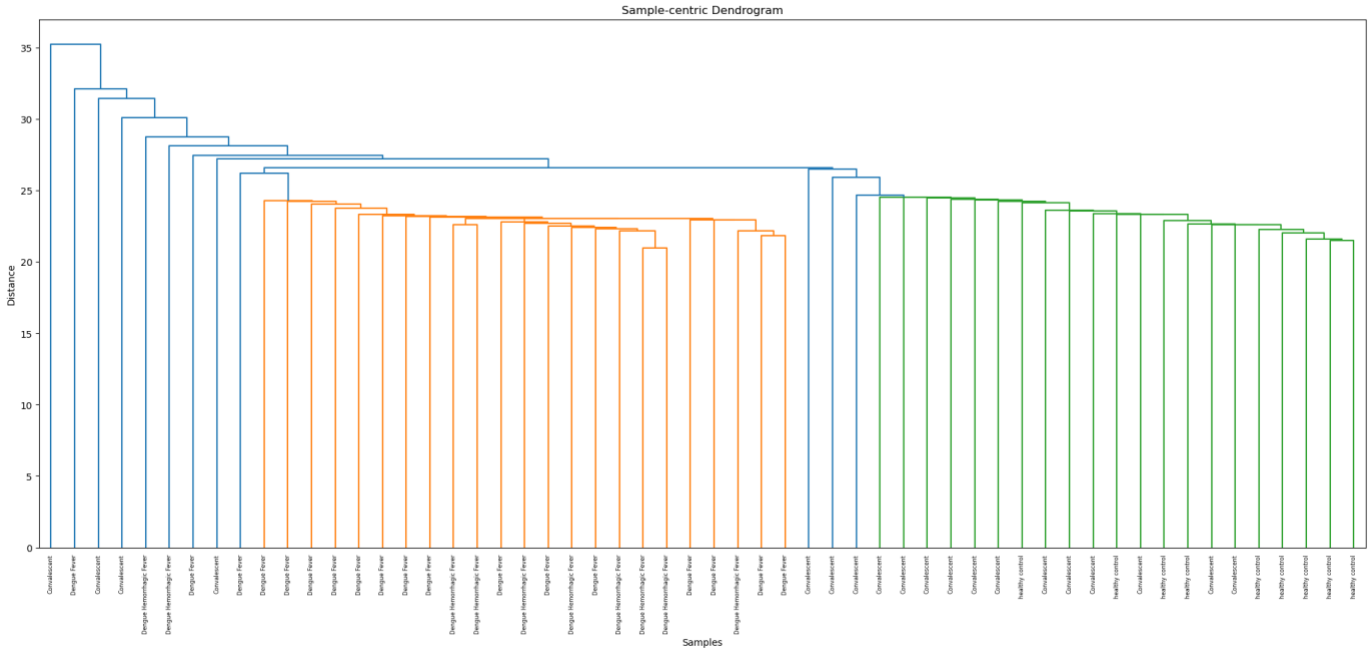


Fig. 2
Sample-centric Dendrogram
Shows two main clusters the green and the orange, green consists entirely of Convalescent and Control patients, and orange is all infected individuals.

Figure 2 is much better at showing the different clusters formed, as the outliers are shown as blue branches. The two groups consist entirely of just convalescent and control or the two types of dengue again supporting the idea we got from the PCA scores, that the gene expression of the recovering patients is returning to normal levels. Even with the dendrogram, we cannot see much difference between the different populations within the two main clusters.

3.3 **Volcano Plots**

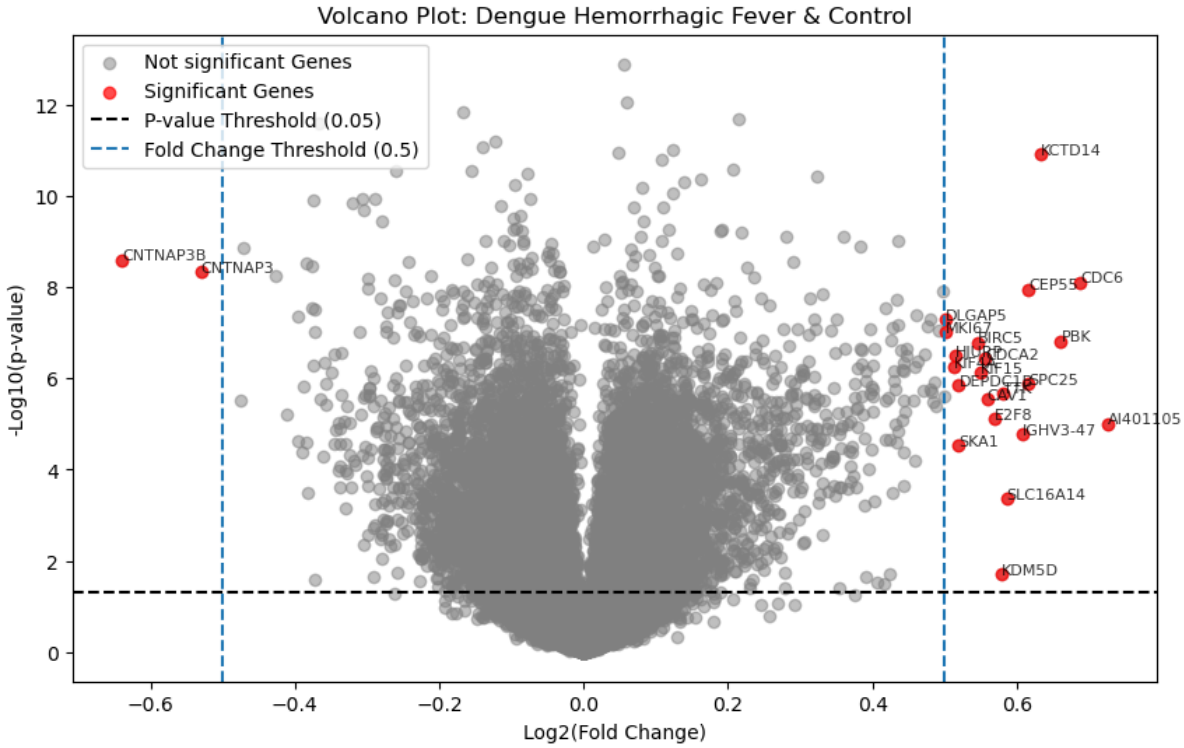


Fig. 3a
Volcano plot DHF against Control
The majority of the genes did not meet the fold change threshold.
Some highlighted genes are KCTD14, CNTNAP3B

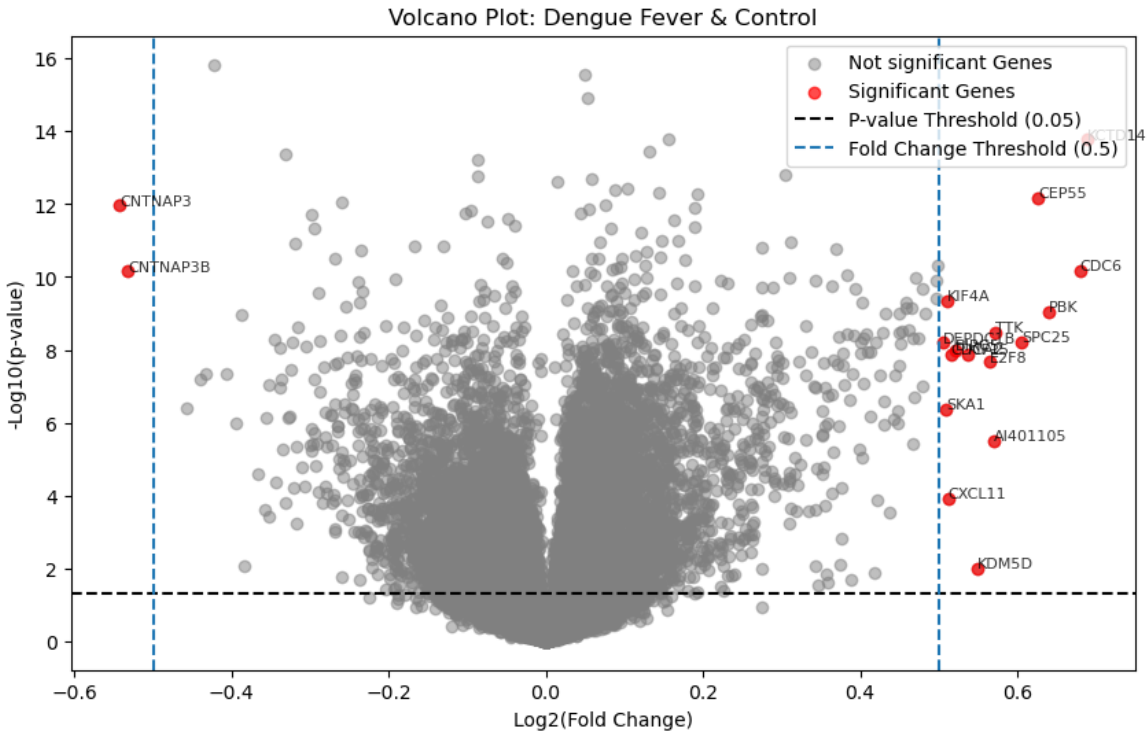


Fig. 3b
Volcano plot DF against control
Very similar to Fig. 3a
The significant genes are almost identical.

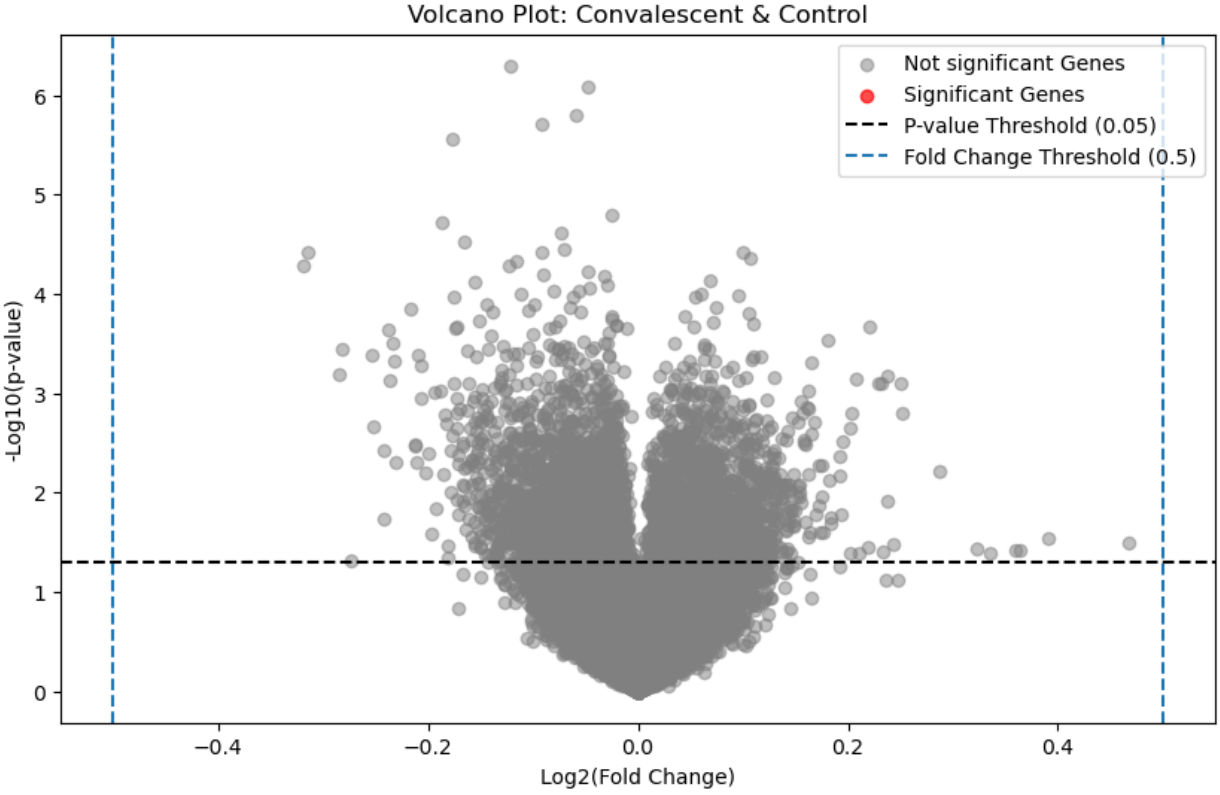


Fig. 3c
Volcano plot Convalescent and Control
No significant genes with the same threshold

Smaller distribution of genes

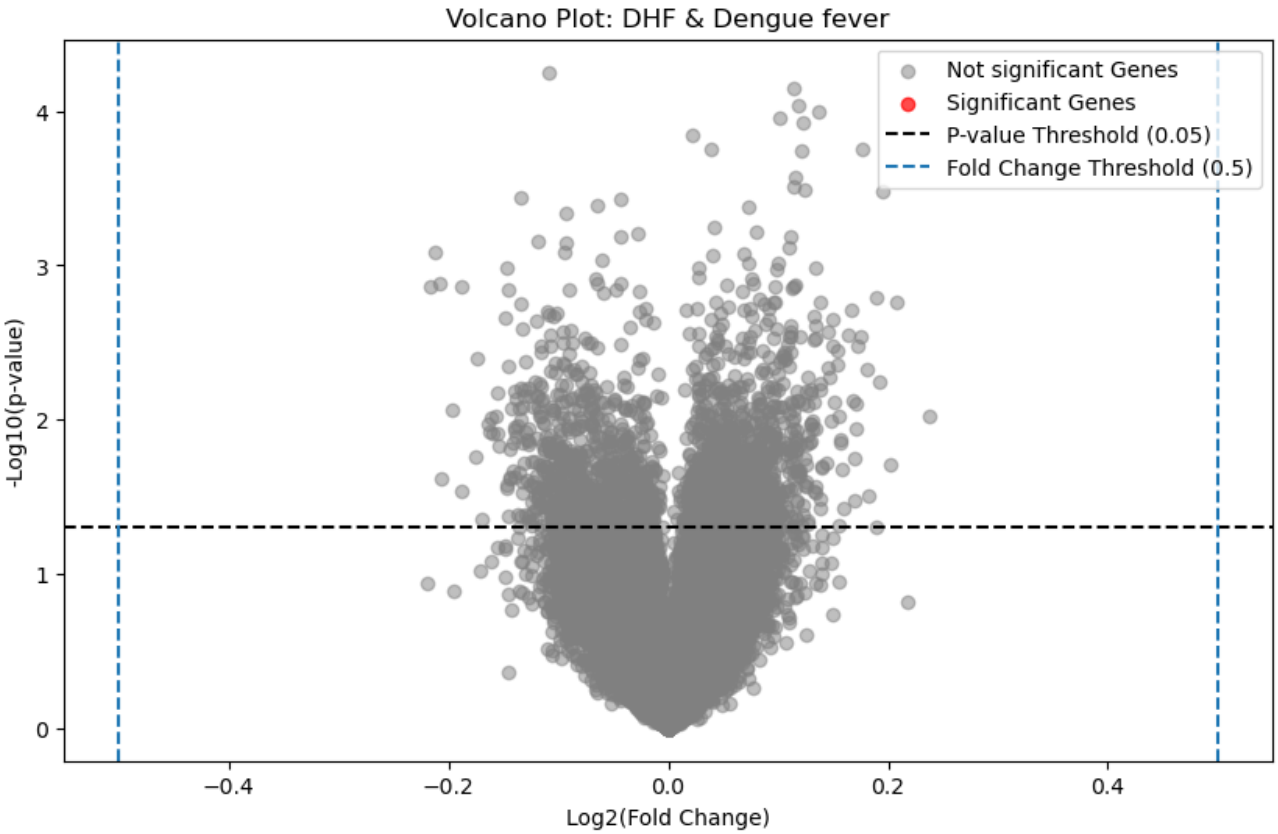
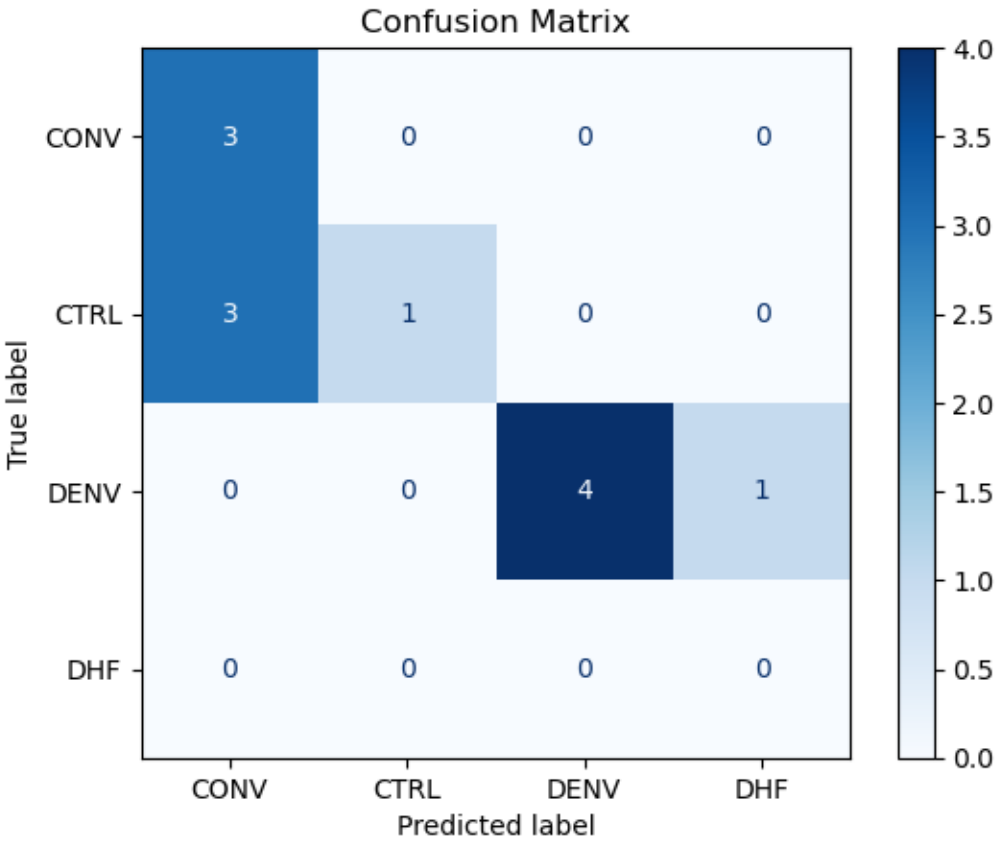


Fig. 3d
Volcano plot DHF against DF
No significant genes

The volcano plots provide very good information on the genes that are expressed more or less than the control, in the case of 3a and 3b. One of the highlighted genes that is significantly under expressed is CNTNAP3 which shows up in both DHF and DF volcano plots. CNTNAP3 is Predicted to be involved in cell adhesion and thought to be an integral component of membrane[3]. We know that one of the signs of DHF is when the patient has thrombocytopenia, which is when the platelet count in the blood is low causing bleeding all over the body. One reason the platelet count of a patient could be low is due to low plasma or water concentration in the bloodstream[4], indicating there may be issues with water uptake in blood vessels, cells, or tissues. CNTNAP3 being expressed less may be causing a cascade effect of cells being unable to take or taking too much water/solubles.

3.4 **Confusion Matrix**



Accuracy: 66.67%

	precision	recall	f1-score	support
CONV	0.50	1.00	0.67	3
CTRL	1.00	0.25	0.40	4
DENV	1.00	0.80	0.89	5
DHF	0.00	0.00	0.00	0
accuracy			0.67	12
macro avg	0.62	0.51	0.49	12
weighted avg	0.88	0.67	0.67	12

Fig. 4a
Confusion matrix
None of the test samples included DHF

Fig. 4b

The confusion matrix was not very accurate due to the small sample size, one class was unable to be tested on, so we switched to bootstrapping to train more models and make them more accurate.

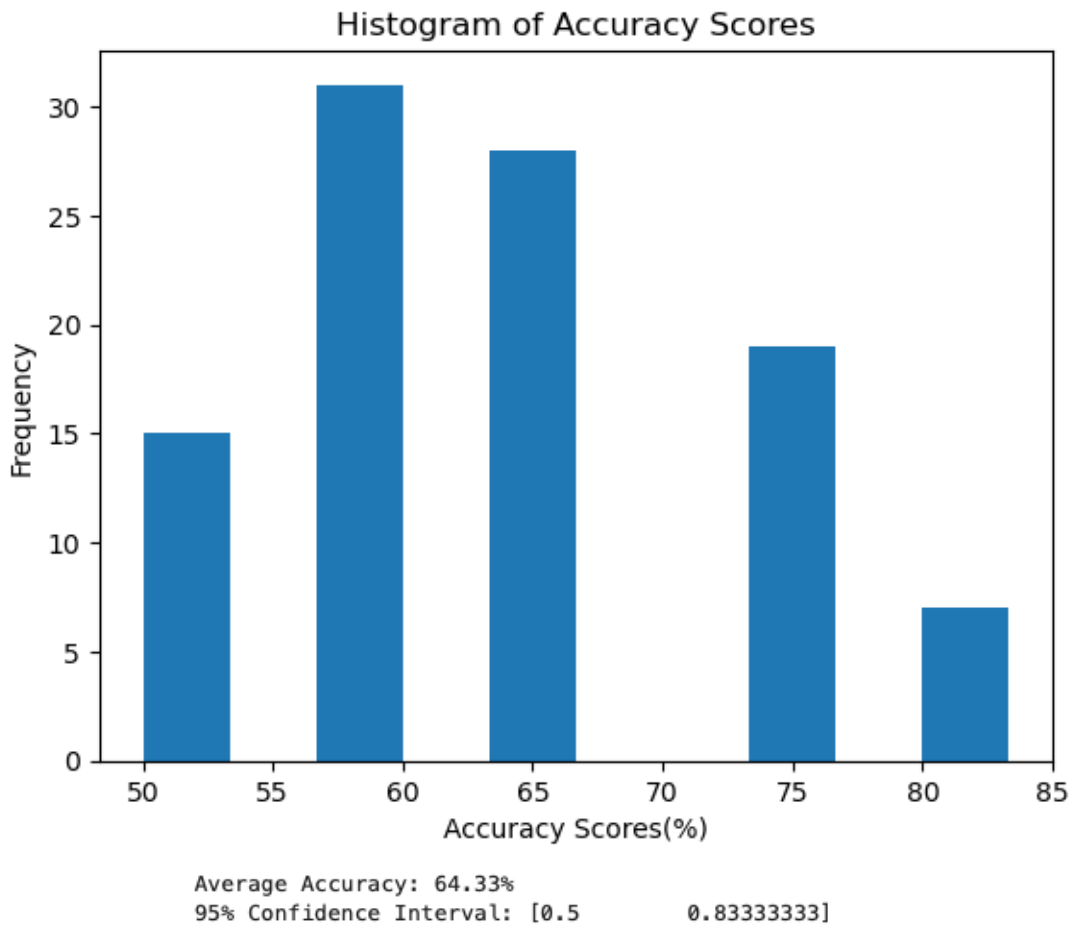


Fig. 4c
Bootstrapped SVM histogram
Shows how the accuracy of the100 bootstrap models
All over 50% - highest was 83%
Around 60% were between 57.5% - 67.5%

The bootstrapped SVMs had less average accuracy score but with 100 of them, we can take the majority classification for each test sample and be able to achieve a more accurate classifier. With this, our model should be able to distinguish between the patients with DHF and DF much better.

Conclusion

This comprehensive analysis of gene expression patterns in patients with dengue fever and dengue haemorrhagic fever reveals a multifaceted landscape with potential implications for understanding the molecular underpinnings of the diseases. The integration of exploratory data analysis methods, including PCA, HCA, and Volcano Plots, alongside machine learning techniques like SVMs, provides a rich tapestry of insights into the complex interplay of genes during infection. The application of Volcano Plots serves as a powerful tool to identify genes significantly deviating from normal expression levels. Highlighted genes offer potential biological insights. Notably, the under expression of CNTNAP3 in both dengue haemorrhagic fever and dengue fever patients raises intriguing possibilities. CNTNAP3's predicted involvement in cell adhesion and membrane integrity suggests a potential link to thrombocytopenia, a hallmark of severe dengue, pointing toward water uptake issues in blood vessels and tissues. The machine learning had some limitations and the method can be improved such as adding permutation testing to test if the model we made is any better than random chance. This is a small step in understanding how DF and DHF affect gene expression and will need much more research and data to understand all the interacting parts.

REFERENCES

1. Kwissa M, Nakaya HI, Onlamoon N, Wrammert J et al. Dengue virus infection induces expansion of a CD14(+)CD16(+) monocyte population that stimulates plasmablast differentiation. *Cell Host Microbe* 2014 Jul 9;16(1):115-27. PMID: 24981333
2. Jolliffe Ian T. and Cadima Jorge 2016 Principal component analysis: a review and recent developments *Phil. Trans. R. Soc. A.*3742015020220150202
3. *Alliance of Genome Resources (2022)*. Available at: https://www.alliancegenome.org/gene/HGNC:32035#disease-associations
4. Gubler, D. J. (1998). Dengue and Dengue Hemorrhagic Fever. *Clinical Microbiology Reviews*, *11*(3), 480-496. https://doi.org/10.1128/cmr.11.3.480