Predicting Case Fatality of Dengue Epidemic: Statistical Machine Learning Towards a Virtual Doctor

S. Chattopadhyay[1], A.K. Chattopadhyay[2,*] and E.C. Aifantis[3]

[1]*Acculi Labs Pvt. Ltd., Bangalore 560098, Karnataka, India*

[2]*Department of Mathematics, Nonlinearity and Complexity Research Group, Aston University, Birmingham B4 7ET, UK*

[3]*Laboratory of Mechanics and Materials, Aristotle University of Thessaloniki, GR 54124, Greece*

Abstract: Dengue fever is a self-limiting communicable viral disease, transmitted through mosquito bites. Its Case Fatality Grade (CFG) varies across population due to variations in viral load, immunity of the patient, early diagnosis, and availability of high-end treatment facility. This study describes an initial effort to automate the process of Dengue CFG predictions. Two established Statistical Machine Learning (SML) algorithms, Multiple Linear Regressions (MLR) and Multinomial Logistic Regressions (MnLR), are combined to substitute the existing Deep Learning methods for clinical decision making. We consider a vector of eleven sign-symptoms (independent variables), each weighted between [0,1] on a 3-point scale - 'Mild' (CFG<=0.33), 'Moderate' (0.33<CFG< 0.66), and 'Severe' (CFG>0.66). Results show that both classifiers are effective in early screening with similar accuracy levels (68% for MLR versus 72% for MnLR) although precision levels are far superior with MnLR (88%) than MLR (61%). This study is a futuristic step towards Machine Learning (ML) aided clinical diagnostic paradigms, as an alternative to computationally intensive Artificial Intelligence.

Keywords: Statistical machine learning, Dengue epidemic, Multiple linear regressions, Multinomial logistic regressions.

1. INTRODUCTION

Dengue is a deadly, communicable, but a self-limiting disease with mortality rate of ca 1% if detected early and attended by proper medical care. However, in the course of morbidity, even with medical care, mortality can reach up to 2-5% and close to 20% when left untreated (Smith, *et al*., 2019). Therefore, early detection by rapid antigen test and optimum therapeutic care is required in preventing the high mortality rate. The root of the medical problem lies in the accuracy of clinical diagnostics that amounts to identifying the correct grade of sign-symptoms which are patient subjective in nature. Moreover, interpretation of the Case Fatality Grade (CFG) also varies with the clinician in charge. All of these often contribute to under or over diagnosis of the illness, be it acute or chronic (Ashish, Chattopadhyay, Gao, & Hui, 2019). The present study is targeting to first automate the CFG mechanism based on self-consistent data training and thereby improve the quality, *i.e*. accuracy of diagnostics using a combination of Statistical Machine Learning (SML) techniques.

SML classifiers are the popular breed of Machine Learning (ML) algorithms and statistics, due to their ease of use, although not particularly high on the prediction accuracy level as also in probabilistic extrapolation of data. The more advanced data analytic trends are now reliant on Deep Learning (DL), which is computationally complex but more accurate for larger samples for training (Naeem, Jamil, Khan, & Nazir, 2020). Based on protracted experience as a practicing clinician (first author), an underlying premise of this work is to accord low weightage to the need for high accuracy levels in diagnosis, screening, prognostic evaluations. An accuracy threshold of 70% has been set in this work that is drawn from the accuracy level registered for a specialist consultant, which is 71% (Richens, Lee, & Johri, 2020), and employ two SML classifiers: Multiple Linear Regressions (MLR) and Multinomial Logistic Regressions (MnLR). With such 'hard' classifiers, model-fitting and prediction accuracy may be a challenging task and data training with small sample size could be an issue (Miguel-Hurtado, Guest, Stevenage, Neil, & Black, 2016). A key aspect of this study is to adapt to this technical challenge.

SML classifiers (SCs) have seen key success in analyzing Mosquito-borne communicable diseases, ranging from prediction to demographic distributions, incidence and prevalence rates in various populations and geographical locations. *Malaria* outbreak rate has been recently predicted as an impact outcome of ambient variables like rainfall, temperature, humidity, and rate of Plasmodium falciparum (causing cerebral malaria) in Maharashtra, India (Comert, Begashaw, & Turhan-Comert, 2020). The study analyzed an

*Address correspondence to this author at the Department of Mathematics, Nonlinearity and Complexity Research Group, Aston University, Birmingham B4 7ET, UK; E-mail: a.k.chattopadhyay@aston.ac.uk

E-ISSN: 2311-8792/21

ensemble of three ML algorithms: Random forest, Gaussian processes and Logistic regressions. The study has demonstrated that each algorithm could accurately predict the occurrence of cerebral malaria without any true or false positive cases at almost 100% accuracy. Detection of Plasmodium parasite in the red blood cell (RBC) images (obtained by microscopy) using Convolutional Neural Network (CNN)-based Deep Learning approach again showed that CNN was able to classify RBCs with and without Plasmodium within it at an accuracy rate of 97% (Narayanan, Ali, & Hardie, 2019).

Recently, weekly cases of Dengue eruption in Columbia have been predicted for a period of 12 weeks beyond the last data study, using Random Forest (RF) and Artificial Neural Network (ANN) (Zhao, *et al*., 2020). Both algorithms demonstrated reasonable accuracy in prediction with RF outperforming ANN. Additional research on Dengue prediction in the Indian scenario was attempted (Kapoor, Ahuja, & Kadyan, 2020), aiming a predictive model to diagnose dengue disease at early stages, ensuring timely intervention. More generalized approaches based on modeling virus propagation (Farafonov & Nerukh, 2019; Tarasova & Nerukh, 2018; Tarasova, et al, 2017) studied the biochemistry of viral capsid environment but again not the propagation itself.

While Machine and Deep Learning based data modeling techniques are the toast of the day, traditional approaches to analyze epidemic propagation have been through continuum modeling (Murray 2002; Murray 2003) and statistical modeling (Chattopadhyay & Chattopadhyay 2021) that led to excellent qualitative understanding of the evolution of epidemics and their

Table 1: Overview of the Methodology Consisted of Approaches and Techniques

Approaches	Techniques
Data Collection and Structuring (rulebase creation)	*Tracking* Dengue cases (N=100) in several hospitals/health centers in Southern part of India for past two years (2018 and 2019). *Clinical validation:* All cases are Dengue positive, tested with rapid kit test (NS1 antigen – single stranded RNA) and ELISA that detects immunogenic response (determined by the levels of IgM – active cases and IgG – chronic cases or past infections) as an immunogenic response to the viral antigen (Nagar, Savargaonkar, & Anvikar, 2020). *Identifying* common/typical Independent variables (sing-symptoms) and its corresponding Case Fatality Grade (CFG). *Assigning weights* [0, 1] to sign-symptoms (independent variables) in a 3-point scales as 'Mild' (values <=0.33), 'Moderate' (values >0.33 to 0.66), and 'Severe' (values >0.66) by a set of five medical doctors with average experience of 10 years. *Assigning class labels* to Dengue CFG (dependent variable) as '0', '1' and '2' as 'Mild', 'Moderate' and 'Severe', respectively by another set of five medical doctors (who had not participated in assigning weights of sign-symptoms) to reduce clinical classification bias. *Data structuring in the form of rulebase:* Each sign-symptom and the corresponding CFG are *weighted* [0, 1] by a set of 10 *medical doctors* having average experience of 10 years in clinical practice to set an *IF-THEN rulebase*. Each case serves as one rule. Therefore, the structure of data is actually a set of 100 rules. It is worth mentioning here that weights [0, 1] are assigned by 5 doctors, while the remaining 5 doctors have assigned the corresponding class labels for Dengue CFG.
Statistical Data Mining	To note the *statistical features* of the data using *Descriptive statistics*, *significance test* to find out the significant or most important sign-symptoms that determine high CFG using *Principal Component Analysis (PCA)* (Swathi & Pothuganti, 2020). *Correlation* and *Covariance* of the variables to note inter-dependency of sign-symptoms (Dacosta-Aguayo, Wylie, DeLuca, & Genova, 2020). *Normal distribution* plot to see the shape (skewness) of the data (Sumair, *et al*., 2020). *Probability distribution (Q-Q plot)* to see how data points are scattered across mean value (Esteves, Caramelo, & Ribeiro, 2020).
Machine Learning (ML)	*Data wrangling* – Consists of two steps, data preprocessing and data scaling. *Model development and model fitting* *Statistical Classifier development* - Two *classifiers* developed for Dengue CFG prediction. *Multiple Linear Regressions (MLR)* (Ahmed, Jalal, & Kim, 2020) *Multinomial Logistic Regressions (MnLR)* (Ahmed, Jalal, & Kim, 2020) Each classifier has been trained using 75% training case data, and validated on the remaining 25% test cases using *stratified k-fold Crossvalidation (k-fC)* (Valavi, Elith, Lahoz-Monfot, & Guillera-Arroita, 2020): *Classifier's performance testing by calculating accuracy* and *precision* (Sanders, Servaas, & Slagt, 2020) Testing classification *accuracy* – both label-wise and overall (see equation 1). Testing *precision* of classification – overall (see equation 2).

probabilistic decay patterns. Of late, these two approaches have been merged such that the patterns derived from the continuum models are also quantitatively accurate, ensured through probabilistic parameterization using statistical modeling techniques (Grela *et al* 2019) and Bayesian Monte Carlo methods (Chattopadhyay *et al* 2021). The present article, though, is exclusively a study in data modeling.

There are two focal points in this study. Our first target is to develop simple statistical classifiers with a small sample to predict Dengue CFG at an acceptable accuracy level that is comparable to that of a specialist clinician, as a groundbreaking extension of telemedicine. Given a set of independent parameters (here, it is sign-symptoms), the classifiers would be able to predict CFG within the aforementioned tolerance limit of that we set at 70% – a threshold value adapted from human clinician accuracy that is averaged at *71.4%* (Richens, Lee, & Johri, 2020). Second, given the fact that about 4 billion population worldwide are eventually at risk of being infected with Dengue virus, early detection could be the key to survival (Nagar, Savargaonkar, & Anvikar, 2020). A clear knowledge of the infection level, that is 'mild', 'moderate' or 'high', could lead to timely therapeutic interventions. Our study is, thus, closely associated with healthcare provisions and policy formulations.

2. METHODOLOGY

An overview of the methodology flowchart (**Approaches** and **Techniques**) is given in Table **1**.

The schematic in Figure **1** represents a flowchart that sequentially traces the methodological outlier presented in Table **1** above. The eventual target is to automate the human learning protocol by mimicking the .functional outliers through a combination of statistical processors, starting with data reading and ending with risk management.

2.1. Data Collection

Collecting clinical data is a difficult task that has multiple challenges due to constraints on *availability/access*, *authenticity/truthfulness*, and underlying *ethical issues* due to sensitivity and privacy of the patients. The present study uses data collected between 2016 and 2019 in multiple hospitals and clinics in India for 100 patients by the first author himself in his faculty of a professional clinician, in active collaboration with other clinician colleagues. The patients were between 20 and 68 years in age (average age 42.6 years). The key novelty of the data profile is in its originality and longitudinal breadth as the data were collected by frontline clinicians who were deputed with the responsibility of first analyzing and then deciding on the therapeutic regime.

2.2. Data Curation and Structuring

A team comprising 10 senior consultant physicians have categorized the data according to eleven common sign-symptoms (variables): *Fever* ***(F)***, *Sore throat* ***(S),***

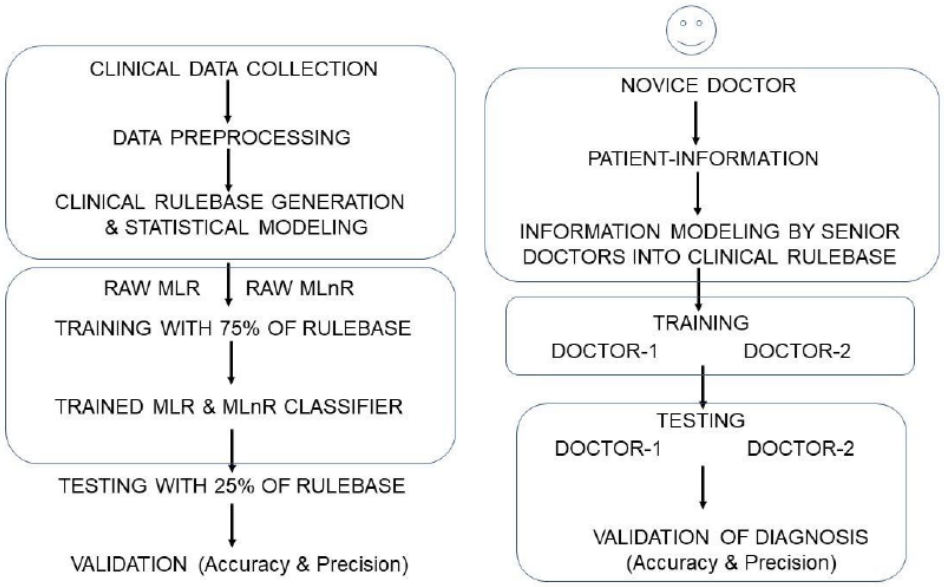

Figure 1: The VIRDOCD Machine Learning Control (MLC) protocol that is trained identically to medical doctors 1 and 2 (novice doctors) who learn under the guidance of senior doctors, the human equivalent of the advanced classifiers, who teach them the clinical rules and help pre-process (reduce noise) within the information received from the patient.

Headache ***(H),*** *Nausea* ***(N),*** *Vomiting* ***(V),*** *Stomachache* ***(S),*** *Myalgia* ***(M****), Rashes* ***(R),*** *Diarrhea* ***(D),*** *Joint pain* ***(J),*** *Bleeding gums* ***(B)***. Each sign-symptom is weighted between [0, 1] and its corresponding *CFG* ***(O)*** is labeled on a 3-point system: [0, 1, 2], where "0" indicates mild outcome, "1" represents moderate outcome and "2" is severe outcome. It is worth noting here that to prevent any diagnostic/prediction or classification bias; five out of the ten expert clinicians involved scored the sign-symptoms category (weight of 0.1) while the remaining five marked the corresponding class labels (mild = 0, moderate = 1, high = 2). Weights are essentially divided into a 3-point scale as values <=0.33 is 'Mild' (denoted by Blue-colored cells), >0.33 - <=0.66 as 'Moderate' (denoted by Green-colored cells) and >0.66 as 'Severe' (denoted by Magenta-colored cells). Therefore, each case is a weighted sign-symptoms vector leading to a labeled CFG as either 0 (as 'Mild'), 1 (as 'Moderate'), or 2 (as 'Severe'), which is shown in Table **2**, below. Similar coding is used to determine the CFG labels, as also seen in Table **2**. It is important to note here that prediction values <=0.74, 0.75-1.24, and >1.24 are labeled as 'Mild', 'Moderate', and 'Severe' by the clinical experts, where class labels are marked as float values in MLR-based classification, whereas in MnLR these predictions are exactly the class labels of '0', '1', or '2' and such label demarcations are not required.

The predictor classifiers are essentially trained with 75% of the rule-base while the remaining 25% of the rule-base have been used for testing individual classifier's performance using *k-fold Cross-validation (k-fC)*. After careful observation, each case represents a rule-base with the portfolio of 100 cases representing a rule-base grid comprising '*100 IF-THEN*' rules, *e.g.* if **'F'** has a weighted value of 0.4235 & **'ST'** has value of 0.9951 & **'H'** is 0.7885 & …& **'B'** has value of 0.2473 THEN **'O'** is '2'. In qualitative term, *IF* **'F'** is 'Moderate' & simultaneously both **'ST'** is 'Severe', **'H'** is 'Severe' & …& **'B'** is 'Mild', *THEN* **'O'** is graded as 'Severe'.

2.3. Statistical Data Mining and Machine Learning (ML)

To identify the *'significant'* sign-symptoms and the degree of their interdependence, Principal Component Analysis (PCA) has been performed. It is worth noting that we have not used PCA for feature (dimension) reduction purpose, as the number of features/sign-symptoms is only 11 and thus well-manageable. *Eigenvectors* and *Eigenvalues* of a covariance/correlation matrix is the core of any PCA, as they determine the 'significant' or 'principal' directions of a given feature space and its magnitude, respectively.

It is important to note that the classifiers used (MRL and MnLR) are trained with 75% of the rule-base obtained from the aforementioned group of experienced doctors and tested on 25% of the cases using *stratified K-fold-Cross-validation (KfC)*. Accuracy (%) in prediction is calculated as Eqn (1):

$$Accuracy = \frac{cp}{tp} * 100 \quad (1)$$

where '*cp'* denotes the number of correct predictions and '*tp'* is the number of total cases, expressed in %.

Precision (%) in prediction is calculated as in Eqn (2) after matching the classifier's prediction against human (10-panel expert) prediction, True positive (TP) cases are for the 'matches', while 'mismatches' lead to False Positive (FP) cases:

$$Precision = \frac{TP}{TP+FP} * 100 \quad (2)$$

(a) Multiple Linear Regressions (MLR): It is a popular statistical classifier used to map relationships between two or more independent variables (eleven Dengue sign-symptoms) and the dependent variable (Dengue Case Fatality Grade, CFG). It also predicts the value or class label (*i.e.*, the severity grade of either '0' or 'Mild', '1' or 'Moderate' or '2' or 'Severe') of the dependent variable at a certain value *i.e.*, the

Table 2: Glimpse of each Case having CFG Labels with a given Set of Weighted Sign-Symptoms

F	ST	H	N	V	S	M	R	D	J	B	O
0.4235	0.9951	0.7885	0.7903	0.9890	0.0129	0.5451	0.4906	0.5829	0.2238	0.2473	2
0.0168	0.2689	0.1986	0.8332	0.6194	0.3411	0.0028	0.7765	0.3140	0.1712	0.4704	1
0.2407	0.0060	0.9032	0.0126	0.7929	0.2719	0.8087	0.3304	0.8788	0.7399	0.0764	0

weight assigned by the doctors as <=0.33 or 'Mild', between 0.33 - 0.66 'Moderate' or >0.66 'Severe'. MLR relies on basic assumptions like homoscedasticity (or homogeneity of variance), independence of observations (low correlation between any two observations), to define linear regression between variables as follows:

$$y = B_0 + B_1 x_1 + B_2 x_2 + \cdots + B_n x_n + \varepsilon \quad (3)$$

where the following notation is adopted: 'y' = predicted CFG; 'B_0' = the y-intercept (*i.e.*, value of 'y' while all the independent variables/sign-symptoms are set to be '0'); 'B_1x_1' = the regression coefficient 'B_1' of first sign-symptom 'x_1' (which is 'Fever' or 'F') on the predicted 'y' value; 'n' = number of sign-symptoms, *i.e.*, 11; and 'ε' = model error or the deviation of the estimated/predicted value from that of the target value.

(b) Multinomial Logistic Regressions (MnLR): This technique, which is also commonly called Softmax regression, generalizes logistic regression to multiclass problems with more than two (*i.e.* multiple) outcomes (here the Dengue CFG grades/categories of 0, 1, or 2). Given a set of independent variables (here, the sign-symptoms each having three categories – 'Mild', 'Moderate', and 'Severe', having weights of <=0.33, between 0.33 - 0.66, and >0.66, respectively), which may have real-value, binary value or categorical value and it is able to predict different possible outcomes of a categorically distributed dependent variable (*i.e.*, the Dengue CFG grades/categories/class labels of 0, 1, or 2) using the Logit function that assumes log distribution. The MnLR model for Probability P(y=j|, for j=1, 2, …k takes the following form:

$$P(\boldsymbol{\alpha}) = \frac{exp[w^T]\boldsymbol{\alpha}}{\sum_j exp[w_j^T]\boldsymbol{\alpha}_j} \quad (4)$$

where 'w' represents the model parameters, and 'w_0' is the bias value:

$$w=[w_0,w_1,\ldots,w_m]^T \quad (4a)$$

The 'm' independent variables or attributes or sign-symptoms are written as a vector

$$\tilde{a} = [1,a_1,\ldots,a_m]^T \quad (4b)$$

The denominator $\sum_j exp[w_j^T]\boldsymbol{\alpha}_j$ normalizes the probabilities over all classes *j*. MnLR classifier development comprises the following steps:

Step 1: Data Preprocessing: Data checked for 'null' values. If a null value is recorded, it is substituted by the median value of that column.

Step 2: Standardization of the Data: Standard scaling on the Data ensuring that standard deviation of each feature column is set at 1.

Steps 1 and 2 together complete Data Wrangling. The multinomial regression function consists of two functional layers- Linear prediction function that is commonly known as the 'logit layer' and Softmax function, which is also known as 'softmax layer'.

Step-3: Development of logit layer: Logit scores calculated for each possible outcome. Random weights (each of 11 sign-symptoms) and biases (each for 3 class labels of the predictor) are then assigned in the model.

Step-4: Development of the softmax layer: This layer comes into play when logit scores of each possible outcome are converted into probability values, that is values between 0 and 1.

Step-5: Measuring accuracy: Eqn (1) quantifies the accuracy of the model.

3. RESULTS AND DISCUSSIONS

In this section, we interpret the statistical modeling results through a clinical lens. Table **3** below shows a representative set of statistical outliers.

3.1. Clinical Correlation of the Values

Count: 99 denotes 0-99, i.e. 100 rule bases/cases

Mean and Median: indicates the average weights of each sign-symptom and the probable class label for any given rule/case. From Table **3**, it is observed that on average weight of each of the sign-symptoms and the class label fall under the grade 'Moderate', which means the tendency of the given weights is around 'Moderate' grade of sign-symptoms leading to 'Moderate' CFG.

Standard deviation (Std) is the data dispersion, which is fairly similar for all sign-symptoms, varying closely between 0.27-0.29. Hence, there's no outlier as such in the data.

Quartile deviation (25%, 50% and 75%) is a measure of central tendency measuring data dispersion (the range) within which 50% of samples lie. Values of sign-symptoms are also quite close to each

Table 3: Descriptive Statistics

	F	ST	H	N	V	S	M	R	D	J	B	O
Count	99											
Mean	0.464	0.483	0.546	0.499	0.517	0.531	0.569	0.481	0.484	0.476	0.519	0.960
Std	0.297	0.276	0.283	0.283	0.311	0.318	0.284	0.285	0.291	0.291	0.288	0.794
Min	0.001	0.006	0.019	0.001	0.011	0.007	0.003	0.002	0.011	0.011	0.014	0.000
25%	0.194	0.257	0.356	0.263	0.250	0.262	0.338	0.292	0.254	0.222	0.256	0.000
50%	0.471	0.487	0.570	0.542	0.518	0.560	0.545	0.481	0.446	0.450	0.539	1.000
75%	0.719	0.735	0.806	0.742	0.805	0.811	0.816	0.732	0.726	0.748	0.760	2.000
Max	0.965	0.995	0.982	1.000	0.998	0.998	0.991	0.980	1.000	0.991	0.983	2.000
Median	0.471	0.487	0.570	0.542	0.518	0.560	0.545	0.481	0.446	0.450	0.539	1.000
Skewness	-0.009	0.048	-0.232	-0.088	-0.077	-0.125	-0.190	0.001	0.228	0.211	-0.095	0.073
Kurtosis	-1.324	-1.152	-1.067	-1.162	-1.393	-1.356	-1.226	-1.120	-1.083	-1.380	-1.195	-1.406
z-scores(avg.)	0.000	0.000	1.306	-0.011	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
p-values(avg.)	0.501	0.498	0.508	0.503	0.503	0.505	0.506	0.500	0.491	0.491	0.504	0.497

other; as shown in the table, 50% weights of the sign-symptoms lie between 0.33 and 0.66 of the 3-point scale, which is labeled as 'Moderate'.

Skewness: values under -1 or over +1 denote over-skewed data. From Table **3**, none of the values point to such over-skewness, confirming a normal univariate distribution. Table **3** indicates that Sore throat (ST), Rash (R), Diarrhea (D) and Joint Pain (J) are positively skewed while the rest are skewed negatively. It is important to note that although 'F' (negative) and 'R' (positive) are oppositely skewed, respectively, their values are close to zero. Thus, on a first approximation, we set zero skewness for these variables.

Kurtosis: values falling under -2 or over +2 satisfy normal univariate distribution, which is noted for this dataset. From Skewness and Kurtosis numbers in Table **3**, it can be stated that the data are symmetrical.

Z-score: it is a statistical measure of a score's departure from the mean value. Z=0 means it lies on the mean. As in Table **3**, apart from Headache (H) and Nausea (N), Z-score values of all other sign-symptoms are '0' that means these fall on the mean. Z-score of 1.306 for H means it is 1 standard deviation above the mean. While Z-score of -0.011 for the sign-symptom 'N' indicates that it is 0.011 standard deviation below the mean. Summarily it can be ascertained that ca 90% of the data fall on the mean and hence symmetrical in nature.

p-value: in Table **3**, none of the values fall under 0.05 and thus it can be stated that there is no significant difference of means of each sign-symptom, thus endorsing null hypothesis.

The available data are visualized in two steps. First, the frequency of 11 sign-symptoms (independent variables) occurrence is extracted as histograms (Figure **2**). Second, the overall probability distribution function is analyzed, which interestingly follows a normal distribution (Figure **3**).

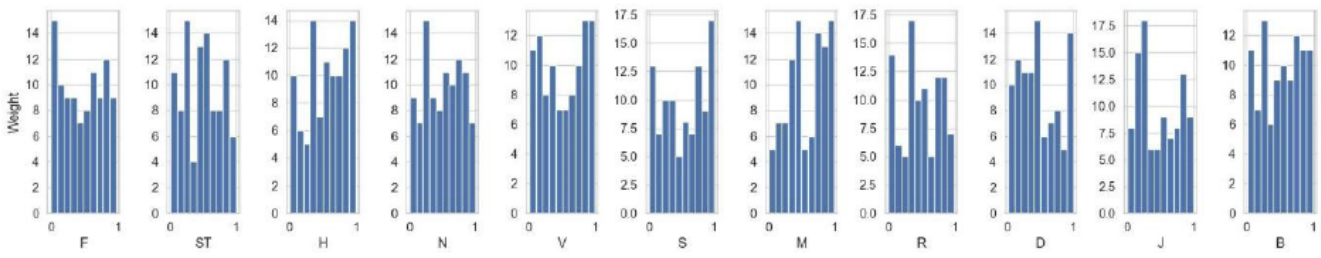

Figure 2: Histogram plots of each sign-symptom (weight vs frequency).

Table 4: Covariance statistics

	F	ST	H	N	V	S	M	R	D	J	B	O
F	0.088	-0.008	0.005	-0.013	-0.005	-0.007	0.015	-0.010	0.014	0.010	-0.010	0.012
ST	-0.008	0.076	-0.008	-0.001	0.001	0.006	0.000	0.004	0.000	-0.003	0.006	-0.034
H	0.005	-0.008	0.080	-0.003	-0.007	0.021	0.002	-0.008	-0.011	0.005	0.007	0.017
N	-0.013	-0.001	-0.003	0.080	0.009	-0.009	0.001	0.002	-0.007	-0.002	0.023	0.044
V	-0.005	0.001	-0.007	0.009	0.097	0.005	0.004	-0.003	0.015	0.004	-0.023	0.010
S	-0.007	0.006	0.021	-0.009	0.005	0.101	-0.003	0.000	0.003	-0.006	0.006	-0.026
M	0.015	0.000	0.002	0.001	0.004	-0.003	0.080	-0.006	0.004	0.010	-0.001	-0.002
R	-0.010	0.004	-0.008	0.002	-0.003	0.000	-0.006	0.081	0.000	-0.005	0.008	-0.024
D	0.014	0.000	-0.011	-0.007	0.015	0.003	0.004	0.000	0.085	0.009	-0.013	0.001
J	0.010	-0.003	0.005	-0.002	0.004	-0.006	0.010	-0.005	0.009	0.085	-0.013	0.033
B	-0.010	0.006	0.007	0.023	-0.023	0.006	-0.001	0.008	-0.013	-0.013	0.083	-0.038
O	0.012	-0.034	0.017	0.044	0.010	-0.026	-0.002	-0.024	0.001	0.033	-0.038	0.631

Table **4** shows covariance matrix between two sign-symptoms at a time. Positive covariance means when value of one sign-symptom increases, the value of the other sign-symptom also increases. Negative covariance is the interpretation when decrement of one sign-symptom's value decreases the value of the other sign-symptom. For *e.g*., we can see Nausea (N) and Headache (H) has negative covariance, while Joint pain (J) and Myalgia (M) has positive covariance. It is interesting to note that except for Nausea (N), Vomiting (V), and Myalgia (M), all other sign-symptoms have positive covariance with the CFG/Outcome (O) and so are clinically significant for interpretation of the CFG.

Table **5** shows the correlation matrix of the sign-symptoms and the CFG. It is interpreted as follows:

- Correlation value of '-1' or very close to '-1' indicates a perfectly *negative linear correlation* between two variables (refer to magenta colored cells in the Table). For *e.g*., in Table **5**, it is observed that Fever (F) has negative linear correlation with Sore throat (ST), Nausea (N), Vomiting (V), Stomachache (S), Rash (R), and Bleeding (B); Sore throat (ST) is negatively correlated with Headache (H), Nausea (N), and Joint pain (J); Headache (H) is negatively correlated with Nausea (N), Vomiting (V), Rash (R), and Diarrhea (D); Nausea (N) is negatively correlated with Stomachache (S), Diarrhea (D) and Joint pain (J); Vomiting (V) has negative correlation with Rash (R) and Bleeding (B); Stomachache (S) has negative correlation with Myalgia (M), Joint pain (J), and CFG/Outcome (O); Myalgia (M) has negative correlation with Rash (R); Rash (R) is negatively correlated with Diarrhea (D), Joint pain (J) and CFG/Outcome(O); Diarrhea (D) and Joint pain (J) has negative correlation with Bleeding (B); while Bleeding (B) has negative correlation with CFG/Outcome(O).

Table 5: Correlation statistics

	F	ST	H	N	V	S	M	R	D	J	B	O
F	1	-0.094	0.054	-0.158	-0.057	-0.079	0.175	-0.122	0.163	0.117	-0.112	0.051
ST	-0.094	1	-0.108	-0.011	0.006	0.067	0.005	0.054	-0.002	-0.032	0.077	-0.155
H	0.054	-0.108	1	-0.035	-0.076	0.233	0.022	-0.096	-0.135	0.061	0.082	0.074
N	-0.158	-0.011	-0.035	1	0.100	-0.104	0.014	0.020	-0.087	-0.023	0.284	0.194
V	-0.057	0.006	-0.076	0.100	1	0.053	0.050	-0.031	0.171	0.044	-0.257	0.041
S	-0.079	0.067	0.233	-0.104	0.053	1	-0.028	0.005	0.031	-0.068	0.062	-0.103
M	0.175	0.005	0.022	0.014	0.050	-0.028	1	-0.078	0.053	0.123	-0.009	-0.010
R	-0.122	0.054	-0.096	0.020	-0.031	0.005	-0.078	1	-0.003	-0.057	0.101	-0.104
D	0.163	-0.002	-0.135	-0.087	0.171	0.031	0.053	-0.003	1	0.104	-0.150	0.004
J	0.117	-0.032	0.061	-0.023	0.044	-0.068	0.123	-0.057	0.104	1	-0.152	0.144
B	-0.112	0.077	0.082	0.284	-0.257	0.062	-0.009	0.101	-0.150	-0.152	1	-0.165
O	0.051	-0.155	0.074	0.194	0.041	-0.103	-0.010	-0.104	0.004	0.144	-0.165	1

- Correlation value of '0' or very close to '0' indicates *no linear correlation* between two variables (refer to blue colored cells in the Table). For *e.g.*, in Table **5**, it is observed that Sore throat (ST) has no linear correlation with [Vomiting (V), Myalgia (M), and Diarrhea (D)]; Stomachache (S) with Rash (R); Myalgia (M) with Bleeding (B) and CFG/Outcome (O); and Diarrhea (D) with CFG/Outcome(O).

- Correlation value of '1' or very close to '1' indicates a perfectly *positive linear correlation* between two variables (refer to green colored cells in the Table). For *e.g.*, Fever (F) has perfectly positive linear correlation with Headache (H), Myalgia (M), Diarrhea (D), Joint pain (J), and CFG/Outcome (O). Similarly, Sore throat (ST) has positive correlation with Stomachache (S), Rash (R), and Bleeding (B); Headache (H) with Stomachache (S), Myalgia (M), Joint pain (J), Bleeding (B), and CFG/Outcome (O); Nausea (N) has positive correlation with Myalgia (M), Rash (R), Bleeding (B), and CFG/Outcome (O); Vomiting (V) is positively correlated with Myalgia (M), Diarrhea (D), Joint pain (J), and CFG/Outcome (O); Stomachache (S) is positively correlated with Diarrhea (D) and Bleeding (B); Myalgia (M) has positive correlation with Diarrhea (D) and Joint pain (J); Rash (R) is positively and linearly correlated with Bleeding (B); and finally Joint pain (J) is positively correlated to CFG/Outcome(O).

- This dataset is clearly non-normal in distribution (Figure **3** in Appendix) with an extended left tail (negative skewness) and slender right tail (positive skewness). The tip is notched and not purely bell shaped as the mean and median do not fall on each other at all time.

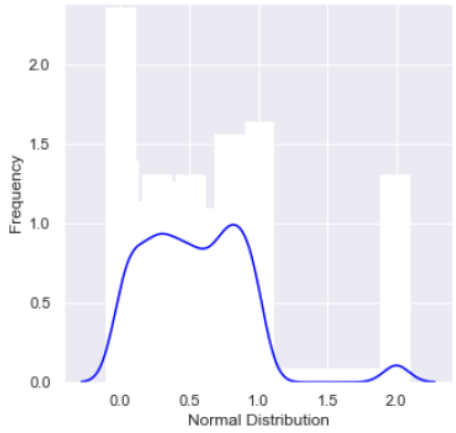

Figure 3: Normal distribution curve of Dengue dataset.

- The Probability distribution plot (Figure **4**) is actually a Quartile-Quartile or Q-Q plot. Its x-axis and y-axis are 'Theoretical Quantiles' and 'Ordered Values' or 'Sample Quantile', respectively, offering a visualization of how data are distributed around the mean value (depicted by the straight line in Figure **3**).

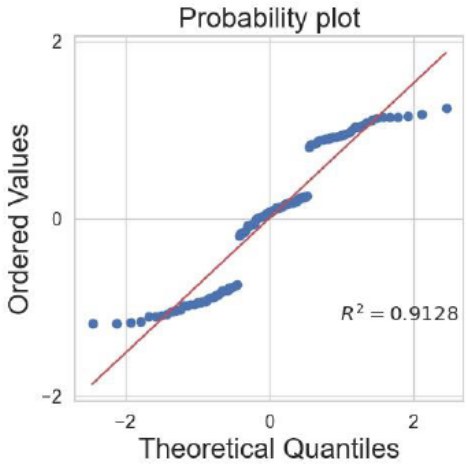

Figure 4: Probability distribution plot.

- Shapiro-Wilk Normality Test has been conducted to examine whether the variables have Gaussian distribution or not based on the p-value. It is seen for this data that p-value is 0.0035, which is <0.05, hence these are not Gaussian. R^2 value denotes goodness of model fit, which is 91%.

- Pearson's Correlation test: to test whether the factors are independent or not based on the p-value, which is 0.435, *i.e.*, >0.05, indicating that the factors are probably independent of each other.

3.2. Results of Principal Component Analysis (PCA)

This statistical technique is commonly used for exploratory data analysis and making predictive models. Another key usage is in dimensional reduction by projecting each data point onto the first few principal components, expressed as Eigenvalues (magnitude) and Eigenvectors (direction). Eigenvalues thus calculated are F = 19.17, ST = 36.99, H = 52.18, N = 63.71, V = 73.53, S = 80.73, M = 86.88, R = 92.06, D = 95.68, J = 98.13, and B = 99.43. Figure **5** shows the PCA plot of '%Variance Explained' (y-axis) vs. '# of Features' (x-axis). High Eigenvalues are found for sign-symptoms 'Stomachache (S)', 'Myalgia (M)', 'Rash (R)', 'Diarrhea (D)', 'Joint pain (J)', and 'Bleeding (B)', when >80% (marked by red vertical line) is considered to be

a high cut-off mark, selected arbitrarily (refer to Figure **5**).

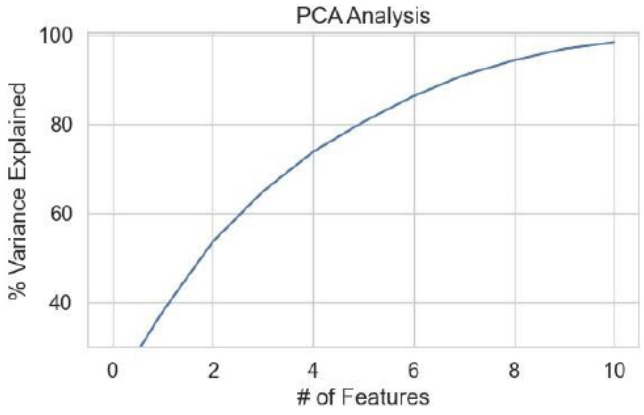

Figure 5: Principal Component Analysis (PCA) plot.

PCA has been carried out as the part of machine learning analysis to understand how percentage variances are placed on the corresponding features and not for feature extraction or elimination (*i.e.*, the dimensionality reduction) as the number of features (sign-symptoms) are only 11 and manageable. PCA suggests that variables/features/sign-symptoms are independent of each other.

3.3. Results of Multiple Linear Regressions (MLR)

Regression coefficients (*B*-values as per equation 2) can be seen in Table **6**, below.

Table 6: Regression coefficients

B0	0
B1	0.088203
B2	-0.32084
B3	0.248072
B4	0.742001
B5	-0.08632
B6	-0.1613
B7	-0.11357
B8	-0.18688
B9	0.000856
B10	0.279788
B11	-0.60162

For prediction (y) coefficient values are multiplied with the corresponding independent variables and the products are summed up as shown in equation (2). B_0 is the intercept/slope and here its value is '0'. Prediction accuracy is calculated using equation (1).

Table **7** shows the actual/target vs. predicted class label for 25 test cases. Class labels have been created as per the opinion of five of the ten experienced doctors and is shown in the last row of the table. Matched class labels are marked as green for easy visualization, while grey cells indicate all True positive cases and red cells are indicative of False positive cases.

Table 7: Actual vs. Predicted results of 25 test cases by MLR Classifier

Test_Case	Pred	Label	Target	Label
0	0.728735	Mild	0	Mild
1	1.300217	Severe	2	Severe
2	1.289668	Severe	2	Severe
3	0.871166	Moderate	1	Moderate
4	1.023017	Moderate	2	Severe
5	0.74853	Moderate	1	Moderate
6	1.14356	Moderate	0	Mild
7	1.430796	Severe	2	Severe
8	0.705385	Mild	0	Mild
9	0.731374	Mild	0	Mild
10	0.943354	Moderate	1	Moderate
11	0.717203	Mild	0	Mild
12	0.672041	Mild	2	Severe
13	1.20287	Moderate	1	Moderate
14	1.021389	Moderate	2	Severe
15	0.982583	Moderate	2	Severe
16	0.882127	Moderate	1	Moderate
17	0.443533	Mild	0	Mild
18	0.743033	Moderate	1	Moderate
19	0.630945	Mild	0	Mild
20	1.071419	Moderate	1	Moderate
21	1.179165	Moderate	0	Mild
22	0.841234	Moderate	0	Mild
23	0.556801	Mild	0	Mild
24	1.230286	Moderate	1	Moderate

From Eqn (1), the overall accuracy thus calculated is:

17/25 *100 = **68%**, closely matching with human doctor's average diagnostic accuracy of *71.4%* (Richens, Lee, & Johri, 2020). Label-wise accuracy can be seen in Table **8**.

Using equation (2), precision of classification has been calculated as follows:

TP = [0, 3, 5, 7-11, 20, 23, 24], *i.e.*, total number of cases = 11

Table 8: Class label accuracy measure of MLR classifier

Class label	No. of cases	%	#Correctly predicted	% Accuracy
Label-1 (Mild)	10	40	5	**50**
Label-2 (Moderate)	8	32	5	**62**
Label-3 (Severe)	7	28	1	**14**
Overall Accuracy	25	100	11	**68**

FP = [1, 2, 16-19], *i.e.*, total = 6

Therefore, precision = $\frac{11}{11+6} * 100$ = ***61%***

It is important to note MLR is unable to predict exactly as the class labels, instead it gives some float values seen in Table **7**. Those predicted values are interpreted by 5 out of 10 expert clinicians who avoided weight scoring for the sign-symptoms (to prevent classification bias): 0.0 - 0.74 is 'Mild' grade, represented by class label '0'; 0.75 - 1.24 is 'Moderate', represented by class level '1'; >1.24 is 'Severe', represented by class label '2'.

3.4. Results of Multinomial Logistic Regressions (MnLR)

As mentioned, MnLR has got two functions; detailed flowcharts below:

1) Logit function

Step-1: Creating random weights and biases for the model, initialized at '0'

Step-2: Calculating logit scores (glimpse of first 5 cases can be seen below) with shape (99, 3, *i.e.*, 3 weights for each of 99 total cases)

Weights Case 1: [[2.80895057 3.31177159 2.8193969]

Weights Case 2: [2.00201906 2.02635133 1.8194785]

Weights Case 3: [1.99955182 2.64174856 2.72071056]

Weights Case 4: [2.41963375 3.33930231 3.12822817]

Table 9: Run Vs. Accuracy (MnLR Classifier)

		RUN-1		
Test_Case	Pred	Label	Target	Label
0	2	Severe	0	Mild
1	2	Severe	2	Severe
2	2	severe	2	Severe
3	2	Severe	1	Moderate
4	2	Severe	2	Severe
5	2	Severe	1	Moderate
6	2	Severe	0	Mild
7	2	Severe	2	Severe
8	2	Severe	0	Mild
9	2	Severe	0	Mild
10	2	Severe	1	Moderate
11	2	Severe	0	Mild

		RUN-2		
Test_Case	Pred	Label	Target	Label
0	1	Moderate	0	Mild
1	1	Moderate	2	severe
2	1	Moderate	2	severe
3	2	Severe	1	Moderate
4	1	Moderate	2	severe
5	2	Severe	1	Moderate
6	2	Severe	0	Mild
7	2	Severe	2	severe
8	2	Severe	0	Mild
9	1	Moderate	0	mild
10	1	Moderate	1	Moderate
11	1	Moderate	0	Mild
12	2	Severe	2	severe

12	2	Severe	2	Severe
13	2	Severe	1	Moderate
14	2	Severe	2	Severe
15	2	Severe	2	Severe
16	2	Severe	1	Moderate
17	2	Severe	0	Mild
18	2	Severe	1	Moderate
19	2	Severe	0	Mild
20	2	Severe	1	Moderate
21	2	Severe	0	Mild
22	0	Mild	0	Mild
23	0	Mild	0	Mild
24	2	Severe	1	Moderate

Overall Accuracy: 36%

13	2	Severe	1	Moderate
14	2	Severe	2	severe
15	1	Moderate	2	severe
16	2	Severe	1	Moderate
17	1	Moderate	0	Mild
18	2	Severe	1	Moderate
19	1	Moderate	0	Mild
20	2	Severe	1	Moderate
21	1	Moderate	0	Mild
22	1	Moderate	0	Mild
23	1	Moderate	0	Mild
24	1	Moderate	1	Moderate

Overall Accuracy: 20%

RUN-3

Test_Case	Pred	Label	Target	Label
0	1	Moderate	0	Mild
1	1	Moderate	2	Severe
2	1	Moderate	2	Severe
3	1	Moderate	1	Moderate
4	0	Mild	2	Severe
5	1	Moderate	1	Moderate
6	1	Moderate	0	Mild
7	1	Moderate	2	Severe
8	1	Moderate	0	Mild
9	1	Moderate	0	Mild
10	1	Moderate	1	Moderate
11	0	Mild	0	Mild
12	1	Moderate	2	Severe
13	0	Mild	1	Moderate
14	0	Mild	2	Severe
15	1	Moderate	2	Severe
16	2	Severe	1	Moderate
17	1	Moderate	0	Mild
18	1	Moderate	1	Moderate
19	1	Moderate	0	Mild
20	1	Moderate	1	Moderate
21	1	Moderate	0	Mild
22	2	Severe	0	Mild
23	2	Severe	0	Mild
24	1	Moderate	1	Moderate

Overall Accuracy: 24%

RUN-4

Test_Case	Pred	Label	Target	Label
0	0	Mild	0	Mild
1	0	Severe	2	severe
2	2	Severe	2	severe
3	1	Moderate	1	Moderate
4	2	Severe	2	severe
5	1	Moderate	1	Moderate
6	1	Severe	0	Mild
7	2	Severe	2	severe
8	0	Mild	0	Mild
9	0	Mild	0	mild
10	1	Moderate	1	Moderate
11	2	Severe	0	Mild
12	0	Mild	2	severe
13	1	Moderate	1	Moderate
14	2	Severe	2	severe
15	2	Severe	2	severe
16	0	Mild	1	Moderate
17	0	Mild	0	Mild
18	2	Severe	1	Moderate
19	0	Mild	0	Mild
20	0	Mild	1	Moderate
21	1	Moderate	0	Mild
22	0	Mild	0	Mild
23	0	Mild	0	Mild
24	1	Moderate	1	Moderate

Overall Accuracy: 72%

Weights Case 5: [2.19083221 2.87746546 2.87449524]]

2) Softmax function:

Step-1: Converting logit scores to probability values

Step-2: Performing Multinomial Logistic Regression on a set with 'features', 'weights', and 'biases' that returns the probability values of the prediction. It is worth noting that the prediction values vary from one run to another run due to inherent stochasticity of the data. Below are few examples of varied accuracy at a given number of run in Table **9**. From the table it can be seen that highest accuracy of **76%** can be viewed in **Run 4**. Based on this observation it can be assumed that accuracy can further be increased with more Runs (ensemble averaging).

Table **9** shows the actual/target vs. predicted class label for 25 test cases, used in MLR classification testing. Class labels have been created as per the opinion of five of the ten experienced doctors and is shown in the last row of the table. Matched class labels are marked as green for easy visualization, while grey cells indicate all True positive cases and red cells are indicative of False positive cases.

Compared to first three Runs, the fourth Run shows much higher accuracy. Therefore, there is a chance that MnLR may render better accuracy in predicting the Dengue CFG on further runs and can outperform the MLR classifier, although their overall accuracies are equal in this experiment. Higher dataspace might be able to fit both the models better. Here instances are only 100. It would be interesting to analyze how these classifiers behave in higher dataspace (on synthetic data). Also, the MnLR classifier algorithm can be further optimized with minimum information loss.

Using equation (2), precision of classification by MnLR has been calculated as follows:

TP = [0-4, 7-9, 13-15, 17, 20, 22-24], *i.e.*, total number of cases = 16

FP = [5, 10], *i.e.*, total = 2.

Here, it is important to note that MnLR is able to predict exactly as the class labels unlike MLR, which predicts CFG as a float value and its interpretation had to be made by the expert panel to match with the class label, according a precision value of $\frac{11}{11+2} * 100 =$ ***88%***

Comparison of Classifiers' performance: Performance of MLR and MnLR has been compared in terms of accuracy and precision, using equation 1 and 2, respectively. Tables **7** to **10** show that the MnLR classifier can predict Mild, Moderate, and Severe Classes/Labels at 70%, 37% and 85% accuracy marks, respectively with overall accuracy of 72%. It is worth noting that there is a possibility that the accuracy can be further increased with more Runs, as evident from the experiment. On the other hand, MLR classifier is able to predict the labels 'Mild', 'Moderate', and 'Severe' with 50%, 62%, and 14% accuracies, respectively with overall accuracy of 68% that is also close to MnLR in clinical setting. MnLR is found to be more precise (88%) compared to MLR (61%).

Prediction of 'Mild' cases with close to 50-*70% accuracy* by both the classifiers are clinically acceptable. Both classifiers can reliably screen ca 50-70% of the 'Mild' cases, which is desired in a clinical set up (human eyes often over or underestimate the future risk) due to its subjective course of morbidity. A key derivative of this study is the ability to predict the longitudinal progression of a case study from 'moderate' to 'severe', or even 'mild' to 'moderate' to nip the health prognosis at its bud.

4. CONCLUSIONS

The first key achievement of this study is the establishment of Statistical Machine Learning (SML) as a reliable diagnostic tool in the early prediction of Dengue CFG, a tool that can subjectively extract information both from epidemiological and sign-symptoms data to quantify the health state of a patient. SML classifiers, such as MLR and MnLR, though computationally 'hard', yet these are able to predict Dengue CFG at an acceptable level of precision and

Table 10: Class label accuracy measure of MnLR classifier (on the best Run *i.e.*, Run 4)

Class label	No. of cases	%	#Correctly predicted	% Accuracy
Label-1 (Mild)	10	40	7	**70**
Label-2 (Moderate)	8	32	3	**37**
Label-3 (Severe)	7	28	6	**85**
Overall Accuracy	25	100	19	**72**

accuracy that is comparable to diagnostics from human clinicians. More importantly, using Statistical classifiers, 'Mild' cases could be screened with 50-70% accuracy, which are often overlooked by human clinicians, leading to either an under-diagnosis or over-diagnosis. Under-diagnosis has the innate peril of ascribing low or insignificant weightage to therapeutic assessments that often lead to silent progression of the disease, with progressive and fast evolution to 'Moderate' and then to a 'Severe' diseased state. Incidences of death due to lack of detection at a sufficiently early stage is actually quite common. This whole chain of silent but fast deterioration can be checked at the very initial stage using SML. Deep learning and other computationally complex ML methods may offer higher predictive accuracy, but these are computationally demanding, that may preclude their implementation in public healthcare domain, both due to lack of appropriate training and disbelief of clinicians on such agents of automation. Over-diagnosis a complementary peril, one of whose critical manifestation is in the form of antibiotic apocalypse that amounts to inefficacy of antibiotic treatment due to over usage and subsequent immunity of the bacteria against such medications. As shown in this study, SML can be an excellent complementary consort in such diagnostic assessment.

In real-world clinical practice, *70% accuracy* is a go by number, because at the end of the day, it is the clinician who takes the final call. This entire decision-making process can hugely benefit through adaptation of unbiased ML-based toolkits, be it in the form of simple SML classifiers or complex Deep Learning based predictors, helping to identify the correct health state and therapeutic regimen of the infected. This is particularly relevant for a rapidly deteriorating infectious disease like dengue for which the accuracy of illness grade prediction is one of the worst. Reasons for such inaccuracy have been attributed to a combination of personal judgment as also the real time involvement of a clinician with a patient, particularly in heavily populated countries like India.

The present study combines clinical judgment with unbiased data profiled diagnostic, thus establishing a mathematical rationale behind the clinical practice number (71%). Another contribution of this work is in establishing a full proof mechanism of cross-validation of the clinical deduction, first through human clinicians concurring on SML predicted infection grades, and then validating the collective opinion of the medical board (involving 10 clinicians for the present data set) through SML classifiers.

The study thus effectively identifies a 'virtual doctor' that aids smart human decision making by complementing human diagnosis with a process of 'adaptive learning' instead of complex 'deep learning' as natural human cognition mechanism. This is a finely debated topic as questions have been asked about the nature of human decision making, whether that is a stochastic process moderated by Bayesian inference, leading to 'on the spar' decisions (Khalavati 2019), or a robust didactic regimen that builds on supervised learning as is used in 'Deep Learning' methods to understand vulnerabilities in the decision-making process (Dezfouli 2020). This is also a finer, faster and robust form of human intelligence that has been modeled here for clinical decision making.

5. FUTURE WORK

The outcomes of this study are undeniably constrained by their own limitations. These have worked on CFG levels with sharp boundaries defined on a 3-point scale. However, in reality, the boundaries defining the severity grades of a patient may not be so precise but highly fuzzy in nature, especially in the transition state as the inherent feature of the course of morbidity. Also, subjective analysis by an expert may inject 'twilight zone' fluctuations in a prognosis. For example, a case noted as 'mild' by an expert could as well be deemed at 'moderate' risk by another or even 'severe' by someone else. A key future target will be to extend the scope of the present SML classifier-based prediction algorithm to accommodate stochastic fluctuations in a clinical decision-making process, combining Bayesian inferencing with clinical second-best diagnostics.

Another dimension that we are presently working on is to add an independent Deep Learning (DL) kernel to the automated prognosis outlined in this work. In particular, the target will be to predict how many of the initial *false negatives* could later turn fatal and vice versa. This is a virgin area of study where we expect to further complete DL modes of learning with stochastic machine learning.

6. ACKNOWLEDGMENT

The authors acknowledge discussions and insightful comments from Prof J P Ioannidis of the University of Stanford, USA. The authors gratefully acknowledge partial financial support from the H2020-MSCA-RISE-2016 program, grant no. 734485, entitled 'Fracture Across Scales and Materials, Processes and Disciplines (FRAMED)'.

7. AUTHOR CONTRIBUTION STATEMENT

SC collected and machine learned the clinical data; AKC contributed to statistical modeling; ECA contributed to the paper writing, alongside SC and AKC.

REFERENCES

[1] Ahmed A, Jalal A, & Kim K. A Novel Statistical Method for Scene Classification Based on Multi-Object Categorization and Logistic Regression. Sensors, 2020; 20(14): 3871. https://doi.org/10.3390/s20143871

[2] Akazawa M, & Hashimoto K. Artificial Intelligence in Ovarian Cancer Diagnosis. Anticancer Research, 2020; 40(8): 4795-4800. https://doi.org/10.21873/anticanres.14482

[3] Anirudh VK. (2020). What is Machine Learning: Definitions, Types, Applications And Examples. Retrieved Feb 9, 2021, from Potential Analytics Artificial Intelligence for Healthcare: https://www.toolbox.com/tech/artificial-intelligence/tech-101/what-is-machine-learning-definition-types-applications-and-examples/

[4] Ashish K, Chattopadhyay S, Gao XZ, & Hui NB. Neural Network-based Diagnostic Tool for Identifying the Factors Responsible for Depression. International Journal of Computational Intelligence and Applications, 2019; 18(2): 195001-14. https://doi.org/10.1142/S1469026819500147

[5] Chattopadhyay AK, Choudhury D, Ghosh G, Kundu B & Nath SK. Infection kinetics of Covid-19 and containment strategy. Scientific Reports 2021; 11: 11606: 398. https://doi.org/10.1038/s41598-021-90698-2

[6] Chattopadhyay AK. & Chattopadhyay S. VIRDOCD: A VIRtual DOCtor to predict Dengue fatality. Expert Systems. 2021;e12796.

[7] Chattopadhyay S. Mathematical Modelling of Doctors' Perceptions in the Diagnosis of Depression: A Novel Approach. International Journal of Biomedical Engineering and Technology, 2013; 11(1): 1-17. https://doi.org/10.1504/IJBET.2013.053702

[8] Chattopadhyay S. A Neuro-fuzzy Approach for the Diagnosis of Depression. Applied computing and Informatics, 2017; 13(1): 10-18. https://doi.org/10.1016/j.aci.2014.01.001

[9] Comert G, Begashaw N, & Turhan-Comert A. (2020, July 21). Malaria Outbreak Detection With Machine Learning Methods. Retrieved Feb 8, 2021, from BioRxiv: https://www.biorxiv.org/content/10.1101/2020.07.21.214213v1.full.pdf https://doi.org/10.1101/2020.07.21.214213

[10] Dacosta-Aguayo R, Wylie G, DeLuca J, & Genova H. Anatomical Covariance Analysis: Detection of Disrupted Correlation Network Related to Clinical Trait Fatigue in Multiple Sclerosis: A Pilot Study. Behavioural Neurology, 2020 Article ID 5807496, 1-9. https://doi.org/10.1155/2020/5807496

[11] Dezfouli A, Nock R. & Dayan P. Adversarial vulnerabilities of human decision-making. PNAS 2020; 117 (46): 29221-29228. https://doi.org/10.1073/pnas.2016921117

[12] Esteves L, Caramelo F, & Ribeiro I. (2020). Probability distribution of copy number alterations along the genome: an algorithm to distinguish different tumour profiles. 10(14868). https://doi.org/10.1038/s41598-020-71859-1

[13] Farafonov VS & Neruhk D. MS2 bacteriophage capsid studied using all-atom Molecular Dynamics. Interface Focus, 2019; 9(3). https://doi.org/10.1098/rsfs.2018.0081

[14] Fuhad KM, Tuba JF, Sarker MR, Momen S, Mohammed N, & T R. Deep Learning Based Automatic Malaria Parasite Detection from Blood Smear and Its Smartphone Based Application. Diagnostics, 2020; 10(5): 329. https://doi.org/10.3390/diagnostics10050329

[15] Grela E, Stich M. and Chattopadhyay AK. Epidemiological impact of waning immunization on a vaccinated population. The European Physical Journal 2018; B 91: 267-279. https://doi.org/10.1140/epjb/e2018-90136-3

[16] Kapoor R, Ahuja S, & Kadyan V. Fuzzy Multiple Rules based Machine Learning Diagnostic Prediction Model for Dengue disease in India. International Journal of Advanced Science and Technology, 2020; 29(11s): 2083-2106.

[17] Khalavati K, Park, *et al*. Modeling other minds: Bayesian infreence explains human choices in group decision-making. Science Advances in Cognitive Neuroscience 2019; 5(11): eaax8783. https://doi.org/10.1126/sciadv.aax8783

[18] Krittanawong C, Virk HU, Bangalore S, Wang Z, Johnson KW, Pinotti R, Tang WH. (2020). Machine learning prediction in cardiovascular diseases: a meta-analysis. Scientific Reports, 10, 16057. https://doi.org/10.1038/s41598-020-72685-1

[19] Miguel-Hurtado O, Guest R, Stevenage SV, Neil GJ, & Black S. Comparing Machine Learning Classifiers and Linear/Logistic Regression to Explore the Relationship between Hand Dimensions and Demographic Characteristics. PLoS ONE, 2016; 11(11): e0165521. https://doi.org/10.1371/journal.pone.0165521

[20] Murray JD. (2002). Mathematical Biology I: An Introduction. Springer. https://doi.org/10.1007/b98868

[21] Murray JD. (2003). Mathematical Biology: Spatial Models and Biological Applications. Springer.

[22] Naeem S, Jamil N, Khan HU, & Nazir S. (2020). Complexity of Deep Convolutional Neural Networks in Mobile Computing. Complexity, 2020 (Article ID 3853780), 1-8. https://doi.org/10.1155/2020/3853780

[23] Nagar PK, Savargaonkar D, & Anvikar AK. Detection of Dengue Virus-Specific IgM and IgG Antibodies through Peptide Sequences of Envelope and NS1 Proteins for Serological Identification. Journal of Immunology Research, 2020; 1-8. https://doi.org/10.1155/2020/1820325

[24] Narayanan BN, Ali R, & Hardie RC. Performance analysis of machine learning and deep learning architectures for malaria detection on cell images. Proceedings Volume 11139, Applications of Machine Learning; 111390W (2019) (pp. -). San Diego, California, United States: Proc. SPIE 11139. https://doi.org/10.1117/12.2524681

[25] Raizada S, Mala S, & Shankar A. (2020). Vector Borne Disease Outbreak Prediction by Machine Learning. 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE) (pp. 213-218). Bengaluru: IEEE. https://doi.org/10.1109/ICSTCEE49637.2020.9277286

[26] Reda C, Kaufmann E, & Delahaye-Duriez A. Machine learning applications in drug development. Computational and Structural Biotechnology Journal, 2020; 18: 241-252. https://doi.org/10.1016/j.csbj.2019.12.006

[27] Richens JG, Lee CM, & Johri S. Improving the accuracy of medical diagnosis with causal machine learning . Nature Communications, 2020; 1-9. https://doi.org/10.1038/s41467-020-17419-7

[28] Salim NA, Wah YB, Reeves C, Smith M, Yaacob WF, Mudin RN, Haque U. Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques. Scientific Reports, 2021; 11(939). https://doi.org/10.1038/s41598-020-79193-2

[29] Sanders M, Servaas S, & Slagt C. Accuracy and precision of non-invasive cardiac output monitoring by electrical cardiometry: a systematic review and meta-analysis. Journal of Clinical Monitoring and Computing, 2020; 34: 433-460. https://doi.org/10.1007/s10877-019-00330-y

[30] Saturi S. (2020). Development of Prediction and Forecasting Model for Dengue Disease using Machine Learning Algorithms. 2020 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER) (pp. 6-11). Udupi: IEEE. https://doi.org/10.1109/DISCOVER50404.2020.9278079

[31] Schrider DR, & Kern AD. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends in Genetics, 2018; 34(4): 301-312. https://doi.org/10.1016/j.tig.2017.12.005

[32] Smith DS, Mariano DJ, Teautwein ML, Shepherd SM, Hinfey PM, & Shoff WH. (2019, May 03). /what-is-the-mortality-rate-for-dengue. (M. S. Bronze, Ed.) Retrieved Feb 7, 2021, from https://www.medscape.com: https://www.medscape.com/answers/215840-43425/what-is-the-mortality-rate-for-dengue#:~:text=Dengue%20fever%20is%20typically%20a,is%20as%20high%20as%2020%25.

[33] Sumair M, Alzed T, Gardezi SA, Bhutta MM, Rehman SM, & Rehman SU. (2020, July). Application of five continuous distributions and evaluation of wind potential at five stations using normal distribution. Energy Exploration and Exploitation. https://doi.org/10.1177/0144598720939373

[34] Swathi P, & Pothuganti K. Overview on Principal Compnent Analysis Algorithm in Machine Learning. International Research Journal of Modernization in Engineering Technology and Science, 2020; 2(10): 241-246.

[35] Tarasova E & Nerukh D. All-Arom Molecular Dynamics Simulations of Whole Viruses. The Journal of Physical Chemistry Letters, 2018; 9: 5805-5809. https://doi.org/10.1021/acs.jpclett.8b02298

[36] Tarasova E, Farafonov V, Khayat R, Okimoto N, Komatsu TS, Taiji M & Nerukh D. All-Atom Molecular Dynamics Simulations of Entire Virus Capsid Reveal the Role of Ion Distribution in Capsid s Stability. The Journal of Physical Chemistry Letters, 2017; 8(4): 779-784. https://doi.org/10.1021/acs.jpclett.6b02759

[37] Vaka AR, Soni B, & Reddy SK. Breast cancer detection by leveraging Machine Learning. ICT Express, 2020; 6(4): 320-324. https://doi.org/10.1016/j.icte.2020.04.009

[38] Valavi R, Elith J, Lahoz-Monfot JJ, & Guillera-Arroita G. blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods in Ecology and Evolution, 2020; 10(2): 225-232. https://doi.org/10.1111/2041-210X.13107

[39] Zhao N, Charland K, Carabali M, Nsoesie EO, Maheeu-Giroux M, Rees E, Zinszer K. Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia. PLOS Neglected Tropical Diseases, 2020; 14(9): e0008056. https://doi.org/10.1371/journal.pntd.0008056

[40] Zhen Sh, Cheng M, Yb T, Wang Yf, Juengpanich S, Jiang Zy, Cai Xj. Deep Learning for Accurate Diagnosis of Liver Tumor Based on Magnetic Resonance Imaging and Clinical Data. Frontiers in Oncology, 2020; 10(680). https://doi.org/10.3389/fonc.2020.00680

Received on 07-09-2021 Accepted on 28-09-2021 Published on 15-10-2021

DOI: http://dx.doi.org/10.12974/2311-8792.2021.07.2