IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 July 2022, accepted 9 August 2022, date of publication 16 August 2022, date of current version 26 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3198937

RESEARCH ARTICLE

Associative Classifier Coupled With Unsupervised Feature Reduction for Dengue Fever Classification Using Gene Expression Data

DIPTARAJ SEN[1], (Student Member, IEEE), SAUBHIK PALADHI[2], JAROSLAV FRNDA[3,4], (Senior Member, IEEE), SANKHADEEP CHATTERJEE[5], SOUMEN BANERJEE[6], (Senior Member, IEEE), AND JAN NEDOMA[3], (Senior Member, IEEE)

[1]Department of Computer Science and Engineering, University of Engineering and Management, Kolkata 700160, India
[2]Department of Computer Science and Engineering, University of Kalyani, Kalyani 741235, India
[3]Department of Telecommunications, VSB—Technical University of Ostrava, 70800 Ostrava, Czech Republic
[4]Department of Quantitative Methods and Economic Informatics, University of Žilina, 010 26 Žilina, Slovakia
[5]Department of Computer Science and Technology, University of Engineering and Management, Kolkata 700160, India
[6]Department of Electronics and Communication Engineering, University of Engineering and Management, Kolkata 700160, India

Corresponding author: Sankhadeep Chatterjee (chatterjeesankhadeep.cu@gmail.com)

This work was supported by the VSB—Technical University of Ostrava, the Ministry of Education, Youth and Sports, Czech Republic, under Grant SP2022/5 and Grant SP2022/18.

ABSTRACT Recent studies have established the potential of classifiers designed using association rule mining methods. The current study proposes such an associative classifier to efficiently detect dengue fever using gene expression data. Labelled gene expression data has been preprocessed and discretized to mine association rules using well-established rule mining methods. Thereafter, unsupervised clustering methods have been applied to the discretized gene expression data to reduce and select the most promising features. The final feature reduced discretized gene expression data is subsequently used to mine rules in order to classify subjects into 'Dengue Fever' or 'Healthy'. Two well-known association rule mining methods, viz., Apriori and FP-Growth, have been used here along with various types of well established clustering methods. Extensive analysis has been reported with performance parameters in terms of accuracy, precision, recall and false positive rate using 5-fold cross-validation. Furthermore, a separate investigation has been conducted to find the most suitable number of features and confidence of association rule mining methods. The experimental results obtained indicate accurate detection of dengue fever patients at an early stage using the proposed associative classification method.

INDEX TERMS Gene expression data, association rules, Apriori algorithm, FP-growth algorithm, clustering.

I. INTRODUCTION

Dengue is one of the deadliest diseases of all times. In the last few decades, dengue affected cases were manifold across many countries worldwide. Several studies [1], [2] have estimated that around 4 billion people globally are at risk of dengue infection. Reported deaths in recent years have also increased drastically. Hence, it is essential to work on this disease to provide an efficient way to detect it. The microarray technologies, along with different tools from the domain of Knowledge Discovery and Data Mining, have been helpful in finding essential and correlated biological information from an excessively colossal gene expression dataset.

Association Rule Mining (ARM) [3], [4] is a data mining technique that is used in finding the relationship and patterns among data items. Meanwhile, clustering is a well-known technique in grouping similar types of data together. A weighted ARM technique is proposed in [5] to mine the itemsets with different weights. A database consisting of only binary attributes is unsuitable for the weighted

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan.

ARM. Hence, a new approach is proposed in [6] to overcome the problem without the requirement of pre-assigning of weights on the itemsets. A biclustering based ARM technique is reported in [7] to establish the superiority of the ARM algorithm over the models based on conventional classifiers. The discovery of a set of newly predicted protein-protein interactions is also mentioned. Another bi-clustering method is introduced in [8] to generate unique types of rules by discarding a large number of insignificant rules generated from maximal frequent closed homogeneous itemsets. There is another biclustering based ARM approach proposed in [9] to predict the standard rules among HIV-1 proteins and human proteins. The study can identify some viral protein interactions with similar biological activity. Statistical analysis is carried out to identify significant genes. Another weighted rule-mining approach is proposed in [10] to reduce the number of generated rules using rank and weight-based measures. It is reported that the proposed algorithm generates a more significant number of essential association rules than the traditional Apriori algorithm. Ranking of the rules is also done by using a Genetic Algorithm based method in [11] to develop a rule-based classifier. In [12], the author focuses on the ranking behaviour of the popular interestingness measures on the generated rules from a large number of different datasets. The analysis can identify less computationally expensive but significant interesting measures which are mentioned in some of the existing literature. Along with Apriori, the FP-Growth algorithm [13] is also used in [14] to search biological patterns in dense microarray data. It is concluded that the process of mining all frequent itemsets is both space and time-consuming. It is challenging to generate effective rules from correlated gene expression data. To counter this problem, a rule discovery method, reported in [15], is found to be highly successful even without the help of prior biological knowledge.

The selection of appropriate clustering algorithms and the number of clusters are the key factors to classify meaningful genes. A supervised clustering method is suggested in [16] to handle gene expression data and to identify significant interdependent genes after removing the redundancy among gene attributes. K-means is a widely used clustering algorithm because of its simplicity and computational speed. A modified and improved version of this algorithm named K-means++ is developed in [17] to enhance the speed and accuracy. A different approach to reduce the number of rules using an Agglomerative clustering algorithm is presented in [18]. The association rules are clustered based on their similarity. The methodology is found to be helpful in extracting a set of most significant genes. A survey on a large number of biclustering approaches is carried out in [19], and the performance of those clustering algorithms are evaluated by using different metrics. An attempt to improve the quality of clustering is made in [20]. The approach is found to be successful in overcoming the dimensionality problem of gene expression data. In [21], various traditional and new clustering techniques are reviewed. It is revealed that the recent clustering methods such as tri-clustering, cluster ensemble, dual-rooted MST are useful to avoid the drawbacks of some traditional clustering methods. A comparative study on several Agglomerative clustering techniques is carried out in [22]. In [23], a semi-supervised cluster ensemble framework is reported and is applied on cancer gene expression data for feature selection. It is shown that the framework can enhance the performance of the clustering algorithms adopted by the authors.

An Artificial Neural Network (ANN), trained by Particle Swarm Optimization (PSO), is adopted in [24] to classify different types of Dengue fevers. The model can achieve more than 90% accuracy while classifying the types. The gene expression data comprised of a large number of gene features in which appropriate feature selection without losing information is a difficult task. To combat this challenge, an approach is adopted in [25]. Here, ARM and differential gene expression analysis are considered to identify essential and correlated associations of genes. It is reported that multiple rules can share a common gene, unlike the clustering technique, where each rule is associated with only one cluster. Discretization of raw data is crucial in making the data suitable for mining. The quality of significant rules and relevant patterns are greatly influenced by the choice of the discretization approach [26].

A significant problem of the traditional ARM technique is the generation of a large number of redundant rules, which not only leads to costly computation but also causes the overfitting of data [27]. By considering this, a new framework of closed frequent itemsets mining is suggested in [28]. It is reported that the model can minimize the number of irrelevant rules in both real and synthetic datasets. Another study is conducted in [29] to identify the relationship of PPIs using a support-confidence framework. The approach uses correlation measures to improve the performance of the framework. In [30], a novel approach based on gene enumeration is proposed to handle gene expression data. Instead of using a matrix, an efficient tree data structure is used to store the gene data in its binary representation format. The proposed method can achieve a promising result in finding association rules with high confidence and reducing memory usage while keeping all the association rules. Another attempt to build a classifier for gene expression data using ARM is made in [31]. A Support Vector Machine (SVM) extracts significant biological features with high accuracy in the study. Several clustering analyses of gene expression data have been conducted in past decades to study the biological functionality of genomics. An investigation is made in [32] to analyse the performance of the different clustering algorithms on gene data. The study concludes that a single clustering algorithm cannot be declared as the best because the uniqueness of each algorithm makes them better than others in a particular experimental setup. In most of the literature, the selection criteria of the clustering algorithm are either static or based on the choice of the researchers. In this context, a framework is developed in [33] to establish it as a guiding tool to evaluate

the comparative performance of clustering algorithms on any datasets.

Managing high dimensional data like gene expression is a difficult task. A study is made in [34] to highlight the challenges and resolution for cluster analysis to combat dimensionality problems. The research on predictive classification and identification of gene markers are also made in [35]. Class imbalance is a significant issue in the domain of data mining. Most of the traditional classifiers do not produce good accuracy for the data which suffers from this problem. In [36], a new framework is proposed by introducing a new measure named Complement Class Support (CCS) for imbalanced data. The model displays the improvement of classification error rates against a regular model based only on a support and confidence framework. Another model for mining pattern of itemsets is proposed in [37]. The study reflects an attempt to generalise the operational platform between the data and the miner. A statistically backed up probabilistic approach is also suggested in [38] to minimise the number of frequent itemsets. It is reported that the model can eliminate all of the redundant itemsets. To accomplish the same objective, another study on self-sufficient itemsets [39] is conducted. Another investigation on different measures is made in [40] for pruning the important association rules. It is reported that not more than 50% rules are required to represent a cluster after pruning operation. An associative rule mining classifier combined with a new measure is introduced in [41] to handle both balanced and imbalanced data. It is posited that the performance of the proposed method is vastly superior to other associative classifiers, especially for imbalanced data.

An evolutionary optimization technique is proposed in [42] for mining biologically significant rules using NSGA-II to maximize the utility and interestingness of the sequential rules from gene expression data. Traditional ARM technique faces challenges while dealing with a large dataset. In [43], a heuristic method is proposed to learn from important gene-disease and gene-gene association rules generated from microarray data. The proposed approach has been reported to be more effective than traditional methods. It is known that an imbalanced dataset carries the risk of generating an overfitted model, which displays unreliable predictability. In [44], an unsupervised gene selection framework is proposed to handle imbalance microarray datasets. The method, at first, performs clustering of the genes and then identifies virtual genes which carry the most similar information about their respective clusters. Another study is conducted in [45] to classify the samples of cancer gene-expression data from several open-source datasets using simulated annealing.

The current work proposes an associative classifier based framework for efficient detection of Dengue fever. Literature survey reveals that reduction of unnecessary feature (gene) from the dataset may be beneficial to find valuable rules. Motivated by this, unsupervised feature reduction has been applied wherein clustering algorithms are used to cluster features and to eliminate similar features from the dataset keeping only the unique ones. The current study investigates some of the most successful clustering algorithms for this purpose. Next, the feature reduced gene expression dataset is used to mine rules required to classify patients into two categories viz., 'Dengue Fever' and 'Healthy'. Two well established association rule mining algorithms viz., Apriori and FP-Growth are used in the current study. After rule mining, top rules with confidence more than a predefined threshold have been selected to build the classifier. The performance is evaluated in terms of accuracy, precision, recall and false positive rate. 5-fold cross validation technique has been used to obtain statistically significant results. Furthermore, a separate box plot based analysis reveals that the proposed associative classifier framework is capable of detecting dengue fever with satisfactory performance. Overall the contributions of the current study are as follows:

1) Unsupervised clustering technique has been used in finding the most promising genes for classifying patients into 'Dengue Fever' and 'Healthy' category.
2) Associative classifier has been built by selecting the most confident rules mined by applying well known rule mining algorithms
3) Extensive experiments have been conducted to under stand the performance of a wide range of clustering and association rule mining techniques in the context of the current study.

The remaining work is organized as follows: The description of the dataset used in the present work is provided in section II. The proposed method is then introduced in section III. While, Section IV reports experimental results and the performance analysis. Finally, the conclusion of the present work is included in section V.

II. DATASET DESCRIPTION

In this article, the experiments have been carried out on the dataset named ''Acute Dengue patients: whole blood''. The original dataset is available in the online link https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5093. It contains a set of 54715 genes with expression values for each of the 56 samples. The samples consist of 4 types of classes: Convalescent Patient, Dengue Haemorrhagic, Dengue Fever Patient, and Healthy Control. Among these four classes, the samples belonging to the 'Dengue Fever Patient' and 'Healthy Control' classes are selected for the experiments. Here, 'Dengue Fever Patient' represents the samples of Dengue affected patients, and 'Healthy control' represents the samples of the patients who are not affected by Dengue. We have 18 'Dengue Fever Patient' class sample and 9 'Healthy Control' samples, thus making a total of 27 samples. Initially, the original dataset is transposed before performing any experiment on it. Thus, the final gene expression data matrix comprises 54715 gene columns, 1 class column, and 27 rows of samples for our experimental purpose. Table 1 depicts the abstract view of the data set used in the current experiment.

TABLE 1. The abstract view of the original data set. Here, i th gene and j th sample are denoted by g_i and s_j respectively. The expression value of the i th gene for the j th sample is denoted by $v_{(i,j)}$. The range of the values of m and n are [1, 27] and [1, 54715] respectively. The class column and the class of a sample s_j is denoted by c and c_j, respectively.

	g_1	g_2	g_3	$\cdots$	g_n	c
s_1	$v_{(1,1)}$	$v_{(2,1)}$	$v_{(3,1)}$	$\cdots$	$v_{(n,1)}$	c_1
s_2	$v_{(1,2)}$	$v_{(2,2)}$	$v_{(3,2)}$	$\cdots$	$v_{(n,2)}$	c_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
s_m	$v_{(1,m)}$	$v_{(2,m)}$	$v_{(3,m)}$	$\cdots$	$v_{(n,m)}$	c_m

III. PROPOSED METHOD

A. DATA PREPARATION

At first each of the gene column data is normalized using Z-score normalization. If μ_i is the mean and σ_i is the standard deviation of gene data g_i then the Z-score value of $v_{(j,i)}$ is calculated as $Z(v_{(j,i)}) = \frac{v_{(j,i)}-\mu_i}{\sigma_i}$. Here j is the index of j th sample.

A 3-level discretization is now performed for each of the Z-score normalized column data Z_{g_i}. Each column is split into three equal ranges depending on the maximum and minimum values in that column. If the Z-score values lie in the lower range, discrete value -1 is set in place of those values. Similarly, 0 and 1 are set for the mid and upper range values. The categorical representation of the discrete values -1, 0 and 1 are 'low', 'mid' and 'high'. The algorithm for the discretization process has been described in the algorithm 1.

Algorithm 1: Algorithm to 3-Level Discretize a Z-Score Normalized Gene Column

Data: A Z-score normalized column data Z_g
Result: A column data V with discretized value
$p \leftarrow max(Z_g)$; /* $max()$ returns the maximum value. */
$q \leftarrow min(Z_g)$; /* $min()$ returns the minimum value. */
$k \leftarrow \frac{p-q}{3}$;
$n \leftarrow |Z_g|$; /* $|Z_g|$ is the cardinality of Z_g */
$i \leftarrow 1$;
while $i \leq n$ **do**
  if $Z_{g_i} \in [q, q+k)$ **then**
    $V_i \leftarrow -1$;
  else if $Z_{g_i} \in [q+k, p-k)$ **then**
    $V_i \leftarrow 0$;
  else
    $V_i \leftarrow 1$;
  end
  $i \leftarrow i+1$;
end

B. UNSUPERVISED FEATURE REDUCTION

As our dataset contains a large number of features(genes), it is essential to reduce the number by discarding the non-significant ones. Hence, the popular clustering algorithms like K-means, Optics, Average Agglomerative, Ward Agglomerative and BIRCH have been chosen for this purpose. The feature columns in the dataset represent the feature vectors which are the data points for the clustering algorithms. These clustering techniques produce a total k number of feature clusters. The significant k genes are chosen by identifying the closest gene to the k cluster centres. The algorithm for the selection of potential genes has been described in the algorithm 2. In this article, the different values of k are chosen to be 5, 10 and 15. In our experiment, the clustering methods have been implemented using the scikit-learn package. From the subsection III-B1 to III-B5, the methodologies of constructing the clusters using the clustering algorithms have been discussed.

Algorithm 2: Algorithm to Find the Significant Genes

Data: A set of feature vectors W
Result: A set of significant genes G
$G \leftarrow \{\}$;
$C \leftarrow f(W)$; /* Function $f()$ implements a clustering algorithm and returns the cluster centres. C is the set of cluster centres. */
$n \leftarrow |C|$; /* $|C|$ represents the cardinality of set C */
$i \leftarrow 1$;
while $i \leq n$ **do**
  $id \leftarrow min(d(C_{i1}, C_i), d(C_{i2}, C_i), d(C_{i3}, C_i), \cdots)$;
  /* C_{ij} are the data points in the i th cluster. C_i is the cluster centre of i th cluster. Function $d()$ retuns the euclidean distance between C_{ij} and C_i. Function $min()$ returns the index of closest feature(gene) to C_i. */
  $G \leftarrow G \cup W_{id}$;
  $i \leftarrow i+1$;
end

1) K-MEANS CLUSTERING

K-means clustering [46] is one of the most efficient and simplistic clustering techniques. It needs the value of K, which denotes the number of clusters to be formed. This method assigns a data point to the nearest cluster by calculating and comparing the Euclidean distances between the data point and each cluster centre. In each iteration, new cluster centres are identified by calculating the mean of the data points in the cluster. The algorithm stops when cluster centres remain the same in two successive iterations.

2) OPTICS CLUSTERING

In OPTICS Clustering [47], OPTICS stands for Ordering Points To Identify Cluster Structure. The algorithm is popular for its flexibility on dense data-set. In this method, the mean of all the feature vector points having the same cluster labels is calculated, and after that, the cluster centres are identified.

3) AVERAGE AGGLOMERATIVE CLUSTERING

It is a type of hierarchical clustering which follows a 'bottom-up' approach to construct the clusters. Initially, this method considers each data point as an individual cluster. The closest clusters are then combined based on their distance. The distance is computed by calculating the average distance between all pairs of data points in those clusters. The optimal set of clusters is formed after a series of unions between the smaller clusters.

4) WARD AGGLOMERATIVE CLUSTERING

Ward Agglomerative clustering [48] is also a type of hierarchical clustering. This method defines the distance between two clusters as the combined error sum of squares. In each stage, the merger of a cluster pair occurs when it produces the minimum change of this error. The smaller clusters are then combined to create larger clusters by following this criterion.

5) BIRCH CLUSTERING

BIRCH (balanced iterative reducing and clustering using hierarchies) is an unsupervised data mining algorithm used to perform hierarchical clustering generally over large datasets [49]. This algorithm initially produces a compact version of the dataset from the original one without losing much of the information. Finally, this smaller dataset is clustered instead of the larger one.

C. FREQUENT ITEMSETS AND RELEVANT ASSOCIATION RULE MINING

After the clustering phase, the experimental dataset contains k gene columns and 1 class column. Each discretized value in a gene column g_i represents the gene's correlation level (high, mid, low) with a particular sample s_i. The proposed methodology considers each gene value s_j as a set of items. Suppose the correlation values of a random sample s_j for the set of genes and the class $\{g_1, g_2, \ldots, g_n, c\}$ are $\{low, high, \ldots, mid, c_1\}$. Then the possible set of items from s_i is $\{(g_1, low), (g_2, high), \ldots, (g_n, mid), c_1\}$.

The set of samples s has been divided into training and test dataset using 5 fold cross-validation. In this article, two different algorithms, namely the Apriori and the FP-growth, have been used on the training dataset to mine frequent itemsets. A fixed support value of 25% has been used to identify the frequent itemsets, and the association rules have been generated based on the different confidence threshold values such as 0.70, 0.80 and 0.90. Among all the generated rules, only the association rules with the gene correlation values in the antecedent part and the class label in the consequent part are selected for the experiments. For example a random association rule can be like $(g_1, low), (g_2, high) \rightarrow c_1$.

D. CLASSIFICATION AND PERFORMANCE MEASUREMENT

The proposed approach applies the association rules on the test dataset to perform the classification. For each test sample, t_i, the algorithm counts the number of rules having the antecedent part as a subset of t_i. After this, the weighted voting is performed to determine the class level of t_i.

Suppose $t : \{(g_1, low), (g_2, high), (g_3, low), \ldots, c_1\}$ is a test sample and $r_1 : \{(g_1, low) \rightarrow c_1\}$, $r_2 : \{(g_2, high), (g_3, low) \rightarrow c_2\}$, $r_3 : \{(g_1, low), (g_3, low) \rightarrow c_1\}$ and $r_4 : \{(g_1, high) \rightarrow c_2\}$ are the rules. In this case the antecedent part of r_1, r_2 and r_3 are the subset of t. As the antecedent part of r_4 is not a subset of t, the rule r_4 is not considered for further processing. It can be observed that the consequent part of r_1 and r_3 is the class c_1 and the consequent part of r_2 is the class c_2. By following the majority voting, the algorithm predicts the class of sample t to be c_1. On the basis of this example, the approach has correctly classified the test data t. In case of a tie in the result of majority voting, the benefit of the doubt is given to the most important class. According to our experimental setup, the 'Dengue Fever Patient' class has been given priority for the benefit of the doubt.

To establish the effectiveness of the proposed classifier, the performance metrics viz. precision, recall, accuracy and false positive rate have been considered. The precision, recall, accuracy and false positive rate (fpr) are defined as below:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$fpr = \frac{FP}{TN + FP}$$

Here TP, TN, FP and FN represents true positive, true negative, false positive and false negative, respectively. In our experimental setup, if both predicted and original classes are 'Dengue Fever Patient', it is considered a true positive. If both classes are 'Healthy Control', it is considered a true negative. If the predicted class indicates 'Healthy Control', but the original class indicates 'Dengue Fever Patient', it is a false positive. Whereas if the predicted class indicates 'Dengue Fever Patient', but the original class indicates 'Healthy Control', it is regarded as a false negative. To avoid the biasness of the model, 5 fold cross-validation has been performed five times to generate training and test datasets. The final result has been obtained after averaging the results of all rounds of experiments.

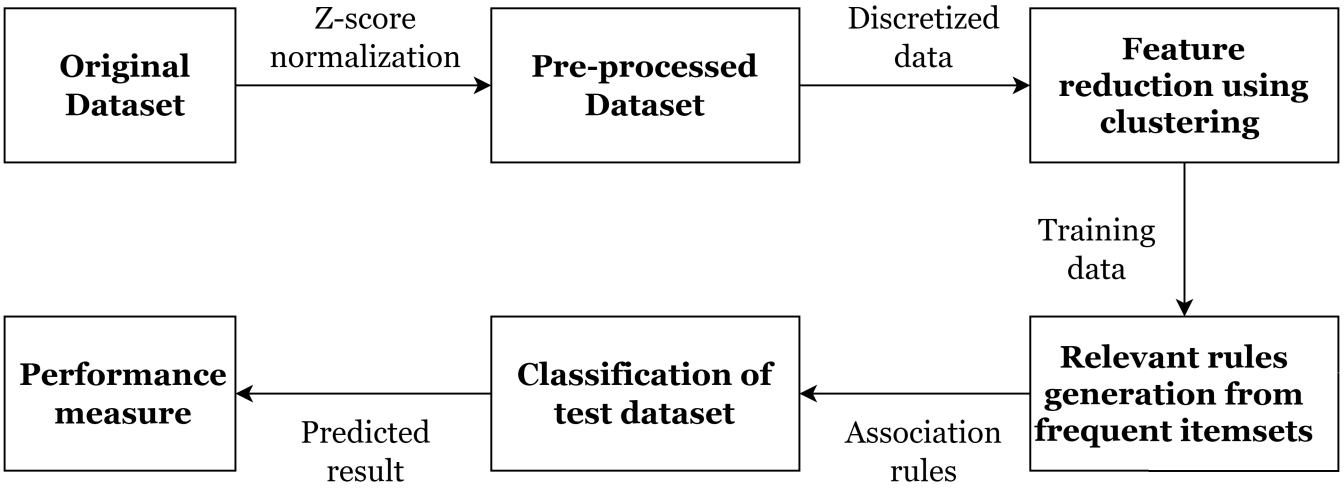


FIGURE 1. Flow diagram for associative rule based classifier.

E. PROPOSED FRAMEWORK FOR ASSOCIATIVE RULE BASED CLASSIFIER

In the current study, we have proposed a framework to predict the class of a data sample. Association rule mining has been employed to train it. The control flow diagram, presented in Figure 1, depicts the modules of this framework.

The first module is dedicated for the preprocessing of data. The column data is initially normalized using Z-score normalization, and 3-level discretization is performed on that data. The second module identifies the potential data columns using a clustering algorithm. In the third module, the data is divided into training and test data using five fold cross-validation. The training data is used to generate the association rules. Only the relevant rules having high confidence values are selected. The fourth module includes the testing phase, where the classes of test data are predicted using the generated association rules. Performance measurement using the metrics precision, recall and accuracy is done in this phase.

The primary purpose of the proposed framework is to classify the data samples into two classes, 'Dengue Fever Patient' and 'Healthy Control', with high accuracy.

IV. RESULTS & DISCUSSION

A. ANALYSIS OF CLUSTERING AND ASSOCIATION RULE MINING ALGORITHMS

In the current study, unsupervised methods like clusterings have been employed to reduce the number of genes. We have used five well-known clustering techniques like K-means, Average Agglomerative, Ward Agglomerative, BIRCH and OPTICS clustering to identify the critical genes. These clustering algorithms have been used separately to form a different number of clusters like 5, 10 (9 for OPTICS), and 15. Only the gene nearest to the cluster centre is selected from each cluster. For OPTICS clustering, as the number of clusters can not be fixed initially, the optimal number of clusters nearest to 10 is coming out to be 9. It has been observed from the current analysis that a lesser number of clusters produces a comparatively more important set of genes. Identifying potential genes is crucial for enhancing the performance of the model. Each of the clustering algorithms returns unique but overlapping sets of genes depending on their clustering characteristics.

In the current framework, the quality of the association rules is entirely dependent on the selected set of genes. While performing the experiments, the frequent itemsets have been mined using Apriori and FP-growth algorithms. The relevant association rules with high confidence values like 0.7, 0.8 and 0.9 have been generated using those itemsets. The comparative analysis of each clustering algorithm with a different number of clusters in terms of accuracy, precision and recall have been plotted in Figure 2, Figure 3 and Figure 4 respectively. Each metric value plotted in these figures is obtained from the average value of the results of all cross-validation rounds of both itemset mining algorithms. Figure 2 depicts the average accuracy of the proposed model while using a different number of clusters and clustering algorithms. In most cases, the increasing number of clusters indicates decreasing accuracy. This plot shows that BIRCH produces the most consistent and Ward Agglomerative produces the least consistent accuracy value when the number of clusters is varied. From Figure 3, it is concluded that the average values of the precision metric remain stable among the different number of clusters if Average Agglomerative and BIRCH algorithms are used for clustering. Whereas after analysing the average values of recall metric in Figure 4, it is concluded that the use of either BIRCH or K-means or OPTICS clustering leads to consistent recall value.

B. EFFECT OF CLUSTER SIZE

In order to understand the effects of the number of genes selected by using clustering algorithms, a separate study

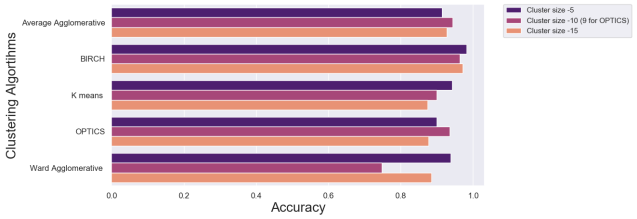


FIGURE 2. Performance of the proposed method in terms of average accuracy while using a different number of clusters and clustering algorithms.

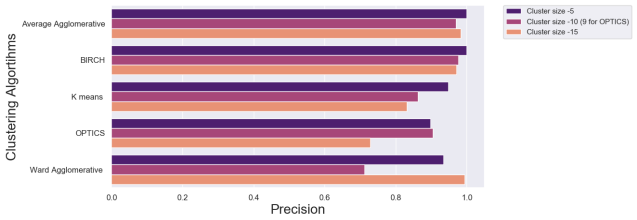


FIGURE 3. Performance of the proposed method in terms of average precision while using a different number of clusters and clustering algorithms.

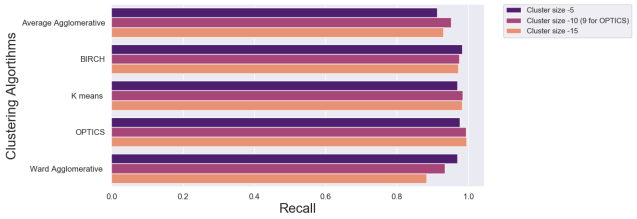


FIGURE 4. Performance of the proposed method in terms of average recall while using a different number of clusters and clustering algorithms.

has been conducted. Figure 5 reports the performance of associative classifier based on Apriori in terms of accuracy, precision, and recall. The best performing clustering algorithm BIRCH has been considered for this study. The study considered size of clusters from 1 itself, however, for cluster sizes 1, 2 and 3 no rules could be mined with desired confidence. Thus, the plot reports performance from cluster size 4. The plot reveals that the associative classifier achieves best performance in terms of accuracy for cluster size 5. However, in terms of precision cluster size 5, 7, 8, 12, 13 and 15 achieves best results. In terms of recall cluster size 5 obtained best performance. Overall, cluster size 5 has been found to be the optimal number of clusters.

C. COMPARISON WITH STATE-OF-THE ART

The proposed associative classifier model has been compared with state-of-the-art XGBoost model in terms of accuracy, precision and recall. The comparison is conducted by considering the dataset after feature reduction. As evident from Section IV-B, the optimal number of cluster should be five. Consequently, five genes have been selected using various

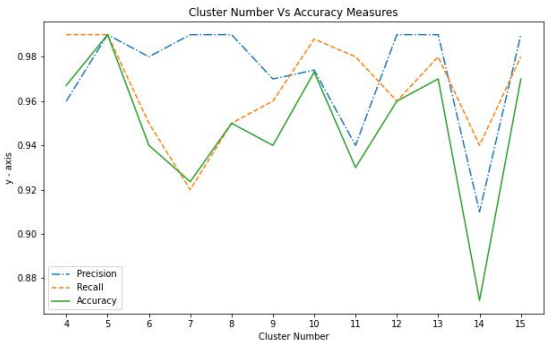


FIGURE 5. Performance of Apriori with BIRCH based gene selection in terms of accuracy, precision and recall for various cluster size.

TABLE 2. Performance of XGBoost Classifier for various clustering algorithms in terms of accuracy, precision and recall.

Clustering Algorithm	Precision	Recall	Accuracy
Average Agglomerative Clustering	0.7181	0.818	0.652
BIRCH clustering	-	0.77	0.54
K means Clustering	0.93	0.96	0.93
OPTICS clustering	0.969	0.939	0.936
Ward Agglomerative Clustering	0.913	0.938	0.887

clustering algorithms and XGBoost classifier is trained and tested with 5-fold cross validation method. The results are reported in Table 2. The experiment reveals that the performance of XGBoost when Average Agglomerative Clustering is used for feature reduction is 0.71, 0.81 and 0.65 in terms of precision, recall and accuracy respectively. From Table 3 it is observed that the performance of the proposed associative classifier is significantly better than XGBoost. In case of BIRCH, with five genes, XGBoost incorrectly classified all test samples to 'Dengue Fever Patient' class thereby no precision value could be calculated whereas the recall and accuracy is much inferior to associative classifier as evident from Table 4. A similar trend can be observed for OPTICS and Ward Ward Agglomerative clustering algorithms however, in case of k-means, the performance of XGBoost is 0.93, 0.96, and 0.93 in terms of precision, recall and accuracy respectively which is close to the proposed associative classifier. Although, Table 5 reveals that the performance of proposed associative classifier is 0.95, 0.96 and 0.94 in terms of precision, recall and accuracy which is still better than XGBoost.

D. PERFORMANCE ANALYSIS OF THE PROPOSED METHOD AFTER APPLYING DIFFERENT UNSUPERVISED FEATURE REDUCTION TECHNIQUES COMBINED WITH APRIORI ALGORITHM

The current section focuses on the performance of the proposed method when different clustering techniques followed by the Apriori algorithm are used. The values of the performance metrics have been tabularized from Table 3 to Table 8 on the basis of different number of genes and confidence

TABLE 3. Performance of the proposed method while using Average Agglomerative clustering and Apriori algorithm.

Confidence of rules	0.7			0.8			0.9			Average		
No. of Genes (Clusters)	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy
5	**1** (±0.00)	0.926 (±0.07)	**93.05%** (±6.03%)	**1** (±0.00)	0.952 (±0.04)	**95.39%** (±4.60%)	N/A	N/A	N/A	**1**	0.939	94.22%
10	0.986 (±0.01)	**0.928** (±0.07)	92.64% (±6.33%)	0.982 (±0.01)	**0.954** (±0.04)	94.84% (±5.00%)	0.976 (±0.02)	**0.986** (±0.01)	**97.02%** (±2.89%)	0.9813	**0.956**	**94.83%**
15	0.982 (±0.01)	0.896 (±0.07)	89.25% (±7.25%)	0.984 (±0.01)	0.926 (±0.06)	92.34% (±6.04%)	**0.988** (±0.01)	0.952 (±0.04)	95.14% (±4.54%)	0.9847	0.9247	92.24%

TABLE 4. Performance of the proposed method while using BIRCH clustering and Apriori algorithm.

Confidence of rules	0.7			0.8			0.9			Average		
No. of Genes (Clusters)	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy
5	**1** (±0.00)	**0.978** (±0.02)	**97.8%** (±2.51%)	**1** (±0.00)	**0.978** (±0.02)	**97.8%** (±2.12%)	**1** (±0.00)	**1** (±0.00)	**100%** (±0.00%)	**1**	**0.985**	**98.53%**
10	0.98 (±0.02)	0.964 (±0.03)	95.29% (±4.64%)	0.978 (±0.02)	**0.978** (±0.01)	96.65% (±3.35%)	0.974 (±0.02)	0.988 (±0.01)	97.31% (±2.19%)	0.9773	0.9767	96.42%
15	0.988 (±0.01)	0.964 (±0.03)	96.29% (±3.38%)	0.988 (±0.01)	0.97 (±0.02)	96.98% (±2.70%)	0.99 (±0.01)	0.98 (±0.02)	97.96% (±1.86%)	0.9887	0.9713	97.08%

TABLE 5. Performance of the proposed method while using K-means clustering and Apriori algorithm.

Confidence of rules	0.7			0.8			0.9			Average		
No. of Genes (Clusters)	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy
5	**0.956** (±0.04)	0.968 (±0.03)	**94.39%** (±5.26%)	**0.958** (±0.04)	0.962 (±0.03)	**95.18%** (±4.56%)	**0.936** (±0.06)	0.966 (±0.03)	92.89% (±7.04%)	**0.95**	0.9653	**94.16%**
10	0.86 (±0.09)	0.968 (±0.03)	88.09% (±9.24%)	0.824 (±0.08)	**0.988** (±0.01)	87.86% (±8.98%)	0.894 (±0.07)	**1** (±0.00)	**93.32%** (±5.92%)	0.8593	**0.9853**	89.75%
15	0.866 (±0.10)	**0.98** (±0.01)	90.18% (±8.22%)	0.816 (±0.08)	0.984 (±0.01)	86.36% (±9.02%)	0.858 (±0.11)	0.99 (±0.01)	90.3% (±8.28%)	0.8467	0.9847	88.94%

values of the association rules. All results in these tables have been reported in (±) standard deviation format. The number of selected genes is the same as the number of clusters. Boxplot of different performance metrics for each clustering methods have been depicted from Figure 6 to Figure 8.

After applying Average Agglomerative clustering, the performance of the proposed framework has been illustrated in Table 3. The average precision, recall and accuracy values are 1, 0.939 and 94.22%, respectively, when the number of genes is 5. Meanwhile, the average values of those performance metrics in the same order are 0.9813, 0.956, 94.83% and 0.9847, 0.9427, 92.24% when the number of genes is 10 and 15, respectively. It has been revealed that the framework has not generated any association rules with a high confidence threshold value of 0.9 for a few genes like 5. Hence, the values of those metrics have been displayed as N/A (Not Applicable) in this case. The experimental result also indicates that for a particular number of genes, the accuracy of the proposed model gradually increases when the confidence threshold of the association rules increases. In terms of accuracy, having the set of 10 genes produces slightly better results compared to having the set of 5 or 15 genes.

Table 4 depicts the performance of the proposed model after employing BIRCH clustering. The average precision, recall and accuracy values are 1, 0.985 and 98.53%, respectively, when the set 5 genes are used. The average precision, recall and accuracy values are 0.9773, 0.9767, 96.42% and

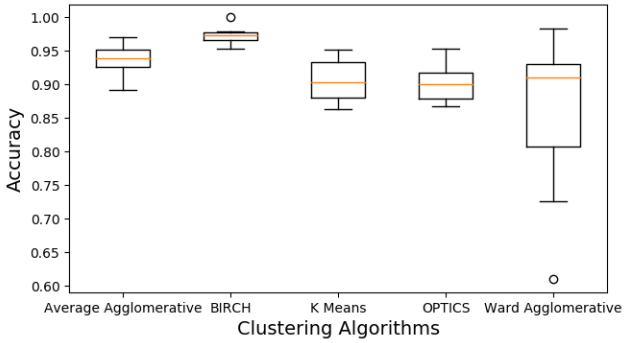


FIGURE 6. Boxplot of the accuracy metric of the proposed framework while using different clustering algorithms and Apriori algorithm.

0.9887, 0.9713, 97.08% when the number of selected genes is 10 and 15, respectively. It has been observed that for a specific number of genes, the accuracy of the proposed model gradually increases when the confidence threshold of the association rules increases. The average result of the performance metrics for the set of 5 genes has been found to be better than the result for a set of 10 or 15 genes.

Table 5 depicts the overall performance of the framework while using K-means clustering. The average accuracy values are 94.46%, 89.75%, 88.94% for the set of 5, 10 and 15 genes, respectively. The best average precision value 0.95 and accuracy 94.46% have been obtained after using the set of five

TABLE 6. Performance of the proposed method while using OPTICS clustering and Apriori algorithm.

Confidence of rules	0.7			0.8			0.9			Average		
No. of Genes (Clusters)	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy
5	**0.896** (±0.08)	0.968 (±0.03)	89.97% (±8.63%)	**0.938** (±0.04)	0.986 (±0.01)	**94.34%** (±4.12%)	0.91 (±0.06)	0.97 (±0.01)	89.92% (±9.38%)	**0.9147**	0.9747	91.41%
9	0.87 (±0.07)	0.988 (±0.01)	**91.46%** (±7.02%)	0.88 (±0.08)	0.992 (±0.00)	91.81% (±8.04%)	**0.938** (±0.05)	0.998 (±0.00)	**95.37%** (±4.19%)	0.896	0.9927	**92.88%**
15	0.708 (±0.18)	**0.99** (±0.00)	87.83% (±11.77%)	0.72 (±0.22)	**0.996** (±0.00)	86.80% (±12.42%)	0.71 (±0.23)	**1** (±0.00)	87.19% (±11.62%)	0.7127	**0.9953**	87.28%

TABLE 7. Performance of the proposed method while using Ward Agglomerative Clustering and Apriori algorithm.

Confidence of rules	0.7			0.8			0.9			Average		
No. of Genes (Clusters)	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy
5	0.926 (±0.07)	**0.896** (±0.03)	**93.07%** (±6.49%)	0.92 (±0.06)	**0.986** (±0.01)	**92.81%** (±6.84%)	0.946 (±0.05)	**0.996** (±0.00)	95.56% (±3.48%)	0.9307	**0.9593**	**93.81%**
10	0.564 (±0.22)	0.864 (±0.13)	60.99% (±22.13%)	0.668 (±0.27)	0.906 (±0.09)	72.59% (±19.69%)	0.924 (±0.07)	0.98 (±0.01)	91.06% (±7.23%)	0.7187	0.9167	74.88%
15	**0.992** (±0.00)	0.804 (±0.12)	80.79% (±12.38%)	**0.996** (±0.00)	0.872 (±0.11)	87.48% (±10.55%)	**0.996** (±0.00)	0.984 (±0.01)	**98.39%** (±1.41%)	**0.9947**	0.8867	88.89%

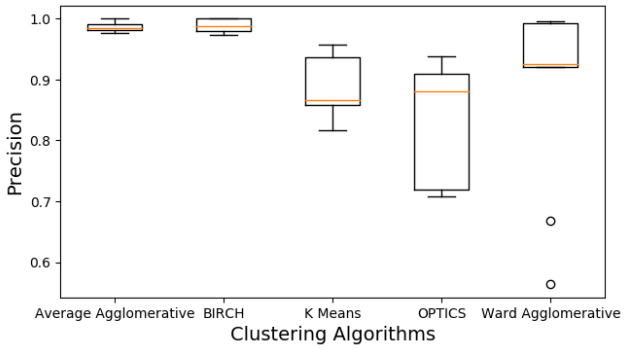


FIGURE 7. Boxplot of the precision metric of the proposed framework while using different clustering algorithms and Apriori algorithm.

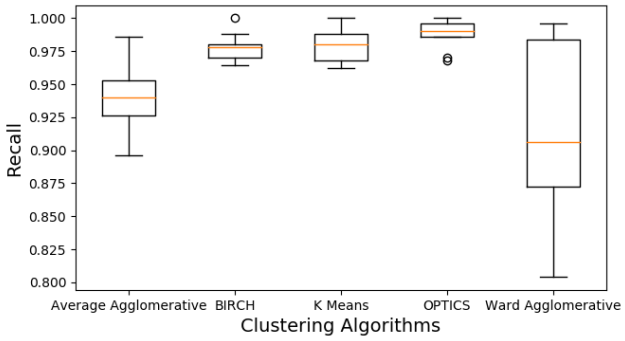


FIGURE 8. Boxplot of the recall metric of the proposed framework while using different clustering algorithms and Apriori algorithm.

genes. Meanwhile, the performance for the set of 10 and 15 genes are more or less identical.

It has been noticed from the result illustrated in Table 6 that after using the OPTICS clustering, the difference of the highest and the lowest average precision value is quite significant. In this case, the best average accuracy, 92.88%, have been obtained after using the set of 9 genes.

Table 7 depicts the performance of the proposed model after employing Ward Agglomerative clustering. It has been observed that the difference in the performance for the set of five genes and the set of ten genes is quite significant. The possible reason for this difference lies in the quality of selected genes. More non-significant genes display poor performance as these genes generate the association rules with low confidence values. In the current result, it has been noticed that the combined performance has been improved when the confidence threshold of the rules has been increased. Table 8 reports the average false positive rate of the proposed method after applying different clustering techniques combined with the Apriori algorithm. It is observed that Ward agglomerative clustering reports the maximum, and Average agglomerative clustering reports the minimum false positive rate among all. Figure 6 depicts the boxplot of the accuracy metric after applying different clustering techniques. The small size boxes for Average Agglomerative, BIRCH, K-means and OPTICS clustering indicate the consistent spread of accuracy values. The distribution in the case of BIRCH clustering indicates higher chances for correct classification than other approaches. For Ward Agglomerative clustering, the accuracy carries a low level of statistical significance. Meanwhile, the distribution of precision and recall have been illustrated in Figure 7 and Figure 8 respectively. The boxes for Average Agglomerative and BIRCH clustering suggests a high level of precision, and the boxes for BIRCH, K-means and OPTICS clustering suggests a high level of recall. Because of high skewness and multiple outliers, the performance of Ward Agglomerative clustering lacks reliability.

E. PERFORMANCE ANALYSIS OF THE PROPOSED METHOD AFTER APPLYING DIFFERENT UNSUPERVISED FEATURE REDUCTION TECHNIQUES COMBINED WITH FP GROWTH ALGORITHM

In this section, the performance of the proposed framework after employing different clustering techniques followed by the FP Growth algorithm has been discussed. The values of the performance metrics viz. accuracy, precision, and recall have been tabulated from Table 9 to Table 14 based on a

TABLE 8. Average false positive rate of the proposed method after applying different clustering techniques combined with Apriori algorithm. Here A.A and W.A indicate average agglomerative and ward agglomerative clustering algoritm.

Confidence of rules	0.7					0.8					0.9				
No. of Genes (Clusters)	A.A	BIRCH	K-means	OPTICS	W.A	A.A	BIRCH	K-means	OPTICS	W.A	A.A	BIRCH	K-means	OPTICS	W.A
5	N/A	N/A	**0.128** (±0.05)	0.18 (±0.07)	**0.16** (±0.03)	N/A	N/A	**0.094** (±0.06)	**0.08** (±0.05)	**0.197** (±0.07)	N/A	N/A	**0.123** (±0.08)	0.184 (±0.05)	**0.116** (±0.07)
10	**0.117** (±0.03)	0.155 (±0.06)	0.158 (±0.06)	0.148 (±0.02)	0.188 (±0.08)	0.117 (±0.06)	**0.125** (±0.04)	0.164 (±0.05)	0.139 (±0.07)	0.228 (±0.05)	0.148 (±0.06)	**0.146** (±0.05)	0.193 (±0.03)	**0.125** (±0.08)	0.248 (±0.04)
15	0.137 (±0.05)	**0.134** (±0.05)	0.149 (±0.03)	**0.143** (±0.08)	0.19 (±0.05)	**0.093** (±0.06)	0.148 (±0.06)	0.156 (±0.08)	0.152 (±0.08)	0.245 (±0.06)	**0.052** (±0.02)	0.193 (±0.05)	0.166 (±0.03)	0.148 (±0.05)	0.255 (±0.06)

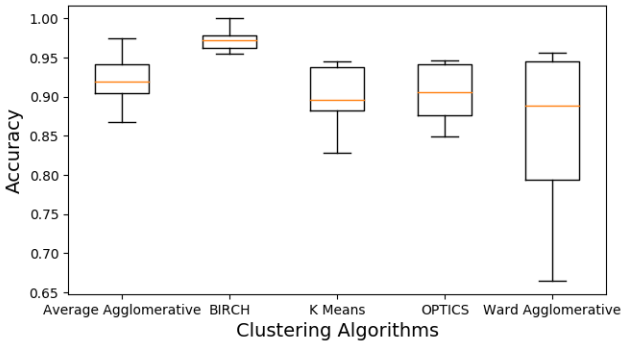


FIGURE 9. Boxplot of accuracy metric of the proposed framework while using different clustering algorithms and FP growth algorithm.

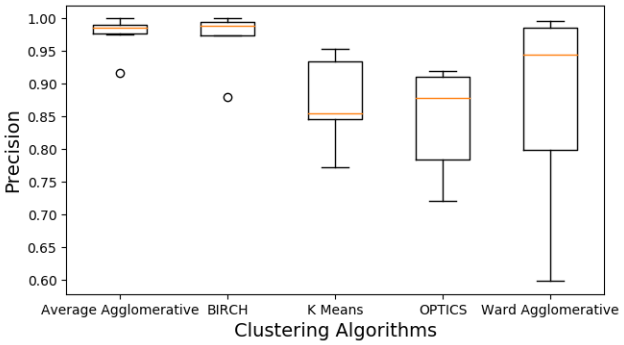


FIGURE 10. Boxplot of precision metric of the proposed framework while using different clustering algorithms and FP growth algorithm.

different number of genes and confidence values of the association rules. All results in these tables have been reported in (±) standard deviation format. Boxplot of these metrics for each clustering methods have been illustrated from Figure 9 to Figure 11.

After employing Average Agglomerative clustering, the performance of the proposed classifier has been presented in Table 9. The best average accuracy, 93.91% and recall 0.9473 are obtained when the selected gene count is 10. The best average precision value 1 is obtained when the set of 5 genes is used. It has been observed that the model has not generated any association rules having a confidence threshold value of 0.9 for the set of 5 genes. Hence, the precision, recall, and accuracy values have been displayed as N/A (Not Applicable). From the experimental result, it is concluded that the rules with high confidence value perform better than those with low confidence.

Table 10 illustrates the performance of the framework after adopting BIRCH clustering. The average precision, recall and accuracy values are 0.998, 0.982 and 98.07%, respectively, when the set 5 genes are used. The average precision, recall and accuracy values are 0.976, 0.9747, 96.29% and 0.952, 0.974, 97.37% when the number of selected genes is 10 and 15, respectively. Similar to Average Agglomerative clustering, the accuracy of the proposed model has been improved when the association rules with high confidence have been applied. In this case, the overall results for different sets of genes are closer to each other.

Table 11 depicts the overall performance of the framework while using K-means clustering. The average accuracy values

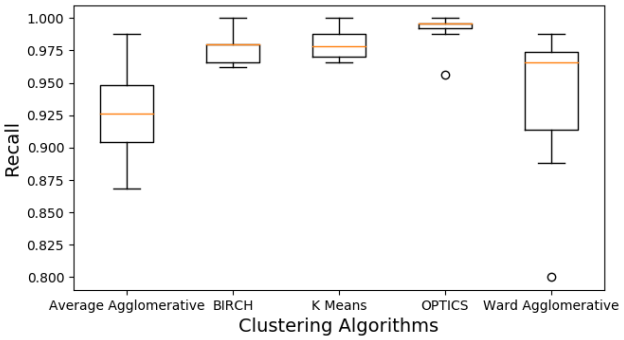


FIGURE 11. Boxplot of recall metric of the proposed framework while using different clustering algorithms and FP growth algorithm.

are 94.26%, 90.24%, 85.98% for the set of 5, 10 and 15 genes, respectively. The best average precision value 0.9467 and accuracy 94.26% have been obtained after using the set of 5 genes. In contrast, the best average recall value of 0.9847 has been obtained after using the group of 10 genes.

The result illustrated in Table 12 indicates a significant gap between the highest and the lowest average precision value when OPTICS clustering is adopted. The highest and lowest precision values have been obtained as 0.912 and 0.7447, respectively. The best average accuracy, 94.35%, has been obtained after using the set of 9 genes, whereas the accuracy percentages are more or less the same after using the sets of 5 and 15 genes.

Table 13 depicts the performance of the proposed model after applying Ward Agglomerative clustering. The best average accuracy, 93.87%, has been obtained when the number

TABLE 9. Performance of the proposed method while using Average Agglomerative Clustering and FP Growth algorithm.

Confidence of rules	0.7			0.8			0.9			Average		
No. of Genes (Clusters)	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy
5	**1** (±0.00)	0.868 (±0.06)	86.80% (±9.91%)	**1** (±0.00)	0.904 (±0.03)	90.40% (±7.33%)	N/A	N/A	N/A	**1**	0.886	88.60%
10	0.986 (±0.01)	**0.912** (±0.06)	**91.01%** (±7.28%)	0.976 (±0.02)	**0.942** (±0.05)	**93.29%** (±5.89%)	0.916 (±0.03)	**0.988** (±0.01)	**97.43%** (±2.35%)	0.9593	**0.9473**	**93.91%**
15	0.986 (±0.01)	0.904 (±0.09)	90.42% (±9.09%)	0.978 (±0.02)	0.94 (±0.04)	92.79% (±6.13%)	**0.986** (±0.01)	0.966 (±0.02)	96.47% (±3.17%)	0.9833	0.9367	93.23%

TABLE 10. Performance of the proposed method while using BIRCH Clustering and FP Growth algorithm.

Confidence of rules	0.7			0.8			0.9			Average		
No. of Genes (Clusters)	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy
5	**0.994** (±0.00)	0.962 (±0.03)	95.80% (±2.28%)	**1** (±0.00)	**0.984** (±0.01)	**98.40%** (±1.54%)	**1** (±0.00)	**1** (±0.00)	**100%** (±0.00%)	**0.998**	**0.982**	**98.07%**
10	0.98 (±0.01)	**0.966** (±0.02)	95.47% (±3.03%)	0.974 (±0.02)	0.978 (±0.02)	96.20% (±3.66%)	0.974 (±0.02)	0.98 (±0.01)	97.21% (±1.83%)	0.976	0.9747	96.29%
15	0.988 (±0.01)	0.962 (±0.03)	**96.49%** (±3.20%)	0.988 (±0.01)	0.98 (±0.01)	97.79% (±1.25%)	0.88 (±0.01)	0.98 (±0.01)	97.82% (±1.84%)	0.952	0.974	97.37%

TABLE 11. Performance of the proposed method while using K-means Clustering and FP Growth algorithm.

Confidence of rules	0.7			0.8			0.9			Average		
No. of Genes (Clusters)	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy
5	**0.954** (±0.03)	0.968 (±0.02)	**93.81%** (±5.28%)	**0.952** (±0.03)	0.97 (±0.02)	**94.48%** (±3.34%)	**0.934** (±0.05)	0.988 (±0.01)	**94.50%** (±5.29%)	**0.9467**	0.9753	**94.26%**
10	0.854 (±0.10)	0.966 (±0.03)	88.20% (±7.06%)	0.846 (±0.09)	**0.988** (±0.01)	89.57% (±8.78%)	0.896 (±0.11)	**1** (±0.00)	92.95% (±6.53%)	0.8653	**0.9847**	90.24%
15	0.822 (±0.11)	**0.978** (±0.02)	86.71% (±9.79%)	0.772 (±0.12)	0.978 (±0.02)	82.78% (±8.03%)	0.852 (±0.11)	0.99 (±0.01)	88.45% (±8.70%)	0.8153	0.982	85.98%

TABLE 12. Performance of the proposed method while using OPTICS Clustering and FP Growth algorithm.

Confidence of rules	0.7			0.8			0.9			Average		
No. of Genes (Clusters)	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy
5	0.85 (±0.10)	0.956 (±0.08)	85.96% (±8.87%)	0.912 (±0.06)	0.994 (±0.00)	91.98% (±7.28%)	0.878 (±0.11)	0.988 (±0.01)	87.57% (±8.52%)	0.88	0.9793	88.50%
9	**0.906** (±0.07)	**0.996** (±0.00)	**94.31%** (±5.19%)	**0.92** (±0.05)	0.996 (±0.00)	**94.62%** (±4.87%)	**0.91** (±0.06)	0.996 (±0.00)	**94.11%** (±4.08%)	**0.912**	0.996	**94.35%**
15	0.72 (±0.12)	0.992 (±0.00)	84.89% (±8.42%)	0.784 (±0.14)	**0.998** (±0.00)	90.60% (±5.54%)	0.73 (±0.12)	**1** (±0.00)	88.59% (±6.47%)	0.7447	**0.9967**	88.03%

TABLE 13. Performance of the proposed method while using Ward Agglomerative Clustering and FP Growth algorithm.

Confidence of rules	0.7			0.8			0.9			Average		
No. of Genes (Clusters)	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy
5	0.92 (±0.07)	**0.98** (±0.01)	**92.31%** (±7.34%)	0.95 (±0.03)	**0.974** (±0.02)	**94.50%** (±4.09%)	0.944 (±0.04)	**0.988** (±0.01)	94.81% (±5.31%)	0.938	**0.9807**	**93.87%**
10	0.598 (±0.18)	0.914 (±0.06)	66.45% (±16.99%)	0.722 (±0.15)	**0.974** (±0.02)	77.89% (±15.49%)	0.798 (±0.17)	0.966 (±0.02)	79.37% (±13.87%)	0.706	0.9513	74.57%
15	**0.986** (±0.01)	0.8 (±0.07)	79.67% (±8.93%)	**0.994** (±0.00)	0.888 (±0.05)	88.84% (±6.91%)	**0.996** (±0.00)	0.956 (±0.02)	**95.56%** (±3.53%)	**0.992**	0.8813	88.03%

of genes is 5. In contrast, the best average precision is 0.992, and the best average recall is 0.9807 when the gene counts are 15 and 5, respectively. The result indicates that the overall performance for the set of 10 genes is relatively poorer than the other set of genes. In the current result, it has been observed that association rules with high confidence positively impact the improvement of the overall performance of the framework. Table 14 reports the average false positive rate of the proposed method after applying different clustering techniques combined with the FP Growth algorithm. It is observed that Ward agglomerative clustering reports the maximum, and Average agglomerative clustering reports the minimum false positive rate among all.

The boxplot of the accuracy metric after applying different clustering techniques is depicted in Figure 9. The smaller box size for Average Agglomerative and BIRCH indicates a greater and more consistent performance than other clustering methods. The distribution of precision and recall have

TABLE 14. Average false positive rate of the proposed method after applying different clustering techniques combined with FP Growth algorithm. Here A.A and W.A indicate Average Agglomerative and Ward Agglomerative clustering algoritm.

Confidence of rules	0.7					0.8					0.9				
No. of Genes (Clusters)	A.A	BIRCH	K-means	OPTICS	W.A	A.A	BIRCH	K-means	OPTICS	W.A	A.A	BIRCH	K-means	OPTICS	W.A
5	N/A	0.13 (±0.05)	**0.142** (±0.03)	0.162 (±0.03)	**0.197** (±0.09)	N/A	N/A	**0.119** (±0.02)	0.141 (±0.03)	**0.127** (±0.04)	N/A	N/A	**0.134** (±0.04)	0.15 (±0.07)	**0.161** (±0.11)
10	0.136 (±0.05)	0.11 (±0.07)	0.177 (±0.05)	**0.137** (±0.07)	0.224 (±0.05)	**0.109** (±0.03)	0.125 (±0.03)	0.154 (±0.02)	**0.096** (±0.08)	0.226 (±0.07)	0.139 (±0.05)	0.107 (±0.04)	0.140 (±0.06)	**0.093** (±0.06)	0.163 (±0.09)
15	**0.093** (±0.05)	**0.116** (±0.06)	0.139 (±0.04)	0.175 (±0.04)	0.212 (±0.08)	0.16 (±0.06)	**0.124** (±0.05)	0.131 (±0.06)	0.106 (±0.08)	0.215 (±0.07)	**0.076** (±0.04)	**0.197** (±0.06)	0.179 (±0.07)	0.132 (±0.08)	0.252 (±0.09)

been illustrated in Figure 10 and Figure 11 respectively. The boxes for Average Agglomerative and BIRCH clustering suggest a high level of precision, and the boxes for BIRCH, K-means and OPTICS clustering suggest a high level of recall.

V. CONCLUSION

The current study proposes an associative classifier framework to efficiently detect Dengue fever using gene expression data. However, not all genes are equally important in detection of dengue fever. To identify most promising genes, an unsupervised feature selection strategy has been adopted by applying well known clustering algorithms. After obtaining the most promising features (genes) the modified dataset is used to mine rules for dengue fever detection. The rules having only the target variable in the body, are kept for classification. To improve the classifier performance, rules with higher confidence value are considered. This ensured that the selected gene have a higher correlation with target variable, thereby making the classifier more confident. A wide range of clustering algorithms have been explored in the current study. Experimental results have indicated that the performance of BIRCH clustering algorithm is most promising in identifying the important genes while using Apriori algorithm for mining rules. In terms of accuracy the performance of Average Agglomerative, K-means and OPTICS have been found to be satisfactory whereas the performance of Ward Agglomerative has been found to be poorest. In case of FP-Growth a similar trend of performance is observed. Optimal number of gene selection plays a vital role in deciding the performance of the proposed classifier. As, in case of average agglomerative algorithm no rules are mined while confidence threshold is set to 0.9 and number of genes selected by feature selection method is 5. Overall, the performance of the proposed model has been found to be extremely satisfactory and statistically significant in detecting dengue fever. Nevertheless, future studies can be focused towards developing multiclass associative classifiers for similar tasks.

AUTHOR CONTRIBUTIONS

Conceptualization: Sankhadeep Chatterjee; Methodology: Saubhik Paladhi, Sankhadeep Chatterjee; Formal analysis and investigation: Saubhik Paladhi, Sankhadeep Chatterjee; Simulation: Diptaraj Sen, Saubhik Paladhi; Writing–original draft preparation: Saubhik Paladhi, Sankhadeep Chatterjee; Writing–review and editing: Sankhadeep Chatterjee, Soumen Banerjee, Saubhik Paladhi, Diptaraj Sen; Resources: Jaroslav Frnda, Jan Nedoma; Supervision: Soumen Banerjee.

REFERENCES

[1] O. J. Brady *et al.*, "Refining the global spatial limits of dengue virus transmission by evidence-based consensus," *PLoS Neglected Tropical Diseases*, vol. 6, no. 8, p. e1760, 2012, doi: 10.1371/journal.pntd.0001760.

[2] S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh, and M. F. Myers, "The global distribution and burden of dengue," *Nature*, vol. 496, no. 7446, pp. 504–507, 2013.

[3] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 1993, pp. 207–216.

[4] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. Int. Conf. Very Large Data Bases*, vol. 1215, 1994, pp. 487–499.

[5] F. Tao, F. Murtagh, and M. Farid, "Weighted association rule mining using weighted support and significance framework," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2003, pp. 661–666.

[6] K. Sun and F. Bai, "Mining weighted association rules without preassigned weights," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 4, pp. 489–495, Apr. 2008.

[7] L. Dey and A. Mukhopadhyay, "Biclustering-based association rule mining approach for predicting cancer-associated protein interactions," *IET Syst. Biol.*, vol. 13, no. 5, pp. 234–242, Oct. 2019.

[8] U. Maulik, S. Mallik, A. Mukhopadhyay, and S. Bandyopadhyay, "Analyzing large gene expression and methylation data profiles using StatBicRM: Statistical biclustering-based rule mining," *PLoS ONE*, vol. 10, no. 4, Apr. 2015, Art. no. e0119448.

[9] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "A novel biclustering approach to association rule mining for predicting HIV-1–human protein interactions," *PLoS ONE*, vol. 7, no. 4, Apr. 2012, Art. no. e32289.

[10] S. Mallik, A. Mukhopadhyay, and U. Maulik, "RANWAR: Rank-based weighted association rule mining from gene expression and methylation data," *IEEE Trans. Nanobiosci.*, vol. 14, no. 1, pp. 59–66, Jan. 2015.

[11] S. Mallik, A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "Integrated analysis of gene expression and genome-wide DNA methylation for tumor prediction: An association rule mining-based approach," in *Proc. IEEE Symp. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Apr. 2013, pp. 120–127.

[12] C. Tew, C. Giraud-Carrier, K. Tanner, and S. Burton, "Behavior-based clustering and analysis of interestingness measures for association rule mining," *Data Mining Knowl. Discovery*, vol. 28, no. 4, pp. 1004–1045, Jul. 2014.

[13] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 1–12, 2000.

[14] S. Alagukumar and R. Lawrance, "A selective analysis of microarray data using association rule mining," *Proc. Comput. Sci.*, vol. 47, pp. 3–12, Jan. 2015.

[15] C. Becquet, S. Blachon, B. Jeudy, J.-F. Boulicaut, and O. Gandrillon, "Strong-association-rule mining for large-scale gene-expression data analysis: A case study on human sage data," *Genome Biol.*, vol. 3, no. 12, pp. 1–16, 2002.

[16] K. Vengatesan, S. Selvarajan, and S. Pragadeeswaran, "The performance analysis of microarray data using occurrence clustering," *Int. J. Math. Sci. Eng.*, vol. 3, no. 2, pp. 69–75, 2014.

[17] S. Vassilvitskii and D. Arthur, "k-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, Jun. 2006, pp. 1027–1035.

[18] S. Alagukumar, C. Vanitha, and R. Lawrance, "Clustering of association rules on microarray gene expression data," in *Advanced Computing and Intelligent Engineering*. Cham, Switzerland: Springer, 2020, pp. 85–97.
[19] B. Pontes, R. Giráldez, and J. S. Aguilar-Ruiz, "Biclustering on expression data: A review," *J. Biomed. Informat.*, vol. 57, pp. 163–180, Oct. 2015.
[20] X. Yu, G. Yu, and J. Wang, "Clustering cancer gene expression data by projective clustering ensemble," *PLoS ONE*, vol. 12, no. 2, Feb. 2017, Art. no. e0171429.
[21] J. Oyelade, I. Isewon, F. Oladipupo, O. Aromolaran, E. Uwoghiren, F. Ameh, M. Achas, and E. Adebiyi, "Clustering algorithms: Their application to gene expression data," *Bioinf. Biol. Insights*, vol. 10, Jan. 2016, Art. no. BBI.S38316.
[22] B. Hossen, H. A. Siraj-Ud-Doulah, and A. Hoque, "Methods for evaluating agglomerative hierarchical clustering for gene expression data: A comparative study," *Comput. Biol. Bioinf.*, vol. 3, no. 6, pp. 88–94, 2015.
[23] Z. Yu, H. Chen, J. You, H.-S. Wong, J. Liu, L. Li, and G. Han, "Double selection based semi-supervised clustering ensemble for tumor clustering from gene expression profiles," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 4, pp. 727–740, Jul. 2014.
[24] S. Chatterjee, S. Hore, N. Dey, S. Chakraborty, and A. S. Ashour, "Dengue fever classification using gene expression data: A PSO based artificial neural network approach," in *Proc. 5th Int. Conf. Frontiers Intell. Comput., Theory Appl.* Cham, Switzerland: Springer, 2017, pp. 331–341.
[25] C. Gakii and R. Rimiru, "Identification of cancer related genes using feature selection and association rule mining," *Informat. Med. Unlocked*, vol. 24, Jan. 2021, Art. no. 100595.
[26] R. G. Pensa, C. Leschi, J. Besson, and J.-F. Boulicaut, "Assessment of discretization techniques for relevant pattern discovery from gene expression data," in *Proc. 4th Int. Conf. Data Mining Bioinf.*, 2004, pp. 24–30.
[27] N. Megiddo and R. Srikant, "Discovering predictive association rules," in *Proc. KDD*, vol. 98, 1998, pp. 274–278.
[28] M. J. Zaki, "Generating non-redundant association rules," in *Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2000, pp. 34–43.
[29] S. S. Sahoo and T. Swarnkar, "A theoretical approach for augmenting association rule mining to predict protein-protein interaction," *Exp. Tech.*, vol. 2, no. 5, p. 8, 2011.
[30] W. Zakaria, Y. Kotb, and F. Ghaleb, "MCR-Miner: Maximal confident association rules miner algorithm for up/down-expressed genes," *Appl. Math. Inf. Sci.*, vol. 8, no. 2, p. 799, 2014.
[31] L. Antonie and K. Bessonov, "Classifying microarray data with association rules," in *Proc. ACM Symp. Appl. Comput. (SAC)*, 2011, pp. 94–99.
[32] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1370–1386, Nov. 2004.
[33] K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo, "Validating clustering for gene expression data," *Bioinformatics*, vol. 17, no. 4, pp. 309–318, Apr. 2001.
[34] M. Steinbach, L. Ertöz, and V. Kumar, "The challenges of clustering high dimensional data," in *New Directions in Statistical Physics*. Cham, Switzerland: Springer, 2004, pp. 273–309.
[35] Y. Wang, D. J. Miller, and R. Clarke, "Approaches to working in high-dimensional data spaces: Gene expression microarrays," *Brit. J. Cancer*, vol. 98, no. 6, pp. 1023–1028, 2008.
[36] B. Arunasalam and S. Chawla, "CCCS: A top-down associative classifier for imbalanced class distribution," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2006, pp. 517–522.
[37] T. De Bie, K.-N. Kontonasios, and E. Spyropoulou, "A framework for mining interesting pattern sets," *ACM SIGKDD Explor. Newslett.*, vol. 12, no. 2, pp. 92–100, Mar. 2011.
[38] A. Gallo, T. D. Bie, and N. Cristianini, "Mini: Mining informative non-redundant itemsets," in *Proc. Eur. Conf. Princ. Data Mining Knowl. Discovery*. Cham, Switzerland: Springer, 2007, pp. 438–445.
[39] G. I. Webb, "Self-sufficient itemsets: An approach to screening potentially interesting associations between items," *ACM Trans. Knowl. Discovery From Data*, vol. 4, no. 1, pp. 1–20, Jan. 2010.
[40] S. Kannan and R. Bhaskaran, "Association rule pruning based on interestingness measures with clustering," 2009, *arXiv:0912.1822*.
[41] F. Verh and S. Chawla, "Using significant, positively associated and relatively class correlated rules for associative classification of imbalanced datasets," in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, Oct. 2007, pp. 679–684.
[42] A. Segura-Delgado, A. Anguita-Ruiz, R. Alcalá, and J. Alcalá-Fdez, "Mining high average-utility sequential rules to identify high-utility gene expression sequences in longitudinal human studies," *Expert Syst. Appl.*, vol. 193, May 2022, Art. no. 116411.
[43] H. Li and P. C.-Y. Sheu, "A scalable association rule learning and recommendation algorithm for large-scale microarray datasets," *J. Big Data*, vol. 9, no. 1, pp. 1–25, Dec. 2022.
[44] M. Rahmanian and E. G. Mansoori, "An unsupervised gene selection method based on multivariate normalized mutual information of genes," *Chemometric Intell. Lab. Syst.*, vol. 222, Mar. 2022, Art. no. 104512.
[45] N. Koul and S. S. Manvi, "Feature selection from gene expression data using simulated annealing and partial least squares regression coefficients," *Global Transitions Proc.*, vol. 3, no. 1, pp. 251–256, Jun. 2022.
[46] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, 1967, pp. 281–297.
[47] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," *ACM SIGMOD. Rec.*, vol. 28, no. 2, pp. 49–60, Jun. 1999.
[48] J. H. Ward, Jr., "Hierarchical grouping to optimize an objective function," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 236–244, 1963.
[49] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," *ACM SIGMOD Rec.*, vol. 25, pp. 103–114, Jun. 1996.

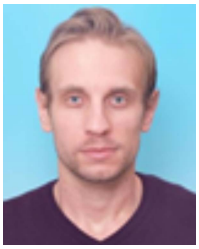
DIPTARAJ SEN (Student Member, IEEE) is currently pursuing the bachelor's degree in computer science and engineering with the University of Engineering and Management, Kolkata, India. He has been doing research on topics, such as association rule mining, imbalance classification, and data-driven challenges. He has international conference publications in reputed conferences as an undergraduate researcher. His research interests include machine learning, data mining, and social network analysis.

SAUBHIK PALADHI received the B.Tech. degree in computer science and engineering from the Maulana Abul Kalam Azad University of Technology, India, in 2015, and the M.Tech. degree in computer science and engineering from the University of Kalyani, India, in 2019. His research interests include data mining and combinatorial optimization. He has published several articles in those domains. He has been awarded the UGC NET Junior Research Fellowship in 2019.

JAROSLAV FRNDA (Senior Member, IEEE) was born in Martin, Slovakia, in 1989. He received the M.Sc. and Ph.D. degrees from the Department of Telecommunications, VSB–Technical University of Ostrava, in 2013 and 2018, respectively. He has been working as an Assistant Professor with the Department of Quantitative Methods and Economic Informatics, Faculty of Operation and Economics of Transport and Communications, University of Žilina, Slovakia, since 2019. He has authored or coauthored 18 SCI-E and nine ESCI articles in WoS. His research interests include quality of multimedia services in IP networks, data analysis, and machine learning algorithms.

SANKHADEEP CHATTERJEE received the B.Tech. degree in computer science and engineering from the Maulana Abul Kalam Azad University of Technology, in 2015, and the M.Tech. degree in computer science and engineering from the University of Calcutta, Kolkata, India, in 2017. He is currently working as an Assistant Professor with the Department of Computer Science and Technology, University of Engineering and Management, Kolkata. He has published and presented more than 60 research papers in reputed international journals/conferences. His current research interests include machine learning, deep learning, metaheuristics, and text data analysis. He obtained the Prestigious Council of Scientific & Industrial Research (CSIR) Senior Research Fellowship from the Government of India in 2019.

SOUMEN BANERJEE (Senior Member, IEEE) received the B.Sc. (Hons.) degree in physics from the University of Calcutta, in 1998, the B.Tech. and M.Tech. degrees in radio physics and electronics from the Institute of Radio Physics and Electronics, University of Calcutta, in 2001 and 2003, respectively, and the Ph.D. degree in engineering from the Indian Institute of Engineering Science and Technology (IIEST), Shibpur, India. He was a Visiting Faculty at the Department of Applied Physics, University of Calcutta. He is currently the Head of the Department of Electronics and Communication Engineering, University of Engineering and Management, Kolkata, India. He has a teaching/research experience of more than 20 years. He has published more than 100 contributory papers in journals and international conferences. He has authored ten books and five book chapters in fields of communication engineering, electromagnetic field theory, microwave, and antenna. He has edited three books published by Springer. His profile is included in IBC, Cambridge, U.K., and Marquis Who's Who in the World, USA. His current research interests include design, fabrication and characterization of wide band gap semiconductor-based IMPATT diodes at D-band, W-band and THz frequencies, SIW technology based antennas, printed antennas and arrays, FSS, dielectric resonator antennas, body wearable antennas, machine learning, fuzzy systems, and evolutionary computation. He is a Senior Member of the IEEE AP Society. He is also a fellow of IETE (New Delhi, India). He is a Reviewer of several international journals, such as the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE, the IEEE TRANSACTIONS ON ELECTRON DEVICES, IEEE ACCESS, IEEE SENSORS LETTERS, *Microwave and Optical Technology Letters* (MOTL-Wiley), *Journal of Electromagnetic Waves and Applications* (Taylor & Francis), *Radioengineering* journal (Czech and Slovak Technical Universities), *Journal of Computational Electronics* (Springer-Nature), *Journal of Renewable and Sustainable Energy* (American Institute of Physics—AIP), *Cluster Computing* (Springer-Nature), and *Journal of Infrared, Millimeter, and Terahertz Waves* (Springer-Nature). He acted as a Convener in international conferences, such as OPTRONIX-2019 (IEEE) and OPTRONIX2020 (Springer); both held at Kolkata, India, and also chaired many technical sessions in several international conferences.



JAN NEDOMA (Senior Member, IEEE) was born in Czech Republic, in 1988. He received the master's degree in information and communication technology from the VSB—Technical University of Ostrava, in 2014. Since 2014, he has been working as a Research Fellow with the Technical University of Ostrava. In 2018, he successfully defended his dissertation thesis. He started working as an Assistant Professor at the VSB—Technical University of Ostrava, in 2018. He has become an Associate Professor of communication technologies, in 2021, after defending the habilitation thesis titled "Fiber-Optic Sensors in Biomedicine: Monitoring of Vital Functions of the Human Body in Magnetic Resonance (MR) Environment." He has more than 150 journal articles and conference papers in his research areas and nine valid patents. His research interests include optical communications, optical atmospheric communications, optoelectronics, optical measurements, measurements in telecommunication technology, fiber-optic sensory systems, data processing from fiber-optic sensors, the use of fiber-optic sensors within the smart technological concepts (smart home, smart home care, intelligent building, smart grids, smart metering, and smart cities), and for the needs of industry 4.0.