V.H.N.SENTHIKUMARA NADAR COLLEGE

(An Autonomous Institution affiliated to Madurai Kamaraj University) Re-accredited with 'A' Grade by NAAC

Virudhunagar

RESEARCH CENTRE IN COMPUTER SCIENCE

NOTIFICATION FOR Ph.D. PUBLIC VIVA-VOCE EXAMINATION

As per the regulations of Madurai Kamaraj University, Madurai, Mr**. P. PANDISELVAM (Reg. No. P4551),** Assistant Professor, Department of Computer Science, Ayya Nadar Janaki Ammal College, Sivakasi, will defend his Ph.D. thesis at a Public Viva-voce Examination through Video Conferencing mode using Google Meet Platform

Title of the Thesis	:	**"Recognition of Dengue Serotypes Using Protein Sequence By Extracting Features By Entropy or Gini-Index and Extracting Rules from Neural Network"**
Date & Time	:	**March 12, 2021 (Friday)** 11.00 AM
Venue	:	Idhayam Rajendran Computer Science Block, Mini Conference Hall, V.H.N.Senthikumara Nadar College, Virudhunagar.
Video Conference Platform	:	**Google Meet**
Meeting ID	:	https://meet.google.com/xrs-qqbw-fwc
External Examiner	:	**Dr. M. Sundaresan,** Professor & Head, Department of Information Technology School of Computer Science and Engineering, Bharathiar University, Coimbatore – 641 046.

The synopsis of the thesis is available in the college website (www.vhnsnc.edu.in) and a copy of the thesis is available in the Research Centre in Computer Science for reference. **Faculty members, Research Scholars, Experts, Students and others who are interested in the subject are invited to attend the Ph.D Public Viva-voce Examination** and take part in the discussion.

Dr.T.Kathirvalavakumar
Supervisor & Convener
Associate Professor & Head,
Research Centre in Computer Science,
VHNSN College, Virudhunagar.

SYNOPSIS

Dengue and its serotype recognition is a big problem in the medical field over the past decades. It is the process of identifying the dengue patients by their protein or DNA sequence or gene expressions. Many laboratory methods Hematology, Serology and Antigen detection, Reverse- Transcription Polymerase Chain Reaction (RT-PCR), ELISA, Platelia, CIMSiM (Container – Inhabiting Mosquito Simulation Model), DENSiM (Dengue Simulation Model), SPSS (Statistical Package for Social Science), SARIMA (Seasonal Auto Regressive Integrated Mixing Average), PCR (Polymerase Chain Reaction) are used to identify the dengue patients by their blood specimen (WHO, 2009). However, there are still many problems after identifying dengue because of the time of fever, recognition of IgG and IgM count, low blood volumes and no vaccines. The burden of dengue is to classify the dengue serotypes. Nowadays, different researchers have focused on classifying and diagnosing dengue. The world is in need of a stable classification method for classifying and diagnosing dengue with low cost and less time consumption.

Dengue infections can be difficult to differentiate from other viral infections. It can be identified as an undifferentiated fever, Dengue Fever (DF), Dengue Hemorrhagic Fever (DHF) and Dengue Shock Syndrome (DSS). These can be identified with distinct serotype namely DENV I (Strain hawaii), DENV II (Strain guineqe), DENV III (Strain H87) and DENV IV (Strain H241). Generally, serotype refers to the subdivision of a virus that is divided based on their surface antigen. Each serotype has its own characteristics and is unique. Each virus and its serotype have toxic proteins that cause diseases in human. Dengue virus has toxic proteins E and M. Recovery from infection provides life-long immunity against the particular serotype. Subsequent infections by other serotypes increase the severity of dengue. If the structure of a protein is known, it would be easier for the biologist to infer the function of the protein. However, it is still costly to decide the structure of the protein (WHO, 2009).

DF is identified within 2-7 days after the bite of the Ades aegypti mosquito by the symptoms of increasing body temperature, fever, headache, pain in muscle and itching. DHF is identified within 3-7 days after the patient is affected by DF by the symptoms decreasing body temperature, minor bleeding from the nose, gums and skins. DSS is identified within 2-3 days after the patient is affected by DHF, it is with the symptoms of fluctuating body temperature, vomiting with the flow of minor blood. These are the stages of dengue infection from the initial stage to critical stages (Iqbal and Islam, 2017).

Different computer based diagnosis methods are proposed in the last decades. Sharma et al., (2013) have proposed a dengue and malaria diagnosis system based on decision support system and fuzzy logic. Rules have been framed for DSS with the symptoms malaria and dengue. The system has assigned one of the classes to the patient from sure dengue, may be dengue, sure malaria, may be malaria and not defined. Salman et al., (2014) have proposed diagnosis method of DHF using fuzzy logic on the mobile. They have assigned fuzzy score for the symptoms of dengue such as fever, skin rash, spontaneous hemorrhaging and

Tourniquet Test (TT) values. At last, their proposed method has assigned final fuzzy score for the patients. If the patient fuzzy score is in the range [90.6, 91] then the patient is highly affected with DHF. Leena Princy et al., (2016) have used informatics tool for analyzing the dataset and to detect dengue-affected patients for early recovery. Dengue serology test may take around 10 days to diagnose the infection in blood specimen (Balasaravanan and Prakash, 2018).

Generally in the field of computer science, dengue diagnosis methods are developed with feature selection and traditional classifiers like neural network (Bin Othman and Yau, 2007), decision tree (Dhamodharan, 2014), J48 (Saravanan and Gayathri, 2018; Solanki, 2014), SMO (Sugandhi et al., 2011), Random forest (Tate et al, 2017), SVM (Mihaila and Ananidou, 2013) and k-NN (Durairaj and Ranjani, 2013) based on the symptoms of dengue. The core objective of feature selection is to enhance the performance of the model towards cost effective and exclude over fitting. Feature selection techniques are used to identify the irrelevant features and focuses on the informative features. It has been active research topic and widely applied in many fields such as genomic analysis, text mining, image retrieval, instruction detection and so on. It well improves classification accuracy (Beniwal and Arora, 2012).

Hanirex et al., (2013) have identified dominating amino acids for DENV I using TDTR (Two Dimensional Transactions Reduction) approach based on an apriori and FP growth algorithm. The utilized dataset hold only 777 amino acids of DENV I which was obtained from GenBank: AAB27904.1. They have identified that Leucine (L), Phenylalanine (F), Lysine (K), Serine (S) and Glycine (G) are the dominating amino acids in DENV I. Changing of climate condition plays a vital role in the spreading of viral diseases. Fatima et al., (2017) have proposed a method for classifying different dengue serotypes. Differences between dengue serotypes are identified using SVM classifier.

Generally, applying feature selection always give benefits to classification as an identification of irrelevant features. Bamakan and Gholami (2014) have used entropy based feature selection for identifying breast cancer. Novakovic et al., (2011) have evaluated the performances of feature ranking methods like entropy based information gain, gain ratio, symmetrical uncertainty, relief-F, one-R and Chi-squared test using the Australian and German credit data set. They have observed that all the ranking methods provide same features on both datasets.

In the proposed methods, feature selection is based on entropy, Gini-index and information gain. Single layer feedforward neural network is used as a classifier model for classifying the dengue serotypes DENV I, DENV II, DENV III and DENV IV. The proposed works have been implemented by Matlab 13. All the proposed works have been tested with DNA, protein sequence and gene expression of dengue patients, which are obtained from NCBI (National Centre for Bio- Informatics) the national resource center for molecular biology information funded by US government.

In the second chapter of the thesis dengue serotypes are identified and classified based on amino acids and components in the protein sequence. Entropy and weighted average are used to identify the dominating and deficiency amino acids and components for the dengue. Relative value of amino acid or component is used for identifying the deviated values from normal patients to dengue patients.

Any type of viral infection spreads the specific amino acids in the protein sequence. The amino acids in the protein sequence may be either dominant or deficient when the person is infected with any type of diseases. The dominant and deficient of particular amino acid is also varying from one disease to another. In the proposed method, dengue is diagnosed by finding the dominant and deficient of amino acids/components of a protein sequence using entropy, relative and weighted averages. Entropy is the quantity of probability of information. Entropy is computed for each and every amino acid and components in the protein sequence. The relative value of amino acid is the deviation of entropy values of infected person from the minimum entropy of normal person. The relative value of components is the deviation of the weighted average of an infected person from the weighted average of components of a normal person.

When an amino acid is considered, the actual value represents the entropy value of an infected person; Expected value represents the minimum entropy value of normal person. When a component is considered, the actual value represents the weighted average of components of an infected person; Expected value represents the weighted average of components of normal person. Relative value of any disease is unique for any patient as the dominant and deficiency of the amino acid and components are unique for the disease.

Dengue served can be identified using amino acids and can be classified. The amino acids with negative relative values are selected. If the person is infected by DENV I then Phenylalanine (F) and Tryptophan (W) are with negative values, if the person is infected by DENV II then Phenylalanine (F), Leucine (K), Valine (V) and Tryptophan (W) are with negative values, if the person is infected by DENV III then Phenylalanine (F) is with negative value. If the person is infected with DENV IV then Phenylalanine (F), Glycine (G), Leucine (K) Valine (V) and Tryptophan (W) are with negative values. From the above conditions, it can be identified that the person is infected with dengue if an amino acid ***Phenylalanine (F)*** is with a negative value.

Dengue served can also be identified using components and can be classified. The weighted average for each component of normal human and diseased person is calculated. Relative values can be calculated for identifying deviation of the weighted average of an infected person from the weighted average of normal person's components. The components can be classified based on the relative value.

Proposed methods uses gene sequence of the dengue patients because gene is a vehicle of genetic information which is used to decide the characteristics (eye color, hair color) of a human. The protein sequence is an organic component composed of amino acids and this sequence of every person is conflicting from another person with only 0.5%. Amino acids are the building blocks of a protein sequence. They are classified into acidic, basic and neutral components based on the amino group and carboxylic group. Neutral components are classified into four sub components Aliphatic, Aromatic, Heterocyclic and Sulfur. Deficiency or dominance of amino acids/components led to disease. This leads to propose a method for identifying the serotypes in dengue using the components.

In the third chapter of the thesis, a classification framework of dengue infection is proposed based on the significant genes in the gene expression of dengue patients. Entropy is used to identify the informative genes. The dengue infections are classified by neural network. The proposed method also identifies the significant genes for dengue by pruning the neural network.

Selection of informative or significant of gene is the most important task in bioinformatics. The entropy-based method is used to filter informative or significant genes from dengue gene expression. Entropy (*H*) is the measurement of information spread in gene expression dataset. In the proposed system, significance based pruning is used. Initially, the trained network has irrelevant genes in the input neurons. The pruning identifies the significant genes from a trained network. For that, the significance measure S_i and threshold value α based method (Augasta and Kathirvalavakumar, 2003) is used. Activation value and the initial weights of a node play a major role in the selection of the significant genes.

Entropy H (gene) for each and every gene in the expression is calculated. Binary value R is set with 1, if H (gene) > 0 otherwise set with 0. Identify the informative genes are identified with the R as 1. The single hidden layer feedforward neural network is trained by informative genes. The significant measurement S_i for each gene from the trained neural network is calculated. Threshold value α is computed from the significant measurement S_i. Set the *sta(gene)* value with S_i and α values. Identify the significant genes using *sta(gene)*. The neural network is retrained with selected significant genes to classify the dengue infection.

The burden of dengue in the world is to classify dengue serotypes. Hence next work suggests a useful and stable method for classifying dengue serotypes based on Discrete Wavelet Transformation (DWT) using dengue gene sequence. *In the forth chapter*, a dengue serotypes classification model is proposed by applying wavelet transformation to the EIIP values of DNA sequences. Entropy based feature selection is used to recognize the significant wavelet coefficients. The neural network is trained with the significant wavelet coefficients for classifying dengue serotypes into DENV I, DENV II, DENV III and DENV IV.

The biological sequences as signals are to be encoded into a suitable format for data analysis and data mining tools. This is usually achieved by assigning a numeral to each symbol that forms the biological sequences. There are two fundamental kinds of biological sequences namely DNA nucleotide sequences and protein amino acids sequences relevant to dengue diagnosis. EIIP value of nucleic acid is a physical quantity that denoting the mean energy of valence electron and also used to find the protein coding regions (hotspots of protein) (Inbamalar and sivakumar, 2012). In this proposed work, the DNA nucleotide sequence is used as the input and is converted into EIIP indicator sequences by replacing the nucleotides with its EIIP values as A=0.1260, G=0.0806, C=0.1340 and T=0.1335.

The proposed system uses the DNA sequence of dengue patients as the gene sequence is responsible for making proteins and the functions of all living things. In the proposed method, the DNA sequence is converted into EIIP indicator sequences based on EIIP values of nucleotides. The EIIP indicator sequence is composed into approximation and detailed coefficients using wavelet transformation. Each approximation coefficients are used in chi- squared based feature selection method for selecting high frequency nucleotides in the EIIP indicator sequence. The selected coefficients are processed using feedforward neural network for classifying the dengue patients as DENV I, DENV II, DENV III or DENV IV.

In this method, the discrete wavelet transformation (DWT) is used to find the highest frequency genomic signals. In this DWT, the approximate LL coefficients in both horizontal and vertical directions, details coefficients in horizontal direction (HL) alone, details coefficients in vertical direction (LH) alone and details coefficients in both horizontal and vertical directions (diagonal edges) (HH) are extracted from the EIIP indicator sequences of dengue patients. This low and high pass filtering of genomic signals require the use of following filter functions through the multiplication of separable scaling and wavelet functions in h1 (horizontal) and v1 (vertical) directions.

$$\Phi(h1, v1) = \Phi(h1)\Phi(v1) \tag{1}$$

$$\varphi^{h}(h1, v1) = \varphi(h1)\Phi(v1) \tag{2}$$

$$\varphi^{v}(h1, v1) = \Phi(h1)\varphi(v1) \tag{3}$$

$$\varphi^{d}(h1, v1) = \varphi(h1)\varphi(v1) \tag{4}$$

where $\Phi(h1, v1), \varphi^{h}(h1, v1), \varphi^{v}(h1, v1)\ and\ \varphi^{d}(h1, v1)$ denote the approximated genomic signals, genomic signal with horizontal details, genomic signals with vertical details and genomic signals with diagonal details respectively. The wavelet coefficients of decomposed genomic signals are calculated using (5) and (6) (Stollnitz et al., 1994).

$$W_{\Phi}\left(j_{0,}k_{1,}k_{2,}\right) = \frac{1}{\sqrt{h1, v1}} \sum_{h1=0}^{H1-1} \sum_{v1=0}^{V1-1} s(h1, v1)\Phi_{j_{0,}k_{1,}k_{2}}\,(h1, v1) \tag{5}$$

$$W_{\varphi}\left(j_{0,}k_{1,}k_{2,}\right) = \frac{1}{\sqrt{h1, v1}} \sum_{h1=0}^{H1-1} \sum_{v1=0}^{V1-1} s(h1, v1)\ \varphi^{i}_{j_{0,}k_{1,}k_{2}}\,(h1, v1) \tag{6}$$

where i denotes the detailed wavelet coefficients of horizontal or vertical or diagonal directions of genomic signals. $s(h1, v1)$ represents separable scaling function of wavelet. In this work, approximation wavelet coefficients are only considered for further processing.

Feature selection is the process of selecting the informative approximation wavelet coefficients using chi-squared test value V from (7). Large value of V indicates there exist association between disease and exposure, small value of V indicates no association exists between disease and exposure. The existence of informative wavelet coefficient R is computed based on the Mc using (8). If chi-squared value is greater than Mc then R is set with 1 otherwise R is set with 0. The wavelet coefficients are informative when R=1. These informative wavelet coefficients are used as features to identify dengue serotypes.

$$V = \frac{(observed\ value - expected\ value)^2}{expected\ value} \tag{7}$$

$$R = \begin{cases} 1 & if\ \ V > Mc \\ 0 & otherwise \end{cases} \tag{8}$$

where $Mc = \sum_{i}^{n} v\,/n$ denotes mean of V, n represents number of approximation wavelet coefficients. R is the binary value for indicating the informative wavelet coefficients. Single hidden layer feedforward neural network is used to classify the dengue serotypes into DENV I, DENV II, DENV III and DENV IV using the selected features.

Rule extraction changes a black box system into a white box system by translating the internal knowledge of a neural network into a set of symbolic rules (Taylor and Darrah, 2005). Usually a rule extraction method is based on sample learning by using some classification method to obtain the classification rules. The techniques used for classification mainly include decision trees, neural networks and the genetic algorithms. An extracted rule (approximately) describes a set of conditions under which the network, coupled with its decision procedure, predicts a given class. In general, there are two types of approaches to extract rules from multilayer networks. One approach is to extract a set of global rules that describe the behavior of the output units in terms of input units. The alternative is to extract local rules by decomposing the multilayer networks into collection of single layer networks. A set of rules is extracted to describe each individual hidden and output unit in terms of the units that have weighted connections to it. The rules for the individual units are then combined into a set of rules that describes the network as a whole. The local rule extraction methods are designed for networks that use sigmoidal transfer functions for their hidden and output units.

In the fifth chapter of the thesis, a rule extraction method is described for the dengue serotypes classification. The proposed method uses the EIIP values of amino acids in the dengue protein sequences. Entropy based feature selection is used to recognize the significant amino acids. The neural network is trained for classifying the dengue serotypes with the selected significant amino acids. In order to improve the performances of the proposed method, the rules are extracted from the neural network by the data range of significant amino acids.

The EIIP values of amino acids are calculated by Eq. (9) and (10). In the method, the energy contribution of each amino acid in the protein sequence is found and is used for classifying dengue serotypes (Inbamalar and Sivakumar, 2012). The protein sequence is converted into EIIP indicator sequence using the energy contribution value (Q) of each amino acid by Eq. (11). Q value is calculated using EIIP value (W) of amino acids.

$$W = 0.25\ \frac{Z^* \sin(1.04\, \pi\, Z^*)}{2\pi} \tag{9}$$

$$Z^* = \frac{1}{N} \sum_{i=1}^{m} n_i\, z_i \tag{10}$$

where Z* is the average quasi valence number, Z_i is the valence number of ith atomic component, n_i is the number of atoms in ith component, m is the number of atomic components in the molecule and N is the total number of atoms.

$$Q_i = \frac{C_i}{T}\, W_i \tag{11}$$

where C_i represents number of particular amino acid in the protein sequence, T represents total number of amino acids in the protein sequence and W_i represents EIIP value of particular amino acid.

An entropy-based method is used for selecting the features. This method selects the amino acids which are used to classify different serotypes of dengue virus. Before applying the feature selection method, the amino acids with the EIIP value as 0 are to be eliminated. Entropy (E) is calculated for the Q values of amino acids in the protein. E value of amino acids ranges from 0 to 1. The Binary value R1 is set based on the E value of particular amino acid using Eq. (12). R1 is set to 1 if E has positive value otherwise R1 is set to 0. The amino acids with R1 value as 1 are selected as features. The selected features are used for training the neural network.

$$R1 = \begin{cases} 1 & if\ \ E > 0 \\ 0 & otherwise \end{cases} \tag{12}$$

The backpropagation algorithm is used to train the neural network for classifying the dengue serotypes. After the training, a data range matrix in Figure. 1 specifies the required data range of each significant attribute I_i to classify the data in a particular class C_k. The required data range of each significant attribute for classifying the data as the target serotype class is derived as follows.

	C_1	C_2	C_k	C_n
I_1	$[L_{11}, U_{11}]$	$[L_{12}, U_{12}]$	$[L_{1k}, U_{1k}]$	$[L_{1n}, U_{1n}]$
I_2	$[L_{21}, U_{21}]$	$[L_{22}, U_{22}]$	$[L_{2k}, U_{2k}]$	$[L_{2n}, U_{2n}]$
I_3	$[L_{31}, U_{31}]$	$[L_{32}, U_{32}]$	$[L_{3k}, U_{3k}]$	$[L_{3n}, U_{3n}]$

I_i	$[L_{i1}, U_{i1}]$	$[L_{i2}, U_{i2}]$	$[L_{ik}, U_{ik}]$	$[L_{in}, U_{in}]$
I_m	$[L_{m1}, U_{m1}]$	$[L_{m2}, U_{m2}]$	$[L_{mk}, U_{mk}]$	$[L_{mn}, U_{mn}]$

Figure 1. Data range matrix

L_{ik} and U_{ik} are selected from the trained neural network. The minimum and maximum values of I_i of the patterns those falls under C_k are considered as L_{ik} and U_{ik}. L_{ik} and U_{ik} are the minimum and maximum value of Q_i of i^{th} amino acid respectively for the target class k. The rules for each target class can be constructed using the non-zero data available in the corresponding column k of data range matrix. In general, rules can be written as

If ((data(I_1)$\geq L_{11}$ ^ data(I_1)$\leq U_{11}$) ^ (data(I_2)$\geq L_{21}$ ^ data(I_2)$\leq U_{21}$) ^……^ data(I_m) $\geq L_{m1}$ ^ data(I_m)$\leq U_{m1}$)) then Class = C_1

Else

If ((data(I_1)$\geq L_{12}$ ^ data(I_1)$\leq U_{12}$) ^ (data(I_2)$\geq L_{22}$ ^ data(I_2)$\leq U_{22}$) ^……^ data(I_m) $\geq L_{m2}$ ^ data(I_m)$\leq U_{m2}$)) then Class = C_2

Else

….

If ((data(I_1)$\geq L_{1n-1}$ ^ data(I_1)$\leq U_{1n-1}$) ^ (data(I_2)$\geq L_{2n-1}$ ^ data(I_2)$\leq U_{2n-1}$) ^……^ data(I_m) $\geq L_{mn-1}$ ^ data(I_m)$\leq U_{mn-1}$)) then Class = C_{n-1}

Else

Class =C_n

The rules can be restructured in the descending order in terms of a number of attributes required for classification (Augasta and Kathirvalavakumar, 2011).

In the sixth chapter of the thesis, dengue serotypes are classified by rule extraction based on apriori algorithm. Gini-index and information gain based feature selection is used to select the significant amino acids. The neural network is trained with significant amino acids for classification process. For improving the performances of proposed work, the most significant amino acids are identified by significant based pruning from the trained neural network. The rules are generated using support and confidence measurements of the most significant amino acids.

The core objective of this work is to classify the dengue serotypes with the most significant amino acids of the protein sequences of dengue patients and also identify the amino acid for main cause of spreading of dengue infections. In this work, the significant amino acids are identified from Gini-index based feature selection and are used in the neural network for classifying dengue serotypes as DENV I, DENV II, DENV III and DENV IV. The most significant amino acids for causing dengue are identified by rule extraction from the trained neural network classifier.

Gini index G(S) is a measure of the probability of information or measure of inequality of protein sequences. G(S) ranges from 0 to 1 where 0 indicates no inequality and 1 indicates maximum possible inequality (Panchatcharam et al., 2018). This method selects amino acids to classify different serotypes of dengue virus. Before applying the feature selection method, the amino acids with EIIP value 0 are to be eliminated. Gini index G(S) is calculated for the Q values of amino acids in the protein sequence using Eq. (13) and (14).

$$G(S) = 1 - \sum p_i^2 \tag{13}$$

$$G(S) = (n1/s)\, G\,(S1) + (n2/s)\, G\,(S2) + \ldots\ldots\ldots + (nm/s)\, G\,(Sm) \tag{14}$$

Where n1, n2....... nm are Q values of amino acids and S1, S2 ...... Sm represents gini-index values of amino acids.

Information Gain 'I' is a measurement of information based on a decrease in Gini index G(S) after a given dataset split on an attribute. Information gain is calculated based on the G(S) values of all the amino acids in the protein sequence using E.q. (15).

$$I = G(overall) - G(amino\ acids) \tag{15}$$

where *G(Overall)* represents Gini index of all the amino acids and *G(amino acids)* represents Gini index of particular amino acid in the specified protein sequence. In the decision tree construction model, the attribute with the highest information gain is selected as the root node as it plays an important role (Panchatcharam et al., 2018). Similarly, in this work, the amino acids with the higher information gain are considered as features for classifying dengue serotypes. The higher information gains are based on some threshold that filters the insignificant amino acids. The selected features are used for training the neural network.

The significant amino acids are used to generate association rules for improving the performances of the proposed method. In market basket analysis, an association rule mining is used to find the items which are frequently purchased by the customers and are used to generate the rules with the relationship on particular items (Kahramanli and Allaverdi, 2009; Raorane et al., 2012). Similarly, in the proposed method an association rule mining is used to identify the frequent patterns and its relationship with dengue serotypes as DENV I, DENV II, DENV III and DENV IV. Rule extraction (Kahramanli and Allaverdi, 2009; Raorane et al., 2012) depends on two factors namely i) Support and ii) Confidence. The support of an item is the percentage of transactions in which that item occurs.

Support (*x*) = *number of times x occurs / total number of transactions*
Support (*x*Uy) = *number of times x and y occurs / total number of transactions*

The confidence (or) strength (λ) for an association rule $X \rightarrow Y$ is the ratio of the number of transactions that contain $X \cup Y$ to the number of transactions that contain X. The association rules are generated with confidence and minimum support.

Confidence (X ⇒ Y) = (*number of times X and Y occurs) / (number of times* X *occurs)*

In the first step, candidate item sets C= { C_1,C_2,...... C_n} is generated with occurrences of particular amino acids in the protein sequence. The candidate item set C_1 is constructed based on the selected amino acid from the pruned neural network. Frequent item sets F= { F_1, F_2.....F_n} are generated based on the value of minimum support, where k represents a number of amino acids in the item set. In the second step, the strong rules are generated based on the highest confidence value of particular amino acids. The dengue serotypes is classified with the generated rules. The experimental results of all the proposed works show that the proposed methods are efficient and obtain higher performances in the recognition of dengue serotypes.

Bibliography

- Augasta, G.M., and Kathirvalavakumar, T.: Pruning Algorithms of Neural Networks – A Comparative Study. Cen. Euro. J. Comp. Sci. Vol. 3(3): pp 105-115, 2003.
- Augasta, M. G. and Kathirvalavakumar, T.: Reverse engineering the neural networks for rule extraction in classification problems. Neural. Process. Lett. Vol. 35(2): pp. 131-150. 2011.
- Balasaravanan, K., and Prakash, M.: Detection of dengue disease using artificial neural network based classification technique. Int. J. Engg. Tech. Vol. 7(1): pp. 13-15, 2018.
- Bamakan, H. M. S., and Gholami, P.: A novel feature selection method based on an integrated data envelopment analysis and entropy model. In 2nd International conference on information technology and quantitative management, (ITQM 2014), Procedia. Computer. Science. Vol. 31: pp. 632-638, 2014.
- Beniwal, S., and Arora, J.: Classification and feature selection techniques in data mining, Int. J. Engg. Res. Tech. Vol. 1(6): 2012.
- Fatima, M., and Pasha, M.: Survey of Machine Learning Algorithms for Disease Diagnostic. J. Intell. Learning Systems and Appl. Vol. 9: pp. 1-16, 2017.

- Hanirex, D. K., and K. P. Kaliyamurthie K.P.: “Finding the dominating amino acids in dengue virus (Type -1) study on mining frequent itemsets”. Int. J. Pharma. Bio. Sciences. Vol. 4(3): pp. 880-889. 2013.
- Inbamalar, Sivakumar.: Filtering Approach to DNA Signal Processing. Paper presented IACSIT Coimbatore Conferences IPCSIT 28, IACSIT Press, Singapore, 2012.
- Iqbal, N., and Islam, M.: Machine learning for dengue outbreak prediction: An outlook. Int. J. Adv. Res. Comp. Sci. Vol. 8: pp. 93-102, 2017.
- Kahramanli, H., and Allaverdi, N., “Rule extraction from trained adaptive neural networks using artificial immune systems”. Expert. Sys. Appl. Vol. 36(2): pp. 1513-1522, 2009.
- Leena Princy, S. S., and Muruganandam, A.: An implementation of dengue fever disease spread using informatica tool with special reference to Dharmapuri district”. Int. J.Inno. Res. Comp. Commu. Engg. Vol. 4(9): pp. 1-10, 2016.
- National Center for Biotechnology Information, http: //www.ncbi.nlm.nih.gov /genomes/ virusvariation/ database/ nph-select.cgi.

- Novakovic, J., and Bulatovic, D. S. P.: Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav. J. Oper. Res. Vol. 21 (1): pp. 119-135, 2011.
- Panchatcharam, P., Varadharajan, R., Mathan, M., and Kumar, P. M.: "A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart diseases". Des. Autom. Embed. Syst. Vol. 22(9): 2018.
- Raorane, R., Kulkarni, V., and Jitkar, B.D.: "Association rule – extracting knowledge using market basket analysis". Res. J. Rece. Sci. Vol. 1(2): pp. 19-27. 2012.
- Salman, A., Lima, Y., and Simon, C.: Computational intelligence method for early diagnosis dengue haemorrhagic fever using fuzzy on mobile device. EPJ Web of conferences. Vol. 68(00003): EDP sciences, 2014.
- Sharma, P., Singh, D. B. V., Bandil, K. M., and Mishra, N.: Decision support system for malaria and dengue disease diagnosis. Int. J. Info. Comp. Tech. Vol. 3(7): pp. 633-640, 2013.
- Stollnitz, E. J., Derose T. D., Salesin D. M.: Wavelets for computer graphics: A primer. Technical report 94-09-11. Department of computer science and engineering, University of Washington, Seatle, Washington, 1994.

- Taylor, B.J., and Darrah, M.A.: Rule extraction as a formal method for the verification and validation of neural networks. IEEE International Joint Conference on neural networks. Vol. 5:pp.2915-2920, 2005.
- World Health Organization, Dengue: Guidelines for Diagnosis, Treatment, Prevention and control. New edition, Geneva. (2009).
- Tate, A., Gavhane, V., Pawar, J., Rajpurohit, B., Deshmwch, G.B.: Prediction of dengue diabetes and swine flu using random forest classification algorithm. Int. R. J. Engg. Tech. Vol. 4: pp. 685-690, 2017.
- Solanki, A.V.: Data mining techniques using weka classification for sickle cell disease. Int. J. Comp. Sci. Info. Tech. Vol. 5(4): pp. 5857-5860, 2014.
- Durairaj, M., and Ranjani, V.: Data mining applications in healthcare sector a study. Int. J. Sci. Technol. Res. Vol. 2(10): 2013.
- Bin Othman, M.F., and Yau, T. M. S.: Comparison of different classification techniques using weka for breast cancer. In 3rd Kuala Lumpur International Conference on bio-medical engineering 2006. Springer Berlin Heidelberg. Pp. 520-523, 2007.
- Sugandhi, C., Ysodha, P., and Kannan, M.: Analysis of a population of cataract patient database in weka tool. Int. J. Sci. Engg. Res. Vol. 2(10): 2017.
- Mihaila, C., and Ananiadou, S.: Recognizing discourse causality triggers in the biomedical domain. J. Bio. Info. Comp. Biol. Vol. 11(6): 2013.

- Dhamodharan, S.: Liver disease prediction using Bayesian classification. Special Issues. 4th national conference on advance computing application technology. Pp. 315-323, 2014.