# Mining of Network Data for Intrusion Detection using Multi-Dimensional Hierarchical K Means Clustering employed with Hybrid ABC-DT

J. Josemila Baby [1]

[1]Department of Computer Applications,
Noorul Islam Centre for Higher Education,
Noorul Islam University, Kumaracoil-629180, India

J. R. Jeba[2]

[2]Department of Computer Applications,
Noorul Islam Centre for Higher Education,
Noorul Islam University, Kumaracoil-629180, India

**Abstract - The interest of Content centric network (CCN) increase tremendously because of its application as a future internet. Since the challenges in CCN also increases with the security and privacy attacks present in the network. An efficient and viable security instrument is required to verify substance and guard against obscure and new types of assaults and inconsistencies. As a rule, clustering calculations would fit the necessities for building a decent Intrusion recognition framework. The customary calculations experience the nearby combination and affectability to determination of the cluster centroids. In this article, we present a narrative multi-dimensional Hierarchical k means-clustering algorithm for Intrusion recognition system. Initially the clustering algorithm is projected to form a numeral of clusters in the CCN and then the optimal clusters are selected by the utilization of Cuckoo search optimization algorithm (CSOA). Finally, we employ an Artificial Bee colony based Decision Tree classifier in order to categorize the customary and anomalous cases present in the network by means of the extract features. The anticipated technique will be implemented on MATLAB working platform and tested on widely used KDD CUP 99 dataset. The consequences of the above implementation will be compared through existing methods. Trial results exhibit that the proposed calculation can accomplish to the ideal number of clusters, well-isolated groups, just as increment the high discovery rate and abatement the bogus positive rate simultaneously when contrasted with some other surely understood clustering calculations.**

**Keywords:** Data mining; Intrusion detection; Hierarchical K means clustering; Hybrid Artificial Bee Colony based Decision Tree (ABC-DT) Classifier.

## 1. Introduction

The concealed vulnerabilities contained in programming applications makes the Intrusion discovery framework as a basic part in light of the disappointment of conventional firewall, get to control and encryption strategies to distinguish the interruption in the systems [1]. Thusly, interruption identification framework (IDS) is required as an extra divider for ensuring frameworks regardless of the aversion methods [2]. Interruption discovery was presented by James Anderson in 1980, from that point forward, interruption location assumes significant job alongside firewall [3]. Interruption identification arrange (IDS) have been created because of the emotional development of assaults by the programmers [4]. The hacking programming additionally created in parallel to assault the IDS. So, there are a few challenges to give defend component to information [5]. Yet specialists are making a decent attempt to create proficient IDS [6].

Firewall is unique in relation to IDS because through firewall we can't ready to anticipate the assault yet if there should be an occurrence of IDS the overseer can distinguish the assault and furthermore we take preventive measures to keep away from those assaults to be happen further [7]. An Intrusion is a technique for involving privacy, uprightness, adaptability and accessibility of system assets [8]. It screens and dissects the client and system traffic, confirms framework arrangements and vulnerabilities and cautions the executive through alerts [9]. The possibility of the Intrusion location framework (IDS) is to keep the PC framework from assault [10]. The IDS is the most fundamental piece of the security foundation for the systems associated with the web in light of the fact that different approaches to bargain the solidness and security of system [11].

IDS incorporate the profiles for client conduct over system and as indicated by those examples it distinguishes the interlopers and passes reaction [12]. IDS are consistently to manage immense measure of information causing moderate preparing and testing process and low identification rate which makes the element choice procedure as one of the key subjects in IDS [13]. IDS incorporate two kinds of discovery approaches specifically abuse and abnormality recognition approaches [14]. Information mining is to extricate important data from huge database or information stockroom by the client [15]. Data can be spoken to as idea, rules law and model [16]. A thought of behind utilizing information mining is to help the leader to separate between information as valuable or insignificant.

Mining procedures are arrangement, Regression, and Deviation discovery are utilized to anticipate obscure or future qualities from another factor [17]. The host put together IDS investigates the exercises with respect to the single PC or host. The principle weakness of the abuse location (signature discovery) technique is that it can't distinguish novel attacks and variety of known attacks [18]. To maintain a strategic distance from these shortcomings we go for inconsistency based location strategies [19]. With this procedure, known and newattacks can be recognized. The problem is that it will produce even more false alerts [20].

With the intention of Intrusion detection in CCN, the involvement of this research exertion is (i) to select an optimal cluster by Cuckoo Search Optimization (CSO) Algorithm, which is formed by means of Hierarchical K means clustering. (ii) To detect intrusion by a narrative Artificial Bee Colony Optimization based Decision Tree (ABC-DT) Classifier. The remnants of this dissertation are prearranged as follows: A summary of some of the literature works related to intrusion detection is carried out in section 2. The design of proposed intrusion detection mechanism is presented in section 3. The simulation results of our proposed work with the performance analysis is exposed in segment4 followed by the finale in section.

## 2. Related Work

*Some recent works associated to anomaly based Intrusion recognition in Content Centric Network (CCN)are listed below.*

D'angelo et al. [21] have displayed another administered AI way to deal with irregularity discovery, whose objective was understanding the elements and practices describing system traffic to create a lot of decides and criteria that can be utilized to adequately segregate atypical occasions in the ordinary traffic stream. Such approach coupled the capacity of deriving unbending decisional structures, spoke to as Boolean equations, from deficient example perceptions, with the adaptability presented by a fluffy based vulnerability the executives procedure. That enabled the identification motor to effortlessly adjust to the altogether different sort of wonders that can be experienced on a genuine system.

Elhag, S et al. [22] have utilized Genetic Fuzzy Systems inside a pair wisewisdom structure for the improvement of such a framework. The benefits of utilizing this methodology aredual: foremost, the utilization of fluffy set, and particularly semantic names, empowered a smoother fringe among the ideas, and permitted a privileged interpretability of the standard locate. Subsequent, the gap and-overcome learning plan, in which they differentiate all conceivable brace of module by way of points, improved the accuracy for the uncommon assault occasions, as it acquires a superior detachability among a "typical movement" and the diverse assault type.

Costa, K. An et al. [23] have projected a nature-enlivened system to deal with gauge the likelihood thickness work utilized for information clustering dependent on the ideal way woodland calculation (OPFC). OPFC deciphers a dataset as a chart, whose nodes are the examples and each example was associated with its k-closest neighbors in each component space (a knn diagram). The nodes of the chart are weighted by their pdf values and the pdf was registered dependent on the separations among the examples and their k-closest neighbors. When the knn chart was characterized, OPFC discovers one example (root) at every limit of the pdf and engenders one ideal way tree (cluster) beginning both root to the rest of the examples of its arch. Grouping viability will rely upon the pdf assessment, and the wished-for methodology effectively registers the finest estimation of k for a certain relevance.

Elhadi M. Shakshuki et al. [24] executed an interruption identification construction named Enhanced Adaptive ACKnowledgment (EAACK) exceptionally anticipated for MANETs. The methodology EAACK was proposed to tackle the inadequacies of Watchdog plot bogus bad conduct, constrained transmission power, and collector impact. Besides, they stretched out their examination to receive a computerized mark conspire throughout the container diffusion progression. As in all affirmation based IDSs, it was indispensable to guarantee the honesty and genuineness of all affirmation clusters. To recognize diverse bundle types in various plans, they incorporated a 2-b packetfooter in EAACK.

Rana Aamir Raza Ashfaq et al. [25] have proposed a narrative fluffiness based semi-directed wisdomadvance by using unlabeled examples helped through regulated wisdom calculation to recover the classifier's presentation for the IDSs. A solitary concealed deposit feed-forward neural classification (SLFN) was prepared to yield a fluffy participation vector, and the example order (squat, average, and high fluffiness classifications) on unlabeled examples was perform utilizing the fluffy amount. The classifier was retrained in the wake of consolidating every classification independently into the first preparing set. The trial results utilizing the method of interruption recognition on the NSL-KDD dataset indicated that unlabeled examples having a place with low and high fluffiness clusters make significant commitments to recover the classifier's exhibition contrasted with accessible classifiers like guileless Bayes, bolster vector machine, irregular woods, and so forth.

## 3. Optimized Cluster Selection And Intrusion Detection with Hybrid ABC-DT

There are different methodologies being used in interruption recognitions, however sadly any of the frameworks so far isn't totally free from abandons. With the goal that information mining method is utilized to locate the intriguing principles from an enormous database relying on the client characterized backing and certainty. A thought of behind utilizing information mining is to help the chief to separate between information as valuable or unessential. Mining systems are grouping, Regression, and Deviation location are utilized to anticipate obscure or future qualities from another factor. Albeit numerous sorts of grouping techniques, for example, Fuzzy C-Means (FCM), K-implies, are broadly utilized in interruption discovery, scarcely any clustering calculations ensure a worldwide ideal arrangement.Based on this intention we developed a new anamoly intrusion detection mechanism employed with multi-dimensional hierarchical k-means algorithm in this research work in instruct to conquer all the beyond issues. The process of proposed Intrusion Detection System (IDS) is exposed in figure 1.

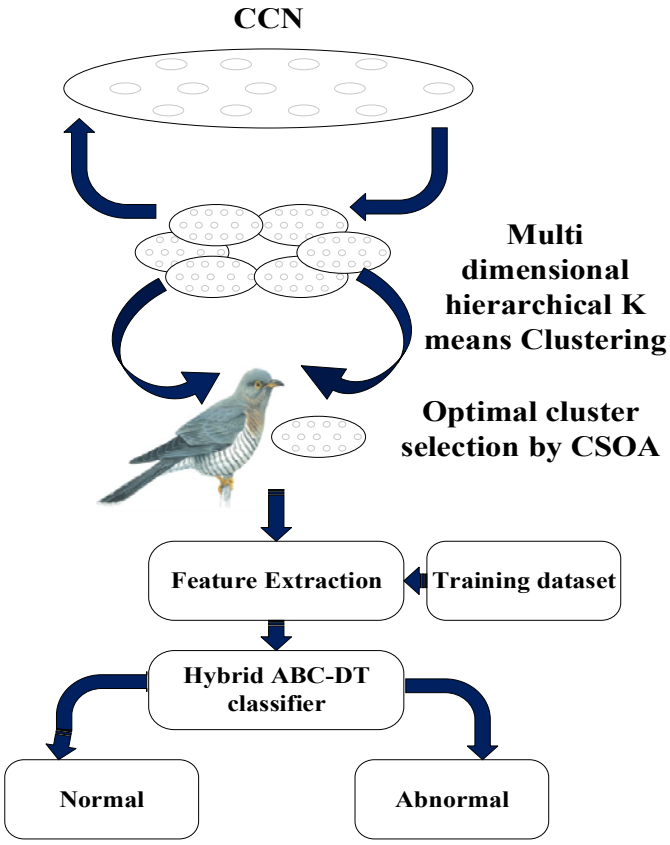


Fig. 1: Process of proposed Intrusion Detection System

Our proposed work starts with the clustering of Content centric networks (CCN) by means of the proposed novel multi-dimensional hierarchical k-means algorithm. The proposed clustering algorithm initially finds the global clusters based on the distance metric and then form an inner clusters based on the similarity metric. And then to find the global optimal clusters instead of local optimal clusters, we propose to use Cuckoo Search Optimization algorithm (CSOA). From the trained datasetthe normal behaviors of a user or a program will be compared with the testing dataset based on feature extraction and similarity matching techniques. Finally, the utilization of the new hybrid Artificial Bee Colony based Decision Tree Classifier (ABC-DT) results the intrusion present in the network.

*3.1 System Model*

It is essential to describe the system model with the appropriate specifications, assumptions, requirements and the threat models present in the network toestimate and prove the significance of our proposed IDS utilized in CCN which is presented in the following sections.

*3.2 Network assumptions*

All interchanges in CCNs are done utilizing two kinds of clusters: Attention and Content. The principle thought in the CCN is that, an attention demand for a substance object is directedin the direction of the area of the beginning substance where it has been distributed. Any switch or center node in transit checks its reserve for coordinating duplicates of the mentioned substance. On the off chance that a stored duplicate of any bit of Interest demand is discovered, it is come back to the requester along the way the solicitation originated from. In transit, back, all the center nodes hoard a duplicate of substance in their reserves to reply to likely identicalconcentration demands commencing consequent clients. Each CCN switch keeps up three significant information structures:

(1) Pending Interest Table (PIT): it comprises all un satisfied Interests that transmits upstream in the direction of potential information sources. Each PIT section holds one or various approaching and active substantial boundaries with correlated Interest packets.

(2) Forwarding Interest Base (FIB): advancing well being to solitary or numerous physical system boundaries dependent on the sending methodologies.

(3) Content Store (CS): incidentally cradles information packet for information recovery efficiency.

*3.3 Attack Model*

In this exploration work DoS assaults in CCNs are measured. There are innovative assault open doors in the types of DoS assaults to compose either contented inaccessible or refuse assistance to clients:

(1) To construct contented inaccessible: a basis can be disturbed by distribution enormous quantities of new and unmistakable Interests or an aggressor can decay the regular store efficiency by over-burdening the reserve when a store get a legitimate traffic. At the point when assailants gain high access power in a switch, they can cause disturbance in directing by to don't forwarding asks for or implement getting out of hand in PIT switches to counteract content recovery.

(2) To help counterfeit reactions: an assailant can cause switches to accept a substantial substance is invalid and answer a "not legitimate" reaction, intentionally. A substance can likewise be caricature by infusing counterfeit reactions that are not marked or marked with an off-base key, trusting that the client acknowledges the reaction in source. An old substance marked with the correct key can be likewise supplanted with the first one, or an assailant might acquire elevated contact to the source's marking explanation to indication substance among the right key.

*3.4 Multidimensional Hierarchical K means Clustering*

Initially the data's in the CCN needs to be clustered in order to make the overall intrusion detection process simple and effective. Information clustering is an information investigation procedure that endeavors to discover gatherings of information dependent on comparable qualities to isolate information objects into significant and sensible gatherings. All together for the estimation of similitude's between information objects, separation metric assumes a significant job. In this exploration work we propose to utilize a novel multidimensional various leveled K means Clustering with the end goal of information clustering in CCN. The proposed clustering algorithm results in number of outer clusters which are formed based on the distance measures. In addition that, it also results in number of inner clusters based on the similarity measure which are more relevant to each other. The formation of inner and outer clusters in the CCN, results the name of K means clustering algorithm as multidimensional hierarchical K means clustering algorithm in our proposed work.

A better result can be achieved by means of the proposed clustering algorithm. Assume two objects represented as $x_i$ , $y_i$ the Euclidean distance between the objects can be measured by:

$$ED(x,y) = \left( \sum_{i=1}^{n} (x_i - y_i)^2 \right)^{1/2} \quad (1)$$

To decide the right and the ideal number of clusters, we should pick the approval criteria which attempt to limit the Mean Square Error (MSE) between information vectors and their group centroid to check the grouping goodness. Be that as it may, MSE isn't sufficient and reasonable measurement for deciding the quantity of the clusters, since it diminishes as the quantity of group increments. Truth be told, the ideal MSE would be the quantity of cluster that equivalents informational index focuses, and the MSE is equivalent to zero. The MSE of the information focuses are determined by:

$$MSE = \frac{1}{N}\sum_{i=1}^{n} d(x_i, k_i)^2 \quad (2)$$

The inter-cluster distance is the distance between the centroids of any two clusters. The new cluster centers for the inter clusters are calculated by:

$$k_i = \frac{1}{p}\sum_{j=1}^{p} m(s_i, j) \quad (3)$$

Where $m(s_i, j)$ is the centroid of m-thcluster. The similarity between the inter clusters can be simply premeditated by the MSE calculated by equation (2) which can be mathematically expressed as:

$$S(k_i, k_j) = \begin{Bmatrix} [d(k_{i,} k_j) \times MSE_i] + \\ [d(k_i, k_j) \times MSE_j] \end{Bmatrix} / d(k_{i,} k_j)^2 \quad (4)$$

where $d(k_{i,} k_j)$ is the detachment between center of the cluster i and cluster j. Little estimation of closeness signifies that the clusters are isolated and a huge worth indicates that the clusters are near one another. The process of clustering in CCN projected in this research employment can be simply explained by the subsequent steps.

*Initialize*
*The amount of substance, utmostquantity of iterations.*
*The number of clusters randomly.*
*Begin*
*Allocateeveryarticle to the cluster with the closest centroid.*
*For (n<max iteration)*
*do*
*Calculate the Euclidian distance between each object which measures the base separation between information objects and each cluster centroid.*
*Calculate the MSE using equation (2).*
*Re-estimate the cluster centroid vector, using equation (3).*
*Calculate the similarity between the objects using equation (4).*
*Recalculate the centroid vectors for the similar objects and update the cluster centers for each iteration.*
*Stop if (n=max iteration)*
*End.*

Algorithm 1: Multidimensional Hierarchical K means Clustering

The number of clusters resulted from the proposed clustering algorithm needs to be optimized in organize to diminish the computational instance and complexity. So that in our proposed work the clusters present in the network with the chance of intrusion is optimally selected by Cuckoo Search Optimization Algorithm (CSOA).

*3.5 Optimal Cluster Selection by CSOA*

CSOA is a Meta heuristic calculation which is motivated by the reproducing conduct of the cuckoo flying creature. For CSOA can be just depicted by the accompanying romanticized convention:

1) Every cuckoo lays each egg in haphazardly picked home;

2) The best homes will continue to the following ages;

3) The quantity of accessible best homes is predetermined, as well as the egg laid by a cuckoo is found by the probability $p_a \in [0, 1]$.

For this circumstance, the congregation winged animal preserve moreover dispose of the egg or leave the nest, and create a original nest. Intended for straightforwardness, this preceding supposition can be approximated by the part father of the n homes are replace by innovative homes (with original self-assertive game plans). For a lift issue, the quality or wellbeing of an answer can basically be comparative with the estimation of the objective work. Various sorts of wellbeing can be portrayed similarly to the health work in inherited counts. For ease, we can use the going with essential depictions that each egg in a home addresses an answer, and a cuckoo egg address another game plan, the fact of the matter is to use the new and possibly enhanced courses of action to

supersede a not extraordinary game plan in the homes. Clearly, this figuring can relate to the more entrapped circumstance where each home has different eggs addressing a ton of plans.

In this research effort, we will exploit the simplest approach where both nest has merely a single egg. The cluster with minimum MSE value is optimally selected by the proposed CSOA which is mathematically expressed as:

$$F(x) = \min[MSE] = \min[\frac{1}{N}\sum_{i=1}^{n} d(x_i, k_i)^2] \quad (5)$$

The optimal cluster selection algorithm using CSOA is summarized as follows.

*Initialize*
*Total number of population, number of host nests, maximum number of iterations.*
*Begin*
*For (t< max iteration)*
*Randomly select a cuckoo by levy flights.*
*Compute the fitness function using (5).*
*Casuallychoose a nest.*
*Find the best solution.*
*If (current best>old best).*
*Update the new best solution.*
*End*
*Built new nests and a fraction of worse nests are abandoned.*
*Remain the finest solutions*
*level the solutions and locate the modern best.*
*End for*
*Return best solutions*
*End.*

Algorithm 2: Optimal Cluster Selection by CSOA

When engendering new solutions $x_i^{(t+1)}$ for, say, a cuckoo *i*, a levy departure is achieved. Which is expressed as:

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus Levy(\lambda) \quad (6)$$

Where $\alpha > 0$ is the progression size which ought to be identified with the sizes of the issue of interests. As a rule, we can utilize $\alpha > 1$. The above condition is basically the stochastic condition for irregular walk. The Levy flight basically gives an arbitrary amble whereas the irregular advance span is strained commencing a toll dissemination of:

$$Levy \sim U = t^{-\lambda}, (1 < \lambda \leq 3), \quad (7)$$

Which has a vast change with a limitless mean. Here the means basically structure an irregular amble progression with a power-law step-length dissemination with an overwhelming tail. A portion of the original arrangements ought to be formed by Levy stroll approximately the best arrangement acquired up until now, this will quicken the neighborhood search. Nonetheless, a substantial portion of the new arrangements ought to be created by a wide margin pasture randomization and whose areas ought to be far adequate commencing the present finest arrangement; this will ensure the framework won't be caught in a neighborhood ideal.

*3.6 Feature Extraction*

From the ideally chosen clusters the interruption present in the system is distinguished dependent on the oddities. We utilized the basic highlights (characteristics) that are extricated from the header's territory of the chose system clusters. These natural highlights are accessible in numerous systems, for instance, the term (length of the association), source have, goal have, source interface, and goal interface. We likewise utilized three highlights interim:

- Number of clusters sent for 2 seconds to the given interface,
- Number of bytes sent for 2 seconds to the specified edge,
- Number of various source-goal sets coordinating the given hostname-interface for 2 seconds.

The quantity of clusters and bytes permits to recognize abnormalities in network, and the third highlights permits to identify the system and the edge examines just as the circulated assaults, which both outcome in an expanded number of source-goal sets. The separated highlights are additionally used by a productive classifier so as to distinguish Intrusion present in the system.

*3.7 Intrusion Detection by Hybrid ABC-DT*

The discovery of interruption in the CCN relies upon the disconnected preparing gave to the classifier. The highlights of the interfered systems are removed from a benchmark dataset and it was prepared to the proposed classifier. In this exploration work we propose to utilize a straightforward Decision Tree (DT) Classifier so as to recognize Intrusion. In any case, the issue emerges with the essential Intrusion identification component, for example,

1. The Intrusion location issue includes numerous numeric traits in gathered review information and different inferred factual estimations. Building models straightforwardly on numeric information causes high discovery blunders.

2. The recognition precision is low, in light of the fact that the limit between the typical and unusual isn't well defined which prompts increment in bogus caution rate.

So as to beat such issues and to improve the discovery precision by include choice procedure we propose an Artificial Bee state Optimization calculation with the combination of DT classifier.

ABC Algorithm is awakened from the sharp sustenance looking through conduct of bumble bee appalling little animals. Bumble bee swarm is solitary of the majority competent swarms exist in nature; which looks for after all out canny strategy, while looking through the sustenance. The bumble bee swarm has different character resembling honey bees can pass on the data, can review the earth, can store and share the data and take choices subject to that. As showed by changes in the earth, the swarm strengthens itself, doles out the errands viably and moves auxiliary by social learning and training. The pursuit strategy for ABC looks for after three essential advances:

1. Send the utilized honey bees to a sustenance resource and figure the nectar eminence;

2. Onlooker honey bees decide on the sustenance sources in the wake of get-together data from utilized honey bees and deciding the nectar eminence;

3. Conclude the scout honey bees and utilize them against conceivable sustenance source.

The district of the sustenance source is discretionarily picked by the honey bees at the concealed juncture and their nectar characteristics are assessed. The utilized honey bees by then offer the nectar data of the sources with the spectator honey bees holding up at the social event area inside the hive. Resulting to sharing this data, each utilized honey bee comes back to the sustenance source checkered throughout the past cycle, as the district of the sustenance source had been studied and in this manner picks new sustenance source utilizing its watched data in the zone of the nearby sustenance resource. At the preceding stage, an observer honey bee utilizes the data recovered commencing the utilized honey bees at the social occasion zone to pick a superior than normal sustenance resource.

*3.7.1 Initialization:*

The ABC count has three parameters: the amount of sustenance masses, the amount of assessment subsequent to which a sustenance resource is honored to get be relinquished and the most extraordinary quantity of sequence. The amount of sustenance sources is proportionate to the business honey bees or passerby honey bees. From the outset, it mulls over a similarly overseen swarm of sustenance sources (SN), wherever both sustenance source xi (I = 1, 2 ...SN) is a D-dimensional vector. Every nourishment resource is produced utilizing subsequent technique:

$$x_{ij} = x_{\min j} + rand[0,1](x_{\max j} - x_{\min j}) \quad (8)$$

Some where r and [0,1] is a capacity that creates an equitably circulated irregular quantity in the range[0,1].

*3.7.2 Employed Bee:*

This stage apprises the present arrangement dependent on the data of distinct encounters and the wellness estimation of the recently discovered arrangement. New nourishment source with sophisticated wellness esteem supplant the current one. The location apprises condition for j-th measurement of i-th up-and-comer during this stage is demonstrated as follow:

$$V_{ij} = x_{ij} + \varphi_{ij}(x_{ij} - x_{kj}) \quad (9)$$

Where $\varphi_{ij}(x_{ij} - x_{kj})$ is the step size, k ∈ {1, 2, ..., SN}, j ∈ {1, 2, ...,D} are two casually chosen indices. k ≠i ensure that step size has some symptomatic enhancement.

*3.7.3 Onlooker Bee:*

The amount of sustenance hotspots for bystander bumble bee is identical as the used. Throughout this stage, all used bumble bee allocate health in sequence of new sustenance source with onlooker bumble bees. Bystander bumble bees find out the decision prospect of each sustenance source made by the used bumble bee. The finest sustenance resource is picked by the bystander. There are quantities of technique for figuring of prospect, yet it ought to consolidate wellbeing. Prospect of each sustenance source is picked by means of its health as seek after:

$$P_i = \frac{fit_i}{\sum_{i=1}^{SN} fit_i} \tag{10}$$

*3.7.4 Scout Bee Phase:*

On the off chance that the domain of a sustenance source isn't strengthened for a predefined number of cycles, by then the sustenance source is accepted to be expelled and scout honey bees sort out is instated. During this stage the honey bee related with the rejected sustenance source changed over into scout honey bee and the sustenance source is dislodged by the discretionarily picked sustenance source inside the intrigue space. In ABC, the predefined number of cycles is a critical control parameter which is called limit for dismissal. Before long the scout honey bees supplant the gave up sustenance source with new one utilizing following condition.

$$x_{ij} = x_{\min j} + rand[0,1](x_{\max j} - x_{\min j}) \quad \forall j = 1,2,...D \tag{11}$$

Thinking about the above depiction, certainly in ABC explore process there are three significant organize parameters: the measure of sustenance sources SN, the most outrageous and the best integer of cycles. The calculation of the ABC is spoken to as looks for after:

*Introduce all parameters;*

*Rehash while Termination criteria isn't meet*

*Employer bee stage for processing new nourishment sources.*

*Onlooker bees stage for refreshing area the nourishment sources dependent on their measure of nectar.*

*Scout bee stage for looking through new nourishment sources instead of dismissed nourishment sources.*

*Remember the best nourishment source recognized up until now.*

*End of while Output:*

*The best arrangement recognized up until now.*

Algorithm 3: ABC Optimization

The ideally chosen highlights are given as contribution to the DT classifier so as to recognize interruption. Given a preparation dataset, D = {x1,x2,...,xn}, each preparation example is spoken to as xi = {xi1,xi2,...,xih} and D contains the accompanying qualities {A1,A2,...,An}. Each property, Ai, contains the accompanying characteristic qualities {Ai1,Ai2,...,Aih}. The preparation information likewise have a place with a lot of classes C = {C1,C2,...,Cm}. A choice tree is a characterization tree related with D that has the accompanying property:

- every inner nodule named with a characteristic, Ai,
- Each circular segment named with a predicate that can be practical to the trait related with the parent,
- Each leaf nodule named with a category, Ci.

When the tree is constructed, it is utilized to arrange both test cases, xi D. The outcome is a characterization for that example, xi. There are two essential strides for the advancement of a DT base relevance:

- Construction the DT from a preparation dataset,
- apply the DT to a test dataset, D.

For the readiness dataset, D, we at primarypertain a crucial NB classifier to mastermind every planning event, xi D. We process the preceding probability, P(Ci), for each class, Ci D and the class unforeseen probability, P(Aij|Ci), for every property estimation (paying little respect to whether it is numeric) in D. By then we bunch every arrangement event, xi D, using these probability. The class, Ci, with the most raised back probability, P(Ci|xi), is picked as the last course of action for the event, xi. By then we eliminate all the misclassified arranging cases from the dataset D. In our assessments, these misclassified cases will when all is said in done be the hazardous preparing models. For instance, a piece of these models either contain conflicting attributes, or pass on important highlights. Acknowledge there is an accessibility dataset with two classes.

We figure the earlier and class unexpected probabilities utilizing this model arranging dataset. By then we figure the P(Class|D) for each model dependent on these probabilities. We have discovered two or three occasions where the probabilities chose utilizing the NB classifier show that they have a spot with "Class = yes". Regardless in the arranging dataset they are separate as "Class = no". Obviously, there is some uproar inside these information, which prompts conflicting outcomes in association with the essential engravings. In that capacity these misclassified occasions are viewed as troublesome models.

*Input:D = {x1,x2,...,xn}/Training dataset, D, which contains a lot of preparing cases and their related class names.*
*Output: T, DT.*
*for each class, Ci D, do*
*Locate the earlier probabilities, P(Ci).*
*end for*
*for each property estimation, Aij D, do*
*Discover the class restrictive probabilities, P(Aij|Ci)*
*end for*
*for each preparation example, xi D, do*
*Locate the back likelihood, P(Ci|xi)*
*if xi is misclassified, do*
*Expel xi from D;*
*end if*
*end for*
*Decide best parting trait;*
*T = Create the root node and name it with the parting characteristic;*
*T = Add circular segment to the root node for each split predicate and name;*
*for each circular segment do*
*D = Dataset made by applying parting predicate to D;*
*if halting point went after this way,*
*T' = Create a leaf node and name it with a suitable class;*
*else*
*T' = DTBuild(D);*
*end if*
*T = Add T' to curve;*
*end for*

Algorithm 4: DT classification

The vicinity of such raucous preparing models will without a doubt prompt a DT classifier to get over fitting, and along these lines rot its precision. In the wake of exhausting those misclassified/irritating cases from the arranging dataset, D, we along these lines hoard a DT for principal organization utilizing the resuscitated preparing dataset D with those totally hullabaloo free information. For the choice tree age, we select the best isolating property with the most exceptional data increment a help as the root center of the tree. Exactly when the root center of DT has been settled, the young center points and its indirect parts are made and added to the DT. The estimation proceeds recursively by adding new sub trees to each reaching out round area. The figuring closes when the models in the reduced arranging set all have a spot with a similar class. This class is then used to check the relating leaf center point. Reality multifaceted nature of a DT calculation relies on the size of arranging dataset, the measure of characteristics, and the size of the made tree.

## 4. Simulation Results

The simulation results of the proposed IDS are obtainable in this segment with the simulation setup and performance analysis.

### *4.1 Simulation Setup*

The proposed Intrusion detection system was implemented on MATLAB working platform and tested on most widely used dataset KDD CUP 99 with the following system configuration.

Operating System: Windows 8

Processor: Intel Core i3

RAM: 4GB

### *4.2 Simulation Results*

The significance of our proposed method has been tested with the above mentioned dataset and the simulation results obtained is tabulated as follows.

Table: 1 Simulation Results of The Proposed Work

| Parameters | Results |
|---|---|
| Accuracy | 96.5247 |
| False Alarm Rate | 0.0106 |
| Specificity | 96.47 |
| Detection Rate | 96.582 |

The computational time required to detect intrusion in CCN by means of our proposed method was tabulated below.

Table: 2 Computational Time of The Proposed Method

| Dataset | Computational time(s) |
|---|---|
| Iris | 780.64 |
| Glass | 778.32 |
| Ionosphere | 793.62 |
| Wine | 790.02 |
| Zoo | 786.91 |

From table 2 it is observed that our proposed method requires 785.902 seconds as an average computational time which is comparatively lower than the existing intrusion detection methods. The proposed CSOA optimally selects the features in order to detect Intrusion in the network which can be represented by the convergence graph shown in figure 2.

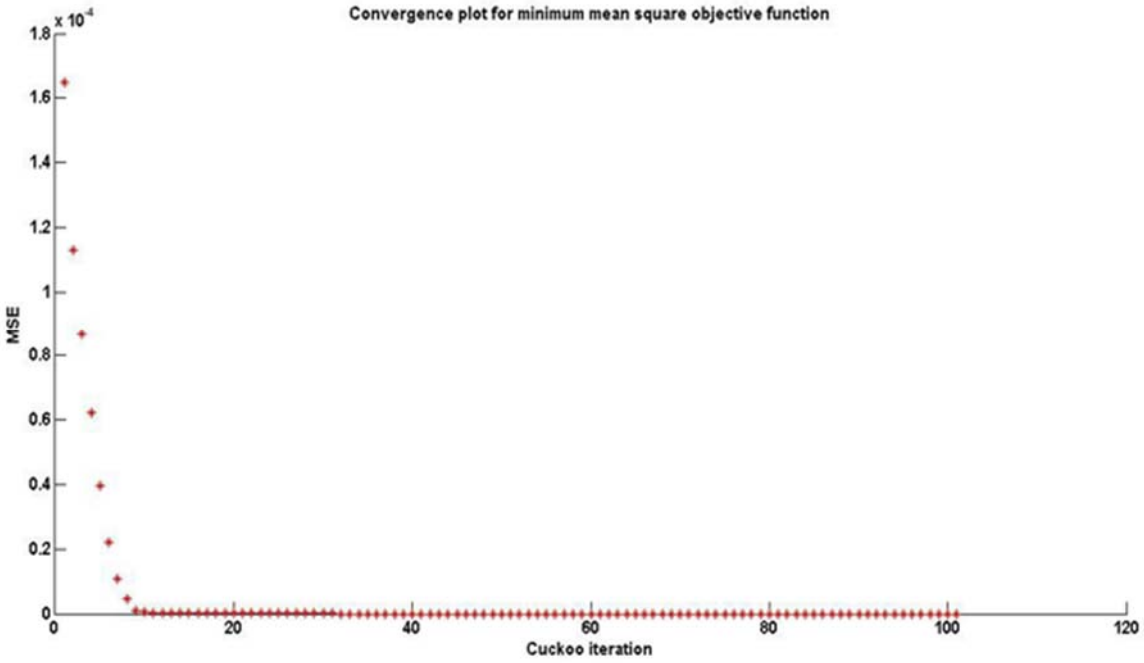


Figure 2: MSE of the proposed method

From figure 2 it is evident that the MSE obtained by the proposed optimization is comparatively very low and the above figure shows the variation of MSE with different k values.

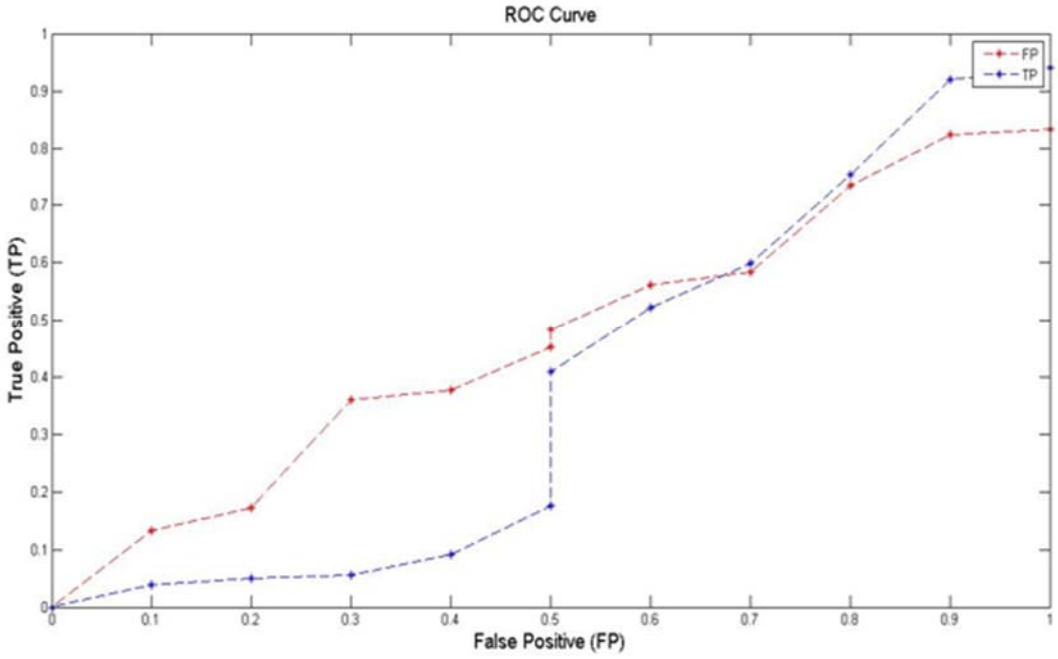


Figure 3: ROC Curve of the Proposed Method

Figure 3demonstrate the ROC curve of the projected method which is the relation between the False Positive and True Negative principles obtained by the Intrusion Detection System.

*4.3 Performance Analysis and Comparison*

These segments present the performance analysis of the proposed Intrusion detection method and the comparison with some existing Intrusion detection methods. The concert of the projected ID system in CCN was evaluate in terms of FPR, Specificity, Accuracy and Detection Rate.

*4.3.1 Specificity*

Specificity can be expressed as:

$$Specificity = \frac{TN}{TN + FP} \tag{12}$$

*4.3.2 Accuracy:*

Accuracy can be expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

*4.3.3 Detection rate (DR):*

The Intrusion detection rate is expressed as:

$$DR = \frac{TP}{TP + FP} \tag{14}$$

*4.3.4 False positive Rate:*

The false positive rate can be expressed as:

$$FPR = 1 - \frac{TN}{TN + FP} \tag{15}$$

By the proposed ID method, we

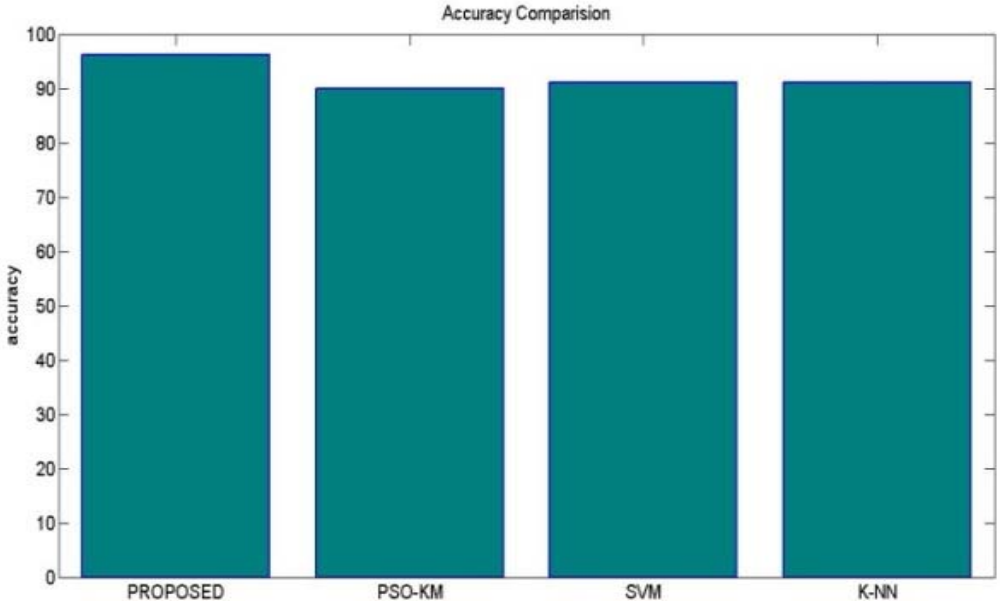


Fig. 4: Comparison of proposed ID method with Existing Methods in terms of Accuracy

In figure 4 the proposed ABC-DT classifier is compared with some of the existing classifiers like SVM, KNN. It is evidently seen that the precision of the projected ID classifier is greatly higher than the offered classifiers. From figure 5 the detection rate of the wished-for manner is compared with the accessible methods in order to show the efficiency of our proposed method. We have achieved 96.582% of detection rate with our proposed method which is comparatively superior than the existing detection rates.

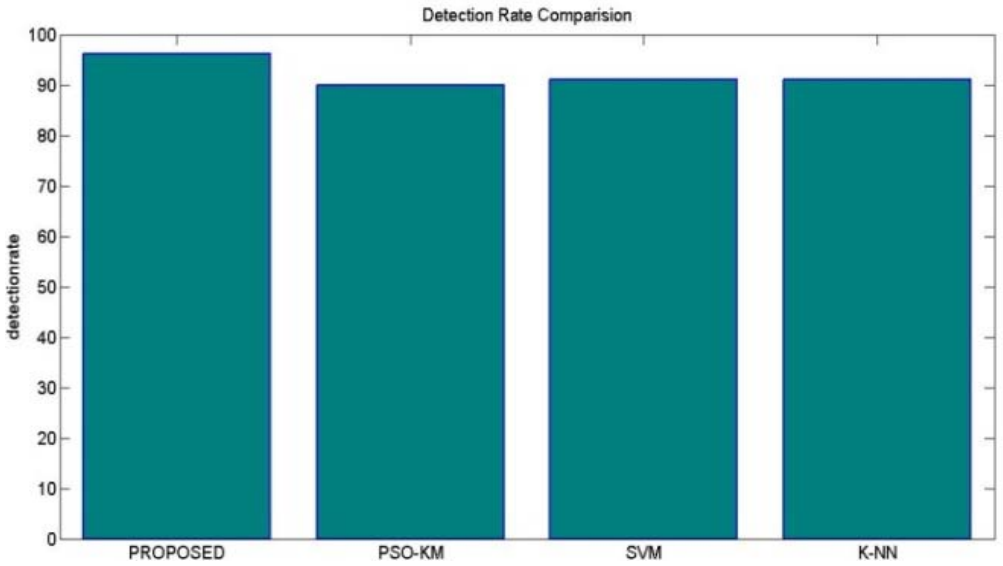


Fig. 5: Comparison of proposed ID method with Existing Methods in terms of Detection Rate

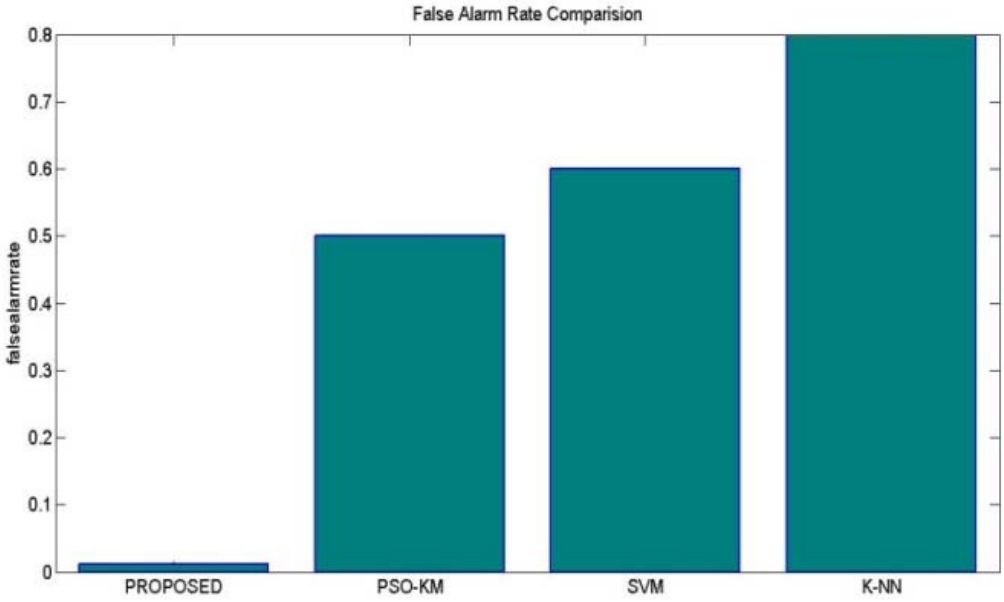


Fig. 6: Comparison of proposed ID method with Existing Methods in terms of False Alarm rate

Figure 6 shows the Impact of our anticipated ID method in requisites of False Alarm Rate. It is obtained that the False Alarm rate of the proposed method is 0.0106 which is greatly lower than the existing methods. Thus it proves that by the utilization of our proposed ID method we can significantly reduce Misclassification or False detection of Intrusion in network. It is essential to calculate the computational time of the overall detection method which is used in the proposed work. Since earlier detection or prediction of Intrusion would highly save the network from damage. So that the Computational time of the anticipated ID technique is estimated and compared with the existing methods which is shown in figure 7.

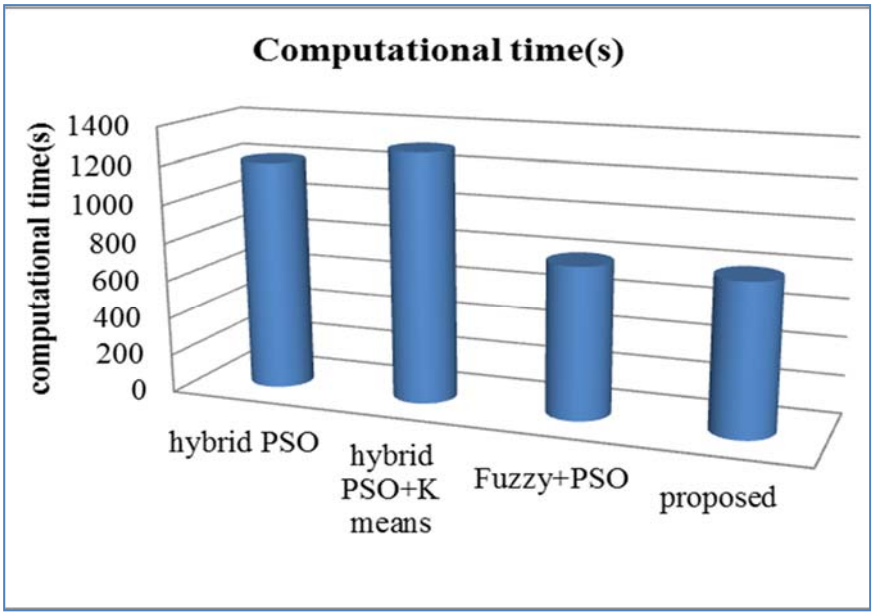


Fig. 7: Computational time Comparison

From figure 7 we can observe that the proposed method can detect intrusion with less computational time. On the whole from the above comparisons and simulation results comparatively our proposed method shows best results in terms of precision and computational time. This shows the superiority and the efficiency of the anticipated technique. Since it can be used as an effective tool on Network Intrusion detection on for coming network trends.

## 5. Conclusion

In this research exertion we have obtainable an IDS base on a novel classification algorithm. The proposed work also depends upon a novel clustering algorithm which improves the detection accuracy. The hierarchical k means clustering is initially utilized to form the number of clusters and then the optimal clusters are determined by means of the cuckoo search optimization algorithm. The features which are essential to train the classifier is then extracted from the optimally selected clusters and finally the optimum features are selected by the ABC optimization algorithm and explored to the DT classifier which produce the results as normal or abnormal by matching the features extracted in the training phase and testing phase. The proposed method will be implemented on MATLAB working platform and tested on most widely used dataset KDD CUP 99 and the simulation results and comparison presented in this article shows the importance and effectiveness of the proposed work than the existing works.

## REFERENCES

[1] Sengupta, Nandita, Jaydeep Sen, Jaya Sil, and MoumitaSaha. "Designing of on line intrusion detection system using rough set theory and Q-learning algorithm." Neurocomputing 111 (2013): 161-168.
[2] Sheikhan, Mansour, Zahra Jadidi, and Ali Farrokhi. "Intrusion detection using reduced-size RNN based on feature grouping." Neural Computing and Applications 21, no. 6 (2012): 1185-1190.
[3] Kim, Gisung, Seungmin Lee, and Sehun Kim. "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection."Expert Systems with Applications 41, no. 4 (2014): 1690-1700.
[4] Kevric, Jasmin, SamedJukic, and AbdulhamitSubasi. "An effective combining classifier approach using tree algorithms for network intrusion detection." Neural Computing and Applications (2016): 1-8.
[5] Lin, Wei-Chao, Shih-Wen Ke, and Chih-Fong Tsai. "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors." Knowledge-based systems 78 (2015): 13-21.
[6] Elhag, Salma, Alberto Fernández, Abdullah Bawakid, Saleh Alshomrani, and Francisco Herrera. "On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on Intrusion Detection Systems." Expert Systems with Applications 42, no. 1 (2015): 193-202.
[7] Kenkre, Poonam Sinai, AnushaPai, and Louella Colaco. "Real time intrusion detection and prevention system." In Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014, pp. 405-411. Springer International Publishing, 2015.
[8] Costa, Kelton AP, Luis AM Pereira, Rodrigo YM Nakamura, Clayton R. Pereira, João P. Papa, and Alexandre Xavier Falcão. "A nature-inspired approach to speed up optimum-path forest clustering and its application to intrusion detection in computer networks." Information Sciences 294 (2015): 95-108.
[9] Shamshirband, Shahaboddin, NorBadrulAnuar, Miss Laiha Mat Kiah, and Ahmed Patel. "An appraisal and design of a multi-agent system based cooperative wireless intrusion detection computational intelligence technique." Engineering Applications of Artificial Intelligence 26, no. 9 (2013): 2105-2127.
[10] Li, Yinhui, Jingbo Xia, Silan Zhang, Jiakai Yan, Xiaochuan Ai, and Kuobin Dai. "An efficient intrusion detection system based on support vector machines and gradually feature removal method." Expert Systems with Applications 39, no. 1 (2012): 424-430
[11] Koc, Levent, Thomas A. Mazzuchi, and ShahramSarkani. "A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier." Expert Systems with Applications 39, no. 18 (2012): 13492-13500.

[12] Feng, Wenying, Qinglei Zhang, Gongzhu Hu, and Jimmy Xiangji Huang. "Mining network data for intrusion detection through combining SVMs with ant colony networks." Future Generation Computer Systems 37 (2014): 127-140.
[13] Satpute, Khushboo, Shikha Agrawal, Jitendra Agrawal, and Sanjeev Sharma. "A survey on anomaly detection in network intrusion detection system using particle swarm optimization based machine learning techniques." InProceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), pp. 441-452. Springer Berlin Heidelberg, 2013.
[14] Eesa, Adel Sabry, ZeynepOrman, and Adnan MohsinAbdulazeezBrifcani. "A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems." Expert Systems with Applications42, no. 5 (2015): 2670-2679.
[15] Baig, Zubair A., Sadiq M. Sait, and AbdulRahmanShaheen. "GMDH-based networks for intelligent intrusion detection." Engineering Applications of Artificial Intelligence 26, no. 7 (2013): 1731-1740.
[16] Nikolova, Evgeniya, and VeselinaJecheva. "Some similarity coefficients and application of data mining techniques to the anomaly-based IDS."Telecommunication Systems 50, no. 2 (2012): 127-135.
[17] Viswanathan, Arun, Kymie Tan, and Clifford Neuman. "Deconstructing the Assessment of Anomaly-based Intrusion Detectors." In International Workshop on Recent Advances in Intrusion Detection, pp. 286-306. Springer Berlin Heidelberg, 2013.
[18] Kuang, Fangjun, Weihong Xu, and Siyang Zhang. "A novel hybrid KPCA and SVM with GA model for intrusion detection." Applied Soft Computing 18 (2014): 178-184.
[19] Shamshirband, Shahaboddin, AminehAmini, NorBadrulAnuar, Miss Laiha Mat Kiah, Ying WahTeh, and Steven Furnell. "D-FICCA: A density-based fuzzy imperialist competitive clustering algorithm for intrusion detection in wireless sensor networks." Measurement 55 (2014): 212-226.
[20] Casas, Pedro, Johan Mazel, and Philippe Owezarski. "Unsupervised network intrusion detection systems: Detecting the unknown without knowledge."Computer Communications 35, no. 7 (2012): 772-783.
[21] D'angelo, G., Palmieri, F., Ficco, M., &Rampone, S. (2015). An uncertainty-managing batch relevance-based approach to network anomaly detection. Applied Soft Computing, 36, 408-418.
[22] Elhag, S., Fernández, A., Bawakid, A., Alshomrani, S., & Herrera, F. (2015). On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on Intrusion Detection Systems. Expert Systems with Applications, 42(1), 193-202.
[23] Costa, K. A., Pereira, L. A., Nakamura, R. Y., Pereira, C. R., Papa, J. P., &Falcão, A. X. (2015). A nature-inspired approach to speed up optimum-path forest clustering and its application to intrusion detection in computer networks. Information Sciences, 294, 95-108.
[24] Shakshuki, Elhadi M., Nan Kang, and Tarek R. Sheltami. "EAACK—a secure intrusion-detection system for MANETs." IEEE Transactions on Industrial Electronics 60, no. 3 (2013): 1089-1098.
[25] Ashfaq, Rana Aamir Raza, Xi-Zhao Wang, Joshua Zhexue Huang, Haider Abbas, and Yu-Lin He. "Fuzziness based semi-supervised learning approach for intrusion detection system." Information Sciences (2016).