MSc ASSIGNMENT BIO720P: AI & DATA SCIENCE IN BIOLOGY 2023

Usama Khan Student No. 231164670

Abstract

This document is an assignment submission in which the genetic profiles of healthy individuals and patients suffering from Dengue Fever are explored. This exploration is done through an exploratory data analysis, advanced statistical methods and finally through a machine learning method. Our data exploration and statistical methods suggest that there are key and explainable differences between the genetic expression of patients in different disease states, however these differences need exploring in much more detail and rigour. Our exercise in Support Vector Machines shows that it is difficult to differentiate between the gene expressions of those with Dengue Haemorrhagic Fever from those with Dengue Fever- but possible. On reflection we feel that our exercise had limitations which did not allow us to fully explore every avenue of this exercise and in some instances sub-optimum decisions were taken.

Introduction

Dengue Fever (DF) is an arboviral infection[1] transmitted by the *Aedes* genus of mosquito to humans. The prominent physiological effects which, in a minority of cases, leads to febrile seizures and dehydration. Dengue Fever is a condition which can be conservatively managed. Dengue Haemorrhagic Fever (DHF) however is a significant sequela of Dengue Fever, characterised by bleeding, and is associated with a higher likelihood of systemic deterioration, developing co-morbidities as well as a having a higher mortality rate[2]. Thus these patients are more likely to have to be escalated for medical attention. Dengue Fever is not considered a significant public health burden in the UK[3], with the disease (due to its mosquito borne nature) being more prevalent in tropical and sub-tropical climates. On a global level the World Health Organisation[4] notes that prevalence is now spreading globally, particularly with Afghanistan now recognised as a new area of vulnerability, with possibly up to 400 million people a year being newly infected.

This assignment submission is an exercise into the exploration of the effects of dengue fever on gene expression. An important reason to study this is to identify biomarkers in dengue fever pathophysiology, which could aid in drug development. Additionally, with DHF being a life-threatening sequelae of dengue fever, bioinformatics might help identify these more vulnerable patients more accurately and usefully than the conventional diagnostic method of clinical judgment alone. Thus, this forms the rationale of this assignment. We will be looking at the gene expression of four disease states in relation to dengue fever. These states are: those infected with Dengue Fever (DENV), those identified with having Dengue Haemorrhagic Fever (DHF), those who are identified as being in convalescence (CONV) and a set of healthy control patients (CTRL). Using a dataset known as GEO dataset GDS5093[5], we will do an exploratory data analysis to look at basic differences in gene expression profiles between patients, we will then follow this up with some statistical analyses (with volcano plots) examining if any individual genes differ significantly in expression between the four disease states. We will then carry out a machine learning exercise to see whether such methods can be used to distinguish between Dengue Fever and DHF patients.

Methods

Question 1

To what extent does the gene expression profile of a patient differ between the four populations?

For this question we are doing an exploratory data analysis.

We initially approached this analysis by generating histograms. Histograms allow for a quick assessment of gene distribution in different disease states. We utilised the package 'ggplot2'. Our histograms would provide a visual representation of gene expression in different disease states. This could allow for comparison of gene expression in different disease states. This initial exploration of this data set can allow us to look at central tendency, skewness and spread. We can also identify any potential outliers.

We then switched to Jupyter notebook (and persisted in a Jupyter notebook for the rest of the exercise) and did a Principal Component Analysis (PCA). By focusing on principle components we can reduce noise and subsequently visualise our data easier. Given the complexity of our data set, the PCA simplified the data allowing us to identify patterns more easily. For the sake of being able to visualise our gene expressions we produced: a PCA bi-plot, a scree plot, a Hierarchical Cluster Analysis and a heat map. These visualisations can assist in pinpointing gene clusters that might reveal interdependent relationships, as well as determining if there's an association between various genes and distinct disease conditions.

Critiquing our methodology, we initially started off working in R and produced our histograms within this software. R was chosen initially because of the ability to more easily produce visualisations. Indeed to this effect, we regret not having persisted in R as we did not manage to colour the branches of his dendrogram (which may have helped visualise relationships between gene expressions) within Python (due to lack of technical expertise) but feel we may have been able to have done this within R. In addition to perhaps having missed out on advantages of R, by switching to Python, we recognise that switching between the two softwares may be a little disorientating to anyone following our exercise and may decrease the reproducibility of this exercise. As well as technical skill, we feel that the time pressures also meant that our ability to revise some of our choices were limited. Had we had more time we would have liked to persisted with working within R.

Question 2

Which, if any, individual genes differ significantly in expression between the four disease states? Volcano plots would be useful here. If there are differences, what do they mean biologically?

As part of our statistical analyses we decided to produce volcano plots. The aim of doing this was to identify any individual genes which differ significantly in expression between diseases states. Again, due to reasons to do with familiarity of producing volcano plots within Jupyter, we continued in Jupyter notebooks. Volcano plots allow for easily visualisation of up and down regulated genes; in addition by combining this with fold changes and p-values, we can easily see significant and extreme values.

We decided to compare our healthy control patients against all other disease states to try and identify any significant gene expressions. Cognisant of the need to compare dengue fever against dengue haemorrhagic fever in the upcoming task we also compared these two groups. Thus we

produced four volcano plots of interest: Dengue Fever vs Healthy Control, Dengue Haemorrgahic Fever vs Healthy Control, Convalescent vs. Healthy Control and Dengue Haemorrgahic Fever vs. Dengue Fever.

After having installed numerous packages our code then performed anova tests. We used a p-value of 0.05. We used this threshold as this is conventionally used. We used a log-fold change value of 0.5, again used as it is a common threshold. With more time we may have better researched and contemplated these parameters. We used the Benjamini-Hochberg method to adjust our p-value in light of controlling for false positives. We calculated log fold changes. Our volcano plots had a blue axis to show values beyond our p–values and green axes showed genes beyond our log fold changes. Our significant genes were highlighted in red and labelled. Although we could have visually inspected what the significant genes were, to reduce errors, our code printed off lists of significant genes.

In hindsight, using *anova* was not a suitable method given we were only using two variables. On investigation we realised that although we had employed an anova method this would function the same as t-testing. On reflection we ought to have used code which would have directly done t testing. Not only would this have been more straightforward for those following our exercise but possibly may have been computationally more economical-we note our code take around an hour to execute at times (although we have observed that given the large dataset this code may not have been swift to execute in any case).

The significant genes which were found were highlighted in red, labelled and also printed. Given that we had so few significant genes produced we decided to tabulate our genes (fig 11). From visual inspection of which gene was the most significant and explored this gene in further detail. We realised that the flaw in doing this is that we may actually be selecting outliers. In addition rather than looking at which genes are the most significant 'by eye' a better way of doing things may be to look at the p-values. Had we had a larger number of significant genes we have have considered doing an enrichR analysis. Indeed due to the aribitrary nature of selecting 'conventional' p-values for our analysis, we may have found many more significant genes had we selected different values on contemplation and research.

Question 3

From a clinical treatment perspective, it would be useful if transcriptomics data could be used to distinguish between patients with dengue fever and dengue haemorrhagic fever. Is it possible to do this using machine learning?

To answer this query we had to embrace a machine learning (ML) method and explore whether a method could be relied on to identify, on a genetic level, the differences between Dengue Fever and Dengue Haemorrhagic Fever.

We chose to use a Support Vector Machine (SVM) because of its good performance with small sample sizes, better overfitting control and our familiarity with this method.

Again we worked in a Jupyter notebook. As for other questions we pre-processed our data, merging our datasets and then filtering only for data related to Dengue haemorrhagic fever and Dengue Fever. We used train_test_split function from scikit-learn to split our data into a training (70%) and testing set (30%). Our training data was standardised/normalised using *StandardScaler.* We had an SVM model (called SVC) trained on this using the fit model. Our trained model was then used for prediction ((X_test_scaled). Accuracy scores were then calculated by comparing to our test set.

To account for the accuracy and significance of our model we also integrated bootstrapping and permutation testing as part of our model. This provides us insights into the reliability of our model which in turn will help us consider the question of whether a machine learning method could be of use in distinguishing between Dengue Fever and Dengue Haemorrhagic Fever.

Results and Discussion

Question 1 Results

Fig 1- A histogram depicting the variability in gene expression amongst healthy individuals.

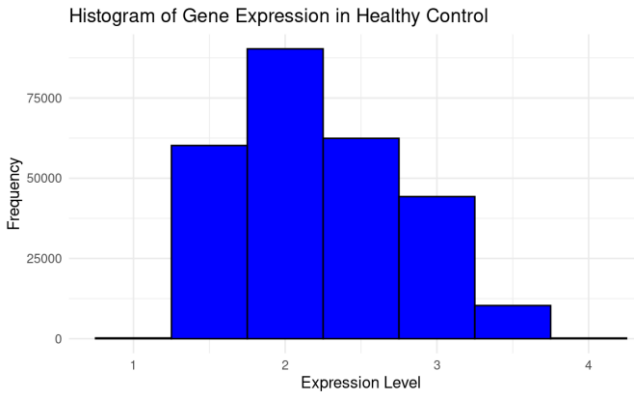
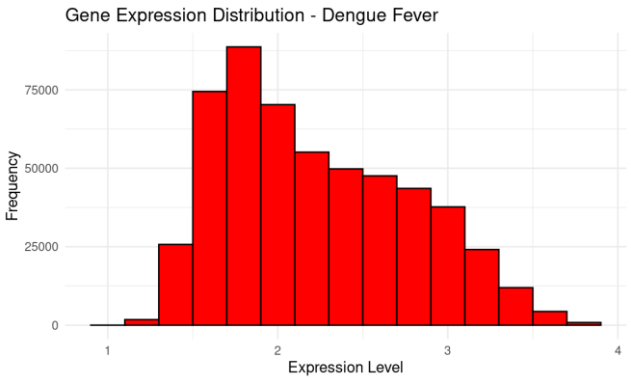
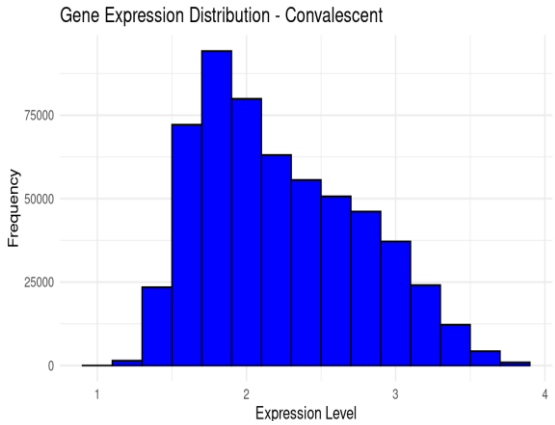
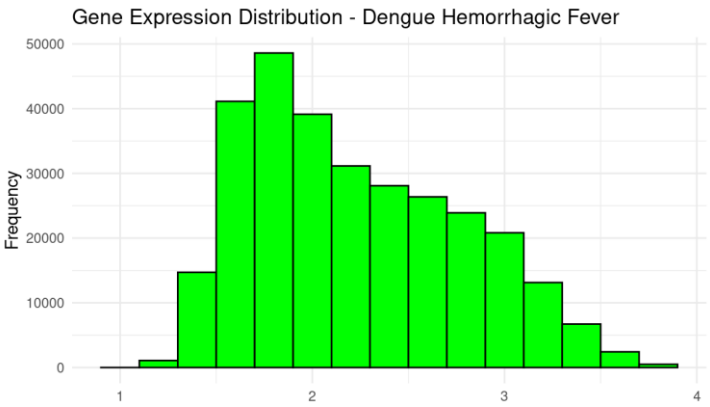


Fig 2- A histogram depicting the variability in gene expression amongst patients suffering from Dengue Fever.



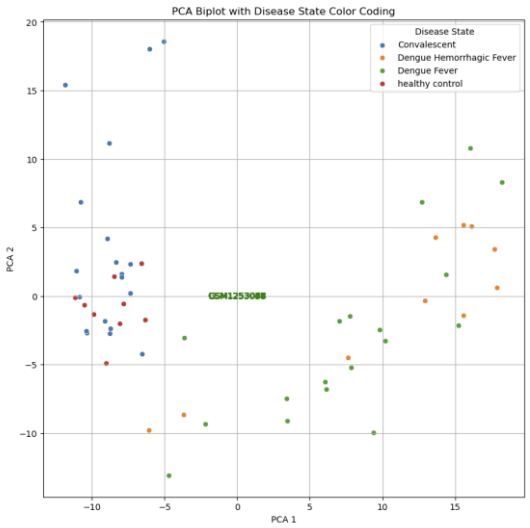




The histograms that we produced within R illustrate gene expression across our four states of disease. Figure 3 which represents our convalescent phase, shows 14 bins/columns, shows the broadest range of expression. This could be due to a broad level of biochemical processes that may be needed in the recovery phase. Figure 1, depicting healthy controls, shows a more consistent expression profile with a narrow range of expression levels- we can use this as a benchmark for other disease states. This distribution could be explained by the patient being relatively more stable haemodynamically, thus limiting the number and type of cellular pathways being active. This contrasts with other histograms which all have more varied expression levels, perhaps reflective of the more complex position the body is in and the need to activate more biochemical pathways.

Figure 4, corresponding to dengue haemorrhagic fever (frequency of less than 50,000), shows a substantial decrease in expression frequency (other graphs such as DF show a frequency of more than 75000), potentially indicative of the profound impact this severe disease state has on systemic biological functions. It must be noted that these explanations for what we can see are simply educated guesses to explain the differences between gene expression levels. If there was a greater luxury of time we would have liked to have tested these explanations by looking at whether our findings align with other datasets and what current literature has to say about such findings.

Fig 5- A PCA Biplot of Gene Expression Data across different disease states.



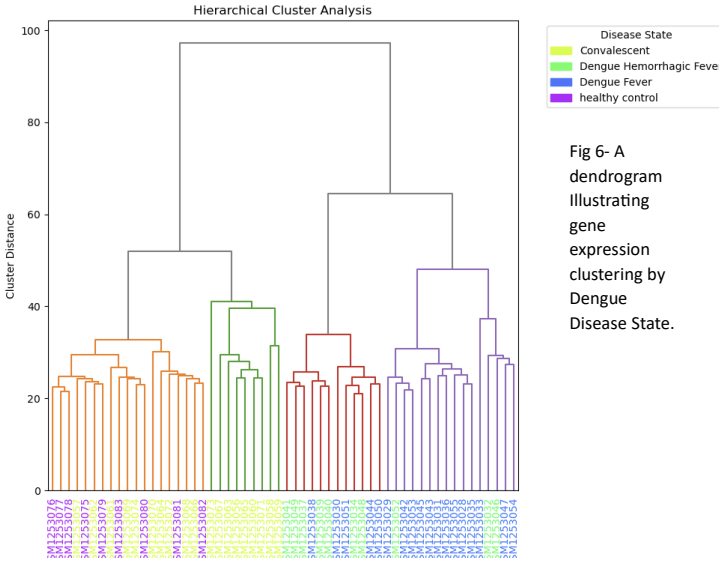


Fig 6- A dendrogram Illustrating gene expression clustering by Dengue Disease State.

In the PCA plot above (figure 5). We can see that healthy and convalescing patients (red and blue coloured points respectively) cluster closely whilst patients suffering from dengue fever and DHF (green and orange points) similarly cluster closely. Our colour coded dendogram similarly shows two clusters of gene expressions and in those two clusters we have two overlapping and clustering disease states. Again if we look at the heatmap we see a similar pattern of two subgroups. This has a very obvious explanation that the recovering patient has a similar gene expression profile to a healthy patient and the same can be said for the patients suffering from the Dengue Fever and DHF.

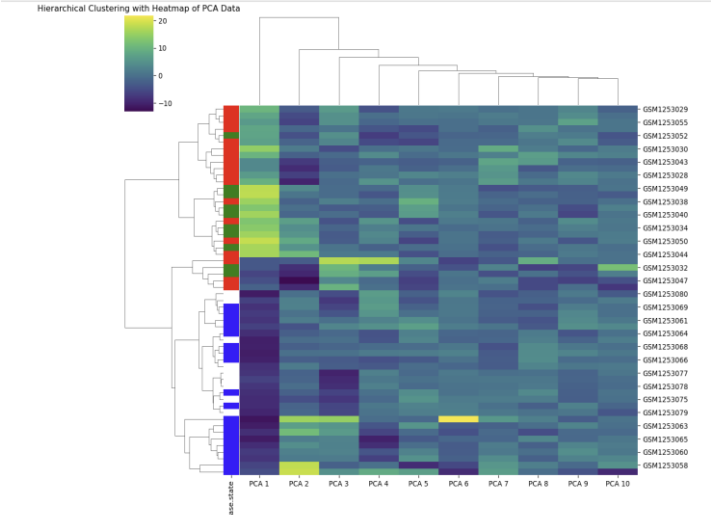


Fig 7- A heatmap with hierarchical clustering based on PCA-Transformed Gene Expression Data

However if we compare these inferences to the inferences we made in our histograms we can also comment in addition that the recovering patient has a larger range of gene expression, perhaps as a result of the more complex processes involved in recovering. At that point in the exercise we also noted a low frequency in gene expression in Dengue Haemorrhagic Fever as compared to Dengue Fever. The similarity in gene expression makes sense in light of DHF being a more severe variant of DF, with both conditions lying on a continuum (being differentiated through clinical judgement) rather than as two clearly distinct pathologies. The lower levels of gene expression in DHF, we feel could be explained by the patients being dehydrated, undernourished leading to decreased physiological activity and thus a decreased frequency of gene expression. Again we stress the need to validate our inferences through a literature review.

Looking ahead to question 3 we can see the need to see if we can use ML methods to differentiate to DHF from dengue fever. By seeing how these two conditions have clustered gene expressions reflecting the need to develop a sophisticated method to develop a method of differentiating between patients through biochemical methods. The close clustering also may highlight the difficulty in doing this, something which we will consider in further detail in the next questions.

Question 2 Results

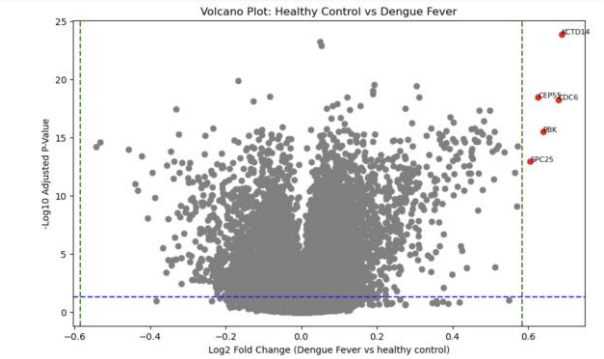


Fig 8- A volcano plot comparing gene expression between our healthy control patients and our dengue fever patients.

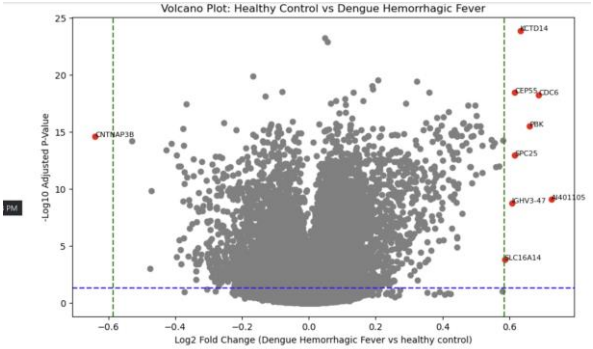


Fig 9- A volcano plot comparing gene expression between our healthy control patients and our DHF patients.

Fig 10- A volcano plot comparing gene expression between our DF and DHF patients

.

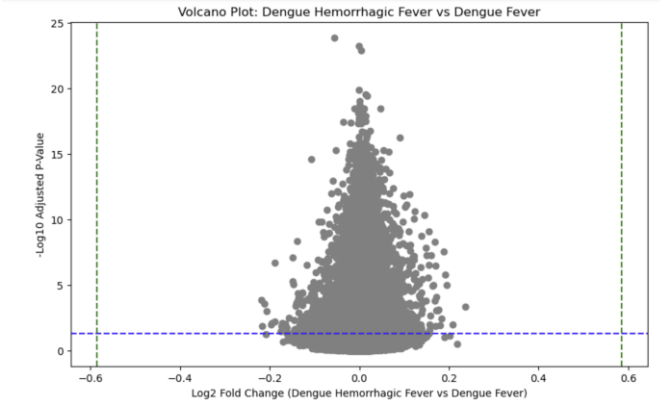


Fig 11 – A table showing the collection of significant genes found in our volcano plots

	Volcano plot	
Row No.	CTRL v DF	CTRL v DHF
1	BIRC5	BIRC5
2	CDC6	CDC6
3	TTK	TTK
4	SPC25	SPC25
5	CEP55	CEP55
6	SKA1	SKA1
7	KIF4A	KIF4A
8	PBK	PBK
9	KIF15	KIF15
10	KCTD14	KCTD14
11	E2F8	E2F8
12	CNTNAP3	CNTNAP3
13	CDCA2	CDCA2
14	DEPDC1B	DEPDC1B
15	AI401105	AI401105
16	CNTNAP3B	CNTNAP3B
17	CXCL11	IGHV3-47
18		SLC16A14
19		CAV1
20		DLGAP5
21		HJURP
22		MKI67

As mentioned in our methods section we only had two volcano plots which showed us any significant genes. Three of our volcano plots were measuring the different disease states with our control. It was in two of these plots, Dengue Fever vs Healthy Control (Fig 7) and Dengue Haemorrhagic Fever vs Healthy Control (Fig 8) that we found any significant genes. On our volcano plots these are shown as red and labelled on our graph. These are those genes which lie beyond our p-value (blue axes) and our log fold changes (green axes). Although they are labelled our code printed these off. These are tabulated. Note that the list of significant genes is mostly similar. As figure 11 shows it is after the 16th gene in our lists (the genes similar in both lists are listed first) that the list of the significant genes diverges. We found 17 significant genes in dengue fever vs control and 22 significant genes in dengue haemorrhagic against our control.

With 16 out of the 17 significant genes found in dengue fever v control analysis also being found in dengue haemorrhagic v control analysis. This would be in line with the assumption that that most of the biological processes in the patient with just dengue fever are also present in dengue haemorrhagic fever (due to DHF being on the spectrum of DF disease). We imagine if further research were to be done these genes may be linked to inflammatory processes.

Having looked at our volcano plot we found that our most significant gene was *KCTD14*. An NCBI[6] search reveals that this gene is linked with protein homooligomerisation. Sadly, as discussed in our methodology, we did not feel that we could explore this subject in any further detail. We feel that an enrichR analysis would have been suitable for this task.

Conversely there are many other biological processes that are in dengue haemorrhagic fever compared to dengue fever, again making sense given that there is the presence of symptoms such as bleeding (as discussed in our introduction) which may not be present in Dengue Fever. Perhaps if we were to look into this further we may find the genes only associated with DHF v control to be associated with haemostatic processes. The finding of significant genes only present in DHF also presents the chance to associate DHF with unique biomarkers, which itself could be a stepping stone to using bioinformatics for DHF assessment and prediction.

One of our other volcano plots (Figure 10) however goes against this suggestion as it does not show any significant genes when we compare the groups Dengue Fever against Dengue Haemorrhagic Fever. Indeed the shape of the plot with a narrow symmetrical shape with no separation in the middle may indicate that there's not extreme differences between the two populations. This is in line with data from our other analyses where these two sub-groups were shown to have an overlap in gene expressions.

The lack of significant genes presents both a need to develop an AI which may differentiate between DF and DHF biochemically but also may present a challenge as we attempt to build an ML off this specific dataset to differentiate between the two conditions. The lack of any significant genes between the two populations may mean we have a sub-optimum set of testing data.

Question 3 results

My results for question 3 are:

Average bootstrap accuracy: **0.64**

Average permutation accuracy: **0.63**

The bootstrap accuracy suggests that the SVM distinguishing between Dengue Fever and Dengue Haemorrhagic Fever 64% of the time.

The permutation accuracy suggests that the models performance of being able to distinguish between the two conditions is only slightly better than chance.

Thus whilst our SVM does display some predictive power, ultimately it looks like our model is highly unreliable and not suitable for use for distinguishing between the two conditions.

To answer the question of whether we can use machine learning methods can be used to distinguish between the two conditions I wish to answer with a conditional 'no'. The condition being that with our current dataset, the quality of data is not good enough to allow for the training of a machine learning model which can distinguish between the DF and DHF. However, our findings of genes which are in DHF but not in DHF highlights the possibility that this may be possible. Perhaps with a large enough dataset there would be enough significant genes which may allow for a better training set and thus a reliable machine learning model.

Reflecting on our exercise with what is available in scientific literature we have found another paper titled '*Classification of Dengue Fever Patients Based on Gene Expression Data Using Support Vector Machine*'[7]. As the title suggests these researchers behind this paper had a similar aim to ours. Using an SVM and a leave-one-out cross validation, these researchers achieved a model accuracy of 96%- far better than our model. They identified the following genes to be of significance, and useful to the success of their ML model, between the two datasets: MYD88, TLR7, TLR3, MDA5, IRF3, IFN-α and CLEC5A. We found none of these genes when comparing the two subgroups to each other or to our control (fig 11). It is beyond the scope of this exercise, due to time limitations, to examine the differences between their study and ours. Of particular interest would be to compare our datasets and examine the presence of these significant genes within our dataset. Their study does leave the door open for the possibility, with the right technical skill and right dataset for a machine learning model to be developed which can differentiate between Dengue Fever and its more insidious variant Dengue Haemorrhagic Fever.

So to reiterate we are saying with *our* limitations it is not possible to use machine learning to distinguish between Dengue Fever and Dengue Haemorrhagic Fever, however given what our literature search has shown us, we do feel that it is possible for others to do so.

Conclusion

In conclusion we have found that, using the dataset GDS5093, there are differences, in terms of gene expression, between the sub-groups of the patient who: is healthy without disease, is convalescing from Dengue Fever, has Dengue Fever or has Dengue Haemorrhagic Fever. The two subgroups of healthy and convalescing patients are closely aligned. Similarly so are the subgroups of Dengue Fever and Dengue Haemorrhagic Fever. We found that there are significant differences in gene expression between those who are suffering from Dengue Fever or Dengue Haemorrhagic Fever and our control population. Whilst it appears that there is not a huge need for identifying those from dengue haemorrhagic fever from healthy patients (as it is either asymptomatic or clinically obvious) we were hoping it is possible to differentiate between those who have or are at risk of dengue haemorrhagic fever from those who have dengue fever. Unfortunately in comparing our two subgroups of DF and DHF we did not manage to find any genes of significance. This may be a factor in explaining why our machine learning model had such a low accuracy (64%)- it may be possible that our testing set is sub-optimum. That being said however a literature search has brough forth another similar exercise which has had an accuracy of 96%- demonstrating that it is possible to develop machine learning model possible of differentiating between variants of dengue fever. Although we have answered a 'no' to whether 'we', with our technical acumen, resources and data, can develop such a model- we definitely think this is possible. We hope others can learn from our exercise to consider their approach and hopefully develop successful models.

Throughout our exercise we have constantly faced limitations in technical expertise and the limitations of time. In addition had we had more time, in hindsight we would have liked to have revised some our approaches. At every step of the way we reflected on the need to validate some of our thoughts with a literature search, which would perhaps validate some of our commentary, however time pressures sadly did not enable us to have done so.

References

1. Dengue fever [Internet]. [cited 2023 Dec 15]. Available from: https://bestpractice.bmj.com/topics/en-gb/1197

2. [Internet]. [cited 2023 Dec 15]. Available from: https://www.cdc.gov/dengue/resources/healthcarepract.pdf

3. England PH. Dengue fever: Guidance, data and analysis [Internet]. GOV.UK; 2008 [cited 2023 Dec 15]. Available from: https://www.gov.uk/government/collections/dengue-fever-guidance-data-and-analysis#:~:text=The%20symptoms%2C%20diagnosis%20and%20epidemiology%20of%20dengue.&text=Dengue%20fever%20(also%20known%20as,infection%20spread%20by%20Aedes%20mosquitoes.

4. Dengue and severe dengue [Internet]. World Health Organization; [cited 2023 Dec 15]. Available from: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue#:~:text=About%20half%20of%20the%20world's,urban%20and%20semi%2Durban%20areas.

5. [Internet]. NCBI; 2014 [cited 2023]. Available from: https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5093

6. KCTD14 potassium channel tetramerization domain containing 14 [homo sapiens (human)] - gene - NCBI [Internet]. U.S. National Library of Medicine; [cited 2023 Dec 14]. Available from: https://www.ncbi.nlm.nih.gov/gene/65987

7. Gomes AL, Wee LJ, Khan AM, Gil LH, Marques ET, Calzavara-Silva CE, et al. Classification of dengue fever patients based on gene expression data using support vector machines. PLoS ONE. 2010;5(6). doi:10.1371/journal.pone.0011267