Modified Genetic Algorithm and Polynomial Distribution Based Convolutional Neural Network for Asthma Disease Diagnosis

Dr. L. Prathiba[a]
[a]
Associate Professor, MIT Art Design and Technology University, Pune.

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: Prevention is considered to be the key solution for asthma rather than treatment for the same. Generally, asthma starts initially at early life, hence primary asthma determination in young children is regarded as significant life saving activity. In prevailing scheme, classification necessitates extra time for task completion as well disease prediction cannot be accomplished precisely using enormous samples for evaluation. Modified Genetic Algorithm and Polynomial distribution based Convolutional Neural Network (MGA+PCNN) algorithm is greatly utilized for mitigating these issues and this methodology encompasses key steps such as normalization, pre-processing, feature selection and classification process. Improved Variance Stabilizing Normalization (IVSN) is exploited for dataset accuracy improvement. Pre-processing is achieved using logistic regression clustering algorithm which helps in increasing asthma patient's attributes significance and these pre-processed features are considered for feature selection process. MGA algorithm is suggested for highest f-measure features selection from specified dataset via best global and local fitness values. Subsequently, PCNN algorithm is deployed for accurate outcome classification for asthma disease identification. It is thereby validated that suggested MGA+PCNN algorithm offers improved True Positive Rate (TPR), False Positive Rate (FPR), F-measure and accuracy than prevailing approaches.

Keywords: Asthma Disease Diagnosis, Modified Genetic Algorithm (MGA), Polynomial Distribution based Convolutional Neural Network (PCNN), Feature Selection, Classification

1. Introduction

The chronic lung disease generally known as asthma instigated due to airway inflammation and considered to be utmost usual pediatric chronic disease. Asthma is regarded as crucial analysis for 1/3 of pediatric emergency branch visits (Vargas et al., 2006), besides utmost recurrent cause for preventable pediatric hospitalization and school nonattendance caused by chronic disorders (Wang et al., 2005). Almost 9.3 billion dollars, or 8% of overall direct healthcare cost for all children, were spent on pediatric asthma in 2008. It is also inferred that symptom persist for 80% of pediatric asthma patients in the age group of six (Hafkamp-de Groen et al., 2013), most of them comes under age group of three (Guilbert et al., 2004). Nonetheless, approximately 1/3 of children with at least one asthmatic symptoms episode by age three tend to suffer from asthma at age six and over. There occurs asthma under-diagnosis in almost 18-75% of asthmatic children. Asthma Over diagnosis is considered to be extensive. About 11% of patients in primary care exploiting inhaled corticosteroids, furthermost intoxicating and reliably effective long-term control medication for asthma possess no medication indication. An accurate model needs to be constructed for predicting whether a child might progress asthma in upcoming days. For predictive models potential, a published predictive model for asthma development has previously been revealed to outclass a physician's asthma diagnosis in young children pertaining to low sensitivity of 29% and a low positive predictive value of 23% offers various advantages.

Asthma arises due to combination of genetic along with environmental aspects (Martinez, 2007), such as air pollution as well as allergens. Additional probable activates are medicines for instance aspirin and beta blockers. Diagnosis mainly relies on indications pattern reaction to therapy over time, and spirometry which is classified with respect to indications frequency, Forced Expiratory Volume in One second (FEV1), and peak expiratory flow rate. , Forced Expiratory Volume in One second (FEV1), and peak expiratory flow rate. The categorization might be done as atopic or non-atopic wher atopy signifies an inclination in implementing a type 1 hypersensitivity reaction (Kumar et al., 2010).

Permanent cure does not exist for asthma but indication reduction can be avoided via activates, such as allergens, annoyances, and inhaled corticosteroids usage. Long-acting beta agonists (LABA) or anti leukotriene agents are exploited apart from inhaled corticosteroids once asthma signs sustain uninhibited. The rapidly deteriorating indications healing are achieved by inhaled short-acting beta-2 agonist for instance salbutamol and corticosteroids taken via mouth. For severe circumstances, magnesium sulfate, intravenous corticosteroids, and hospitalization are necessitated.

Lungs position along with airways in body is described in figure 1. A. The usual airway cross-section is illustrated in figure 1.B. The airway cross-section for asthma indications is depicted in figure 1.C.Sometimes asthma signs are minor besides be off on their own or subsequently trivial treatment with asthma medication. Nevertheless from time to time signs persist to deteriorate. Asthma attacks are also termed as as flare ups or exacerbations (eg-zas-er-BA-shuns). Treatment indications while primarily is very substantial which avoids

inhibiting indications from attaining worse and creating a serious asthma attack for which emergency care is necessitated, and might become dreadful.

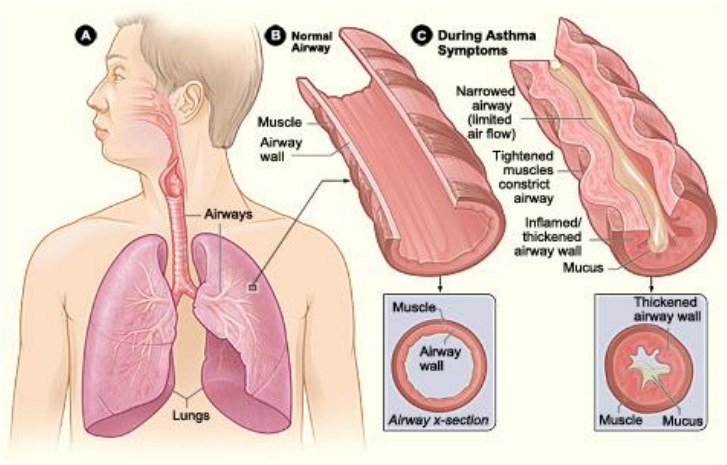


Figure 1. Asthma Disease

The chief sources of asthma is caused due to breathing tubes inflammation (swelling) , which carries air in and out of the lungs, breathing tubes are made very subtle, consequently thinness may occur owing to Asthma. This might ensue randomly, or successively revelation to a trigger. Even if asthma might characteristically be sustained under control, it is yet a severe state, and thereby various difficulties arose. Subsequently, it is very substantial in following treatment strategy and indications should not be disregarded when they are getting worse (Vial Dupuy et al., 2011).

Data mining is considered to be promising solution for obtaining predictive information from massive databases with greater capability for aiding companies' attention on most substantial information in their data warehouses (Larose & Larose, 2014). Data mining tools predict imminent developments and conducts, permitting businesses for dynamic, knowledge-driven ranges creation.Biologists are accelerating their fortitudes in expressing biological processes, owing to disease paths in medical perspectives. A biological and medical data flood is obtained from genomic and protein series, protein interactions, DNA microarrays, biomedical images, to disease paths along with electronic health records (Li et al., 2013).

Since the earlier system consists of lesser amount of health-care data accompanied by several missing values regarding patients, it is considered being inefficient. The performance of the presented approach may get hampered due to the inconsistent features exist in asthma diagnosis method. Hence, preventing the presented technique from these issues is significantly focused by controlling them, thereby asthma diagnosis model can be advanced.

This research work predominantly tends to facilitate the patients with an appropriate and flexible asthma diagnosis system. Through this framework, the patients can be prevented from the vulnerable causes by detecting their asthma level. In this work, a novel classification approach is introduced that helps enhancing the accuracy level of asthma disease prediction, through which the asthma symptoms of several patients can be learnt. MGA algorithm is employed to carry out the feature selection process by producing optimal fitness values. Provided asthma dataset is gone through this classification task with the help of PCNN algorithm.

2. Related Work

Robbie-Ryan & Brown, (2002) comprehensively described the mast cells as follows. They are vital effect/cells as regards adaptive immune responses. Besides, they actively play a crucial role in the pathophysiology of asthma in terms of the secretion of wide-ranging mediators, immediately after the activation done by allergens. Furthermore, they are significantly abundant in lymphoid tissue, uterus, skin, mesentery, thymus, tongue, nasal mucosa, synovia, urinary bladder, conjunctiva, and lung; in the surface of small and large blood vessels, beneath the skin, in submucosal and subserosal layers of the gastrointestinal tract and airway; in the connective tissues of each organ that exist across the blood vessels, excluding brain. Significantly, a group of membrane receptors, high-affinity IgE receptor and cytokine receptors (IL-3R, IL-4R, IL-5R, IL-9R, IL-10R), GM-CSF, chemokine receptors and nerve growth factor receptors are indicated by them. It is worth noticing the recent occurrence, in which the mast cells penetrated through airway smooth muscle have a mutual association with the degree of AHR in asthma.

Zhu et al., (2019) tend to diagnose and predict diabetes in the initial stage by including the Pima Indians Diabetes dataset, for which they confer data mining based framework in this work. K-means is a clustering technique that is widely regarded for its simplistic approach as well as wide-ranging application towards different kinds of data. However, it is pretty sensitive to cluster centers' initial positions that detect the outcome of a final cluster, through which either effectively clustered and adequate dataset can be obtained or small volume of data can be acquired due to inappropriate clustering of the original dataset. Consequently, the performance of the

logistic regression model getting constrained. Determining the methods to increase the accuracy of the logistic regression and k-means clustering is a core objective of this work.

Princy & Sivaranjani, (2016) intended to detect the specificity, sensitivity, time second, and accuracy of asthma prediction data using classification techniques. According to compare with MLP, the SVM method attains greater accuracy rates. Besides, it is capable to deliver 98.90% specificity, 97.50% sensitivity, and 98.50% accuracy, whereas other algorithms attain lesser rates. The breathing tests, namely FEV1, FVC and FEF are predominantly considered for predicting asthma, during which the detection of lung capacity is primarily focused. Identifying various breathing sounds of an individual is named Lung capacity. Ultimately, this work concludes that the Support Vector Machine (SVM) method is adequate to outperform Multilayer Perception (MLP) method as well as other classification methods, based on its accuracy level.

Várkonyi & Buza, (2019) conferred that Extreme learning machine (ELM) is a special single-hidden layer feed-forward neural network (SLFN) that solely accompanies a single hidden layer and arbitrarily selected weights that exist within the input layer and the hidden layer. In ELM, solely the weights amid hidden and output layers must be trained, which is considered being its significant advantage. Thereby, reduced the expenses for computation and attained moderate training time. In this study, the performance of ELM is compared with several regularization methods, such as L1, L2, and No Regularization in case of a binary classification process associated with gene expression data. In the context of several methods, L1 regularization may lead to sparse structures (most of the learned weights are zero), hence this study investigates the distribution of the learned weights and the ensuing structure's sparsity under ELM context.

Sachnev et al., (2015) tends to detect biomarkers accompanied by the capability of indicating various cancer types by using microarray gene expression cancer data. Providing a multi-class cancer classifier is a major objective of the authors, which is capable of differentiating the cancer types and identifying the type-specific biomarkers by applying a neural network-based Extreme Learning Machine (ELM) algorithm and Binary Coded Genetic Algorithm (BCGA). For interpreting the type of molecular features that might be centric for various cancer types, this method utilized gene expression analysis. The hallmark features introduced by the selected genes drive the tasks, through which the following events may occur, i.e. initiation of tumors, participation in cell migration, and implementation of invasive properties that cause metastasis.Arslan & Ozturk, (2019) intended to resolve the classification problems over four different data sets by introducing artificial bee colony programming (ABCP) to feature selection. In accordance with the total number of classes in the data sets, the sensitivity fitness function is exploited to obtain the optimal models, which are later compared with the frameworks built through genetic programming (GP). During the feature selection process, identification of valuable properties that consists of class information through eradicating unnecessary and noisy features exist in the data sets, and facilitation of classifiers are considered as primary objectives. Empirical findings depict the proficiency and accuracy of the proposed approach to outperform the GP, as regards classification accuracy and selection of critical features on familiar benchmark problems.

Strumberger et al., (2019) introduced the method, namely tree growth algorithm to build convolutional neural network architecture. Most often, convolutional neural networks are regarded as a special class of deep neural networks, in which numerous convolution, pooling and fully connected layers are involved. Besides, it is known for its efficient handling of numerous classification processes concurrently. However, identifying the network architecture that possesses the optimal performance for a particular application is the significant challenge involved in this domain. The set of hyper-parameter values, namely the number of convolutional and dense layers, the number of kernels per layer and kernel size are the driving factors for the performance of the network. A new tree growth algorithm from the group of swarm intelligence metaheuristics optimizes the hyper parameters. Taken the familiar dataset called MNIST to evaluate the solutions' quality, robustness, and the performance of the proposed model.

Liu & Yang, (2017) focused on the generic issue involved in the cell detection task, for which they introduced an innovative algorithm. Initially, the generation of cell detection candidates set is initiated by applying various algorithms accompanied by different parameters. Then, a trained Deep Convolutional Neural Network (DCNN) is involved to assign a score for each candidate. At last, the outcomes of the cell detection process is made up of a subset among optimal detection results that are chosen from all candidates. Through formalizing the subset selection task, it is transformed as a maximum-weight independent set problem that tends to identify the heaviest subset of mutually nonadjacent nodes in a graph. Empirical findings depict the efficiency of the proposed general cell detection algorithm to deliver the optimal detection results, which are superior to all other individual cell detection algorithm.

3. Proposed Methodology

This research study proposed the algorithm, namely Modified Genetic Algorithm and Polynomial distribution based Convolutional Neural Network (MGA+PCNN) to procure highly accurate results in asthma

disease diagnosis. the general block diagram of the proposed approach is demonstrated in figure 2.

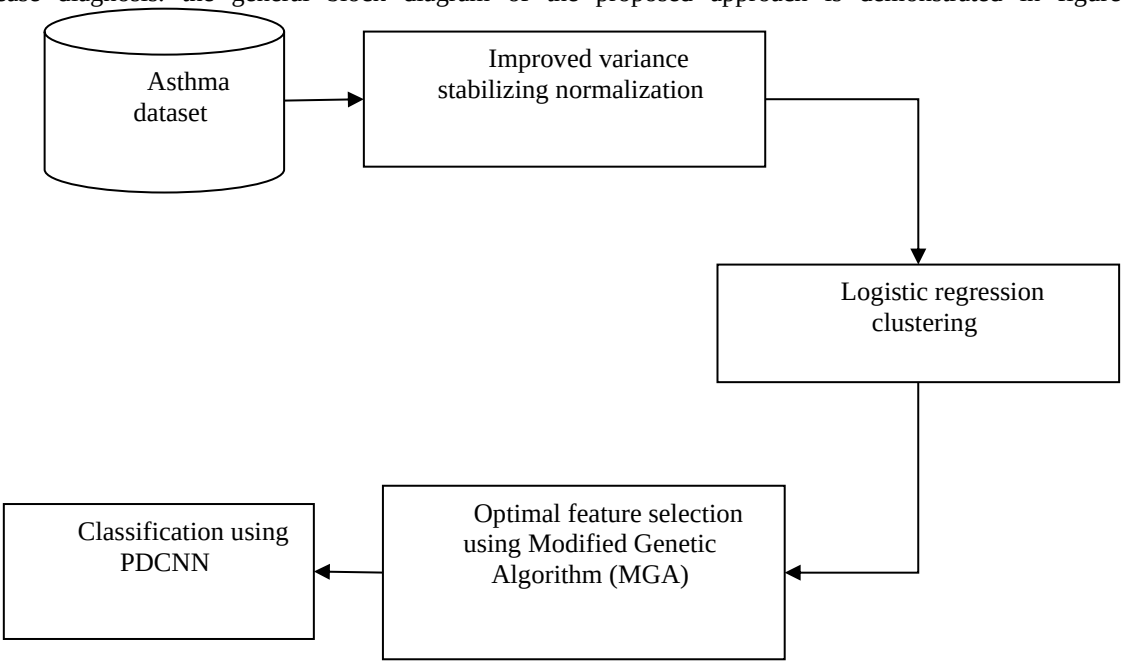


Figure 2. Overall Block Diagram of the Proposed System

1. 3.1 Improved Variance Stabilizing Normalization (IVSN)

Some kind of sources that impact the deliberate gene expression levels can be evacuated in the way called Normalization. In the previous step of microarray data analysis, it assumes a critical part. And, the following examination comes about are profoundly reliant on Normalization. For the rearrangement of information, Variance Stabilizing Normalization (VSN) approach is utilized that modifies the information, through which remaining difference constantly accomplishes the overall intensity spectrum.

With each gathering array elements kept by a single spotting pen (often referred to as a pen gathering/sub grid), the local normalization is connectedfrequently, concerning spotted arrays. Since the Local normalization possesses the desirable position, it is capable of making the modifications in terms of systematic spatial variation in the array, including inconsistencies among the spotting pens used to make the array, variability in the slide surface, and slight local differences in hybridization conditions over the array. Here, each condition and suspicions that causes the legitimacy of the approach needs to be satisfied, if particular normalization algorithm is locally connected. In other words, the components belong to all pen groups need not to be selected to state thedifferentially expressed genes, instead a sufficiently large volume of components need to be amalgamated into each pen group in order to get legitimate.

Even though normalization is capable of modifying the estimations of the $\log_2$ (ratio), yet the difference of the measured $\log_2$ (ratio) qualities may occur through stochastic procedures, which is to vary starting with one locale of an array then onto the next or between arrays. This issue can be resolved through modifying the $\log_2$ (ratio) measures that maintains the fluctuation in same level. On the off chance, a single array accompanied by specific sub gridsis considered that has been gone through neighborhood local normalization, during which a factor for measuring the majority of the estimations within that sub grid is anticipated for each sub grid. For all sub grids, a fitting scaling factor is taken as thevariance for a specific sub grid that is separated by geometric mean of the changes.On the off chance, take that each sub grid possesses M elements due to the efficient balancing of the mean of the $\log_2$ (ratio) values to be zero in each sub grid, besides the estimation of the difference in the n^{th} sub grid is given by,

$$\sigma_n^2 = \sum_{j=1}^{M}\left[\log_2(T_j)\right]^2 \quad (1)$$

Here, the summation goes through overall elements within the sub grid. In context of N_{grids} that denotes the number of sub grids in the array, the following equation expresses the appropriate scaling factor that is for the elements of the kth sub grid on the array,

$$a_k = \frac{\sigma_k^2}{\left[\prod_{n=1}^{N_{grids}} \sigma_n^2\right]^{\frac{1}{N_{grids}}}} \quad (2)$$

Subsequently, each elements inside the kth sub grid is measured through classifying it with the same value a_k that corresponds to the sub grid.

$$\log_2(T_i) = \frac{\log_2(T_i)}{a_k} \quad (3)$$

It is considered as equal as taking the a_k^{th} root of the individual intensities in the k^{th} sub grid,

$$G_i' = [G_i]^{\frac{1}{a_k}} \text{and} R_i' = [R_i]^{\frac{1}{a_k}} \quad (4)$$

Since other variance regularization factors are suggested, a same process needs to be applied for regularizing the variances from normalized arrays.

Nevertheless, the existing transformation parameters estimation approach has the possibility of inadequacy due to its imperfect management of outliers. In this work, an advanced parameter estimation model is designed with the help of information normalization approach, through which highly straightforward outlier exclusion is enabled in a statistical way. It is capable of providing efficient performance regardless of the sample size.

Besides, this approach is applicablein context of $N_{ij} = 1$, where V_{ij} is switched by V_i that is is the unbiased variance of the i[th] gene, and N_{ij} by the number of conditions. Here, a little higher *p*can be considered as optimal to apply due to the appearance of differentially expressed genes.

This computation is accomplished through an information normalization method by considering the instance of small N_{ij}. Even though the instance of small N_{ij} is generic, yet Eq. (5) triggers two issues, i.e. i) in context of small N_{ij}, the data variance converted by this approach always needs to be lesser than 1; ii) If, N_{ij}= $2\ or\ 3$, this approach cannot be applied, since the probability density function of $\hat{V}$ has a single peak at $\hat{v}$ = 0.

$$\hat{\theta} = \arg\max_{\theta} \prod_{i,j} P_{V_{ij}}(\hat{V}_{ij}; N_{ij}, 1) \quad (5)$$

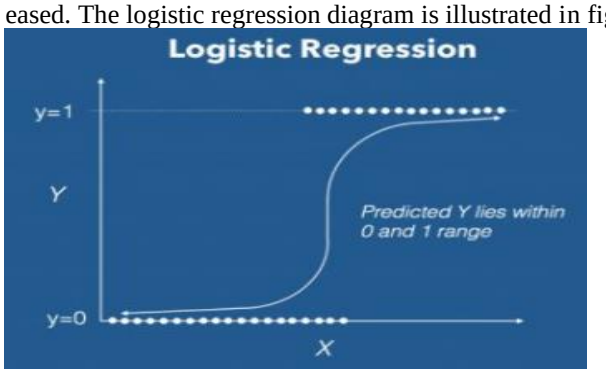
Here,$P_{V_{ij}}$ represents the probability density function of the unbiased variance of N_{ij} independent sample values from a normal distribution with variance $\sigma^2 = 1$,

The variance stabilization performance is efficiently enhanced through this IVSN model that enables maximum probability estimation, thereby optimal outcomes can be generated.

2. 3.2. Logistic Regression Clustering Algorithm for Pre-Processing Dataset

Logistic regression is known as an effective regression predictive analysis algorithm. In context of the dependent variable of a dataset is dichotomous (binary), the application of this algorithm is proved to be effective. To define the relationship amid one dependent binary variable and one/more independent variables, the description and analysis of data is applied with Logistic regression. Through executing this improved data processing task, a huge volume of useable data can be generated, besides the classification algorithm can be improved. Consequently, for the provided asthma dataset, the speed can be accelerated as well as the runtime complexity can be reduced.

In the health data of asthma patients, elimination of inappropriate data can efficiently be carried out with this algorithm that greatly helps to obtain accurate prediction (Jones, 2008); (Sánchez et al., 2009). In various fields, especially the biological sciences, the application of Logistic regression system becomes crucial. In case of the data items need to be classified into classes, the Logistic regression algorithm is applied. In general, the target variable is binary in the logistic regression, i.e. 1 or 0. In this work, it is used to represent that a patient is either of positive or negative as regards diabetes. The proposed logistic regression algorithm predominantly tends to identify the best fit, which is of diagnostically sensible, through which defining the association within our target variable and the predictor variables is eased. The logistic regression diagram is illustrated in figure 3.



Figreu 3. Logistic Regression Diagram

The following expression defines the logistic regression algorithm that bases on a linear regression model

$$y = h_\theta(x) = \theta^T x \quad (6)$$

Since Equation (6) is inadequate to predict the binary values ($y^{(i)} \varepsilon \{0, 1\}$), the Equation (7) is presented in this study that can be applied for predicting the probability, i.e. a given patient (with given attributes) belongs to the "1" (positive) class versus the probability that it belongs to the "0" (negative) class.

$$P(y = 0|x) = 1 - P(y = 1|x) = 1 - h_\theta(x) \quad (7)$$

The application of sigmoid function [Equation (14)] enables to keep the value of $\theta^T x$ that sets within the range of $[0, 1]$. In search of the value of θ, the probability $P\ (y = 1|x) = h_\theta(x)$ is obtained that is considered aslarge if x belongs to the "1" class. Whereas, If x belongs to the "0" class, the probability is small, that is to say .$P\ (y = 0|x)$is large.

$$\sigma(t) = \frac{1}{(1+e^{-t})} \quad (8)$$

In the case of the dataset with unlabeled data, the Logistic regression clustering algorithm proves its significance as it is capable of preprocessing the dataset. To support the dataset that accompanies either labeled or unlabeled data, this technique is presented in this work. Besides, it supports streaming dataset.

3. 3.3. Optimal Feature Selection using Modified Genetic algorithm (MGA)

For enhancing the feature selection process of the provided asthma dataset, this research proposes the method, namely MGA. Being a type of inductive learning strategy, the genetic algorithm can perform the feature selection in an efficient manner, through which the near-optimal solutions can be derived even in complex, and nonlinear search spaces can be obtained that accompany time-efficiency. In context of attribute selection, GA proves to be effectivefor exploring large search spaces, hence it is recognized as a stochastic general search technique. Generally, GAs are capable of performing a global search, which is lacked by many of the search algorithms and they rely on performing local/greedy search. There are three operators involved significantly in a genetic algorithm, i.e. reproduction, crossover, and mutation. Among that, the selection of good string (subset of input attributes) is performed by Reproduction; Crossover tends to create better offspring's through combining optimal strings; in mutation, optimal string generation is endeavored by locally modifying a string. The string is made up of binary bits, in which 1 is to denote attribute selection, or else 0 is to drop that attribute. The population of each generation is assessed and verified to terminate the algorithm. In case of the non-convergence with termination criterion, the population is operated towards the three GA operators, besides revalidated. Iteration of this process continues up to certain number of generation. The population in genetic algorithm is illustrated in figure 4.

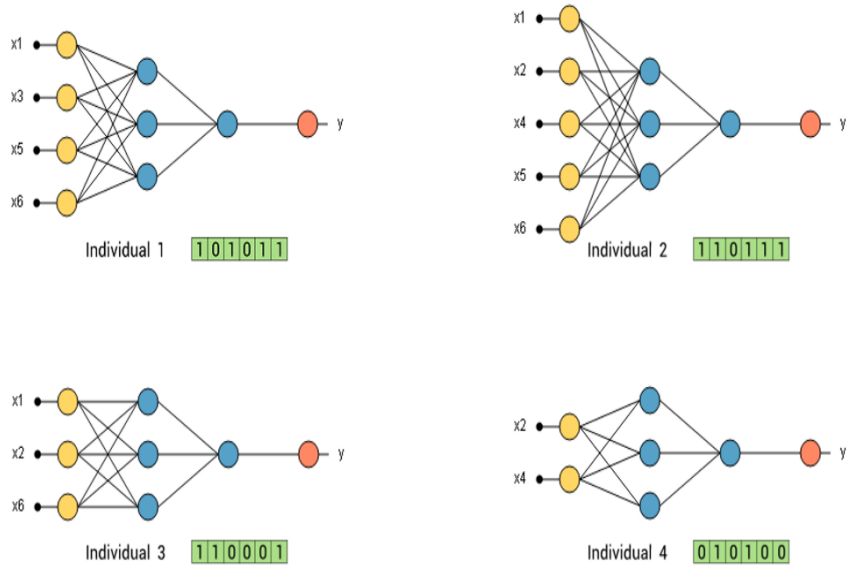


Figure 4. Illustration of the Population in Genetic Algorithm

The fitness of each individual is evaluated through GA by employing a fitness function, where the fitness signifies the quality of solution. The maximum probability is in favor of fitter chromosomes either for being saved in the next generation or for being chosen into the recombination pool with the help of tournament selection methods. The convergence inadequacy of the chromosome/fittest individual may leads to reproduction of subsequent populations for providing more alternating solutions. Being the core operators, the crossover and mutation functionstransform the chromosomes randomly, and influence their impact their fitness value. The evolution continues up to the attainment of appropriate outcomes. Genetic Algorithm is one that can effectively handle large spaces since it accompanies the features of exploitation and exploration search. As such, the chances of getting into local optimal solution is greatly reduced, when compared to other algorithms.

As a heuristic process of natural selection, GA is an inspiration of the evolution procedure in nature. The Darwin's "Survival of fittest" theory plays a major role in this algorithm that is driven by inheritance, mutation, selection, and crossover. In comparative terminology to human genetics, gene represent feature, chromosome are bit strings and allele is the feature value (Zhao, 2011). From algorithm point of view, chromosomes represent the population of individuals that are the binary strings arrangement, where each bit (gene) signifies a particular feature inside a Chromosome (bit strings). Moreover, the Objective function (fitness function) is utilized to evaluate Chromosomes, during which each chrome is being ranked in accordance with their corresponding numerical value (fitness) inside a population. In this study, f-score value is taken as the fitness values as regards optimal feature selection. Accordingly, for each feature, F-score is computed and N numbers of feature's F-Score is categorized in an descending order. Subsequently, the feature subset is generated accompanied by single/multiple features.

The extracted feature vectors are evaluated through Fisher ratio, besides the selection of optimal feature subsets takes place with the help of dynamic searching strategy-based genetic algorithm based on fitness function

maximization. During this process, a dynamic searching method is exploited through changing the chromosomes' length, and changing in real-time the range of the feature candidates. Subsequently, the optimized feature subset is identified from the extracted statistical features using the developed dynamic searching algorithm. The nature of genetic algorithm is demonstrated in figure 5.

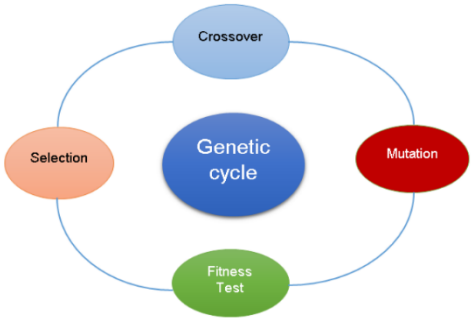


Figure 5. Nature of Genetic Algorithm

4. Process of GA:

Step 1 (Generation starts): By using population size n and number of features m, generation of a random population $(a_{11}, a_{12}, \ldots . a_{mn})$ matrix p of size $n\ x\ m$ is initialized

Step 2 (Tournament): The best-fit individuals is selected in this step, concerning reproduction. During cross over, two chromosomes (parent chromes) accompanied by highest fitness will be participated.

Step 3 (Cross Over): Analogous to biological crossover, it is the exchange of bits inside the chosen parents to produce offspring. From parent P_n, the number of bits b selected, in which parameter: $0 < k < 1$.

$$b = K * P_n \quad (9)$$

Step 4 (Mutation): It define the modification (growth) in the genome of chromosome, flipping of bit strings (genes) of chromosome

Step 5 (Fitness Evaluation): Analogous to "survival of fittest", chromosomes with a particular level of fitness will survive for next generation. Whereas, the others with poor fitness which islower than the threshold value will be rejected.

5. MGA Algorithm Pseudo Code:

Input: Asthma dataset
Objective function: F-measure value
Output: Optimal F-score value

1. Begin
2. Initialization population by selecting random individuals from the feature F
3. For the specified number of generations do
4. For the size of the population (asthma patient) do
5. Assign two individuals (accompanying equal probability) asparent1and parent2 (attributes of asthma patients)
6. Initiate crossover to produce a new individual (child).
7. Execute mutation to child.
8. Estimated1, and d2. Here,the distance within child and parent1 is denoted by d1, and the distance within child and parent2 is signified by d2.
9. Determine the fitness ofchild, parent1 and parent2, which are represented by f,f1, and f2, correspondingly
10. $if\ (d1 < d2) and\ (f > f1) then$
11. Switch parent1 with child
12. Otherwise
13. $if\ (d2 \leq\ d1) and\ (f > f2) then$
14. Swap parent2 with child.
15. Feature modified step:
16. Redo
17. Population ←feature subset F_m
18. Generation =0;
19. Loop for i from 1 to size population (do) using (9)
20. S1←selection (population, fitness)
21. S2←selection (population, fitness)
22. Child ←crossover (s1, s2) check feasibility of n element

23. Child ← mutate (child) check feasibility of n element
24. Fitness (F-score)
25. Generation = generation +1
26. Until generation<max_generation
27. M=m+1
28. End if
29. End if
30. End for
31. End for
32. Attain best individual solution as optimal F-score value

6. 3.4 Polynomial Distribution based Convolutional Neural Network (PCNN) Approach for Classification

For the classification of test data, this study presents Polynomial Distribution based Convolutional Neural Network (PCNN) that classifies the data into yes (or) no classes. By using the proposed deep learning approach, an optimized accuracy rates are obtained. In the standard CNN, an input layer, an output layer, and several hidden layers are involved. Besides, the convolutional layers, pooling layers and fully connected layers are involved in the hidden layers. Convolutional layers process the input through convolution function, through which the result is transmitted to the subsequent layer. Then, the response from an individual neuron is emulated to visual stimuli by the convolution. Figure 6 depicts the overall architecture of PCNN.

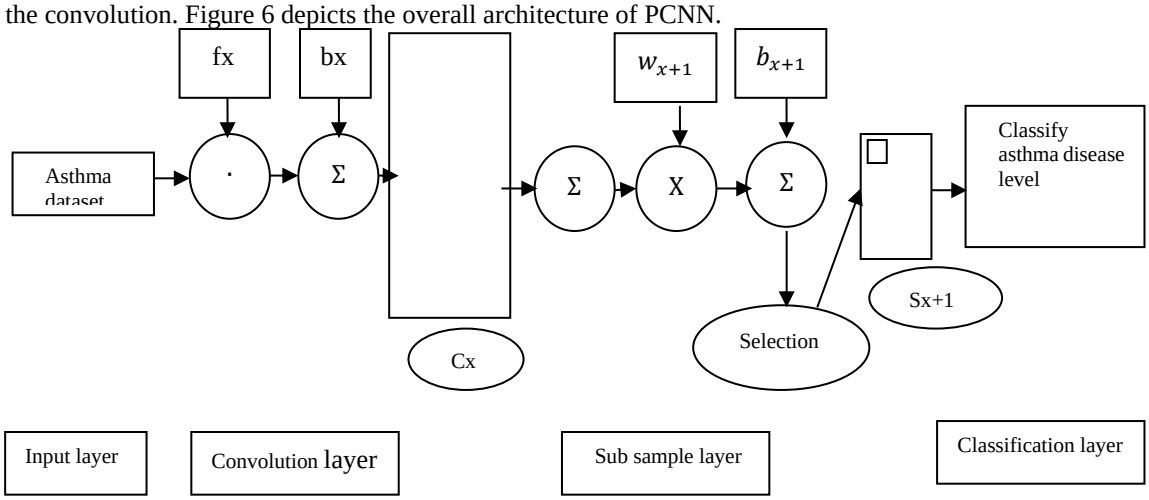


Figure 6. Architecture Diagram of PCNN

The outputs of neuron clusters are merged into a single neuron in the following layer, namely local or global pooling layers, which may have been included by convolutional networks. Whereas, the average value from every individual neurons cluster in the previous layer is utilized by mean pooling. Each neuron from one layer is connected to each neuron in another layer by using fully connected layers. Similar to the conventional multi-layer perceptron neural network, the CNN also follows the same procedure (Tripathi et al., 2017). Accordingly, there are for layers, such as input layer, convolutional layer, sub-sampling layer and classification layer involved in the proposed PCNN. Apparently, the proposed approach includes the same benefits as regards high-dimensional data analysis. In addition, the number of parameters are controlled and reduced through convolutional layers by using a parameter sharing procedure.

From training samples, asthma features are received and converted into a unified form by input layers, through which the data can appropriately be delivered to the subsequent layer. Besides, this layer signifies the initial parameters, namely scale of the local receptive fields and different filters.

The input data is progressed through convolution layer (Cx) using convolution algorithm, by which a feature map is generated which is of numerous layers that includes the estimation resultsof convolution from the previous layers. It predominantly helps in the extraction of key features and reduction of network's computational complexity.

Each convolutional layer is applied with an activation function, through which an output is mapped to a set of inputs that helps to form a non-linear network structure. The overall given feature values are assigned with initial connection weights. Subsequently, applied a new input pattern, and the estimation of output is performed through,

$$y(n) = f(\sum_{i=1}^{i=N} w_i(n)x_i(n) \quad (10)$$

$$\text{Where } f(x) = \begin{cases} +1 & if\ x \geq 0 \\ -1 & if\ x < 0 \end{cases} \quad (11)$$

In accordance with the following expression, updation of connection weights is carried out

$$w_i(n+1) = w_i(n) + \eta(d(n) - y(n)x_i(n), \quad i = 1,2,..N \quad (12)$$

In which, the gain factor is denoted by η

Subsequently, the standard deviation is applied as follows,

$$\sigma = \sqrt{\frac{1}{n} f_i (x_i - \overline{x)^2}} \quad (13)$$

These obtained features of weighted asthma dataset are fed into the proposed PCNN network, by which highly accurate classification results are derived. In addition, the primary results obtained through the analysis on same set of data is ensured by polynomial distribution function.All feature maps from the preceding convolution layer is sub-sampled in this layer. As depicted in Fig 6, $Sx + 1$is summing the informative features.

7. Algorithm 2: Steps in PCNN

1. Initialize Asthma disease dataset
2. Defineasthma patientattribute∈ Asthma dataset for each input attribute,
3. Transform the input into sub layers
4. Identifyasthma patient features using (12) and (13)
5. Choose more informative and relevant features
6. Accomplish training and testing process for the dataset provided
7. Copy predefined asthma feature label for each feature as per the input dataset

Highly accurate asthma disease results can be classified through learning the asthma symptoms of various patients.

4. Experimental Results

As an initial step, the dataset of asthma patients who belongs to various age groups, and accompanied by several symptoms of asthma is collected. In a repository, the collected information of these patients' past and their corresponding symptoms are archived. Besides, each asthma patient is gone through pulmonary function tests for determining breathing level of their lung, from which the attribute values of pulmonary function tests are collected and accumulated to the information of the respective patient. Post generation of datasets, the diagnosis can be carried out in an accurate manner based on the different possibilities of asthma symptoms. Since the collection of information is performed across different patients from various hospitals, the analysis and identification of risk factors get simplified regardless of the types. In accordance with the prediction values, the research model's prediction accuracy is matched up with the changing previous classification algorithm.

Throughout this segment, the performance of the proposed asthma diagnosis method is evaluated based on various ages and different symptoms of asthma. With regard to the prediction values, the prediction accuracy of the proposed and the existing approaches are compared in order to prove the efficiency of the proposed method. During the performance evaluation, True Positive Rate, False Positive Rate, Accuracy and F Measure are taken as the metrics.

8. True Positive Rate (TPR)

TPR defines the ratio of actual positive, which are appropriately classified as spam subjects class. Besides, it represents the sufficiently classified ratio of spam and non-spammer detection. Estimation of TPR is expressed as follows,

$$\text{True Positive Rate (TPR)} = \frac{T_p}{(T_p + F_n)} \quad (14)$$

9. False Positive Rate (FPR)

FPR is also known as false alarm ratio that represents the probability of falsely rejecting the null hypothesis for a specific test.

$$\text{False Positive Rate(FPR)} = \frac{F_p}{(F_p + T_n)} \quad (15)$$

10. Accuracy

As a significant performance metric, Accuracy refers to system's overall appropriateness, which can be estimated through the following expression,

$$\text{Accuracy} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \quad (16)$$

11. F-Measure

Through amalgamating the metrics, namely precision P and recall R, the F-measure ratio can be obtained

$$F = 2.\frac{PR}{P+R} \quad (17)$$

Generally, the classification algorithms are evaluated on the basis of F-measure value due to its standard measurement approach that summarizes both P and R.

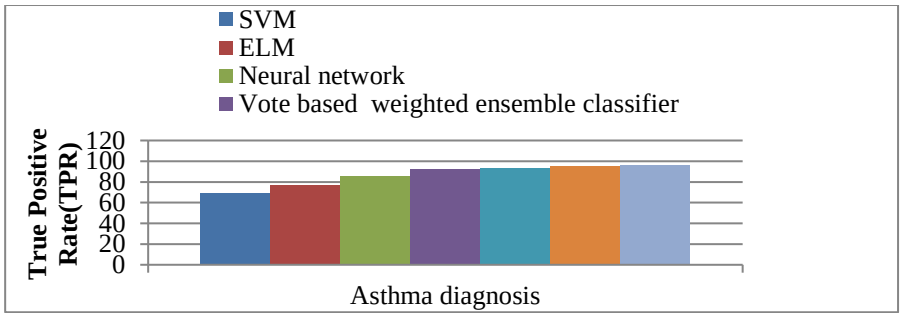


Figure 7. True Positive rate for Features

Figure 7 measures the True Positive Rates obtained by various classifiers with different features designed for classifying asthma health data collected from patients. In the figure, the executed methods lie on X-axis, and Y-axis stands for corresponding TPR values. The graphs depict the efficiency of the proposed MGA+PCNN algorithm to optimize the TRP by selecting the best feature values, as regards asthma diagnosis.

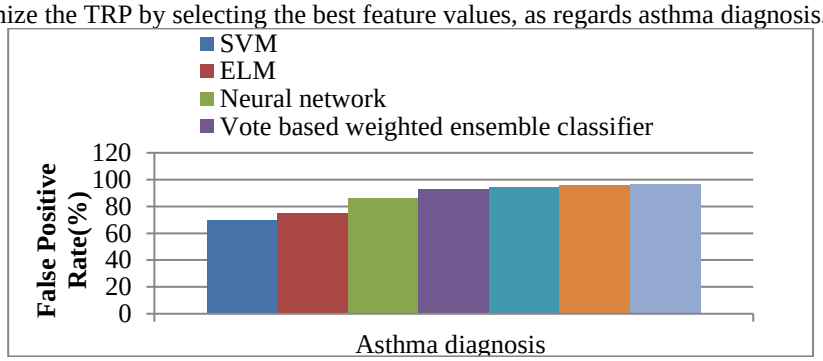


Figure 8. False Positive Rate for Features

Figure 8 assesses the False Positive Rates obtained by different classifiers with diverse features designed for classifying asthma health data collected from patients. In the figure, the executed methods lie on X-axis, and Y-axis stands for corresponding FPR values. The graphs depict the efficiency of the proposed MGA+PCNN algorithm to optimize the TRP by selecting the best feature values, in terms of asthma diagnosis.

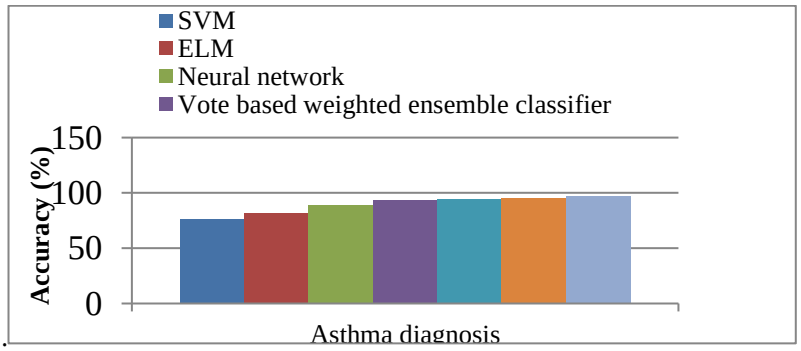


Figure 9. Classification Accuracy for Features

In figure 9, the classification accuracy of the proposed and existing techniques are compared. In the figure, the implemented methods lie on X-axis, and Y-axis stands for corresponding accuracy rates. For provided asthma dataset, the proposed MGA+PCNN algorithm proves to be optimal in terms of accuracy metric, since it provides comparatively higher accuracy than the present methods, namely SVM, ELM, ADF-LCT, neural network, Vote based weighted ensemble classifier, and IAD-HABCDE-NBC algorithm. The reason is that the proposed method improves the classification accuracy by selecting the best features, concerning asthma diagnosis.

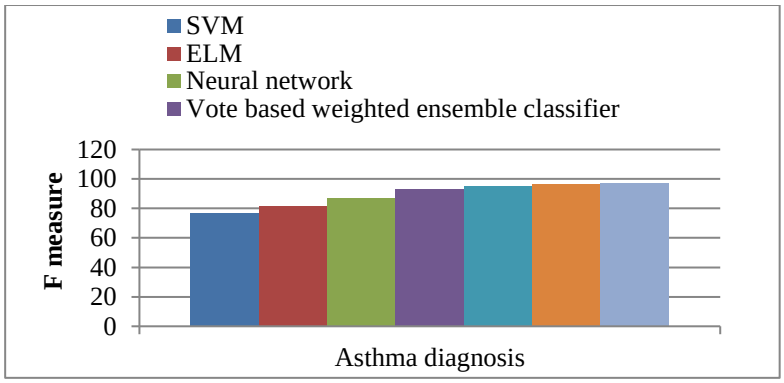


Figure10. F-Measure Accuracy for Features

In figure 10, the F-measure values of the proposed and existing techniques are compared. In the figure, the methods lie on X-axis, and Y-axis represents corresponding F-measure rates. For provided asthma dataset, the proposed MGA+PCNN algorithm procures optimal F-measure rate which is superior to the present methods, namely SVM, ELM, ADF-LCT, neural network, Vote based weighted ensemble classifier, and IAD-HABCDE-NBC algorithm. The reason is that the proposed method improves the classification accuracy by selecting the best features, as regards asthma diagnosis.

5. Conclusion

In recent decades, the occurrence of Asthma turned out to be a common disease across the people around the world. Since the remedy for asthma is not identified yet, it is crucial to follow the appropriate steps for avoiding the vulnerable impacts of the disease. Besides, identification of disease stage is highly significant,through which the treatment can be prioritized for the patientwho suffer from higher level regardless of delay. Moreover, based on the stage of disease, the patients can be provided with suitable treatments. For enhancing the classification accuracy in the results of asthma diagnosis, this research study proposes the method, namely MGA+PCNN that includes normalization, pre-processing, feature selection and classification as a primary steps. At first, the normalization dataset is presented with IVSN technique for enhancing the efficiency of variance stabilization. Through that, optimal results can be generated as it enables estimating the highest probability. In addition, logistic regression clustering algorithm is applied to carry out the pre-processing task, which enhances the accuracy of classification. By using the optimal objective function values, chosen by MGA algorithm chooses these features. Consequently, in the asthma dataset, the redundancy and relevance features are optimized. Besides, PCNN algorithm is involved further for executing the classification process, through which highly accurate results of asthma disease diagnosis are obtained. Empirical findings depict the efficiency of proposed MGA+PCNN algorithm to deliver great FPR, TPR, F-score and accuracy ratios, which is superior to the present methodologies.

References

1. Vargas, P. A., Simpson, P. M., Bushmiaer, M., Goel, R., Jones, C. A., Magee, J. S., & Jones, S. M. (2006). Symptom profile and asthma control in school-aged children. *Annals of Allergy, Asthma & Immunology*, *96*(6), 787-793.
2. Wang, L. Y., Zhong, Y., & Wheeler, L. (2005). Peer reviewed: Direct and indirect costs of asthma in school-age children. *Preventing chronic disease*, *2*(1).
3. Hafkamp-de Groen, E., Lingsma, H. F., Caudri, D., Levie, D., Wijga, A., Koppelman, G. H., & Raat, H. (2013). Predicting asthma in preschool children with asthma-like symptoms: validating and updating the PIAMA risk score. *Journal of allergy and clinical immunology*, *132*(6), 1303-1310.
4. Guilbert, T. W., Morgan, W. J., Krawiec, M., Lemanske Jr, R. F., Sorkness, C., Szefler, S. J., ... & Martinez, F. D. (2004). The Prevention of Early Asthma in Kids study: design, rationale and methods for the Childhood Asthma Research and Education network. *Controlled clinical trials*, *25*(3), 286-310.
5. Martinez, F. D. (2007). Genes, environments, development and asthma: a reappraisal. *European Respiratory Journal*, *29*(1), 179-184.
6. Kumar, V., Abbas, A. K., Fausto, N., & Aster, J. C. (2010). Robbins and cotran pathologic basis of disease. 8th. *Philadelphia: Ed. Saunders Elsevier*. pp. 1-12
7. Vial Dupuy, A., Amat, F., Pereira, B., Labbe, A., & Just, J. (2011). A simple tool to identify infants at high risk of mild to severe childhood asthma: the persistent asthma predictive score. *Journal of Asthma*, *48*(10), 1015-1021.
8. Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining John Wiley & Sons*, pp. 1307-1309
9. Li, X., Ng, S. K., & Wang, J. T. (Eds.). (2013). *Biological data mining and its applications in healthcare World scientific*, pp. 3-417.

10. Robbie-Ryan, M., & Brown, M. (2002). The role of mast cells in allergy and autoimmunity. *Current opinion in immunology*, *14*(6), 728-733.
11. Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, pp. *1-7*.
12. Princy, J. C., & Sivaranjani, K. (2016). Asthma Prediction Using Classification Technique. *IJCTA*, *9*, 27.vol. 9, no. 27, pp. 415-421
13. Várkonyi, D. T., & Buza, K. (2019). Extreme Learning Machines with Regularization for the Classification of Gene Expression Data. In *ITAT*, pp. 99-103.
14. Sachnev, V., Saraswathi, S., Niaz, R., Kloczkowski, A., & Suresh, S. (2015). Multi-class BCGA-ELM based classifier that identifies biomarkers associated with hallmarks of cancer. *BMC bioinformatics*, *16*(1), 1-12.
15. Arslan, S., & Ozturk, C. (2019). Feature Selection for Classification with Artificial Bee Colony Programming. In *Swarm Intelligence-Recent Advances, New Perspectives and Applications*. *IntechOpen*, pp. 1-101
16. Strumberger, I., Tuba, E., Bacanin, N., Jovanovic, R., & Tuba, M. (2019). Convolutional neural network architecture design by the tree growth algorithm framework. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8.
17. Liu, F., & Yang, L. (2017). A novel cell detection method using deep convolutional neural network and maximum-weight independent set. In *Deep Learning and Convolutional Neural Networks for Medical Image Computing*, pp. 63-72.
18. Jones, M. A. (2008). Asthma self-management patient education. *Respiratory care*, *53*(6), 778-786.
19. Sánchez, C. I., Hornero, R., Mayo, A., & García, M. (2009). Mixture model-based clustering and logistic regression for automatic detection of microaneurysms in retinal images. *Medical Imaging 2009: Computer-Aided Diagnosis International Society for Optics and Photonics*, pp. 1-8.
20. Zhao, M., Fu, C., Ji, L., Tang, K., & Zhou, M. (2011). Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes. *Expert Systems with Applications*, *38*(5), 5197-5204.
21. Tripathi, S., Acharya, S., Sharma, R. D., Mittal, S., & Bhattacharya, S. (2017). Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 4746-4752.