# An Efficient Anomaly Detection Based On Optimal Deep Belief Network in Big Data

**Priyanka Dahiya, Devesh Kumar Srivastva**

***Abstract:*** ***Nowadays, the internet and network service user's counts are increasing and the data generation speed also very high. Then again, we see greater security dangers on the internet, enterprise network, websites and the network. Anomaly has been known as one of the effective cyber threats over the internet which increasing exponentially and thus overcomes the commonly used approaches for anomaly detection and classification. Anomaly detection is used in big data analytics to recognize the unexpected behaviour. The most commonly used characteristics in network environment are size and dimensionality, which are big datasets and also impose problems in recognizing useful patterns, For example, to identify the network traffic anomalies from the large datasets. Due to the enormous increase of computer network based facilities it is a challenge to perform fast and efficient anomaly detection. The anomaly recognition in big data sets is more useful to discover fraud and abnormal action. Here, we mainly focus on the problems regarding anomaly detection, so we introduce a novel machine learning based anomaly detection technique. Machine learning approach is used to enhance the anomaly detection speed which is very much useful to detect the anomaly from the large datasets. We evaluate the proposed framework by performing experiments with larger data sets and compare to several existing techniques such as fuzzy, SVM (Support Vector Machine) and PSO (Particle swarm optimization). It has shown 98% percentage of accuracy and the false rate of 0.002 % on proposed classifier. The experimental results illuminate that better performance than existing anomaly detection techniques in big data environment.***

***Keywords: Big data, Anomaly detection, deep learning, high dimensionality, k-means, classification.***

## I. INTRODUCTION

Big data contains huge amount of structured and un-structured data which overflows the business on regular basis. This big data produce large volume of data at small interval of time from connected devices like vehicle fleets or industrial machinery, mobile phones, etc. [1]. Construction and examining of this data is done to reduce several issues, one among them was the issue raised while detecting the anomalous behaviour. Big data analysis was often done for anomaly detection which was recognised as a research problem, so this range of data was not handled by mainstream computing technology.

The problems obtained in big data were overcome by Machine Learning (ML) algorithm and this algorithm is also used to obtain valuable information from hugely available data. ML algorithm frequently appeals prohibitive computational resources to face huge data [3]. The prospects of computation were not accomplished by the toolkit which support the progress of ML software, because the problems of demand and challenge were increased in some cases obstinate by traditional CPU architecture [4].

For the past few years, the applications of big data were developed and a number of researchers from various stream realized the advantage of extracting knowledge in this type of issue. Now a day's big data is used to gather and process a large number of data having capacity to process and receive data rapidly and efficiently with less computational time [5]. The big data also made its entry in scientific, business, engineering streams and in a common network to observe the operator's action and location to provide an enhanced plan, junk and fraud recognition [6].

Data mining and data warehousing are the most advanced techniques in data analysis and also used to determine the type of mining and recovery. Computing units analyse data and evaluates according to the mining procedures. The data scale is large in single personal computer as well as the data processing framework which depends on cluster computers with high performance can handle big data mining [7]. The correlation among the different database and the data from the ML was identified by data mining and the difficulties that obtained during computation were reduced by traditional mining algorithm [8].

In ML, intelligent decisions are made spontaneously and the automatic text classifier learning is produced by an inductive process to achieve accuracy and power saving from a set of pre-classified document [9]. Anomalies are configurations obtained from the data which were not implemented by the certain idea of ordinary performance. After understanding the pre-processing of the real world, it explains that, the raw data was equipped for another processing procedure and perfectly transfer the data into desired format (e.g. neural network) [10].

In bigdata, the dataset is classified into structured and unstructured data. In this dataset feature extraction, and feature reduction are the two important attributes [11].Sampling, normalization, denoising and transformation are done by this data set to obtain noise removal, feature extraction and single input [12]. For large datasets, the sampling process is done by stratified sampling and random sampling [13]. Some issues may occur while separating the trained dataset, so multi class problem decaying method is used to overcome this issue. This method can be done by splitting the main problem into a number of sub problems based on feature or tuple based sample and space [14].

Based on decomposition feature selection in data mining, the classifiers and the innovative feature set is separated into a volume of

**Revised Manuscript Received October 05, 2019**

**Priyanka Dahiya,** Research scholar, Manipal University, Jaipur, Rajasthan

**Devesh Kumar Srivastva,** Professor with the department of School of Computing & Information Technology,Manipal University Jaipur, Rajasthan, India.

subsets [15]. If enormous amount of datasets are available, then it is difficult to identify, store and maintain, so the dataset is decomposed into a number of subsets by concealing the connections between the datasets [16].The clustering technique used in big data is divided into two classification as single machine clustering and multiple machine clustering. In these two classifications the single machine selects the perfect division of variable whereas the multiple machines carries out parallel computing to increase the calculation speed and scalability to achieve better results [17].

The performance of classification is improved by the classifiers like support vector, multilayer perceptron and neural network. The multi-classifier system is produced by the classifiers to improve the classification performance. A new training dataset is obtained by combining the classified dataset with original data and from that maximum shortage details is also identified [18]. Based on future queries a new classifier will be developed and can be widely used in medical field by the data collected from the patient who consumes large time and occupies vast area [19]. The artificial intelligence based neural network technique provides a step to make a final decision and the information received from neurons which is an interconnected nerve cells in surrounding [20].In an epoch of big data, the anomaly detection should be well-organized to process the large volume of data at real time without any loss of dynamic packet flow. There are no ideal solution exist to provide better accuracy. So there is a need to propose an efficient anomaly detection approach for securing the valuable resources from malicious or unauthorized actions. The major requirement of anomaly detection is system accuracy and efficiency.

To address the above challenges, the proposed system uses high speed anomaly detection in a parallel environment Hadoop which detects any network anomalies with more accuracy. We use deep learning algorithm and the main contributions of this paper are listed below:

- To identify abnormal behaviours and attempts caused by intruders in the network and computer system using Hadoop architecture.
- To illustrate the effectiveness of deep learning in an attack detection system is better than the traditional ML in big data analysis using evaluation.
- We use an optimization algorithm to improve the classification accuracy with minimum error rate in big data.

*Organization:* The paper residue is scheduled as follows. An overview of related work is provided in section 2. The methodology is presented in Section 3. Experimental setup and evaluation are described in Section 4. Discussion and performance evaluation of the proposed methodology are explained in section 5. Section 6 conclude the proposed work and describe few future works.

## II. RELATED WORK

Abdulla Amin Aburomman *et al* [21] projected a new method for interruption recognition by generating collaborative classifier with high accuracy by employing PSO generated mass. In this, the parameters for PSO behaviour was identified by Local Unimodal Sampling (LUS).From KDD99 dataset five arbitrary subsets were selected and the collaborative classifiers were obtained by using some new technique and weighted Majority Algorithm (WMA). WMA was replaced by WMV (Weighted Majority Voting) as it provides a high rate of ordering accuracy.

A strong irregularity recognition technique called F-CBCT (Fuzzied Cuckoo based Clustering Technique) was suggested by Sahil Garg et al [22] to minimize the increasing diffusion of security threats. The F-CBCT method works both in training and detection phase where the K-means Clustering Algorithm, Multi-Objective Cuckoo-Search Optimization (CSO) Algorithm and Decision Tree Criterion (DTC) were used to perform the training phase. CSO provides more robust and accurate outcomes than the PSO in terms of accuracy and performance. In the detection phase, fuzzy detection approach was used to detect anomalies based on distance functions and input data calculated in training phase.

Sarah M. Erfani et al [23] has introduced cluster anomaly recognition technique for large-scale high-dimensional unlabelled dataset. It is a mixture of both DBN (Deep Belief Network) and first-class Super Vector Machine (SVM). Non-linear manifold was produced to track the DBN in reduction algorithm of dimensionality and transform the data into low-dimensional feature set. An efficient anomaly detection method which was accurate and scalable was applied on large- dimensional and high-scale domain and also for training the hybrid DBN-1SVM to execute 3 times faster and it was 1000 times faster for training.

Mohammed A. Ambusaidi et al [24] has introduced Flexible Mutual Information Feature Selection (FMIFS) algorithm to minimize redundant and unrelated features in dataset. Hybrid feature selection algorithm (HFSA) based common information was handled in both direct and indirect based dataset. The features that are obtained from the feature selection algorithm were used to construct Interruption Detection System (IDS) called as Least Square Support Vector based IDS (LSSVM-IDS). There are two stages in HFSA they are upper and lower stage. The upper stage performs initial search to deduce the redundant and unrelated features from the innovative dataset and the details obtained in the upper phase was transferred to the lower stage to reduce searching time. The performance was estimated by three IDS datasets they are Kyoto 2006+, KDD Cup 99 and NSL-KDD and the result outclassed the other state-of-art model.

N. Pandeeswari et al [25] who proposed Hypervisor detector which is used to detect anomaly in hypervisor layer. Hybrid algorithm was obtained by grouping both FCM-ANN (Fuzzy C-Means Clustering algorithm and Artificial Neural Network) and was used by hybrid detector to obtain high IDS accuracy. It was performed in three phases. The first phase was performed with fuzzy clustering element, next phase contains different ANN component and the final phase was performed with Fuzzy aggregation element. In this the proposed method was applied in relation with ANN classic algorithm and Naïve Bayes classifier. The result obtained after comparison was used to obtain low rate false

alarm and high accuracy in detection.

### *A. Problem statement*

In this the provided dataset contains illustrations with features that are obtained from different source of datasets like huge number of computing system with high performance, network of initiative and also from the unlabelled datasets. The main aim of this method is to obtain an effective learning method to detect some examples of anomalous dataset with minimum false alarm rate and time constraint data that produced in large amount.

## III. BIG DATA ANALYTICS FOR ANOMALY DETECTION

This work is mainly concentrated in detection of anomaly datasets in bigdata and also to increase the anomaly data detection speed by using ML. For that, first the datasets are divided into training and testing samples and then the dimensionality reduction in dataset is obtained by Generalized Discriminant Analysis (GDA). The dimensionality reduced datasets are then clustered and the appropriate features are obtained from the dataset. At last the splitting of different classes of anomalies are done by Deep Belief Network based fruit fly optimization method and some of the objectives of the proposed method are,

- To identify an anomaly vector with assistance of comprehensive system observations.
- To handle maximum number of data using machine learning algorithm
- To enhance the classification accuracy and reduce the false alarm rate
- To validate the proposed technique with large data set.

### A. System framework

The main aim of this method is to develop deep learning classifier by increasing the anomaly detection accuracy. We describe a Hadoop based distributed big data processing framework. Figure 1 illustrates the process flow of proposed anomaly detection. At first, the raw data is inputted to IDS and pre-processing is done for this system. The pre-processing includes dimensionality reduction, feature selection. Then the features are clustered and loaded in HDFS. Then the anomalies predicted by using RBM layers with weight updation to minimize the classification error rate.

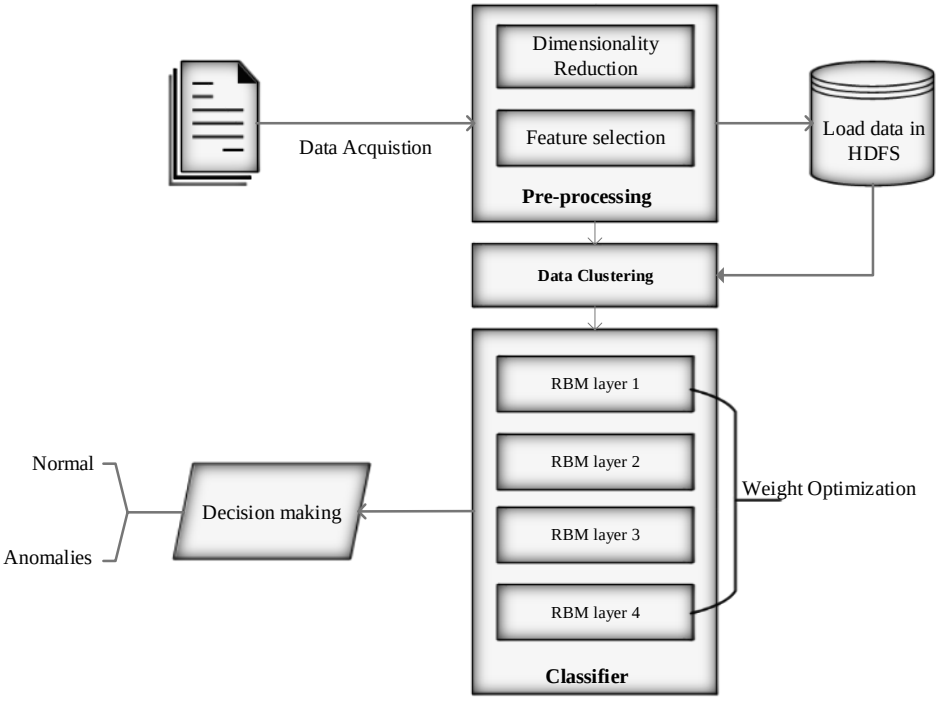


**Fig. 1: Design of Proposed model**

### B. Data pre-processing

GDA [26] is used for the problems in multi-class classification. In the feature space some overlaps are obtained among the patterns of different attacks due to huge difference in that particular classes and the scatter among the classes are reduced by using feature transformation mechanism. GDA is proposed for non-linear classification to transfer the original Z space to the high dimensional new feature space $X : \chi : Z \rightarrow X$ based on kernel function $\chi$ . The scatter that obtained inside and outside the cluster of indirectly mapped data was shown as,

$$C^{\chi} = \sum_{b=1}^{B} N_b n_b{}^{\chi} (n_b{}^{\chi})^T \quad (1)$$

$$V^{\chi} = \sum_{b=1}^{B} \sum_{z \in Z_b} \chi(z) \chi(z)^T \quad (2)$$

Where, $n_b^{\chi}$ represent the class mean $Z_b$ in X and $n_b$ represent the number of samples belongs to $Z_b$. The GDA aims to identify projection matrix $W^{\chi}$ which increase the ratio,

$$W_{opt}^{\chi} = \arg\max \frac{|(W^{\chi})^T C^{\chi} W^{\chi}|}{|(W^{\chi})^T C^{\chi} V^{\chi}|} = [w_1^{\chi}, \ldots w_M^{\chi}] \quad (3)$$

The vectors, $w^{\chi}$ is identified as the resolution for the complications produced by general Eigen value that is, $C^{\chi} w_i^{\chi} = \lambda_i V^{\chi} w_i^{\chi}$.

The vectors of training data are assumed as centred in the feature space X. The reproducing kernels based theory provides any solution $w^{\chi} \in X$ should place all training samples span X, that is,

$$w^{\chi} = \sum_{b=1}^{B} \sum_{i=1}^{N_b} \alpha_{bi} \chi(z_{bi}) \quad (4)$$

Where, some real weights are defined as $\alpha_{bi}$ and the i[th] sample of the class B is defined as $x_{bi}$. The solution is found from the following condition,

$$\lambda = \frac{\alpha^T * KDK * \alpha}{\alpha^T * KK * \alpha} \quad (5)$$

Where, $\alpha = (\alpha_b), b = 1 \ldots B$ represents vector weights with $\alpha = (\alpha_{bi}), i = 1 \ldots N_b$. The $K(M \times M)$ kernel matrix contains nonlinearly mapped data dot products, i.e.

$$K = (K_{kl})_{k=1\ldots B, l=1\ldots B} \quad (6)$$

Where, $K_{kl} = (k(z_{ki} z_{lj})), i = 1 \ldots N_k, j = 1 \ldots N_l$. The $D(M \times M)$ matrix is a block diagonal matrix such that,

$$D = (D_b)_{b=1\ldots B} \quad (7)$$

Where, the B[th] integer in the diagonal part has all elements equal to $1/N_b$. After resolving the Eigen value issues, the coefficient vector $\alpha$ was obtained and was used to define the projection vector $w^{\chi} \in X$. A projection of a testing vector $z_{test}$ was computed as,

$$(w)^T \chi(z_{test}) = \sum_{b=1}^{B} \sum_{i=1}^{N_b} \alpha_{bi} k(z_{bi}, z_{test}) \quad (8)$$

GDA reduces the original features into valuable features.

## C. Clustering

Adaptive K-means algorithm is used for large number of datasets clustering. The aim of clustering is to gather objects and each object should contain relevant object. K-means along with Hadoop [27] calculation contains two major parts, the first part concentrates on the cluster centres and then the sample data is divided into specific size of data blocks. The second part minimizes the task needed for parallel processing algorithm and it is done till the cluster output obtained. Mapper is initiated in the second part which maps the input data to the transitional value sets. The middle road values are reduced by the reducer as it shares the key to lower classification values.

***a) Map function***

In the map function, the input data of each line is given to the mapper function and the data obtained by the mapper function is used to form a small fragment of information.

***b) Reduce function***

The important work of reducer is to manage the information that is produced by the mapper function. Here the reduce function and shuffle organize are mixed for the relevant operation. After completing the handling process another section of output is produced, this made safe in Hadoop DFS.

**Algorithm**

**Step1:** Initially centroid is chosen randomly to view data**.**
**Step2:** Details about initial centroid and data are made available in input record.
**Step3:** In mapper class we use "configure" function to open the record initially and read the centroids which are stored in the data structure in an Array list format.
**Step4:** The information record is read by the mapper and then release adjacent centroid with a mark to the reducer.
**Step5:** The whole data is gathered by the reducer and the new matching centroid is identified and emitted.
**Step6:**Based on the job structure, the files are read and validated
if the variance between original and old centroidis less than 0.1 then
convergence is arrived
else
Use the current centroids and repeat step 2.

## D. DBN classifier

Recently DBN [28] become popular as it provides high advantage by performing the learning process of data in layer-by-layer manner and some of its merits are speed interference and rapidly encode the higher order network structure. DBN is a feed forward neural network (FFNN) using input layer, number of hidden layers and one output layer. The hidden layer permits the DBN to model the complicated relationship with the dataset in rapid manner. From the figure 1 it is identified that the DBN building block is a probabilistic design known as restricted Boltzmann machine which contains two layer neural networks where binary inputs of stochastic is connected to the output by symmetric weighted connection. The visible neurons are placed in first layer whereas the hidden layers placed in second layer and the communication between the two neurons is constrained. The distribution of condition among the data of another layer depends on the data of some layer.

Let us consider the state of hidden neurons

$H = \{h_1, h_2, h_3, h_4, ....h_n\}$ and the visual neurons are represented as $V = \{v_1, v_2, v_3, v_4, ....v_m\}$. The connection matrix between two layers is represented as $W_{m \times n}$. Energy for thr visible and hidden neurons of joint configuration (V, H) was defined as,

$$Energy(V,H) = -\sum_{i=1}^{m} a_i v_i - \sum_{j=1}^{n} b_j h_j - \sum_{i=1}^{m}\sum_{j=1}^{n} v_i w_{ij} h_j \quad (9)$$

Where $V_i$ represents the visible layer of i[th] neuron and $H_j$ represents the hidden layer of j[th] neuron properly. The biases are represented as $a_i \; and \; b_j$ and the bidirectional weight between the i[th] neuron and j[th] neuron is defined as $w_{ij} = w_{ji}$.

After identifying all the constraints, the energy function of the joint probability distribution (V, H) is shown in equation (9),

$$P_r(V,H \mid \theta) = \frac{1}{Z(\theta)} e^{-Energy(V,H|\theta)} \quad (10)$$

Where, $\theta = (w_{ij}, b_j, a_i)$ are the RBM parameters, and $Z(\theta) = \sum_{V,H} e^{-Energy(V,H|\theta}$ is a partition function. Since there are no associations between the two layers, the neurons conditional probability distributions can be calculated as,

$$P_r(v_i = 1 \mid H) = \frac{1}{1 + \exp\left(-a_i - \sum_j h_j w_{ij}\right)} \quad (11)$$

$$P_r(h_j = 1 \mid V) = \frac{1}{1 + \exp\left(-b_j - \sum_j v_j w_{ij}\right)} \quad (12)$$

***a) Optimization approach***

In this paper, a new enhanced DBN is developed to detect the anomalies from the large dataset. This design contains three parts such as DBN training procedure, expansion of trained DBN by Fruit fly, and design for complete process to detect anomalies by DBN optimization.

***b) DBN training***

DBN is designed with stacked RBM and the training for DBN is done by training individual RBM successively. An objective function that is expressed as sum of differentiable function is optimized by Stochastic Gradient Descent (SGD). This optimization is done rapidly and effectively by identifying the gradients from a few examples rather than the whole training set. The visible neurons are first provided to produce $v_i$ with training data and then the hidden layers $h_j$ are sampled based on the probabilities shown in (12). This process is repeated one more time so the visible neurons get updated and buried neurons provide another one-step 'rebuilt' states $v_i$ and $h_j$.

Next the rise of joint likelihood task of data, the describe rule for the visible to hidden weights $w_{ij}$ was given as shown in equation.13,

$$\Delta w_{ij} = \eta((v_i h_j)_{data} - (v_i h_j)_{recon}) \quad (13)$$

Where, $\eta \in (0,1)$ indicates the learning rate, $(\cdot)_{recon}$ denotes to the expectation over the rebuilt data and $(\cdot)_{data}$ denotes the expectation among training data. $\alpha \in [0,1]$ Is a momentum which updates the weights for further stabilization in RBM training process? The weight updated to the current epoch is related to the updated weight of previous epoch with such momentum which formulated as,

$$\Delta_{w_{ij}}^{k+1} = \eta((v_i h_j)_{data} - (v_i h_j)_{recon}) + \alpha \Delta w_{ij}^{\;k} \quad (14)$$

After the RBM the DBN was trained. During the starting stage of training process the first part v of RBM (1) was trained and produce parameter h(1) for initial hidden layer along with the RBM(2) visible layer and the same procedure is followed to identify the parameters h (2), h (3) respectively. All this processing steps come under pre-processing stage and it was shown in fig. 2.

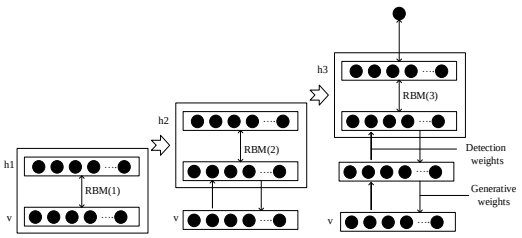


**Fig. 2: Pre-training of DBN**

This layer-by-layer learning process was repeated for several times by combining an additional layer to increase the log probability lower bound which was assigned to the training data. Some of the inputs like number of layers, maximum number of iteration and the pre-processed data are provided for training DBN before initializing the training process.

***c) DBN optimization using Fruit fly***

After constructing DBN, the optimization is done where the selection of optimal amount of hidden layers become difficult. In this paper, DBN with three layer shows better results and fruit fly is used to obtain the improved design of trained DBN with appropriate number of neurons in several hidden layer, momentum and suitable learning rate.

Let us consider, the hidden layer of RBM (1), RBM (2) and RBM (3) has n1, n2, n3 neurons and the learning rate $\eta \in (0,1)$ and momentum $\alpha \in [0,1]$. Here, a 5-dimensional vector $G(n1, n2, n3, \eta, \alpha)$ is designed in Fruitfly. Population of Fruitfly $G_i$ is prepared with enough size M (i=1, 2,…M) and the fruitfly is used to design the

optimization DBN and apply it to detect anomalies.

The algorithm which is shown below is used to obtain the finest neuron numbers in several hidden layers, momentum and the learning rate:

*Step 1:* The training samples are given as an input and trained DBN is prepared.

*Step 2:* Suitable iteration number M is decided and the position of the population gets modified.

*Step 3:* Adjust the positions $X_i, Y_i$ so that the distance of every fruit fly **lies** in a given range.

*Step 4:* Each fruit fly is estimated by misclassification errors from the misclassified samples as well as the training samples. Then we identify the best smell and index of the fruit fly from its history.

*Step 5:* Update the position and distance of all the fruit flies using the below condition:

$$X_i = X\_axis\_Rvalue, \&\& Y_i = Y\_axis\_Rvalue \quad (15)$$

$$Distance = \sqrt{x_i^{\,2} + y_i^{\,2}}; \quad (16)$$

$$S_i = \frac{1}{Distance} \quad (17)$$

Where, R value is the random value.

*Step 6:* The optimization process is stopped when the misclassification error rate of training samples is small enough.

***d) Anomaly detection using optimization***

The initial step in this optimization method is to obtain the regular and irregular behaviour of the dataset and extracting the time-domain features of every sample from several states. In the classification process extra features with good results for the lower level is obtained by pre-processing the input data. The training and testing samples are prepared from pre-processed data which is further used to analyse the rationality of optimized DBN. At first the DBN parameters are initialized and then the fruit fly develops optimization DBN on training sample. The optimization DBN obtained from training is given to testing and it is authorised using misclassification error.

## IV. EXPERIMENTAL SETUP AND EVALUATION

The proposed anomaly detection system is implemented using java in Hadoop system with the Hadoop APIsHadoop -pcap-lib, Hadoop-pcap-input and Hadoop -pcap-serde to process the real time packets. Machine learning classifiers are developed in java as a decision maker when Hadoop processes sequence file and calculates parameters' values for each incoming packet flow. Table 1, Simulation parameters are explained in parameter and values.

**Table 1: Simulation parameters**

| Parameter | Value |
|---|---|
| Neurons of input layer | 9 |
| Neurons of output layer | 2 |
| RBM number | 3 |
| Iteration of every RBM | 80 |
| N1 | 22 |
| N2 | 11 |
| N3 | 5 |
| $\alpha$ | 0.90 |
| $\eta$ | 0.11 |
| Iteration of Fruitfly | 40 |
| Population size of Fruitfly | 15 |

### A. Datasets

In this section number of studies is made using different datasets and the datasets are divided into two parts as dataset by institution and dataset by individuals. Datasets are provided into the public domain by the dataset owners and it is a difficult task for every researcher to obtain a dataset for anomaly detection, but some work use the datasets from the second type. Most famous datasets available in public domain are KDD-99 and DARPA-99 and for estimation UNB ISCX 2012 interruption detection dataset is used and the obtained dataset size is 90.9 GB. There are four types of attacks available in this dataset are penetrate the network from inside, distributed denial of service (DDoS), brute force SSH, by HTTP denial of service (DoS) and IRC Botnet and. Finally the data collection of UNB ISCX 2012 needs to be evaluating the anomaly detection method.

### B. Evaluation measurement

KDD-99 dataset is taken for test evaluation. Two types of files there in the dataset which are 10% dataset known as s KDDCUP.Data.10.percent.correceted. The one is complete dataset known as kddcup. data. corrected. 41 fixed feature attributes are presented in each data record. The requirements of anomaly detection are high accuracy in detection and reduced false alarm rate. The label of the network performance is shown by the identifier feature as Normal or Attack. Some of the features from the label file to CSV file is extracted by Java program and the extracted features are total Source Packets, total Source Bytes, source protocol Name, total Destination Bytes, source Port, destination, total Destination Packets, destination Port, start Date Time, stop Date Time and Tag. A 119.9 MB file is produced from this result and isused to analyse the correctness of classification by the detection system. Below the Table 2, are define the General behaviour anomaly detection. The classification accuracy, classification error rate and detection rate of anomaly detection is identified by using the following formula:

$$Accuracy = Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (18)$$

$$DetectionRate = DR = \frac{TP}{TP + FP} \quad (19)$$

$$ErrorRate = \frac{FP}{TN + FP} \quad (20)$$

**Table 2: General behaviour anomaly detection**

| Actual case | Normal prediction | Attack prediction |
|---|---|---|
| **Normal behavior** | TN | FP |
| **Anomalies** | FN | TP |

True negative Rate (TNR): The rate between the normal labels with the packets and the same label with the packets noticed by the system.

True positive Rate (TPR): the rate between the attack labels with the packets and the same label with the packets identified by the system.

False positive Rate: the rate between the packets documented as normal and the packets identified by the attack class of the system.

False negative Rate: the rate between the packets recognized as attack and the packets noticed by the normal class member of the system.

## V. RESULTS AND DISCUSSION

Results and discussion of this paper is deliberated in this section. Large number of tools is used to develop this system and one of the main problems discussed in this paper is storing and reading pcap data's from the dataset and also it is important to fulfil the scalable and flexible requirement. The accuracy results for various machine learning classifier are compared with proposed classifier is illustrated in table 3.

**Table 3: performance analysis**

| Classifier | KDDCUP.Data.10.percent.correceted | | kddcup.data.corrected | | Overall | |
|---|---|---|---|---|---|---|
| | TP (%) | FP (% | TP (%) | FP (%) | P (%) | FP (%) |
| Fuzzy | 94.1 | 0.0002 | 95 | 0.0015 | 4.6 | 0.0012 |
| SVM | 95.8 | 0.0001 | 94.3 | 0.0001 | 5.93 | 0.001 |
| PSO | 82.2 | 0.005 | 77 | 0.0001 | 9.03 | 0.038 |
| DBN | 98.9 | 0.0004 | 98.9 | 0 | 8.9 | 0.0002 |

The anomaly detection with the presented classifier provided improved performance with the selected features. Different metrics are considered while evaluating the presented approach includes detection rate, accuracy, F-measure and false positive rate. The DBN based anomaly detection technique earned better accuracy on detection and lower false positive rate; it is described in table 3. Figure 3 to 6 represents the rate of detection and false alarm of proposed and three other existing classifiers. Figure 3 describes that the presented detection system provided better accuracy on detection rate for frequent attacks. Figure 4 describes the detection rate obtained on both proposed and existing classifiers (fuzzy, PSO, and SVM). From figure 4, during the detection of Probe and DoS attack with DBN got similar detection rate under different models. However, higher detection rate is achieved with DBN on low frequent attacks than the existing ones. Also, there is very lower false alarms were produced by presented approach than the existing ones; it is shown in figure 5. From obtained f-measure rates, the DBN classifier earned better results than other classifiers and it is shown in figure 6.

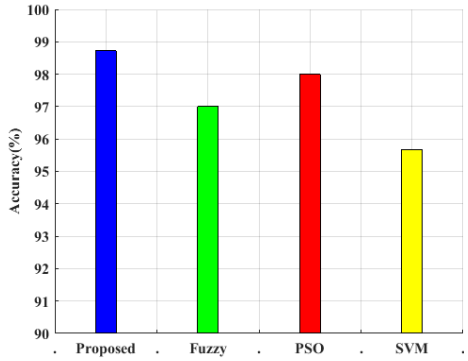


**Fig. 3: Accuracy analysis**

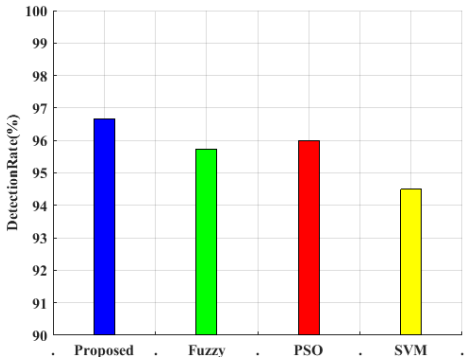


**Fig. 4: Detection rate analysis**

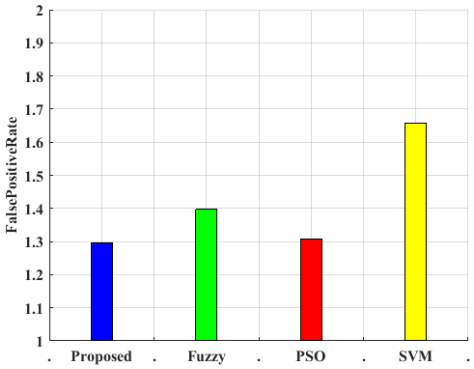


**Fig. 5: Analysis of False Positive Rate**

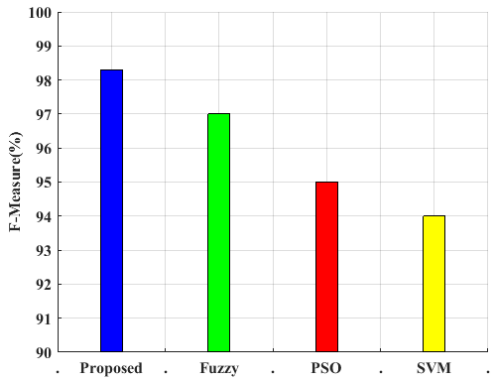


**Fig. 6: F-measure analysis**

In this method, by using deep learning algorithm and GDA high accuracy is obtained with small FPR.

## VI. CONCLUSION

Afast and efficient anomaly detection is a major recent research challenge due to the increasing usage of internet based services. Anomaly detection can effectively help in finding the fraud, finding abnormal action in huge and complex Big Data sets. In this method anomaly based IDS is proposed and in short time interval large amount of network packets are analysed by Apache Hadoop. Hadoop groups are used to maintain large datasets and GDA is introduced to decrease the feature dimension and the anomaly behaviours in dataset are detected using DBN. This method is proposed to obtain less false positive value and high accuracy in practical intelligent IDS and the obtained result shows the accuracy of 98% and the 11% false positive rate.

## REFERENCES


1. M. Bilal, L.O. Oyedele, J. Qadir, K. Munir, S.O. Ajayi, O.O. Akinade, H.A. Owolabi, H.A. Alaka, M. Pasha, "Big Data in the construction industry: A review of present status, opportunities, and future trends", Advanced Engineering Informatics, vol. 30, no. (3), 2016, pp.500-21.
2. Zhang, Ji, H. Li, Q. Gao, H. Wang and Y. Luo, "Detecting anomalies from big network traffic data using an adaptive detection approach", Elsevier, Information Sciences, vol.318, 2015, pp.91-110.
3. T. Huang, H. Sethu and N. Kandasamy, "A new approach to dimensionality reduction for anomaly detection in data traffic", IEEE Transactions on Network and Service Management, vol.13, no. (3), 2016, pp.651-665.
4. X. Wu, X. Zhu, G.-Q. Wu, W. Ding, "Data mining with big data", IEEE transactions on knowledge and data engineering, vol. 26, no. (1), 2014, pp.97-107.
5. Y. Wang, X. Li and X. Ding, "Probabilistic framework of visual anomaly detection for unbalanced data", Elsevier, Neuro computing, vol. 201, 2016, pp.12-18.
6. A.B. Hernández, M.S. Perez, S. Gupta and V. Muntés-Mulero, "Using machine learning to optimize parallelism in big data applications", Elsevier, Future Generation Computer Systems. 2017.
7. S. Suthaharan, "Big data classification: Problems and challenges in network intrusion prediction with machine learning", ACM SIGMETRICS, Performance Evaluation Review, vol.41, no. (4), 2014, pp.70-73.
8. G. Quellec, M. Lamard, A. Erginay, A. Chabouis, P. Massin, B. Cochener and G. Cazuguel, "Automatic detection of referral patients due to retinal pathologies through data mining", Elsevier, Medical image analysis, 29, 2016, pp.47-64.
9. R. Ali, S. Lee and T.C. Chung, "Accurate multi-criteria decision making methodology for recommending machine learning algorithm", Elsevier, Expert Systems with Applications, vol. 71, 2017, pp. 257-278.
10. A.P. Vela, M. Ruiz and L. Velasco, "Distributing data analytics for efficient multiple traffic anomalies detection", Elsevier, Computer Communications, vol.107, 2017 pp.1-12.
11. W. Li, G. Wu and Q. Du, "Transferred Deep Learning for Anomaly Detection in Hyperspectral Imagery", IEEE Geoscience and Remote Sensing Letters, vol. 14, no. (5), 2017, pp.597-601.
12. M. Uchida, "Human error tolerant anomaly detection based on time-periodic packet sampling", Elsevier, Knowledge-Based Systems, vol.106, 2016, pp.242-250.
13. R.A.R. Ashfaq, X.-Z. Wang, J.Z. Huang, H. Abbas and Y.-L. He, "Fuzziness based semi-supervised learning approach for intrusion detection system", Elsevier, Information Sciences, vol. 378, 2017, pp.484-497.
14. S. Fong, R. Wong and A.V. Vasilakos, "Accelerated PSO swarm search feature selection for data stream mining big data", IEEE transactions on services computing, vol. 9, no. (1), 2016, pp.33-45.
15. A. Karami and M. Guerrero-Zapata, "A fuzzy anomaly detection system based on hybrid pso-kmeans algorithm in content-centric networks", Elsevier, Neurocomputing, vol. 149, 2015, pp.1253-1269.
16. M. Muja and D.G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. (11), 2014, pp.2227-2240.
17. Z. Deng, X. Zhu, D. Cheng, M. Zong and S. Zhang, "Efficient kNN classification algorithm for big data", Elsevier, Neurocomputing, vol. 195, 2016, pp.143-148.
18. R. Al, M. Mohamad, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani and R.R. Yager, "Deep learning approach for active classification of electrocardiogram signals", Elsevier, Information Sciences, vol. 345, 2016, pp.340-354.
19. S. Kanarachos, S.-R.G. Christopoulos, A. Chroneos and M.E. Fitzpatrick, "Detecting anomalies in time series data via a deep learning algorithm combining wavelets, neural networks and Hilbert transform", Elsevier, Expert Systems with Applications, vol. 85, 2017, pp. 292-304.
20. A.R. Revathi and D. Kumar, "An efficient system for anomaly detection using deep learning classifier", Springer, Signal, Image and Video Processing, vol. 11, no. (2), 2017, pp.291-299.
21. A.A. Aburomman and M.B.I. Reaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system, Elsevier", Applied Soft Computing, vol. 38, 2016, pp.360-372.
22. S. Garg and S. Batra, "Fuzzified Cuckoo based Clustering Technique for Network Anomaly Detection", Elsevier, Computers & Electrical Engineering. 2017.
23. S.M. Erfani, S. Rajasegarar, S. Karunasekera and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning", Elsevier, Pattern Recognition, vol. 58, 2016, pp.121-134.
24. M.A. Ambusaidi, X. He, P. Nanda and Z. Tan, "Building an intrusion detection system using a filter-based feature selection algorithm", IEEE transactions on computers, vol. 65, no. (10), 2016, pp.2986-2998.
25. N. Pandeeswari and G. Kumar, "Anomaly detection system in cloud environment using fuzzy clustering based ANN", Springer, Mobile Networks and Applications, vol. 21, no. (3), 2016, pp.494-505.
26. Wang, Hao, Yuanyuan Fan, Baofu Fang, and Shuanglu Dai. "Generalized linear discriminant analysis based on euclidean norm for gait recognition." International Journal of Machine Learning and Cybernetics 9, no. 4 (2018): 569-576.
27. Pandey, Vaibhav, and Poonam Saini. "An Energy-Efficient Greedy MapReduce Scheduler for Heterogeneous Hadoop YARN Cluster." In International

Conference on Big Data Analytics, pp. 282-291. Springer, Cham, 2018.
28. Wang, Shitong, Fu-Lai Chung, and Xiongtao Zhang. "An Interpretable Fuzzy DBN-based Classifier for Indoor User Movement Prediction in Ambient Assisted Living Applications." IEEE Transactions on Industrial Informatics (2019).

## AUTHORS PROFILE

**Priyanka Dahiya** received the Master's degree from Sikkim Manipal Institute of Technology, Sikkim, India, in 2011 in Computer Science and Engineering. She is currently pursuing the Ph.D. degree with Manipal University Jaipur, Rajasthan, India. She is currently working in Mody University of Science and Technology. Her current research interests include data mining, Operating System and big data.

**Dr. Devesh Kumar Srivastava** he is working as a Professor with the department of School of Computing & Information Technology, Manipal University Jaipur, Rajasthan, India**.**