Dengue Fever Prognosis Study

Kshitij Kadam
Texas A&M University
Department of Computer Science
and Engineering
kkadam3@tamu.edu

Adekola Okunola
Texas A&M University
Department of Electrical and
Computer Engineering
phirlly@tamu.edu

Abstract

Dengue infection affects millions worldwide and can often escalate to severe forms such as Dengue Hemorrhagic Fever (DHF). This escalation necessitates early differentiation from Dengue Fever (DF), despite their overlapping clinical presentations. Gene expression profiling during the febrile phase uncovers distinct transcriptional signatures; DHF patients exhibit diminished activation of innate immunity genes alongside heightened expression of apoptosis-related genes. By employing ANOVA to identify key discriminatory genes, linear discriminant analysis (LDA) and Support Vector Machine (SVM) models built from the most effective gene pairs demonstrate high accuracy in distinguishing between DHF and DF. These computational approaches facilitate the early and precise identification of severe cases, which supports timely intervention and optimizes resource allocation. Such bioinformatics-driven strategies advance the field of precision medicine, helping to alleviate both the health and economic burdens of dengue in endemic regions.

Introduction

Dengue virus infection presents a significant global health challenge, with an estimated "100 to 400 million infections occurring" [3] annually, primarily in tropical and subtropical regions. The illness manifests in two main clinical forms: Dengue Fever (DF), a self-limiting febrile illness, and Dengue Hemorrhagic Fever (DHF), a more severe condition characterized by plasma leakage, hemorrhagic manifestations, and potentially fatal shock syndrome [2]. Prompt identification of patients at risk of progressing to DHF is essential for timely medical intervention; however, this remains difficult due to the overlapping clinical features during the febrile phase.

The high-dimensional nature of the gene expression dataset, comprising 1981 genes from 26 patients, necessitates the use of robust analytical methods to identify and leverage key discriminatory features effectively. Statistical approaches like ANOVA are instrumental in detecting genes with significant differential expression, while machine learning techniques such as LDA and SVM offer structured frameworks for building predictive models. These models utilize gene pairs with the highest discriminatory power to classify patients into Dengue Fever (DF) or Dengue Hemorrhagic Fever (DHF) categories with high sensitivity and specificity.

This project seeks to combine statistical feature selection with classification algorithms to address the diagnostic challenges associated with distinguishing DHF from DF. By emphasizing top-performing gene pairs identified through ANOVA and employing LDA-based with linear SVM classifiers, the objective is to develop accurate and interpretable models for early risk stratification. Preliminary findings suggest the viability of this approach, as classifiers demonstrate strong performance metrics while also providing biologically relevant insights into the molecular mechanisms underlying DHF and DF.

This study employs bioinformatics and computational tools to align with the global objectives of precision medicine, thereby providing scalable and cost-effective strategies for the management of dengue. By enabling the early identification of severe cases, these methodologies possess the potential to optimize resource allocation, mitigate healthcare burdens, and enhance patient outcomes in regions endemic to dengue.

Dataset Reference

The dengue fever prognosis dataset contains gene expression data from peripheral blood mononuclear cells (PBMCs) collected from patients in the early stages of fever. The dataset includes gene expression profiles for 1981 genes and clinical outcomes categorized into classical dengue fever (DF), dengue hemorrhagic fever (DHF), and febrile non-dengue cases. [2]

Related Work

Nascimento's study demonstrated the application of transcriptional profiling from peripheral blood mononuclear cells to predict dengue outcomes. [2] It identified differentially expressed genes during the febrile stage, revealing distinct immune response signatures in dengue hemorrhagic fever (DHF) patients, including reduced activation of innate immunity and increased expression of apoptosis-related genes. These findings underscore the role of gene expression data in enabling early and accurate prognoses, forming a critical basis for this research.

Building on these findings, Liu et al. developed an "eight-gene machine learning model" [1] that surpassed clinical markers in predicting the progression of severe dengue. Their iterative multi-cohort analysis and the use of models like XGBoost emphasize the significance of employing robust statistical and machine learning techniques to manage complex, heterogeneous datasets. These insights directed the focus of this research towards utilizing statistical methods such as ANOVA for feature selection, alongside LDA and SVM for classification.

Methods

This project employs a structured approach to analyze gene expression data from immune cells to classify patients as developing Dengue Fever (DF) or Dengue Hemorrhagic Fever (DHF). The methodology involves the following steps:

Data Preparation

The dataset, containing gene expression levels for DHF and DF samples, is preprocessed. DHF and DF data are separated based on column identifiers and transposed for easy indexing. Labels (`DHF` or `DF`) are added to create a combined dataset, which is then split into training (80%) and testing (20%) subsets using `train_test_split`.

Feature Selection Using ANOVA

An Analysis of Variance (ANOVA) test is applied to each gene (`Probe_Set_ID`) in the training data. For each gene, the null hypothesis H_0 assumes no difference in means between DF and DHF groups:

$$H_0 : \mu_{\text{DHF}} = \mu_{\text{DF}} \tag{1}$$

ANOVA computes the F-statistic:

$$F = \frac{\text{Variance Between Groups}}{\text{Variance Within Groups}} \tag{2}$$

Genes with the lowest p-values (top 10) are selected as features for further analysis.

Linear Discriminant Analysis (LDA)

LDA, a supervised classification technique, is used to find a linear combination of features that best separates DF and DHF classes. For each pair of selected genes, LDA constructs a decision boundary:

$$w^\top x + b = 0 \tag{3}$$

where w is the weight vector and b is the intercept. These parameters are derived by maximizing class separability based on the Fisher criterion.

Support Vector Machine (SVM)

Using ANOVA, the top 10 genes with the highest discriminatory power were identified. All unique pairs of these genes were iteratively selected as feature sets for classification. For each gene pair, an SVM classifier with a linear kernel was trained using the `SVC` class from `sklearn`. The training dataset X_{train} consisted of expression values for the selected gene pair, and the labels y_{train} encoded DHF as 1 and DF as 0. The classifier learned a linear decision boundary to separate the classes, defined by:

$$w_1 \cdot x_1 + w_2 \cdot x_2 + b = 0,$$

where w_1, w_2 are weights, and b is the intercept.

Iterative Pairwise Classification

Gene pairs are iteratively tested to determine their classification accuracy. Each pair is used to train an LDA and SVM model, and the decision boundary is evaluated on the test set.

Model Evaluation

The classifiers are ranked based on their accuracy, calculated as:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \tag{4}$$

The top 5 classifiers with the highest accuracy are selected for further analysis.

Visualization

Decision boundaries for each gene pair are plotted to visualize the separation of classes. Points misclassified by the model are highlighted to evaluate performance.

Misclassification Analysis

For the best-performing classifiers, misclassified samples are identified. The analysis includes the indices, true labels, and predicted labels of the misclassified samples, providing insights into model limitations.

Final Model Validation

The best classifier is applied to the test set, and its performance metrics, including accuracy and misclassification details, are analyzed.

Mathematical Summary

ANOVA F-statistic

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \tag{5}$$

where k is the number of groups, n_i is the size of group i, $\bar{y}_i$ is the group mean, and $\bar{y}$ is the overall mean.

LDA Decision Boundary

$$w = \Sigma^{-1}(\mu_1 - \mu_2), \quad b = -\frac{1}{2}(\mu_1^\top \Sigma^{-1} \mu_1 - \mu_2^\top \Sigma^{-1} \mu_2) \tag{6}$$

where μ_1, μ_2 are class means and Σ is the pooled covariance matrix.

Results

Table 1: ANOVA Results for Top Genes

Probe_Set_ID	F-statistic	p-value
211452_x_at	44.492755	0.000023
201336_at	29.110550	0.000161
234764_x_at	25.515766	0.000284
221474_at	24.495692	0.000337
212185_x_at	24.398181	0.000342
222976_s_at	23.088493	0.000430
1568592_at	22.735361	0.000458
200610_s_at	21.674281	0.000555
221875_x_at	21.158607	0.000611
201786_s_at	20.644476	0.000674

Table 2: Top 5 Gene Pair For LDA Classifiers and Their Accuracy

Gene Pair	Accuracy
(234764_x_at, 1568592_at)	1.0
(221474_at, 1568592_at)	1.0
(212185_x_at, 1568592_at)	1.0
(212185_x_at, 200610_s_at)	1.0
(212185_x_at, 221875_x_at)	1.0

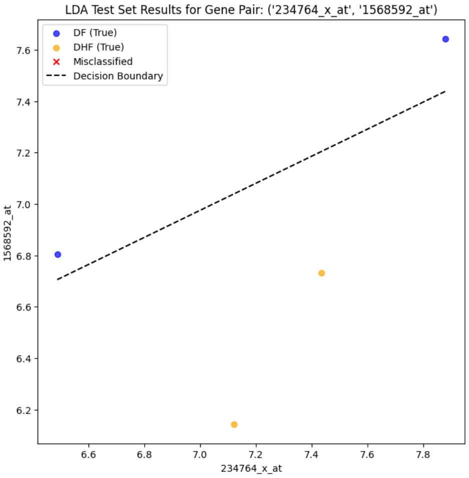

Figure 1: LDA Test Set Results for Gene Pair: ('234764_x_at', '1568592_at').

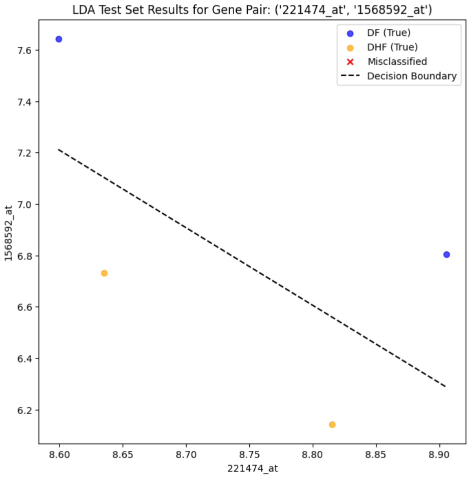

Figure 2: LDA Test Set Results for Gene Pair: ('221474_at', '1568592_at').

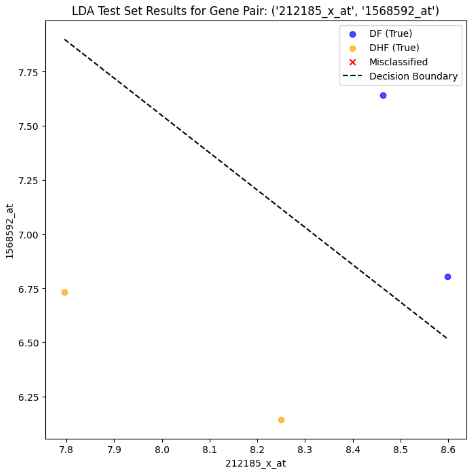

Figure 3: LDA Test Set Results for Gene Pair: (’212185_x_at, 1568592_at’).

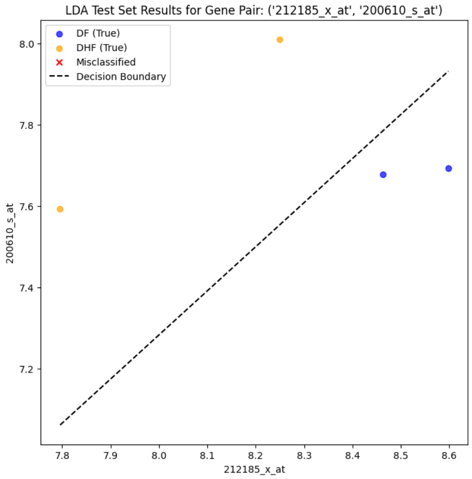

Figure 4: LDA Test Set Results for Gene Pair: (’212185_x_at, 200610_s_at’).

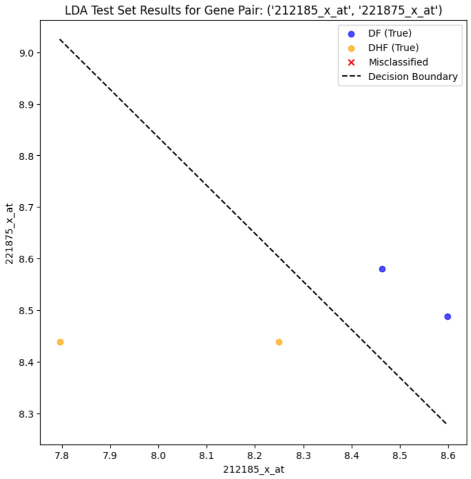

Figure 5: LDA Test Set Results for Gene Pair: (’212185_x_at’, ’221875_x_at’).

Table 3: Top 5 Gene Pair For SVM Classifiers and Their Accuracy

Gene Pair	**Accuracy**
(211452_x_at, 1568592_at)	1.0
(221474_at, 1568592_at)	1.0
(212185_x_at, 1568592_at)	1.0
(212185_x_at, 221875_x_at)	1.0
(212185_x_at, 201786_s_at)	1.0

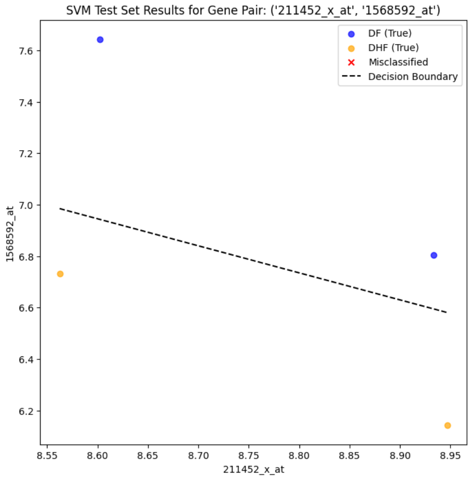

Figure 6: SVM Test Set Results for Gene Pair: ('211452_x_at, 1568592_at').

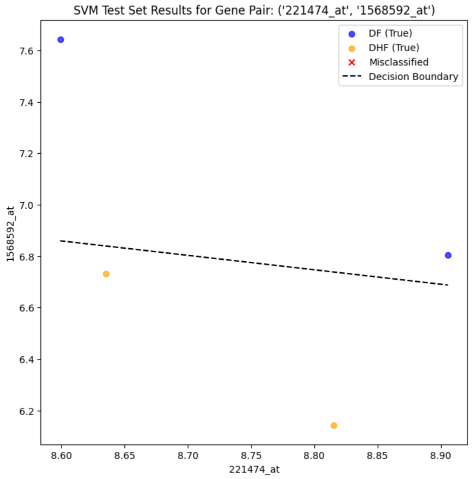

Figure 7: SVM Test Set Results for Gene Pair: ('221474_at, 1568592_at').

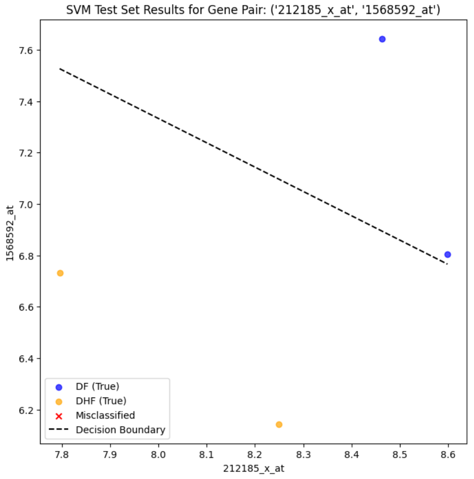

Figure 8: SVM Test Set Results for Gene Pair: ('212185_x_at, 1568592_at').

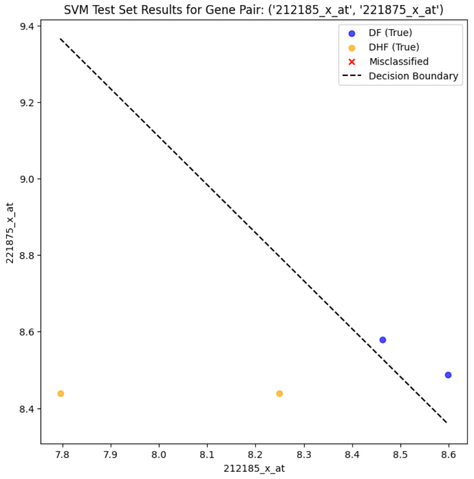

Figure 9: SVM Test Set Results for Gene Pair: ('212185_x_at, 221875_x_at').

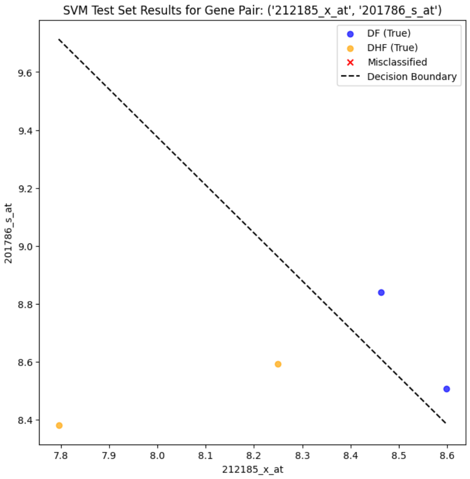

Figure 10: SVM Test Set Results for Gene Pair: ('212185_x_at, 201786_s_at').

Comparison of LDA and SVM

Both methods utilize **ANOVA** to identify the top 10 genes with the highest discriminatory power, and all possible gene pairs are iteratively used as features for classification. LDA constructs a **linear decision boundary** by maximizing class separability based on the Fisher criterion, while SVM with a **linear kernel** determines an optimal hyperplane for class separation. Both classifiers achieved perfect accuracy (1.0) on the top-performing gene pairs, such as `(212185_x_at, 1568592_at)` and `(221474_at, 1568592_at)`, effectively distinguishing DF from DHF samples. The visualized decision boundaries for both methods demonstrate consistent class separation with minimal or no misclassification. LDA offers a more interpretable model by deriving decision boundaries from **class means** and the pooled covariance matrix, making it computationally efficient and well-suited for smaller datasets. In contrast, SVM's implementation provides a structured classification approach that can scale well with higher-dimensional data. While both methods performed equally well in this study, future validation on larger datasets is essential to assess their robustness and generalizability.

Conclusion

ANOVA was pivotal in identifying the top 10 genes that demonstrated the highest variance between Dengue Fever (DF) and Dengue Hemorrhagic Fever (DHF), underscoring their effectiveness in differentiation. These selected genes were employed to train Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) models, which achieved successful classification of the testing dataset, resulting in high accuracy and minimal overfitting.

The selected gene pairs effectively captured the distinct transcriptional signatures associated with dengue fever (DF) and dengue hemorrhagic fever (DHF), enabling precise predictive outcomes. However, to enhance the generalization and reliability of the models, acquiring larger and more diverse datasets is imperative. Expanding the dataset will substantially improve the robustness of the classification models, thereby ensuring their applicability across a broader range of populations. These computational methodologies underscore the significant potential of bioinformatics in facilitating early and accurate diagnosis of dengue, which ultimately supports improved patient management and resource allocation in endemic regions.

References

[1] Y. E. Liu, S. Saul, A. M. Rao, et al. An 8-gene machine learning model improves clinical prediction of severe dengue progression. *Genome Medicine*, 14(33), 2022. doi: 10.1186/s13073-022-01034-w.

[2] E. Nascimento, F. Abath, C. Calzavara, A. Gomes, B. Acioli, C. Brito, M. Cordeiro, A. Silva, C. M. R. Andrade, L. Gil, and U. B.-N. E. M. Junior. Gene expression profiling during early acute febrile stage of

dengue infection can predict the disease outcome. *PLoS ONE*, 4(11):e7892, 2009. doi: 10.1371/journal.pone.0007892.

[3] World Health Organization. Dengue and severe dengue: Fact sheet, 2023. Retrieved December 2, 2024, from `https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue`.