Prognosis of Dengue Hemorrhagic Fever Utilizing Human Genome Data and Machine Learning Technology

by

Autumn Toki

BS Biochemistry, Indiana University of Pennsylvania, 2020

Submitted to the Graduate Faculty of the

Department of Infectious Diseases and Microbiology

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Autumn Toki

It was defended on

April 21, 2022

and approved by

Dr. Jeremy Martinson, Assistant Professor, Infectious Diseases and Microbiology

Dr. Ernesto Marques, Associate Professor, Infectious Diseases and Microbiology

Dr. Yesim Demirci, Associate Professor, Human Genetics

Thesis Advisor: Dr. Jeremy Martinson, Assistant Professor, Infectious Diseases and Microbiology

Prognosis of Dengue Hemorrhagic Fever Utilizing Human Genome Data and Machine Learning Technology

Autumn Toki, MS

University of Pittsburgh, 2022

Dengue is of worldwide public health concern as it has become one of the most significant arthropod-borne diseases, affecting an estimated 4 billion individuals globally. Dengue infections range from asymptomatic, to the debilitating febrile illness dengue fever (DF), and to dengue hemorrhagic fever (DHF), which is severe and potentially life-threatening. The precise immunopathology of DENV infection has yet to be elucidated, however, it has been shown that host genetics influence infection and disease development. Single nucleotide polymorphisms (SNPs) offer early clinical identification and better understanding of disease progression. Our study utilizes a cohort of patients from Brazil, who are infected with Dengue, and who have developed either DF or DHF. Individuals were genotyped for 20 immune-relevant SNPs, including a core set of 13 SNPs recognized to be most influential in disease progression. A machine learning (ML) algorithm together with human genotyped data was utilized to identify the combination of SNPs that have the greatest utility in predicting risk for developing the severe dengue phenotype. From this study, dengue severity can be predicted based solely on human genomic markers.

Table of Contents

List of Tables

List of Figures

Preface

This thesis is ultimately based on a previous study conducted by Caio Davi and colleagues, although is of my own original work, and was written to fulfill the graduation requirements of the Infectious Diseases and Microbiology Department in the University of Pittsburgh's School of Public Health. None of the text included in the thesis have been taken from previously published articles. I was engaged in researching and writing this thesis from January 2021 to April 2022.

The TaqMan genotyping and statistical analyses are of my original work, whereas the design and implementation of the ANN was done in collaboration. We were fortunate enough to work with Ernesto Marques and Caio Davi, who were members of the previous study. The ANN described was designed by Caio who is of faculty in the Department of Electrical and Computer Engineering at Texas A&M University. Sections 1.12, 2.6, and 3.0 describe the composition and results of the neural network. My genotyped results from TaqMan PCR were given to Caio to analyze via the ANN, where the dengue outputs were shared with me as well as how to implement the ANN in Python through utilization of shared workbooks. The analyses and performance evaluations of the ANN runs are of my own work.

I would like to thank my supervisor Jeremy Martinson for the excellent guidance and support offered during the process. I would also like to thank the collaborators Ernesto Marques and Caio Davi. Without their cooperation I would have not been able to conduct essential analyses.

1.0 Introduction

Dengue has become one of the most significant arthropod-borne diseases and is of global public health concern. Dengue virus (DENV) is a member of the genus Flavivirus in the Flaviviridae family (1). DENV is a single-stranded, enveloped, positive-strand RNA virus that has four antigenically related serotypes, including DENV-1 through DENV-4. The genome of the virus is approximately 11 kilobases in length, which encodes for its three structural proteins: nucleocapsid, envelope, and membrane, and its seven nonstructural proteins: NS1, NS2a, NS2b, NS3, NS4a, NS4b, and NS5 (2). Infections range from asymptomatic, to the debilitating febrile illness dengue fever (DF), and to dengue hemorrhagic fever (DHF), which is severe and potentially life-threatening. Individuals generate immunologic memory upon infection to one serotype, however, immunity does not protect against other DENV serotypes, meaning that a person can be infected with virus up to four times in their lifetime.

1.1 Dengue Transmission

The *Aedes* species of mosquito, specifically, infected *Ae. Aegypti* mosquitos are responsible for the transmission of the dengue virus. These arthropods are most active and prevalent in tropical climates, where their eggs are typically laid near standing water. Mosquitos become infected after feeding on a susceptible host infected with DENV. Infected mosquitos can then spread the virus to other humans through a blood meal. An estimated 4 billion individuals live in high-risk areas of dengue infection, with 400 million infected each year (3) Outbreaks affect

many countries globally, including those in America, the Middle East, Pacific Islands, Asia, and Africa.

1.2 Dengue in Brazil

The risk for Dengue currently reported in Brazil is frequent and continuous (3) The first epidemic was reported in Rio de Janeiro, in 1845 (1), and the virus has been present since. From 1995 to 2009, Brazil had the fifth highest incidence rate of dengue fever among Latin American and Caribbean countries and rose to the top in 2014 (1). Figure 1 exhibits the geographic composition of Brazil, including the Southeast, Midwest, Northeast, North, and South. Of these five, the Southeast region represents the highest number of cases, exceeding 300,000 in 2014 alone. The city of Recife is located in the Pernambuco municipality which resides in the Northeast region of Brazil. Recife was home to the highest number of cases per 100,000 inhabitants between 2015 and 2017 (4) The climate of Recife is also home to the highest rainfall levels out of the Pernambuco municipality (4), employing perfect conditions for mosquito habitation, thus increasing DENV transmission and infection.

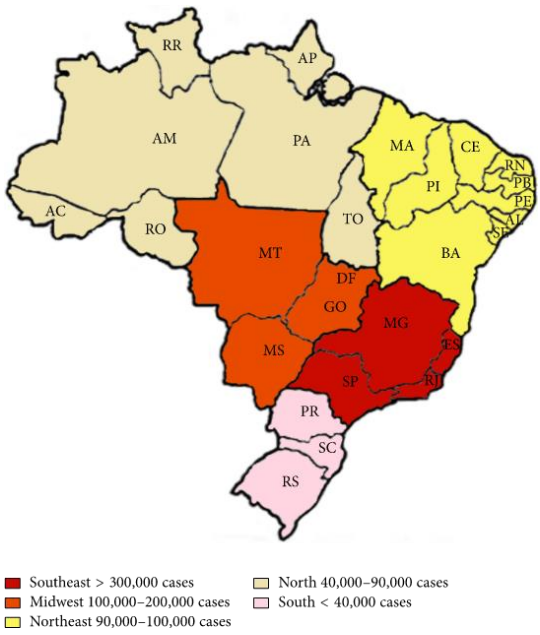

Figure 1. Prevalence of dengue in Brazil based on location

1.3 Dengue in Recife

Given the humid, tropical climate of Recife, Brazil, frequent infection transmitted via mosquitos is common, and many individuals are at risk of developing DF or DHF. DENV-1 through DENV-4 serotypes have a 65-70% nucleotide sequence homology (5), indicating that each dengue genotype has a different virulence. Individuals with primary infection are mainly asymptomatic or develop a mild febrile illness which includes fever, muscle aches, joint pain, rash, nausea and vomiting, and other clinical signs and symptoms. After infection with one serotype, an individual possesses immunity to that specific serotype, however, dengue infection with another serotype increases an individual's chance of developing more serious clinical manifestations including hemorrhaging, plasma abnormalities, coagulation, and increased vascular permeability

(4). These are the signs and symptoms of the life-threatening DHF and can progress to dengue shock syndrome (DSS), which includes hypovolemic shock and organ failure.

1.4 Dengue Classification

There are several methods used by clinicians to classify dengue severity and determine diagnosis. Testing for viral RNA, non-structural protein 1 (NS1) viral antigen, anti-dengue IgG, and anti-dengue IgM are some of these virological and serological methods utilized (6). From patient serum or plasma samples, viral RNA and NS1 antigen levels can be detected. As referenced in Figure 2, the other applied markers of anti-dengue IgG and anti-dengue IgM are widely used because in primary infection, anti-dengue IgM antibodies slowly increase after the day of infection, and anti-dengue IgG antibodies increase only after the increase in IgM (7).

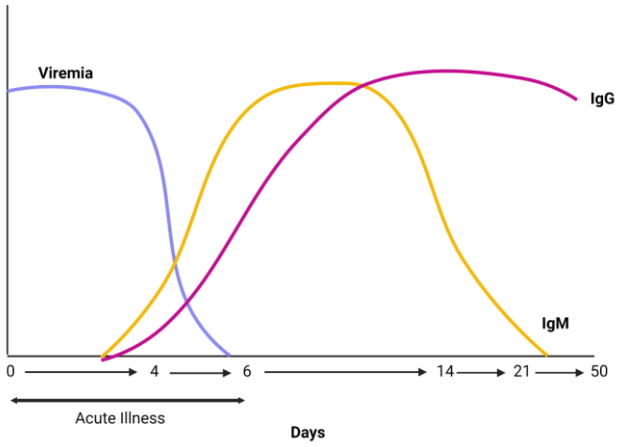

Figure 2. Dengue course of viremia

In secondary infection, both IgM and IgG antibodies increase simultaneously, meaning that the presence of anti-dengue IgM or the increased amount of anti-dengue IgG are used as helpful biomarkers for diagnoses (7). One of the largest struggles regarding dengue infection to this day is evaluation of a patient's risk for developing severe dengue or dengue hemorrhagic fever. Typically, clinicians utilize an individual's clinical presentation along with laboratory blood test results. In 2009, the World Health Organization (WHO) announced new criteria for risk evaluation. Based on the WHO's criteria, patients are divided into three categories: dengue fever without warning signs, dengue with warning signs, and severe dengue as visualized in Table 1 (3). Another major challenge clinicians face is the amount of time needed to perform additional blood tests to rapidly classify patients with a high risk of developing severe dengue symptoms, to ensure the necessary medical treatment or hospitalization is provided. Because at risk individuals progress to severe symptoms at different time points, a new classification system for disease evaluation is urgently required, especially during major outbreaks. Figure 3 references the infection timeline that is subject to change from person to person. The critical phase is indicative of DHF symptoms, which can occur at any time point in infection, but usually begins when initial fever subsides.

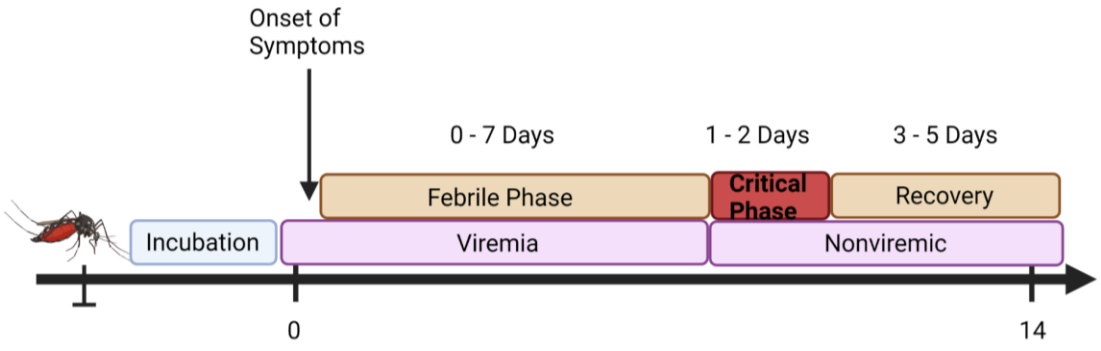

Figure 3. Infection timeline for dengue

Due to the wide range of clinical manifestations and often sudden onset of severe symptoms, the mechanism of DENV infection is an object of intense research.

Table 1. World Health Organization (WHO) criteria for evaluation of dengue risk

DF with no warning signs	DF with warning signs	Dengue Hemorrhagic Fever
Fever, headache, eye, muscle and join pain, nausea, vomiting, skin rash.	More severe abdominal pain, persistent vomiting, rapid breathing, bleeding gums, fatigue, hematemesis	High fever, platelet depletion, secondary infections, rhabdomyolysis, hemorrhaging

1.5 DENV Pathogenesis

The pathogenesis of DENV infection remains complex and not completely understood, although, there are several immunologic factors known to contribute to disease progression, including antibody dependent enhancement (ADE), proinflammatory cytokines, complement activation, viral or host genetics, non-structural protein 1 (NS1) viral antigen, cross-reactive T-cells, autoimmunity, and others.

The innate immune system provides the host with the first line of defense against infection. When an infected mosquito feeds on a human, the virus is presumably injected into the bloodstream with some spillover in the epidermis and dermis (8), initiating local infection. DENV enters host cells via receptor-mediated endocytosis. The virus recognizes ubiquitous cell surface molecules and can utilize multiple receptors to gain entry, including interaction of the viral envelope with heparan sulfate, lipopolysaccharides, C-type lectins, and others (2). Following

receptor binding, fusion of the viral and host vesicular membranes occur and the nucleocapsid is released into the cytoplasm of the host cell. By mechanisms not yet understood, local viral replication then takes place. Innate immune cells at the site of DENV infection include immature Langerhans cells; the dendritic cells (DCs) of the epidermis, that are hypothesized to play a role in local viral replication and dissemination (8). Once infected with virus, the DCs display the viral antigens on their surface and migrate from the site of infection to the lymph nodes, recruiting and activating monocytes and macrophages. This recruitment ultimately amplifies infection, and the virus is disseminated through the lymphatics. Cells susceptible to infection from primary viremia include innate immune cells of the mononuclear lineage; monocytes, myeloid DCs, and spleen and liver macrophages. Primary viremia ultimately leads to the clinical manifestations of dengue fever including fever, headache, nausea and vomiting, and others as described previously. Other factors contributing to viral pathogenesis include NS1 viral antigen, ADE, autoimmunity, and cross-reactive T-cell activation. Development of dengue hemorrhagic fever utilizes host immune mechanisms to its advantage, although the exact mechanisms of disease progression remain unclear.

1.6 Role of NS1 Antigen

Of the seven non-structural viral proteins, NS1 is known to play a large role in dengue pathogenesis. The protein is observed in multiple oligomeric forms while present in the host, including m-NS1 which is found on the cell surface, and s-NS1 which serves as a soluble, secreted lipoparticle (5). Results from many research experiments exhibit a positive correlation between s-NS1 levels and disease severity. As s-NS1 levels increase in the host, disease severity tends to

increase as well during the acute phase of infection (5). The NS1 antigen aids in DENV pathogenesis because it has a direct impact on disruption of the endothelial cell monolayer by eliciting inflammatory cytokine production. NS1 can activate macrophages and peripheral blood mononuclear cells (PBMCs) by utilizing their Toll-like receptors, specifically, TLR-4 to increase vascular endothelial permeability (9). NS1 antigen also elicits shedding of heparan sulfate proteoglycans on the host cell surface. This loss of sialic acid leads to disruption of the host glycocalyx layer in pulmonary vascular endothelial cells (10), contributing to increased vascular permeability, thus increased DENV pathogenesis. The more severe clinical manifestations of dengue hemorrhagic fever leading up to shock syndrome include plasma leakage and accumulation of fluid, which can also be attributed to NS1 antigen. The viral protein also directly triggers the complement activation pathway, leading to the production of inflammatory cytokines.

1.7 Role of Complement System

The complement system/cascade is a mechanism employed by the host immune system for defense against viral infection by opsonizing viral particles for elimination. Complement proteins C5b through C9 create a membrane attack complex (MAC) which is formed on the surface of the target cell where lysis ultimately ensues. The complement system also stimulates robust yet disorganized expression of inflammatory cytokines such as IFN-γ, IL-2, TNF-α, IL-6, and others that are associated with the development of dengue hemorrhagic fever (5). Although there is not a current understanding as to why, a positive correlation has been seen between NS1 levels and C5b-C9 levels, thus increased complement activation, generation of cytokine storm and disease progression, leading to plasma leakage and thrombocytopenia as seen in DHF. Researchers have

also noticed that during severe disease manifestations, specifically plasma leakage, the activation products C3a and C5a of the complement system are present in plasma. Patients that have progressed to dengue shock syndrome, have decreased C3a and C5a levels in their plasma (2). Ultimately, NS1 aids in DENV pathogenesis through activation of a variety of host immune factors.

1.8 Antibody-Dependent Enhancement

Extensive research and evidence regarding dengue studies have shown that Antibody-Dependent Enhancement (ADE) is a contributing factor to the development of DHF, although the exact mechanisms of ADE remain unknown. As previously stated, infection with one dengue serotype elicits immunologic memory, but does not provide long-term cross-protective immunity to the remaining dengue serotypes. Heterotypic DENV infection after primary infection results in the production of non-neutralizing pre-existing antibodies that can bind to the virus but do not achieve successful neutralization (11). Various immune cells such as phagocytes recognize the antibody bound to the virus through their Fcγ receptors, which facilitates viral uptake, resulting in infection of a healthy immune cell. Viral replication is then amplified, which in turn generates increased viral load, driving an immunopathogenic cascade. Exaggerated cytokine production is induced in response to ADE, which then affects vascular permeability, causing the clinical manifestations associated with DHF including fever, bleeding, and lymphopenia.

Another proposed mechanism which explains the increase in viral replication is known as intrinsic ADE. As stated previously, Fcγ receptors facilitate viral uptake of DENV. The intrinsic ADE mechanism proposes that during viral uptake, necessary antiviral responses such as type I

IFN production, are inhibited through suppression of antiviral genes (5). IFN-1 responses are critical in driving antiviral immune responses through various mechanisms, including inhibition of viral replication, activating antigen presentation, and triggering adaptive immune responses through interaction with T cells and B cells. Inhibition of the genes necessary for IFN-1 response generate a very negative and harmful downstream effect. In response to ADE and intrinsic ADE, IL-10 is produced in abundance, with mechanisms as to why remain unknown. IL-10 is a cytokine with anti-inflammatory properties, which plays a significant role in immunoregulation, limiting host immune response to pathogens, and maintaining tissue homeostasis (2). IL-10 production also downregulates expression of Th1 cytokines, which are produced to stimulate immune and proinflammatory responses (7). This shifts the immune response towards Th2, which is involved in the humoral response and has limited antiviral effects, thus increasing dengue pathogenesis.

1.9 Role of Cross-Reactive T cells

The main immune cells responsible for producing IL-10 and other cytokines include T-cells, monocytes, macrophages, and dendritic cells. T cells include cytotoxic T cells and helper T cells. The main roles associated with these crucial immune cells include directly killing infected host cells, activation of other immune cells, cytokine production, and immune response regulation. Studies have shown that cross-reactive T cells play a role in DHF pathogenesis (5). Relating to cross-reactive antibodies, cross-reactive T cells are also active only during secondary heterotypic dengue infection. As part of the immune process, T cells are activated by interaction with antigen-presenting cells, with dendritic cells acting as the best APCs. This interaction occurs when the T cell receptor (TCR) on either CD4+ Th cells, or CD8+ cytotoxic T cells, binds to the presented

antigen through recognition of the antigen-MHC complex (MHC class I for CD8+, MHC class II for CD4+). The binding between the TCR specific for one antigen and the antigen-MHC complex sets the immune response in motion. The next step in T cell activation includes co-stimulatory signals in the form of CD28 on T cells, and CD80/CD86 on dendritic cells (9). This signal ensures that the recognized antigen is presented by a pathogen-activated APC and eliminates the possibility of binding self-antigens. The last step in T cell activation is differentiation through cytokine signals in the surrounding environment at the time of activation. The cytokines produced are essential for the generation of correct immune responses and activation of specific subtypes of Th cells.

In the case of dengue infection, CD8+ T cells recognize non-structural proteins NS3 and NS5 as epitopes (10). It has been shown that activated CD8+ T cells involved in heterotypic secondary infection exhibit increased cross-reactivity and are not able to eliminate the cells infected by the new dengue serotype (9). The cross-reactive T cells produce higher levels of cytokines and chemokines, which are positively associated with DHF. The high number of effector T cells produced in response to dengue infection results in tissue damage because of increased inflammation from cytokine production, which correlates with severe dengue symptoms. In summary, the mechanisms surrounding dengue infection and leading up to dengue hemorrhagic fever are not yet understood but are likely to be multifactorial (5). Host genetic factors have been known to affect immune response and disease progression in a variety of infectious diseases; dengue included. The research outlined in this report focuses on the relationship between the host genetic factors and dengue severity directly.

1.10 Host Genetics

Host genetic factors have been shown to have physiologic influence during dengue infection and pathogenesis. How the host battles infection is dependent on genetic factors, as both innate and adaptive immune pathways rely on genomic composition. There is a plethora of essential genes involved in innate responses. Human Leucocyte Antigen (HLA) is shown to play a role in DENV pathogenesis. Interestingly enough, HLA class-I alleles have been associated with severe dengue pathogenesis in secondary infection, whereas HLA class-II alleles have been observed to have a protective effect on the host (12). My research focused on various innate immune genes, including 11 genes hypothesized to be most influential in disease severity: *CLEC4C, IRF1, IFIT1, MYD88, TLR8, MX1, OAS2, VEPH1, IFNγ, OAS3,* and *IRAK4*. C-type lectin domain family 4 member C *(CLEC4C*) encodes a member of the C-type lectin superfamily. Properties of this family include a diverse set of cell functions, such as cell-to-cell signaling, cell adhesion, dendritic cell function, inflammation, and immune response (13). Dengue can utilize CLEC4C gene function to its advantage by displaying CLEC4C ligands on the viral surface which are recognized by the DC-SIGN receptor on dendritic cells (13). The ultimate downstream effect of this interaction prevents the dendritic cells from successfully processing and presenting viral peptides to activate adaptive immune responses. Thus, IFN-1 responses are suppressed and DENV spreading is facilitated. Interferon regulatory factor 1 (*IRF1*) regulates host transcription by activating genes involved in innate and adaptive response. Encoded proteins activate essential genes involved in host response to viruses, playing a role in immune response, DNA damage response, cell proliferation, and others (14). Regarding dengue, *IRF1* is essential in transcriptional activation of type I interferons in infected host hepatocytes to aid in antiviral defense (14). Interferon induced protein with tetratricopeptide repeats 1 (*IFIT1*) encodes a protein induced upon

interaction with interferons that inhibits viral replication and translation. *MYD88* is an innate immune signal transduction adaptor. The encoded protein is valuable in both innate and adaptive responses by transducing TLR and interleukin-1 signaling pathways, which ultimately provide regulation of various proinflammatory genes (15). Research has shown that dengue infection enhances expression of inflammatory molecules in dendritic cells via the MYD88/TLR2 pathway, ultimately aiding in the development of ADE in secondary infection (9). Th2 responses are favored in this environment, advocating disease progression towards DHF. Toll-like receptor 8 (*TLR8*) is a member of the Toll-like receptor family involved in pattern recognition and innate immunity. These receptors recognize pathogen-associated molecular patterns (PAMPs) that are expressed on viral surfaces and elicit necessary cytokine production in response to this interaction (15). The polymorphism we studied in the *MX1* gene is an intron variant. The gene itself encodes a GTP-metabolizing protein that aids in antiviral response through activation by type I/II IFNs (5). Once activated, the protein disrupts viral replication. 2'-5'-oligoadenylate synthetase 2 (*OAS2*) belongs to a gene family that encodes proteins essential to innate responses. Interferons activate the encoded protein, which ultimately utilizes 2'-5'-oligoadenylates to activate latent RNase L (16). Once activated, the downstream effect of RNase L is to degrade viral RNA and inhibit viral replication. Ventricular zone expressed PH domain containing 1 (*VEPH1*) encodes a protein predicted to be active in the plasma membrane and induce phosphatidylinositol-5-phosphate binding activity (16). Interferon gamma (*IFNG*) encodes a cytokine part of the type II interferon class. Immune cells involved in both innate and adaptive responses secrete this cytokine to trigger various cellular responses to viral infection (5). 2'-5'-oligoadenylate synthetase 3 (*OAS3*) also encodes an enzyme in the 2'-5'-oligoadenylate synthase family. Similar to OAS2, the enzyme of OAS3 is induced via interferons and binds and activates RNase L (15). Lastly, Interleukin 1

receptor associated kinase 4 (*IRAK4*) encodes a kinase enzyme that activates the NF-κB pathway upon TCR and TLR stimulation (9). NF-κB proteins are transcription factors that control various cellular processes including immune and inflammatory responses and apoptosis. IRAK4 function is essential to innate immune response to viral infection (16).

Table 2. Genes referenced and their corresponding function

Gene	Function
CLEC4	Cell signaling/adhesion Dendritic cell function Inflammation
IRF1	Host transcription regulation
IFIT1	Inhibitions viral replication/translation
MYD88	Transduces TLR and interleukin-1 signaling pathways Regulation of proinflammatory genes
TLR8	Pattern recognition Cytokine production
MX1	Disrupts viral replication
OAS2	Activate RNase L Degrade viral RNA/inhibit replication
VEPH1	Induce phosphatidylinositol-5-phosphate binding activity

IFNG	Generates *IFNγ* cytokines belonging to Type II IFN response to infection
OAS3	Binds/activates RNase L
IRAK4	Host transcription activator Activates NF-κB in TCR/TLR pathways

1.11 Single-Nucleotide Polymorphisms

Single-nucleotide polymorphisms (SNPs) can be described as a single nucleotide variation in a DNA sequence. For example, SNPs occur when there is a single base pair change in a shared sequence between members of the same species or paired chromosomes in individuals (16). Most common SNPs have only two alleles, with one of these alleles dictated as being the minor allele in a population, corresponding to the lesser observed allele at a specific locus (16). Minor allele frequencies are important biomarkers because they vary from population to population, indicating that alleles that are rare in one population may be common in another. Types of SNPs include substitutions, deletions, or insertions that can occur in coding, non-coding, or intergenic regions of the genome. The SNPs used in this study have all shown to be substitutions. There are many advantages of studying SNPs, including early clinical identification, risk of developing diseases, drug response, epidemiology, gene tracking, susceptibility to environmental factors, and others. Regarding dengue, single-nucleotide polymorphisms are known to influence disease progression (12). Twenty immune-relevant SNPs from previous studies were utilized here, including a core set of 13 SNPs previously identified as being the most influential. These SNPs have direct influence on the innate genes listed previously as well as others including Vitamin D receptor (*VAD*), Acute

plasma glycoprotein mannose-binding lectin (*MBL*), and dendritic cell-specific intercellular adhesion molecule 3 (*ICAM-3*)-grabbing nonintegrin (*DC-SIGN*) (12).

Table 3. All SNPs utilized in the study with the core set of 13 shown in red. Corresponding allelic frequency taken from UCSC Genome Browser database with reference allele shown in the second column.

SNP	Allelic Frequency
rs17199006	A (0.832867) G (0.167133)
rs17256081	T (0.705430) C (0.294570)
rs2069718	A (0.616813) G (0.383187) T (Unknown)
rs2069727	T (0.725040) A (Unknown) C (0.274960)
rs2070729	C (0.393970) A (0.606030) T (Unknown)
rs2072137	T (0.676518) C (0.323482)

rs2072138	C (0.780950) A (Unknown) G (0.219050) T (Unknown)
rs2240188	C (0.715855) G (Unknown) T (0.284145)
rs303212	T (0.494209) C (0.505791)
rs3737399	C (0.482228) T (0.517772)
rs3911403	T (0.741014) A (0.258986)
rs4251580	C (0.87780) T (0.112220)
rs4988457	C (0.958067) G (0.041933)
rs2296414	C (0.841054) T (0.158946)
rs3132468	C (0.194888) T (0.805112)
rs3213545	G (0.686901) A (0.313099)

rs3740360	A (0.917133) C (0.082867)
rs486907	C (0.769369) T (0.230631)
rs606231248	G (0.999987) A (0.000013)
rs7277299	C (0.979433) A (0.020567) T (Unknown)

1.12 Machine Learning Applications

Due to the difficulty in disease evaluation conducted by clinicians and the exact mechanism of dengue infection still not well understood, a new and efficient prognostic tool is needed. Various artificial intelligence and machine learning (ML) applications have been produced to offer patient phenotyping and disease outcome prediction with high efficiency and accuracy (6). Advantages to these methods include but are not limited to, improved medical treatment and the rapid diagnosis of high-risk patients after first examination to ensure proper care is provided, if necessary. There are a substantial amount of research projects relating to dengue that utilize Machine Learning technologies to provide DF and DHF diagnoses. Classification algorithms have been used such as Classification and Regression Trees (CART), Linear Discriminant Analysis (LDA), and Support Vector Machines (SVM) (12). My research utilizes the Machine Learning approach proposed in a previous study completed by Davi and colleagues (12). This approach consists of a SVM-RFE

(SVM recursive feature for elimination) for the identification of the 13 core SNPs stated previously and artificial neural networks (ANN) for dengue prediction (12). The second step consisting of the ANN was composed containing three hidden layers and nine neurons to classify patients into DF or DHF (12). Data Acquisition is the first objective in this methodology, followed by Data Preprocessing, Feature Selection, and finally, Patient Classification. The genes and 13 core SNPs listed above in Tables 2 and 3 were generated from the SVM and ANN systems together. Patients from Recife, Brazil, with dengue-related symptoms were screened to generate a sample size of 102 individuals, with 27 diagnosed with DF and 75 with SD (12). In Data Preprocessing, the individuals were genotyped according to polymorphisms at 322 loci and classified as homozygous dominant, heterozygous, or homozygous recessive. Next, the Feature Selection is where the SVM-RFE algorithm in Python was utilized to generate the list of genes and core set of 13 immune system-related SNPs described previously (12). Utilizing this information, the ANN algorithm was implemented to classify patients with DF or SD. Overall, the approach displayed above 86% accuracy, over 98% sensitivity, and over 51% specificity when compared against the clinical results (12).

1.13 AIMS

The work outlined here uses a similar methodology to validate this approach in a different cohort of samples from Brazil. We have a cohort of 148 genotyped individuals from Recife, Brazil, who are positive for DF or DHF infection. Individuals were genotyped for a set of 20 SNPs shown in Table 3, further testing the efficiency of the proposed Machine Learning technology. In conclusion, the composition of the host genome along with ML technology offers utility in

predicting pathogenesis towards dengue hemorrhagic fever. This proves extremely useful in outbreak situations, and to ensure proper medical care is readily available even before infection is initiated.

2.0 Materials and Methods

2.1 Patient Cohort

Individuals with dengue-related symptoms were screened from hospitals in Recife, Brazil, and clinicians classified patients as positive for infection with DF or DHF. An explanation of the study was reiterated to patients, and those who gave informed consent were enrolled. The samples utilized in this study were made available in five 96-well plates that contained the DNAs extracted from PBMC samples from the cohort of patients from Recife, Brazil, as well as Yellow Fever (YF) samples and controls. After DNA was extracted, the concentration was evaluated via NanoDrop analysis. Yellow Fever samples were utilized from previous studies done in the laboratory because it is thought that similar SNPs are associated with YF infection as well.

2.2 Selection of SNPs

The list of 13 core SNPs was generated from the previous study done by Davi and colleagues. An SVM-RFE algorithm was implemented in Python and used the scikit-lean library to produce the 13 SNPs and their corresponding innate genes. The list of seven other SNPs used in the study are generated from various literature reviews where results exhibited the importance of these SNPs in DENV pathogenesis. In total, 20 SNPs were analyzed in this study.

2.3 Sequenom MassARRAY

The five 96-well plates were first sent to the Genomics Research Core Lab at the University of Pittsburgh for SNP genotyping through utilization of Sequenom instrumentation. The platform for Sequenom MassARRAY utilizes a homogeneous reaction format and a single extension primer. Allele-specific products are generated each with its own distinct mass. First, a locus-specific PCR reaction takes place, followed by a locus-specific primer extension reaction. This step in the protocol allows an oligonucleotide primer to anneal upstream from the polymorphism site. After primer extension in the PCR, MALDI-TOF is used to evaluate the mass of the extended primer, which indicates the sequence mass. From mass spectrometry, the alleles present in the polymorphic site are determined. All SNPs, except for rs303212, were genotyped and results were sent back to the research laboratory for analysis.

2.4 TaqMan Assay

TaqMan Analysis was done utilizing the StepOnePlus instrumentation. The assay utilizes two probes with different fluorescent reporters that are specific to the single-nucleotide polymorphism under investigation. During the annealing step in PCR, the probes hybridize to the targeted SNP site. The first probe is labeled with VIC fluorescent dye, and detects the first allele sequence, whereas the second probe is labeled with FAM fluorescent dye and detects the second allele. During extension in PCR, Taq polymerase binds and begins creating the complementary strand to the desired sequence. If the sample sequence is homozygous for the first allele, the result is VIC fluorescence. On the other hand, if the sample is homozygous for the second allele dictated

by the second probe, the result is FAM dye fluorescence. Heterozygous samples yield results that contain roughly equal signals of both fluorescent dyes. Double-stranded DNA template, forward and reverse primers specific to the sequence, and Taq polymerase are also required. 500 µL of Master Mix, 15 µL of the assay mix, and 485 µL of ddH_2O were combined in an Eppendorf tube, and 10 µL was dispensed into all wells of a 96-well plate. Next, 2 µL of the sample DNA from one of the five 96-well sample plates as stated above, was transferred into each of the wells of the genotyping plate. The plate for genotyping was then loaded into the StepOnePlus and run using the "Genotyping Assay" run method. Each of the five sample plates was genotyped for rs303212, rs4988457, rs486907, and rs17199006 following the same methodology as described above. After the run was completed, the fluorescence signals for the samples were visualized on an allelic discrimination plot. All genotypes were called manually as either homozygous or heterozygous and exported to an excel file. The exported data was combined with the previously generated Sequenom data to create a full set of genotypes. The missing data in the Sequenom results were replaced with the TaqMan result data, and vice versa. All results that did not agree on the genotype result were replaced with the TaqMan data, although, there was almost identical overlapping of data between the two assays.

2.5 Statistical Analysis

Statistical analyses were performed in STATA. The input file included the full set of genotypes generated as described above. The data available included Patient ID, Sample Number, Age, Sex, Dengue (DF or DHF), and the 20 SNPs with their corresponding genotypes. First, sample numbers that were linked back to the same Patient ID were eliminated, to ensure a unique

set of genotypes from different individuals were to be included in the final dataset. Next, results were tabulated and visualized. The diagnoses of DHF and DF were determined as well as the frequency of each genotype generated for each SNP. Visually, bar charts were created to compare the frequency of genotypes between DF and DHF. Linear regression was also performed on Dengue and the 20 SNPs together. Data was first coded into quantitative data where DF = 0, DHF = 1, Male = 0, Female = 1, and the lowest allelic frequency for the genotype was designated as 0, the second highest allelic frequency as 1, and so on. The regression was performed again on the three SNPs and Dengue that had the lowest p-values from the first regression. Matrix plots were generated as well as correlations between Dengue and Age and Dengue and Sex to ensure that the two were not related. Regression was also performed between Age, Sex, and Dengue, Dengue, Sex, and the 20 SNPs, Dengue, Age, and the 20 SNPs, and finally, Age, Sex, and the SNPs.

2.6 Advanced Neural Network (ANN)

The Advanced Neural Network (ANN) was implemented using the same methodology as the previously conducted study by Davi and colleagues. The ANN was crafted using scikit-learn library and was coded in Python. The algorithm follows a set of inputs that have a weighted line associated with them. Each input with its associated weighted line was connected in combination to a set of five neurons. Each neuron contained an activation bias threshold, and the combination of inputs that surpassed the threshold moved on to the next set of three neurons where the same principle was implemented. The last step was the activation function which generated the output of either DF or DHF. The activation function was based off the logistic sigmoid function, and the weighted optimization was based off the limited-memory BFGS. The ANN was utilized to classify

patients. Outputs of the ANN were compared to the original clinician diagnoses conducted in Brazil. Accuracy, specificity, and sensitivity were conducted based on this comparison of results to evaluate performance of the ANN.

3.0 Results

The methodology followed is outlined in Figure 4. The DNA from the cohort of individuals was extracted and genotyped. Following this procedure, statistical analyses were performed on the entire genotyped dataset to evaluate the frequency of patients diagnosed with DHF and DF. Data points were then eliminated that were traced back to the same Patient ID, to generate a final dataset consisting of a unique set of genotyped individuals. Other statistical analyses were then performed again, and the information was subsequently implemented into the ANN.

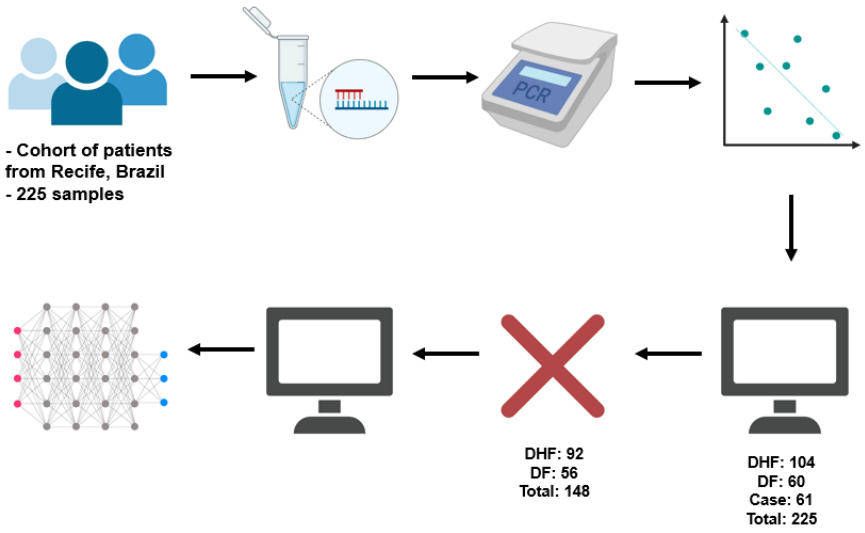

Figure 4. Methodology

The cohort of patients from Brazil included 225 samples that were genotyped via Sequenom for each of the 20 SNPs, except for rs303212. An example of Sequenom output can be visualized in Figure 5, where homozygous, heterozygous, and undetermined results are exhibited.

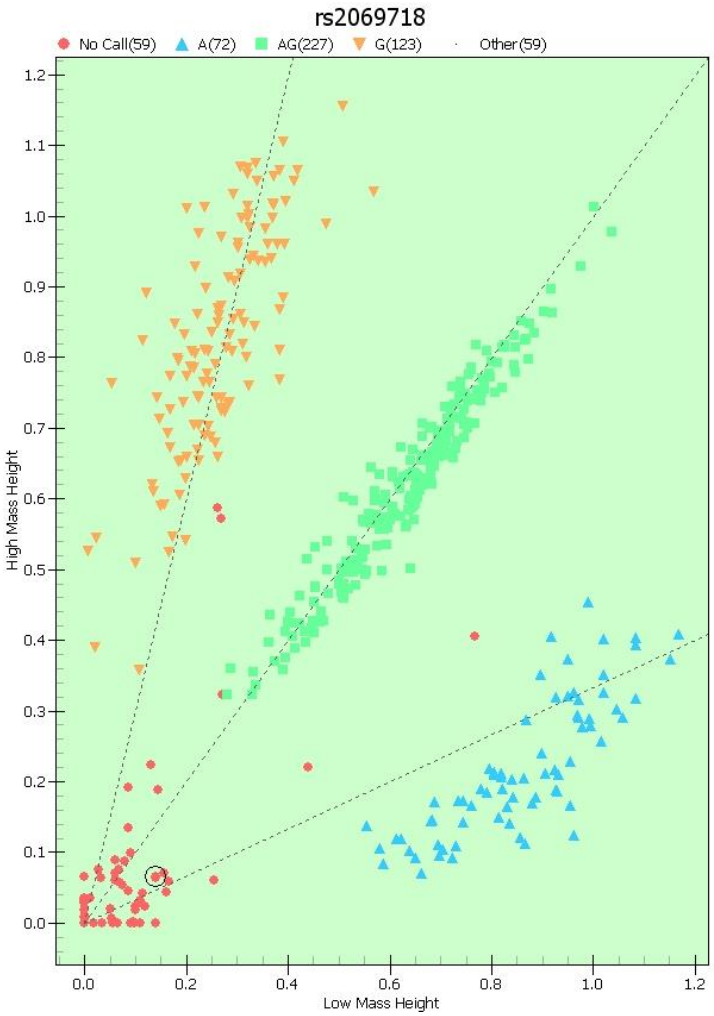

Figure 5. Sequenom results for SNP rs2069718

TaqMan genotyping was conducted to minimize undetermined results generated from the Sequenom, as well as to produce more genotyped information. An example of TaqMan genotyping is pictured in Figure 6, where an allelic discrimination plot was produced as output of the assay. The genotypes were manually called as undetermined, homozygous, and heterozygous for each assay run. SNPs rs303212, rs4988457, rs486907, and rs17199006 were analyzed via TaqMan genotyping. It was noticed that plates four and five produced worse results compared to the other plates and proved more difficult to manually call. This observation could be contributed to the

possibility of degraded sample DNA, low concentration of sample DNA (possibly ocer-estimated by NanoDrop), or most of the sample DNA in those plates consisting of controls. Once genotyped information was completed for TaqMan analysis, the results were compared with Sequenom results. TaqMan results were utilized for results that did not agree between the two analyses. Missing data in each of the analyses were replaced by the results of the other, to generate a complete dataset of genotyped individuals, including 104 DHF, 60 DF, and 61 case samples.

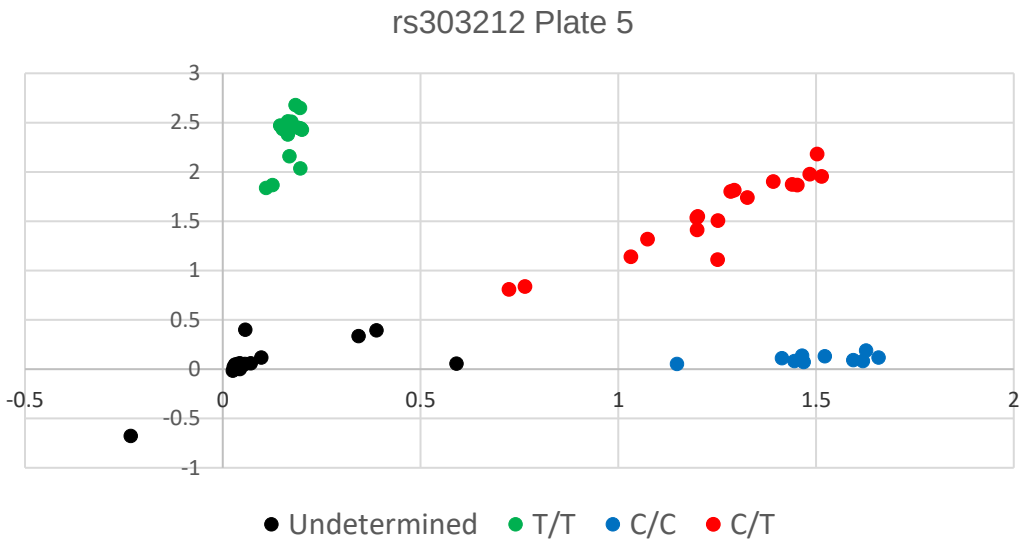

Figure 6. TaqMan genotyped results for SNP rs303212

Samples labeled as "case" alone were eliminated, as they did not correspond to DHF or DF classifications. Samples that were linked back to the same Patient ID were also eliminated, to ensure genotyped information was based off a unique, non-repeating set of individuals. The final sample set consisted of 92 DHF and 56 DF patients to create 148 samples in total to be analyzed.

Statistical analyses through STATA allowed us to visualize the genotyped information. An example of SNP genotype frequency is exhibited in Figure 7. Statistical analyses were performed

on all 20 SNPs, and the observation that at least one genotype was much more prevalent in DHF patients was seen throughout all samples. The hypothesis that SNPs have an impact on DENV pathogenesis was validated throughout this visualization, as certain genotypes were skewed towards DHF, such as in the case of rs7277299. In DHF patients, heterozygous C/A genotype accounted for approximately 0.7% of individuals, whereas homozygous C/C was present in 61% of patients. On the other hand, the C/C genotype was present in only 33% of DF patients, inferring that the C/C genotype in rs7277299 aids in pathogenesis to DHF. The A allele for rs7277299 is relatively rare, so no A/A homozygotes were seen in this dataset.

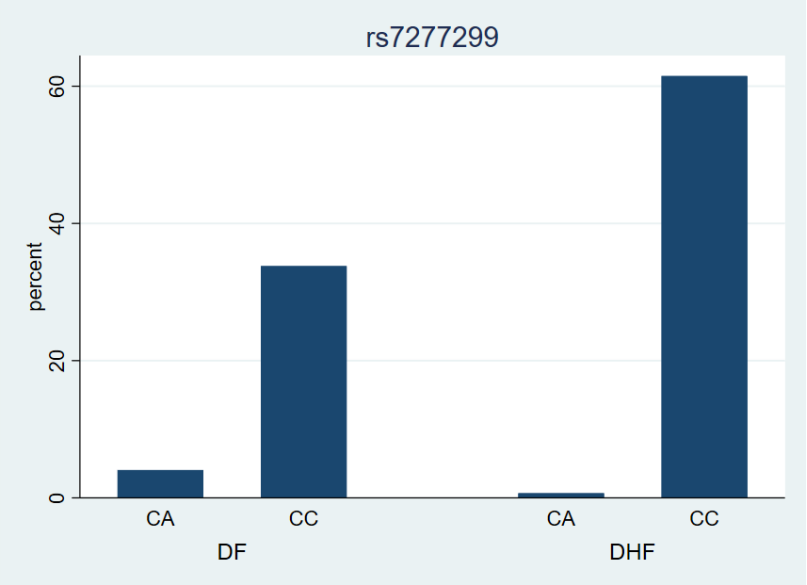

Figure 7. Frequency of genotypes for rs7277299

As stated previously, the observation that at least one genotype is far more prevalent in DHF patients holds true for all 20 SNPs. In the case of rs3740360, the A/A genotype accounted for approximately 57% of DHF patients, and 27% of DF patients, stating that for this SNP, the A/A genotype is most associated with DHF.

Other statistical analyses performed included linear regression. Regression takes the relationship of each individual SNP into account separately with Dengue, not in combinations as is done in the ANN. However, regression was performed to elucidate if any SNPs had a major effect over the other in dengue pathogenesis. The linear regression focusing on the 20 SNPs and their relationship with dengue produced an R^2 value of 0.2896, stating that 28.96% of the variation in dengue can be explained by its relationship with SNPs, as seen in Figure 8.

Source	SS	df	MS	Number of obs	=	90
				F(20, 69)	=	1.41
Model	5.79271987	20	.289635994	Prob > F	=	0.1495
Residual	14.2072801	69	.205902611	R-squared	=	0.2896
				Adj R-squared	=	0.0837
Total	20	89	.224719101	Root MSE	=	.45376

Dengue	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
rs17199006	-.1047669	.1252971	-0.84	0.406	-.3547278	.1451939
rs17256081	.0352602	.0660782	0.53	0.595	-.0965623	.1670826
rs2069718	-.0162883	.068275	-0.24	0.812	-.1524932	.1199166
rs2069727	.0232119	.0812354	0.29	0.776	-.1388483	.185272
rs2070729	.0383212	.0695316	0.55	0.583	-.1003905	.1770329
rs2072137	.041557	.0976817	0.43	0.672	-.1533127	.2364268
rs2072138	-.1134555	.1284524	-0.88	0.380	-.369711	.1428001
rs2240188	.0876913	.1021227	0.86	0.393	-.1160378	.2914205
rs303212	-.0036679	.070264	-0.05	0.959	-.1438407	.136505
rs3737399	.0844193	.1152421	0.73	0.466	-.1454825	.314321
rs3911403	.0876961	.1085596	0.81	0.422	-.1288743	.3042665
rs4251580	-.0724784	.1135324	-0.64	0.525	-.2989693	.1540125
rs4988457	-.0472597	.1467597	-0.32	0.748	-.3400373	.245518
rs2296414	-.0882491	.098345	-0.90	0.373	-.2844421	.1079438
rs3132468	.1811024	.0895792	2.02	0.047	.0023968	.359808
rs3213545	-.1600647	.0947311	-1.69	0.096	-.3490481	.0289187
rs3740360	.3436848	.1368364	2.51	0.014	.0707036	.616666
rs486907	-.0751154	.0821555	-0.91	0.364	-.2390111	.0887803
rs606231248	-.1051836	.075582	-1.39	0.168	-.2559656	.0455985
rs7277299	1.021106	.3588188	2.85	0.006	.3052817	1.73693
_cons	-.2004276	.5819025	-0.34	0.732	-1.361291	.9604361

Figure 8. Linear regression results of 20 SNPs and their relationship to DENV

The resulting p-values for each SNP were also calculated from the regression, with rs7277299, rs3740360, and rs3132468 shown to be the most statistically significant, with p-values of 0.006, 0.014, and 0.047 respectively. To further evaluate the significance of these SNPs, linear regression was performed again, only referencing these three SNPs and their relationship with

dengue. From Figure 9, the results depicted an R^2 value of 0.1324, stating that 13.24% of the variation in dengue can be explained by its relationship with these three SNPs. Another factor of note is the variation in p-values from the previous regression. rs7277299 was the most statistically significant SNP from the first linear regression, but results from Figure 5 show that rs3740360 has become the most significant in its relationship with dengue. This observation can be explained by the drastic change in SNP input. The regression model has changed between the two runs from the deletion of 17 input SNPs. Given this, the change in p-values is a function of the interrelationships among each independent SNP and dengue. Nevertheless, rs7277299, rs3740360, and rs3132468 remain statistically significant.

Source	SS	df	MS			
				Number of obs	=	148
				F(3, 144)	=	7.32
Model	4.60749272	3	1.53583091	Prob > F	=	0.0001
Residual	30.2033181	144	.209745265	R-squared	=	0.1324
				Adj R-squared	=	0.1143
Total	34.8108108	147	.236808237	Root MSE	=	.45798

Dengue	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
rs7277299	.5668886	.1782036	3.18	0.002	.2146557	.9191214
rs3740360	.3750148	.1026365	3.65	0.000	.172146	.5778837
rs3132468	.0469579	.0658015	0.71	0.477	-.0831037	.1770194
_cons	-.3059486	.2255762	-1.36	0.177	-.7518168	.1399196

Figure 9. Linear regression with 3 most statistically significant SNPs and their relationship with dengue

Another important observation is that these three statistically significant SNPs are not part of the core set of 13 SNPs generated from SVM-RFE algorithm done by Davi and colleagues. These three SNPs were implemented because of literature review, suggesting that since previous publication, other novel single-nucleotide polymorphisms have been discovered, with some having greater effects on DHF pathogenesis and susceptibility. Another hypothesis for this observation is

based off the performance of the SVM-RFE algorithm and the patient cohort utilized in the previous study. Nonetheless, these three SNPs have shown to be individually significant in their influence on DHF susceptibility. In addition to the regression performed on the SNPs, regression was also performed on sex and age variables. Tables 4 and 5 exhibit the frequency of males and females with DF or DHF and their age categories, respectively.

Table 4. Distribution of Males and Females with DHF and DF

Sex	DHF	DF
Male	31	53
Female	25	37

Table 5. Distribution of Patient Age

Age	DHF	DF
0 – 15	21	9
16 – 20	3	7
21 – 25	5	8
26 – 30	4	12
31 – 35	4	14
36 – 40	3	18
41 – 45	2	7
46 – 50	7	2
51 - 78	7	10

The regression revealed that neither age nor sex were linked to dengue pathogenesis, as the R^2 values were not significant. The results are shown below. These results are of importance because it allows us to focus solely on genomic information.

- Age + Sex + Dengue: $R^2 = 0.0264$
- Dengue + SNPs + Sex: $R^2 = 0.2468$
- Dengue + SNPs + Age: $R^2 = 0.2749$
- Age + Sex + SNPs: $R^2 = 0.2761$

As referenced in Table 2, there are a myriad of innate genes involved in this research, with each SNP present in a different gene, ensuring a unique set of SNPs that are not present on the same loci. From the regression analysis, one of the most significant SNPs is rs7277299, which is an intront variant present in the MX1 gene. This SNP is an intron variant in a gene that encodes for a GTP-metabolizing protein. This protein aids in antiviral response through activation by type I/II IFNs. Once activated, the protein disrupts viral replication. SNP rs3740360 is found in *PLCE1*, otherwise known as phospholipase C epsilon 1. This gene is ultimately responsible for regulation of cell growth and secondary messengers. Mutations in *PLCE1* have been shown to cause early-onset nephrotic syndrome, meaning mutations are associated with some of the clinical manifestations of DHF, including edema and kidney malfunction. The kidney is one of the organs that is heavily affected by dengue in severe situations, with waste products from the blood unable to be filtered, causing acute renal failure, hematuria, and others DHF manifestations. The last significant SNP rs3132468 is present in *MICB,* otherwise known as MHC class I polypeptide-related sequence B. This gene encodes a protein that is an essential ligand for the NKG2D type II receptor. Ultimately, binding of this ligand activates the responses of NK cells, CD8 alpha beta T cells and gamma delta T cells. The process of ligand binding is stress-induced and is like MHC I molecules. Through the three statistically significant SNPs and their associated genes, one can see how variations in these innate genes can enhance dengue pathogenesis in favor of DHF. Individually, these SNPs seem to have a significant impact on DHF progression, however, it is

imperative to consider the combinations of different SNPs that are ultimately needed for individuals to be classified as susceptible to DHF infection.

The Advanced Neural Network (ANN) was utilized to identify patients as positive for infection with either DF or DHF, and its methodology can be visualized in Figure 10. The algorithm was implemented in Python as previously described in the study conducted by Davi and colleagues. The neural network was trained and introduced with the 13 core SNPs, meaning that there are 13 inputs and one output (DF or DHF). Once outputs were generated, they were compared against the clinicians' original classifications. Out of the 148 samples, the ANN correctly described 74.

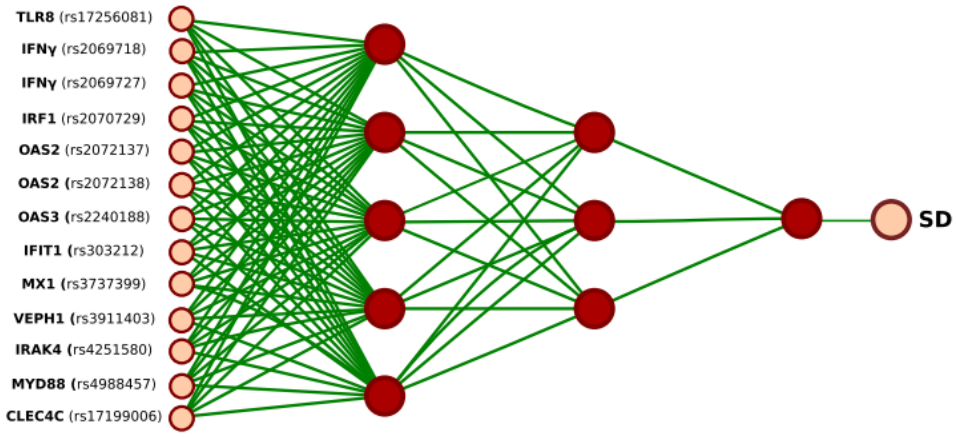

Figure 10. Schematic of the ANN

As stated previously, our genotyped samples included 92 DHF patients and 56 DF patients. The ANN described only 36 DHF patients and 112 DF patients. The performance of the first run of the ANN is summarized in Figure 7, where sensitivity, specificity, positive-predictive value (PPV), negative-predictive value (NPV), and accuracy were calculated.

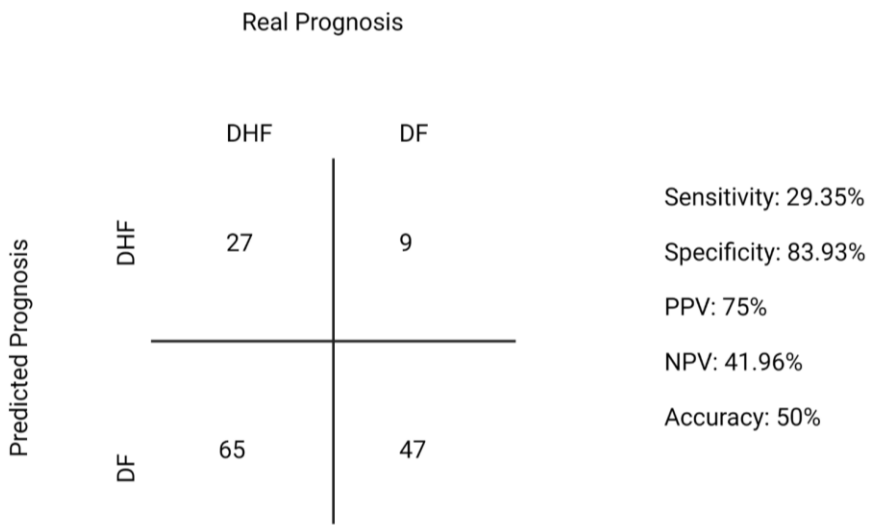

Figure 11. ANN performance evaluation for the first run

Referencing Figure 11, the chart sections can be divided into true positive (top left), false positive (top right), false negative (bottom left), and true negative (bottom right). True positive references 27 individuals who were diagnosed as positive for DHF infection by both the ANN and clinicians. True negative exploits the 47 individuals who were diagnosed as positive for DF infection by both the ANN and clinicians. 65 individuals represent the false negative, where the ANN classified the patients as positive for DF when the clinician diagnosis was DHF. Lastly, the false positive represents 9 individuals who were classified as positive for DHF via the ANN, when clinician diagnosis was DF. These variables were utilized to evaluate performance of the ANN. Sensitivity is described as the capacity of the ANN to correctly identify those with DHF infection, and specificity is known as the capability of the ANN to correctly identify those without DHF. From Figure 7, one can note the difference between sensitivity and specificity. The ANN proved to be specific, however, sensitivity was not up to par, suggesting that the algorithm performed better when diagnosing DF. This can also be seen when referencing the ANN diagnoses of 36 DHF individuals and a large 112 DF diagnoses. The ANN analysis was completed again due to the low

performance of the first run and due to the lack of column re-ordering after encoding. Results of the second run are presented in Figure 12. Accuracy remains the same at 50%, however, the other performance variables changed overall. The ANN correctly described 74 results out of the 148 samples. The correctly described results included 54 DHF diagnoses and 20 DF diagnoses. The sensitivity of the second run increased to approximately 59%, suggesting that the program has correctly described more DHF patients as compared to the previous run. Although performance remains low, the increase in sensitivity is of positive observation because the algorithm has correctly detected a greater percentage of DHF patients.

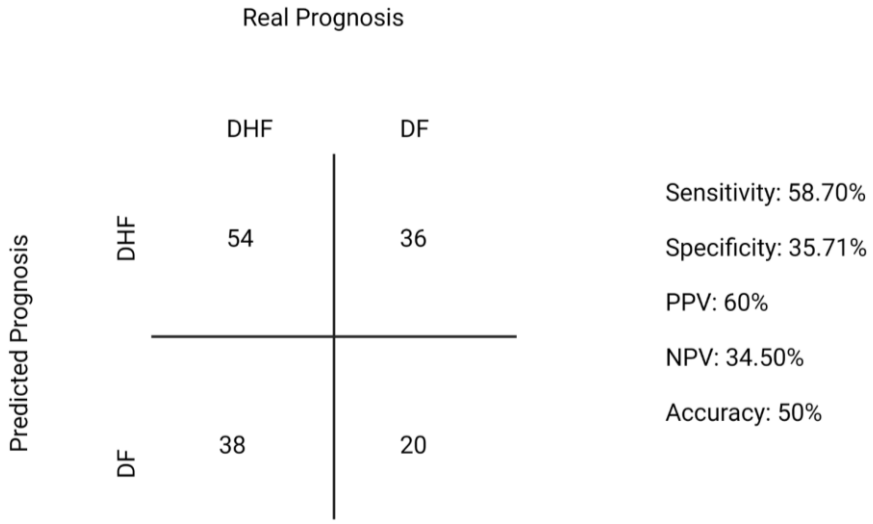

Figure 12. ANN performance evaluation for the second run

It is possible for aspects of the ANN to be changed to better fit the desired output. An example of this includes changing the weights of the input lines leading up to the activation bias. Another possibility would be to re-train the neural network, as it has previously been created and adapted to the SNPs and outputs from the previous study conducted by Davi and colleagues. The

goal is to generate a classification method for DHF that can be universally applied to all population subsets and geographic locations without having to change or enhance the ANN. Artificial Intelligence is an extraordinarily useful tool that can be applied to early classification, as in the case of DHF. The ANN exhibits positive performance results for the first two runs utilizing our SNP genotype data, however, much needs to be evaluated and changed to generate a successful and highly sensitive program.

4.0 Discussion

Dengue is of global public health concern as it has become one of the most well-known arthropod-borne diseases. Infection types range from asymptomatic, to dengue fever (DF), and to the life-threatening dengue hemorrhagic fever (DHF). Clinicians face challenges regarding dengue classification, as onset of DHF can occur at any point in the infection timeline. Individuals who progress to DHF are susceptible to an array of symptoms, including plasma leakage, edema, increased vascular permeability, hemorrhaging, and even shock (5). These severe clinical manifestations and sudden onset of symptoms are reasons as to why early classification of DHF is needed. This is especially true for individuals who reside in high-risk areas of dengue infection, and when outbreaks occur.

The exact mechanisms of DENV pathogenesis remain unknown. Factors such as cross-reactive T cells, antibody dependent enhancement (ADE), the complement system, and NS1 antigen have been shown to play a role in the progression to dengue hemorrhagic fever. The likelihood of progressing to DHF increases with each dengue infection (17). During primary infection, clinicians can diagnose patients based on physical manifestations in conjunction with laboratory testing for biological markers such as presence of viral RNA, NS1 antigen, and IgG and IgM antibodies. IgM antibodies offer short-term protection, and increase before tapering off, whereas IgG increases only after an increase in IgM, and offers long-term protection. Because of the 65-70% sequence homology shared by dengue serotypes, repeated infection is common. During secondary infections, IgG and IgM antibodies increase simultaneously (7). These biomarkers are useful in laboratory testing and DENV diagnoses. However, laboratory testing takes time, and susceptible individuals may not have that time available if they were to progress

to DHF. Early clinical identification of at-risk individuals would ensure that proper care and treatment would be provided if infection would occur.

Artificial intelligence has been implemented to offer early clinical identification for dengue in various forms. Some examples include utilization of Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), Support Vector Machine (SVM), and Logistic Regression to name a few (6). However, many early classification methods utilize laboratory testing for the biomarkers described previously in conjunction with artificial intelligence. The study outlined in this report is based off a previous study conducted by Davi and colleagues, where laboratory testing was not implemented. Instead, the methodology was based off solely human genome data.

The human genome has been shown to influence dengue infection and pathogenesis to DHF. Both innate and adaptive immune pathways are dependent on genomic composition, suggesting that the framework of each individual's genetic makeup has a direct influence on disease progression. Specifically, single-nucleotide polymorphisms (SNPs) in various loci/genes are associated with infection phenotypes. Advantages of SNP evaluation include gene tracking, drug response, evaluation of risk, and in the case of this study, early clinical identification (12). Common SNPs have a major and minor allelic frequency. Tracking of the minor allele is useful because the variant may be rare in one population, but common in another. In the previous study, a SVM-RFE (Support Vector Machine recursive feature elimination) was implemented to identify essential innate genes along with their corresponding SNPs that are involved in progression to DHF (12). The SVM-RFE is a form of artificial intelligence that is the first to be applied to solely genomic information. The results of the machine learning application generated a set of 13 essential SNPs in 11 innate genes, as described in Table 2. In conjunction with the SVM-RFE, an

Advanced Neural Network (ANN) was used for dengue diagnosis. The ANN utilized the 13 SNPs as inputs followed by 3 hidden layers, which included five, three, and one neuron respectively, which ultimately led to the output of DF or DHF (12). The sample set of the study included 75 DHF patients and 27 DF patients, and performance of the ANN was evaluated. Overall, the ANN performed exceptionally well, with an accuracy of 96%, and sensitivity and specificity of 100% and 85.71% respectively (12). The purpose of the experimentation outlined in this document is reflective of reproducing the ANN experimentation done by Davi and colleagues.

The 13 core SNPs generated from the previous study were utilized, in addition to seven others identified throughout literature reviews to entertain a total of 20 SNPs evaluated in the research project. The SNPs are present in a variety of innate genes, particularly those of the IFNγ pathway, which has been shown to play a role in DHF pathogenesis (12). The linear regression performed pointed out three SNPs in particular that seemed to share a significant relationship with dengue. SNPs rs7277299, rs3740360, and rs3132468 are found in genes *MX1, PLCE1*, and *MICB*, respectively. The murine myxovirus resistance 1 (MX1) gene encodes a protein that inhibits viral entry and replication by blocking incoming virus particles that have penetrated the host cell via endocytosis. This gene is also known to reduce viral titers of the DENV-2 serotype by its inhibition mechanisms through interferon stimulated genes, although the exact mechanisms are still not understood (7). Mutations present in this gene as in the case of rs7277299, have a direct effect on failure to eliminate viral entry and replication, thus enhancing pathogenesis of DHF. Mutations present in Phospholipase C epsilon 1 (PLCE1) are associated with early nephrotic syndrome, which include characteristics of proteinuria, swelling in the legs, ankles, and feet, and excess fluid retention (12). Given these clinical manifestations, it is suggested that rs3740360 has a direct effect on endothelial integrity due to its presence in *PLCE1*. The SNP rs3132468 is found in the MICB

gene, otherwise known as the MHC class I polypeptide-related sequence B gene. *MICB* encodes a protein that binds to the NKG2D type II receptor, which activates both NK and CD8 cells (14). Although exact mechanisms remain unclear, the link between *MICB* and DHF suggest that mutations present in this gene contribute to the dysfunction of NK and CD8 cells which lead to disease progression. From the regression, it was observed that the most statistically significant SNP was rs3740360, which resides in *PLCE1*. However, the regression considers each SNP individually and its relationship with dengue, not in combination as the ANN does. These three SNPs do have a slightly profound effect individually, as the R^2 produced from the linear regression stated that 13.24% of the variation in dengue can be explained by its relationship with just these three SNPs. Another observation noted was the absence of these three novel SNPs in the core set of 13 that was generated from the previous study. This suggests that perhaps these three SNPs have been recently discovered regarding their relationship with dengue, however, it remains unclear as to why the SVM-RFE algorithm did not detect these three. Another possible explanation includes the sample set which was utilized in the previous study. It is possible that these three SNPs are less prevalent in the genotyped individuals from Recife, Brazil.

As previously stated, the composition of the ANN remained the same for implementation in our analyses. Although we utilized 20 SNPs in our study, the input of the original 13 SNPs was kept, leading to the one output of either DF or DHF. One would expect similar performance of the ANN between the two studies, although discrepancies were noticed. With the first run of the ANN, sensitivity was evaluated at approximately 30%, specificity at 84%, and accuracy at 50%. Out of the 148 genotyped samples, 74 were described correctly. The second run of the program exhibited sensitivity at 58%, specificity at 36%, and accuracy again at 50%. Again, the ANN described 74 samples correctly out of the 148. There are several hypotheses to explain these varying

observations between experiments. First, our study is utilizing a different sample size with a variety of individuals. Brazil is home to a wide variety of ethnic groups subsets, suggesting different genetic makeups. As previously stated, certain minor alleles rare in one population may be common in another, which could be the case in this scenario. This factor would have a direct effect on the results, as the ANN was first trained to the cohort utilized in the previous study. However, the purpose of the machine learning algorithm is to implement a universal procedure to detect individuals susceptible to developing DHF, regardless of ethnic origin. This is an example of one of the current problems that must be evaluated with the ANN. Another factor to consider is that the distribution of dengue patients utilized in our ANN analysis included 92 DHF and 56 DF patients, which is not indicative of the general population of Brazil, which could indeed bias the accuracy and precision of the instrument. On the other hand, this observation would not affect the specificity and sensitivity results, which are independent of dengue prevalence. As reported in the previous study, the sensitivity of the ANN was 100%. Our study generated a sensitivity of 58%, which is not up to par considering the severity of DHF and the need for accurate classification. A possibility of increasing the performance for our cohort to gain more accurate results would be to change the weights of the inputs. Although we are still using 13 inputs as was done previously, we do have 20 SNPs in total that need to be accounted for. To use all 20 SNPs as the inputs, the neural network would need to be trained again. This would impact the results, although it is not known up to this point if performance would be enhanced or not.

Ultimately, machine learning technology in the form of the advanced neural network has the potential to offer early clinical identification for individuals susceptible to dengue hemorrhagic fever based solely on their genetic composition. The results highlighted in this study exhibit sensitivity and accuracy in the beginning stages of development that have the capacity to be

utilized as a predictive tool. It has been previously shown that the ANN is able to predict dengue hemorrhagic fever based on a set of single nucleotide polymorphisms. Key advantages to this methodology include application at any disease stage, especially before infection, as well as utilization of a broad choice of human sample tissue for DNA extraction. The lack of laboratory and clinical testing needed to perform this analysis is also an advantage, as results can be generated quickly. This is especially of importance in areas of high dengue prevalence, or when outbreaks occur, to ensure at risk individuals are properly cared for before they develop life-threatening conditions. Lastly, this methodology can be extended to other genetically influenced diseases. From this, target genes can be determined which are of importance in therapeutics. Overall, the Advanced Neural Network displays promising implementation for the early classification of dengue hemorrhagic fever.

Appendix A Important Abbreviations Used in This Report

DHF	Dengue hemorrhagic fever
DF	Dengue fever
ML	Machine Learning
ANN	Advanced Neural Network
ADE	Antibody Dependent Enhancement
SNPs	Single-nucleotide polymorphisms
SVM-RFE	Support Vector Machine Recursive Feature Elimination
DENV	Dengue virus

Appendix B Supplementary Figures

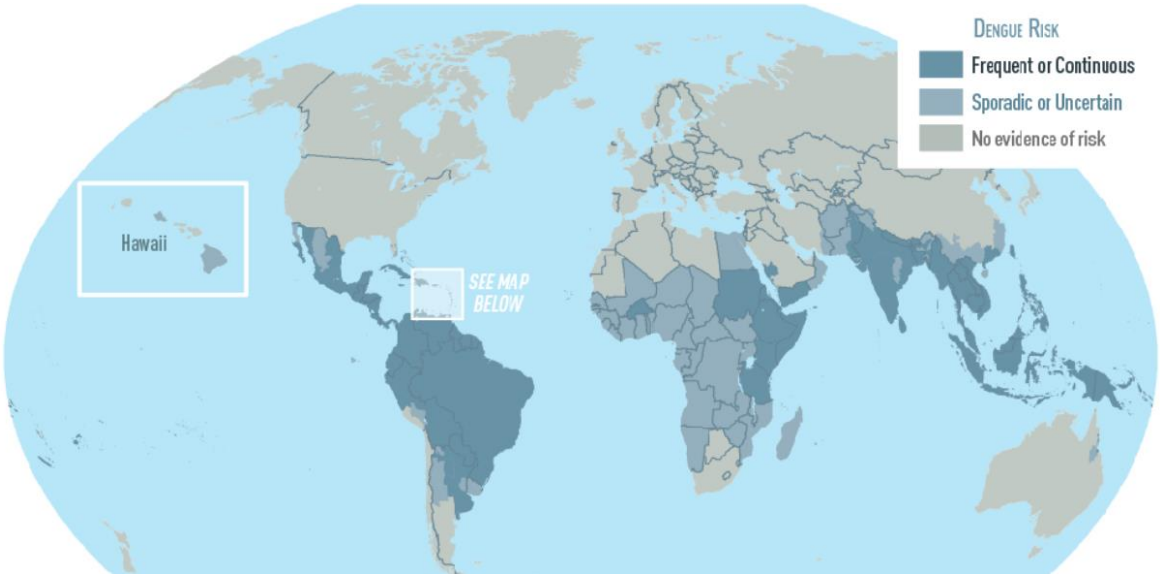

Figure 13. Prevalence of dengue globally

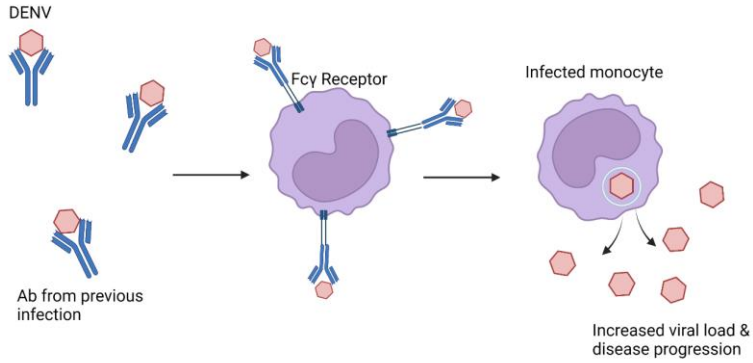

Figure 14. Mechanism of Antibody Dependent Enhancement (ADE)

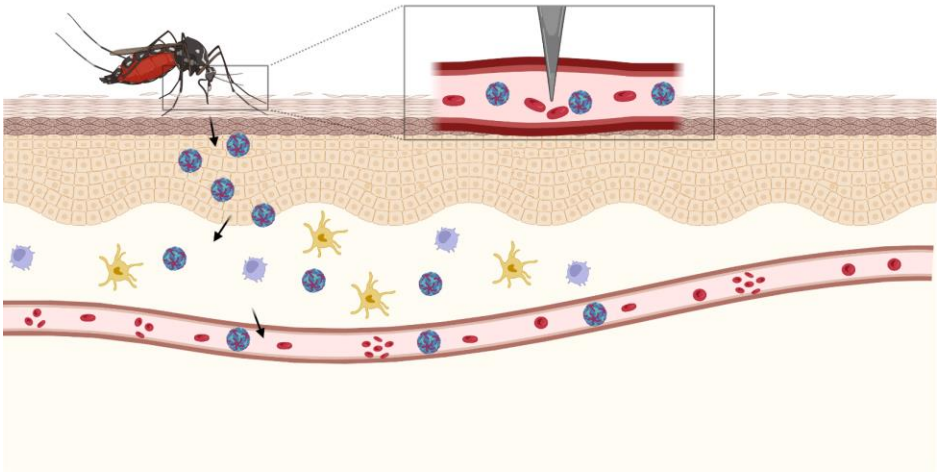

Figure 15. Initial DENV infection and pathogenesis

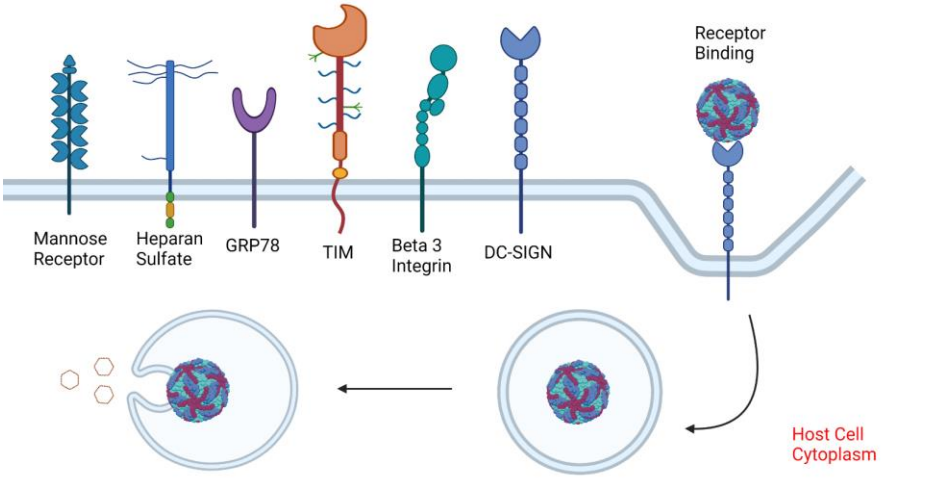

Figure 16. Examples of ubiquitous host cell receptors utilized by DENV to facilitate entry

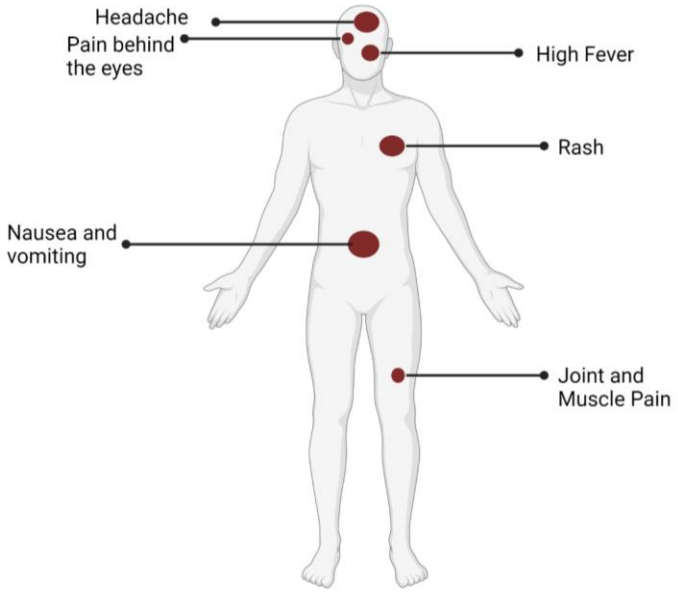

Figure 17. Symptoms of DF

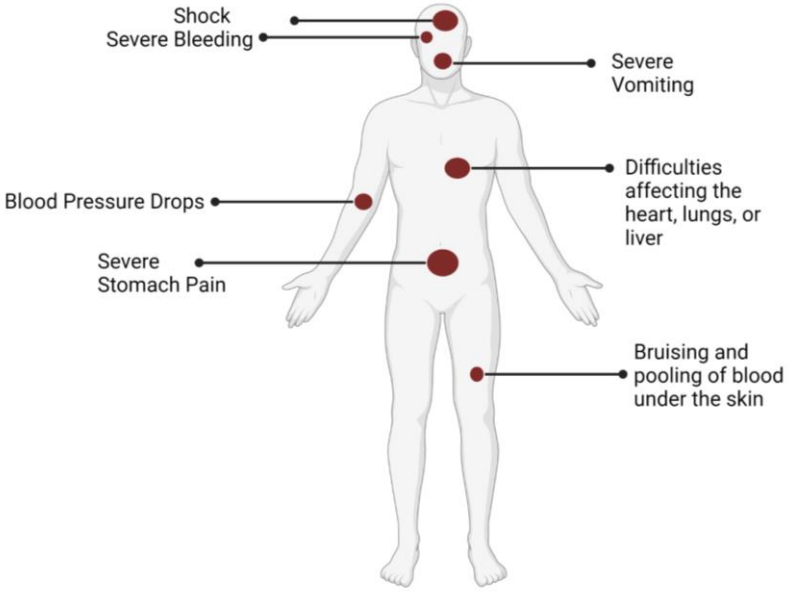

Figure 18. Symptoms of DHF

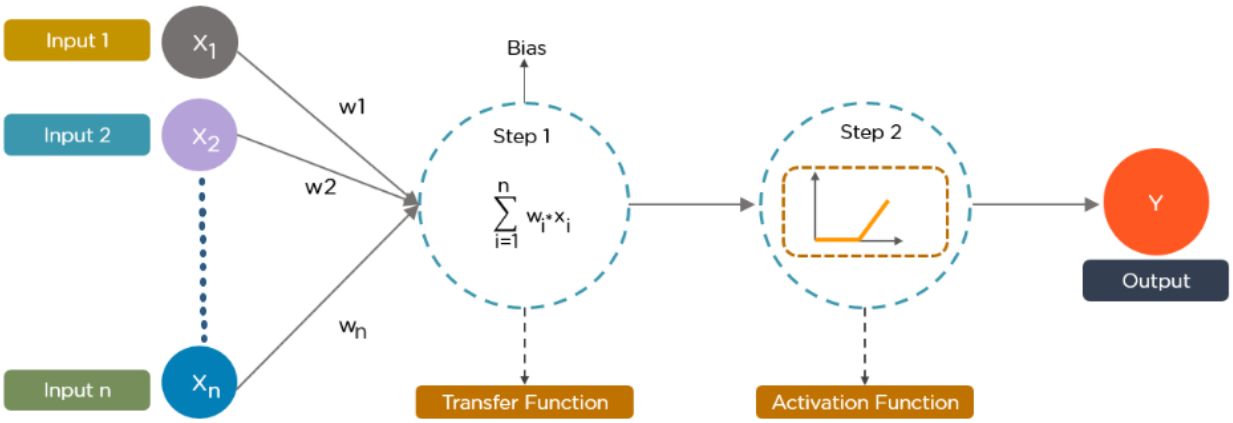

Figure 19. Visualization of ANN components

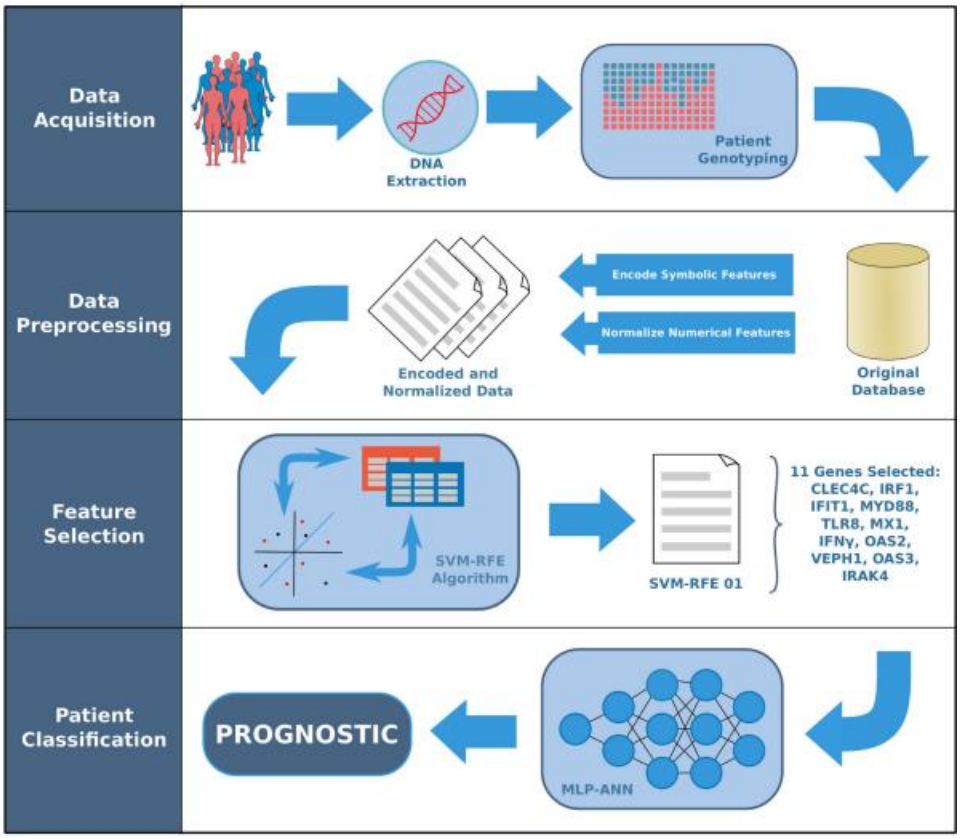

Figure 20. Methodology implemented from previous study conducted by Davi and colleagues

Bibliography

1. Fares RC, Souza KP, Anez G, Rios M. 2015. Epidemiological Scenario of Dengue in Brazil. Biomed Res Int 2015:321873.

2. Murugesan A, Manoharan M. 2020. Dengue Virus, p 281-359, Emerging and Reemerging Viral Pathogens doi:10.1016/b978-0-12-819400-3.00016-8.

3. CDC. Dengue. https://www.cdc.gov/dengue/index.html. Accessed March 3, 2022.

4. do Nascimento IDS, Pastor AF, Lopes TRR, Farias PCS, Goncales JP, do Carmo RF, Duraes-Carvalho R, da Silva CS, Silva Junior JVJ. 2020. Retrospective cross-sectional observational study on the epidemiological profile of dengue cases in Pernambuco state, Brazil, between 2015 and 2017. BMC Public Health 20:923.

5. Bhatt P, Sabeena SP, Varma M, Arunkumar G. 2021. Current Understanding of the Pathogenesis of Dengue Virus Infection. Curr Microbiol 78:17-32.

6. Huang SW, Tsai HP, Hung SJ, Ko WC, Wang JR. 2020. Assessing the risk of dengue severity using demographic information and laboratory test results with machine learning. PLoS Negl Trop Dis 14:e0008960.

7. Silva MM, Gil LH, Marques ET, Jr., Calzavara-Silva CE. 2013. Potential biomarkers for the clinical prognosis of severe dengue. Mem Inst Oswaldo Cruz 108:755-62.

8. Martina BE, Koraka P, Osterhaus AD. 2009. Dengue virus pathogenesis: an integrated view. Clin Microbiol Rev 22:564-81.

9. Tremblay N, Freppel W, Sow AA, Chatel-Chaix L. 2019. The Interplay between Dengue Virus and the Human Innate Immune System: A Game of Hide and Seek. Vaccines (Basel) 7.

10. Wang WH, Urbina AN, Chang MR, Assavalapsakul W, Lu PL, Chen YH, Wang SF. 2020. Dengue hemorrhagic fever - A systemic literature review of current perspectives on pathogenesis, prevention and control. J Microbiol Immunol Infect 53:963-978.

11. Pang X, Zhang R, Cheng G. 2017. Progress towards understanding the pathogenesis of dengue hemorrhagic fever. Virol Sin 32:16-22.

12. Davi C, Pastor A, Oliveira T, Neto FBL, Braga-Neto U, Bigham AW, Bamshad M, Marques ETA, Acioli-Santos B. 2019. Severe Dengue Prognosis Using Human Genome Data and Machine Learning. IEEE Trans Biomed Eng 66:2861-2868.

13. Riboldi E, Daniele R, Parola C, Inforzato A, Arnold PL, Bosisio D, Fremont DH, Bastone A, Colonna M, Sozzani S. 2011. Human C-type lectin domain family 4, member C (CLEC4C/BDCA-2/CD303) is a receptor for asialo-galactosyl-oligosaccharides. J Biol Chem 286:35329-35333.

14. Feng H, Zhang YB, Gui JF, Lemon SM, Yamane D. 2021. Interferon regulatory factor 1 (IRF1) and anti-pathogen innate immune responses. PLoS Pathog 17:e1009220.

15. Poonpanichakul T, Chan-In W, Opasawatchai A, Loison F, Matangkasombut O, Charoensawan V, Matangkasombut P, Thailand D. 2021. Innate Lymphoid Cells Activation and Transcriptomic Changes in Response to Human Dengue Infection. Front Immunol 12:599805.

16. Tuiskunen A, Monteil V, Plumet S, Boubis L, Wahlstrom M, Duong V, Buchy P, Lundkvist A, Tolou H, Leparc-Goffart I. 2011. Phenotypic and genotypic characterization of dengue virus isolates differentiates dengue fever and dengue hemorrhagic fever from dengue shock syndrome. Arch Virol 156:2023-32.

17. Zhao Y, Amodio M, Vander Wyk B, Gerritsen B, Kumar MM, van Dijk D, Moon K, Wang X, Malawista A, Richards MM, Cahill ME, Desai A, Sivadasan J, Venkataswamy MM, Ravi V, Fikrig E, Kumar P, Kleinstein SH, Krishnaswamy S, Montgomery RR. 2020. Single cell immune profiling of dengue virus patients reveals intact immune responses to Zika virus with enrichment of innate immune signatures. PLoS Negl Trop Dis 14:e0008112.