Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July, 2021:1802 – 1819

# Synthetic Minority Oversampling And Levy Flight Grey Wolf Optimization For Big Data Classification

S.Saravanabavanandam, Dr.S.Duraisamy

## Abstract

Primary disease detection, patient care and community services are greatly benefited through precise medical data exploration which in turn is greatly attained due to tremendous big data evolution in biomedical and healthcare communities. Thereby data volume, velocity, and variation also increase promptly. Henceforth, massive data storing, processing and visualising through classic approaches is a challenging issue. For mitigating these issues, improved frame work for big data classification where feature selection is performed via adaptive cuckoo search besides classification by weighted convolutional neural network. Conversely, class imbalance problem is yet another challenge for dataset considered in this research which is not concentrated in prevailing research, in addition classifier generalisation ability are also affected. Also, local optimal solution is yet another difficulty faced due to cuckoo search. Synthetic Minority Oversampling Technique (SMOTE) is greatly utilized for mitigating class imbalance problem. The dataset are balanced where numbers of instances are increased in minority class along with suggested model error rate reduction. Levy flight grey wolf optimization is greatly involved in feature selection for computation time reduction and thereby classification accuracy is improved. Lastly big data classification is achieved through weighted convolutional neural network. The suggested model is validated through experimental outcomes where effectiveness is demonstrated pertaining to precision, recall and accuracy for Covtype, ECBDL14-S and Poker database.

**Keywords**: HealthCare, Weighted Convolutional neural network, Cuckoo search, Synthetic Minority Oversampling, grey wolf optimization, levy flight and feature selection.

---

[1]Ph.D Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore- 641 046
Bavanandam@yahoo.com

[2]Assistant Professor, Department of Computer Science, Chikkanna Government Arts College, Tirupur, Tamil Nadu. Sdsamy.s@gmail.com

## 1. INTRODUCTION

Big data is viewed as a massive data group which is complicated in processing by means of classical data processing methods as well as platforms. Alternatively, massive data storing, processing and visualising through classic approaches is a challenging issue. Presently, data generation bases are enlarged intensely, for instance sensor networks, streaming machines,

telescopes, streaming machines, large throughput instruments, etc., and these massive data are necessitated by these platforms[1,2].

Recently Big Data finds its application in various fields like healthcare, scientific research, natural resource management ,business organization, industry, social networking and. Massive data storage are difficult for classical database systems not only in terms of dimension however also variation in data types like Video, Text, Audio and some other format [3,4].

ANN has grabbed many researcher attentions recently in big data classification, pattern recognition and regression. ANNs learning process and parameters optimization extremely have sensitive impact on its performance. Most extensively used ANNs is multilayer perceptron (MLP) neural networks in which two central supervised training techniques categories are present: gradient-based and stochastic methods. The back-propagation algorithm and its variants are deliberated as gradient-based methodswhich is a classic instance.

Nonetheless, gradient-based approaches have three key disadvantages: highly reliant on initial values, effortlessly trapped in local optima, and comparatively slow when it moves toward to convergence. Local optima avoidance is yet another challenge for furthermost machine learning algorithms. Local optimum mentions an optimal solution contained by candidate neighbors set that might mistakenly deliberate as global optimum. The number of iterations with dissimilar preliminary solutions an optimizer requires for attaining global optimum[5,6,7] is another issues due to high dependency on initial solutions.

For mitigating these issues, improved frame work for big data classification where feature selection is performed via adaptive cuckoo search besides classification by weighted convolutional neural network. Conversely, class imbalance problem is yet another challenge for dataset considered in this research which is not concentrated in prevailing research, in addition classifier generalisation ability are also affected. Also, local optimal solution is yet another difficulty faced due to cuckoo search.

Synthetic Minority Oversampling Technique (SMOTE) is greatly utilized for mitigating class imbalance problem. The dataset are balanced where numbers of instances are increased in minority class along with suggested model error rate reduction. Levy flight grey wolf optimization is greatly involved in feature selection for computation time reduction and thereby classification accuracy is improved. Lastly big data classification is achieved through weighted convolutional neural network.

The paper is organized in five sections. An overview of big data classification in healthcare system is given in section I. Various methods for big data classification is offered in section II. Design approach for suggested big data classification model is elucidated in Section III. Experimental analyses along with manifold outcomes evaluation is explained in section 4. Conclusion along with future scope is discussed in section V.

## 2. Literature Review

An reviews about various approaches for big data classification is outline here.

Duan et al [8] utilized Spark framework (SELM) for designing a proficient ELM encompassing three parallel sub algorithms for big data classification. Almost all computations are performed locally through conforming data sets partitioning via algorithm such as hidden layer output matrix, matrix Û decomposition, and matrix V decomposition. Simultaneously, intermediate outcomes are retained in disseminated memory besides diagonal matrix are cached as broadcast variables rather than numerous copies for every task for costs reduction, and these activities

supports in strengthening SELM learning ability. As a final point, SELM algorithms are implemented for massive data sets classification. The suggested algorithm effectiveness is substantiated through extensive experimentation. SELM attains an 8.71× speedup on a cluster with ten nodes, 13.79× speedup with 15 nodes, 18.74× speedup with 20 nodes, 23.79× speedup with 25 nodes, 28.89× speedup with 30 nodes, and 33.81× speedup with 35 nodes.

Nair, et al [9] established distant health status prediction system in real time constructed around open source Big Data processing engine namely Apache Spark, positioned in cloud highlighting machine learning model on streaming Big Data. Health attributes are tweeted by user in this scalable system as well as the same are received by applications, attribute extraction and utilizing machine learning model for user's health status prediction which is then directly messaged directly for making appropriate action.

Hadi et al [10] suggested a scheme for Out-Patient (OP) centric Long Term Evolution-Advanced (LTE-A) network optimization. Big data gathered from OPs' medical histories, accompanied by recent observation from their body-connected medical IoT sensors processing are done and investigated for life-threatening medical condition probability prediction, such as, an impending stroke. This forecast is utilized for ensuring OP assignment of an optimal LTE-A Physical Resource Blocks (PRBs) is done for their acuteinformation transmission to their healthcare source with nominal delay. Weighted Sum Rate Maximization (WSRMax) and Proportional Fairness (PF) methodology are greatly utilized in this system. OPs' average SINR is increased by 26.6% and 40.5%, correspondingly using this system. The system's total SINR is increased to a greaterlevel by WSR Max methodology than PF methodology, conversely, PF approach described higher SINRs for OPs, superior fairness and a lesser error margin.

Galicia et al [11] predicted big data time series by suggesting ensemble models which comprises three techniques (decision tree, gradient boosted trees, random forest) since these techniques ensues better outcomes in earlier application. The ensemble weights computations are done through weighted least square technique. A static or dynamic ensemble model is developed by considering two strategies correlated to weight update. Every ensemble member predictions are found by fragmentation of forecasting problem into h forecasting sub-problems, corresponding toevery prediction horizon value. Machine learning algorithm are greatly utilized for solving these sub-problems from big data engine Apache Spark, confirming system scalability. The suggested model performance evaluation is done on Spanish electricity consumption data for 10 years estimated with a 10-minute frequency. The outcomes reveal that both dynamic and static ensembles are executed properly, outclassing distinct ensemble members by combination.

Wan et al [12] suggested big data solution for active defensive maintenance in manufacturing atmospheres by providing system architecture initially. Based on data characteristics, technique is analysed for manufacturing big data collection. Later, data processing is perfomed in cloud comprising cloud layer architecture, real-time active maintenance mechanism, along with offline prediction and analysis technique. Lastly, a prototype platform is examined through comparison of conventionally used technique with suggested dynamicdefensive maintenance technique. Thereby Industry 4.0 implementation is greatly accelerated by manufacturing big data schemedeployed for dynamicdefensive maintenance.

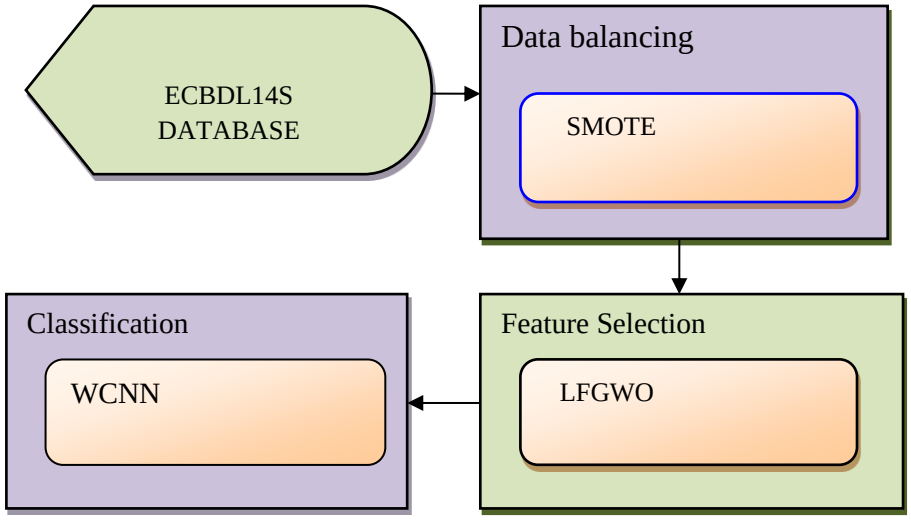
Yassine, et al [13] suggested learning as well as discovering human activity patterns system for health care uses by utilizing smart home big data. Also frequent pattern mining, cluster exploration, and prediction are deployed for estimating and analysing energy usage changes sparked through occupants' behavior. People's habit identification are done routinely by

discovering these customs allows in identifying anomalous activities which might signify people's problems in taking self-care, such as no food preparation or not via a shower/bath. Alsoemporal energy consumption patterns are analysed at appliance level which is correlated to human activities in direct way. The suggested mechanism assessment is performed using U.K. Domestic Appliance Level Electricity data set-time series data of power consumption collected from 2012 to 2015 with 6 s time resolution for five houses with 109 appliances from Southern England. Recursive mining of data from smart meters in 24 h quantum/data slice and outcomes are sustained across successive mining exercises. In addition to suggested method outcomes, shortand long-term predictions accuracy is also provided.

Kumar, et al [14] utilized condition-based maintenance (CBM) optimization for designing a big data analytics structure for maintenance optimization in addition also enhances prediction accuracy for quantifying left over life prediction uncertainty. By condition monitoring along with prediction information effective utilization, CBM might boost equipment reliability leading to drop in maintenance cost. A new linguistic interval-valued fuzzy reasoning technique for information prediction is achieved using this CBM optimization technique. Experimental outcomes are achieved on a big dataset which is produced from gas turbine propulsion plant simulator. It is substantiated through relative examination that suggested method performs in a better way than conventional approaches pertaining to classification accuracy and other statistical performance assessment metrics.

## 3. PROPOSED METHODOLOGY

The suggested framework for big data classification is deliberated here. The model mainly comprises three phases. First one is data balancing using Synthetic Minority Oversampling Technique (SMOTE), second one is feature selection on levy flight grey wolf optimization basis and third one is classification through weighted convolutional neural network. Overall proposed model architecture is revealed in figure: 1.




### 3.1. CLASS BALANCING USING SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

ECBDL14S cancer dataset is the input which is categorized under imbalanced data and oversampling is greatly utilized in pre-treatment process. Oversampling in data analysis are techniques mainly meant for adjusting dataset class distribution which is utilized for increasing number of minority class samples by arbitrarily copying minority samples, with the purpose of balancing minority class and the majority class example size. SMOTE (Synthetic Minority

Oversampling Technique) is widely used oversampling methods for solving imbalance issues, generating synthetic training examples via linear interpolation for minority class. SMOTE algorithm mainly involves two stages [15, 16, 17].

During initial stage, k nearest neighbors are obtained through Euclidean distance calculation from every minority data pertaining to all other minority data and ascending order arrangement is done, following which k lowest distance data are considered as nearest neighbors (kNN). The definition for Euclidean distance among one minority data (x) and another minority data (y) from first attribute to n (maximum number of attributes) is given below

$$D(x, y) = \sqrt{\sum_{a=1}^{n}(x_a - y_a)^2} \quad (1)$$

In second stage, synthetic data generation are done through interpolation technique amid two minority data. Its kNN will be randomized to be candidates in synthetic data generation process [18,19]. Then, original minor data (x) and one chosen candidate (y) is utilized for generating new synthetic data amid x and y. Synthetic data between x and y for a-th attribute definition is given in (2)

$$\text{SyntheticData}\quad a(x, y) = x_a + r \cdot (x_a - y_a) \text{ for } 0 \le r \le 1 \quad (2)$$

Where,

r- random number amid 0 and 1

This is useful for n attributes. The process repetition is done until preferred synthetic data amount is attained.

### 3.2. **FEATURE SELECTION USING LEVY FLIGHT GREY WOLF OPTIMIZATION**

Once after data balancing important features are greatly necessitated since dataset possesses number of features as well as more time for computation is consumed. Levyflight grey wolf optimization is chiefly utilized for feature selection by mitigating those issues.

One among meta-heuristic algorithm on basis of leadership hierarchy mathematical model and gray wolves hunting process is the Grey Wolf Optimization (GWO). By and large, grey wolves live in a 5-12 groupsize and tend to possess firm social dominant hierarchy.Based on dominance, there are four groups specifically; alpha ($\alpha$), beta ($\beta$), delta ($\delta$), and omega ($\omega$) as[20,21,22,23].

Group leaders meant for hunting are called as alpha grey wolves and all other group members ensue their orders. Subsequent hierarchylevel encompasses grey wolves who are subordinates to alpha supports in attaining decisions. Delta is succeeding level comprising grey wolves following α and β groups and leading lowest level group termed as omega. Delta group members may function as scouts, sentinels, elders, hunters, and caretakers. Omegas are submissive to α, β, and δ wolves and hence α, β, and δ are significant wolves in group hunting process. Group hunting is yet othergrey wolves social activity including key phases such as tracking, encircling, and attacking prey[24].

GWO helps in mathematical modelling of social hierarchy and Grey wolves hunting procedure for optimization issue as conferred in subsequent segment.

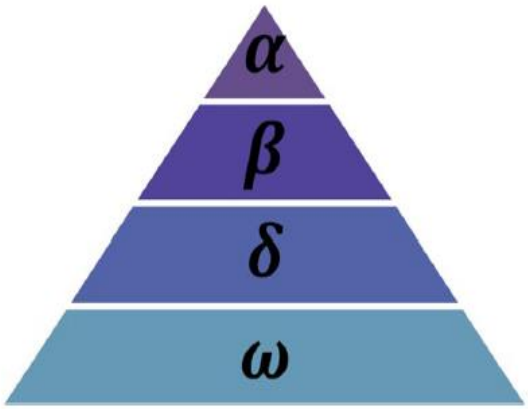



**Figure: 2** Grey wolf Dominance hierarchy, increasing from bottom to top

**Social hierarchy**

Grey wolves social hierarchy is mathematically modeled by taking into account first fittest, second fittest, and third fittest solution as alpha ($\alpha$), beta ($\beta$), and delta ($\delta$) correspondingly. And remaining are named as omega ($\omega$).

**Encircling Prey**

Encircling process mathematical modelling is given by

$$D = \left|C.X_p(t) - X(t)\right|$$

$$X(t+1) = X_p(t) - A.D.$$

where, $\mathbf{X_p}(t)$ and $\mathbf{X}(t)$ represents prey and grey wolf position at iteration $t$ . **A** and **C** denotes Coefficient vectors, computed using (4) and (5) correspondingly.

$$A = 2a.r_1 - a$$

$$C = 2.r_2$$

Here $\mathbf{r_1}$and $\mathbf{r_2}$signify random vectors in [0, 1]. **a**refers to factor which decreases linearly from 2 to 0 over iterations course. This factor is utilized for controlling Grey wolf step size (**D**).

**Hunting**

Generally, hunting guidance is given by alpha besides facilitated by beta and delta. Consequently, for this behaviour simulation, first three best solutions ($\alpha$, $\beta$, $\delta$) in an iteration are utilized by other search agents (omegas) for their positions update through subsequent equations:

$$\vec{D}_\alpha = \left|\vec{C}_1 . \vec{X}_\propto - \vec{X}\right| , \vec{D}_\beta = \left|\vec{C}_2 . \vec{X}_\beta - \vec{X}\right|, \vec{D}_\delta = \left|\vec{C}_2 . \vec{X}_\delta - \vec{X}\right|$$

$$\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1 . \left(\vec{D}_\alpha\right), \vec{X}_2 = \vec{X}_\beta - \vec{A}_2 . \left(\vec{D}_\beta\right), \vec{X}_3 = \vec{X}_\delta - \vec{A}_3 . \left(\vec{D}_\delta\right)$$

$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3}$$

**Attacking prey (exploitation)**

Grey wolves hunting is accomplished through attacking prey and they stop moving.This phenomena implementation is done by decreasing **A** value which is determined by **a**. The **'a'** value decreases linearly from 2 to 0 and it has a significant role for exploiting search space whereas trivial exploration leads to trapping into local optimum.

**Search for prey (exploration)**

For enhancing GWO divergence behavior, **A**value is taken either greater than 1 or less than -1 which in turn increases GWO exploration capability. Additionally, C **C**utilizearbitrary values in all iterations revealing superior exploration not only in the course of initial iterations however also in final iterations.

The two significant optimization algorithm features are exploration and exploitation and their good compromise supports in attaining precise solution via escaping local optima. In GWO, search agent step size linearly decreases with iterations. **A** refers to parameter deployed for controlling this step size. Nonetheless, it is revealed that at far ahead iterations GWO tend to possess trapping into local optimum restriction because of poor divergence.

Levy flight is suggested for mitigating this issues for value modification which also improves GWO exploration and exploitation capability concurrently.

Levy flight utilized Levy probability distribution function which is function of power-law and meant for jump size computation. Levy distribution is as follows.

$$\mathrm{L}(s,\gamma,\mu) = \begin{cases} \sqrt{\frac{\gamma}{2\pi}} \exp\left[-\frac{\gamma}{2(s-\mu)}\right] \frac{1}{(s-\mu)^{3/2}} if\ 0 < \mu < \infty \\ 0 \qquad\qquad if\ s \le 0 \end{cases}$$

Where $\mu$, $\gamma$, and $s$ notates position parameter, scale parameter for controlling distribution scale and samples collection in this distribution correspondingly.

LFGWO pseudo code is offered in Algorithm:1.

**Algorithm: 1. Levy flight Grey Wolf Optimization (GWO)**
**Input: N** search agents $X_i$ $(i = 1,2,\ldots,n)$ having dimensions.
**Ouput :**Best solution ($X_S$ ).
Arbitrarily initialize N search agents initial population
Initialize a, A, and C value
Assess each search agent fitness
$X_S$= best search agent
$X_\gamma$= second best search agent
$X_\mu$= third best search agent
**while**stopping criteria is not satisfied **do**
**for**i=1 to N **do**
Update $X_i$ positionby (8)

**end for**
Modify a, A, and C
Estimate all search agents fitness
Modify $X_S X_\gamma \; and \; X_\mu$
**end while**
return

### 3.3. CLASSIFICATION USING WEIGHTED CONVOLUTIONAL NEURAL NETWORK

Post feature selection process, Convolutional neural network (CNN) is employed to carry out big dataclassification. Unlike typicalArtificial Neural Network (ANN) structure, a CNN varies, since the input is flatten to a vector in traditional ANN, whereas the layers of a CNN can be selected, through which the input data can be matched spatially. In standard CNN, single/multiple blocks of convolution and sub-sampling layers are involved, besides one/multiple fully connected layer(s) and an output layer.

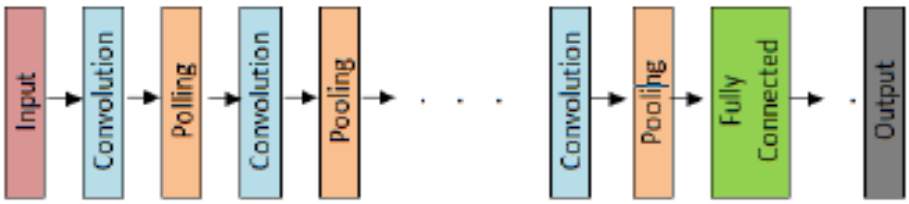



**Figure: 3.Typical CNN**

**Drawbacks of Traditional CNN**

Information loss may occur frequently through dimension reduction, since traditional CNN architecture employs pooling operation during the process of dimension reduction. Hence, weighted CNN is utilized in this work to resolve this problem.

**Weighted CNN**

There are three kinds of layers involved in CNN, such as convolution layer, sub sampling layer as well as fully connected layer. In Figure 1, a standard CNN architecture is depicted, besides the following segments summarize each layer type.

**Convolution layer**

The selected features are considered as an input in this proposed study. During the process of convolution layer, an input features are convolved with a kernel (filter). The *n* output features maps are generated through input feature and kernelconvolution. In general, a kernel of the convolution matrix is signified as a filter, whereas output features derived through convolving kernel and e input is defined as size i*ifeature maps.

CNN is capable of including numerous convolutional layers, besides inputs and outputs of following convolutional layers are feature vectors. In every convolution layer, a group of n filters are included that are convolved with input, besides generated feature maps(*n**)depth and number of filters functional during convolution operation are equal. Every filter map is taken as a unique feature at specific inputlocation .

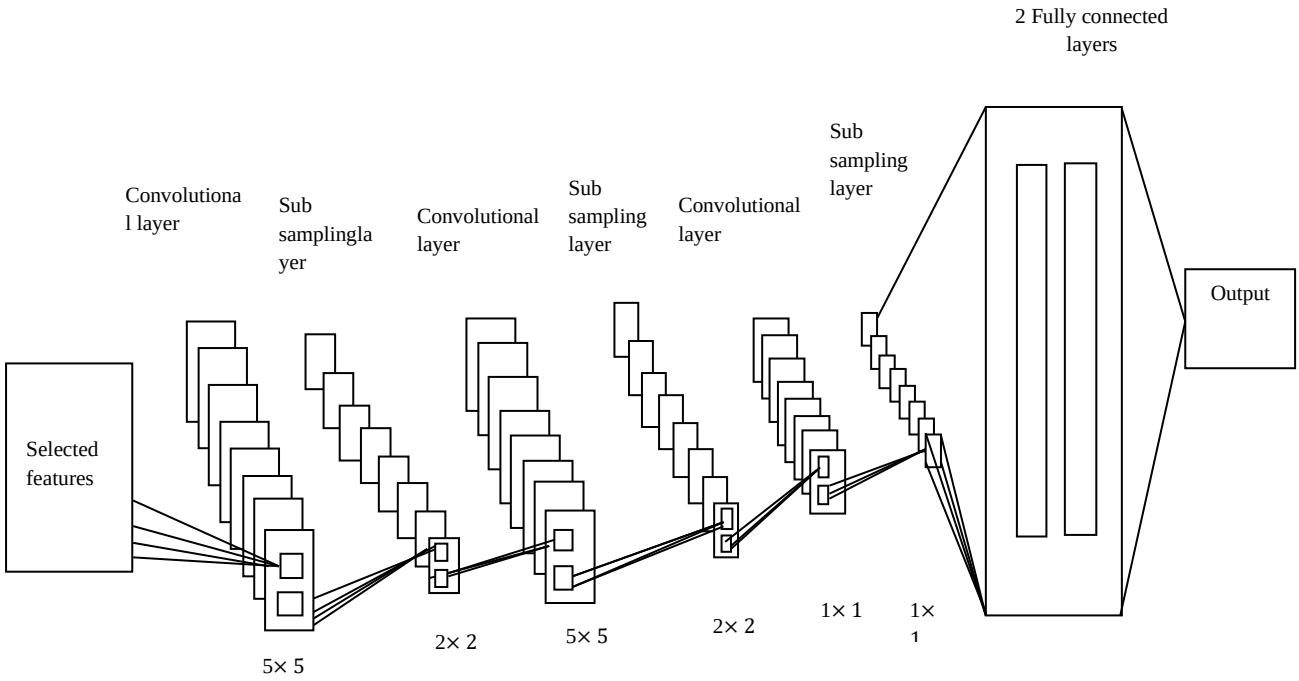



**Figure.4: Convolutional Neural Network architecture**

Here, $C_i^{(l)}$ signifies the *l*-th convolution layer output that includes feature maps. It can be estimated as follows,

$$C_i^{(l)} = B_i^{(l)} + \sum_{j=1}^{a_i^{(l-1)}} K_{i,j}^{(l-1)} * C_j^{(l)} \quad (9)$$

In which, the bias matrix is signified by $B_i^{(l)}$; a convolution filter/kernel of size *a*a*is denoted by $K_{i,j}^{(l-1)}$, through which the *j*-th feature map in layer $(l-1)$ is connected with *i*-th feature map in same layer.

In output $C_i^{(l)}$ layer, feature maps are comprised. In Equation 10, first convolutional layer $C_i^{(l-1)}$is input space, which can be notated as $C_i^{(0)} = X_i$. The feature map is generated by the kernel. Post process of convolution layer, a nonlinear transformation of theconvolutional layer outputs can be carried out by applying the activation function.

$$Y_i^{(l)} = Y(C_i^{(l)}) \quad (10)$$

Here, the output of the activation function is denoted by $Y_i^{(l)}$, the input received by the activation function is signified by $C_i^{(l)}$.

**Sub sampling or pooling Layer**

This layer tends to spatially decrease the features map dimensionality that isderived from the preceding convolution layer. Subsequently, sub sampling operation is carried out amid the mask and feature maps. Even though various sub sampling approaches, like averaging pooling, sum pooling, and maximum pooling have been proposed, application of max pooling

is widely recognized, in which each block's highest value is taken as the corresponding output feature. The convolution layer is eased using a sub sampling layer in toleration of rotation and translation amid input images.

**Fully Connected layer**

The last CNN layer is a conventional feed forward network with one/multiple hidden layers. The following Softmax activation functionis applied by the output layer.

$$Y_i^{(l)} = f(z_i^{(l)}),$$

$$\text{Where } z_i^{(l)} = \sum_{i=1}^{m_i^{(l-1)}} w_{i,j}^{(l)} y_i^{(l-1)} \quad (11)$$

Here, $w_{i,j}^{(l)}$ indicates the weights which needs to be adjusted by the completely connected layer to form every classdepiction, whereas transfer function is denoted by $f$, through which the nonlinearity is represented. Unlike convolutions and pooling layers, the nonlinearity is constructed between its neurons in the fully connected layer, not in different layers.

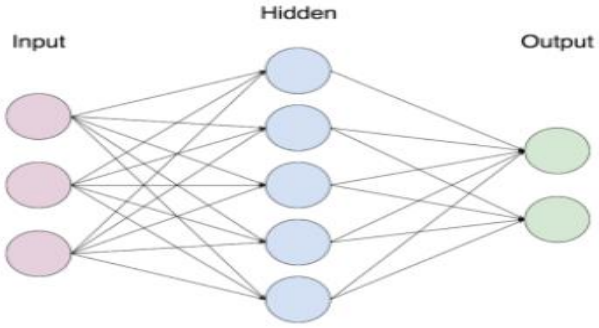



**Figure:5.** Fully Connected Layer

The fuzzy membership function can be applied to estimate the aforementioned equation weights, i.e. $w_1 = 0.3, w_2 = 0.4, w_3 = 0.5, w_4 = 0.7$, in addition the following equation can be applied to estimate them.

$$\boldsymbol{o^2 = u_i^{(j)}\left(a_i^{(2)}\right)}$$

Here, a membership function $\boldsymbol{u_i^{(j)}(.): R \longrightarrow [0, 1]}$, i=1,2,…,M, j =1,2,….,N is signified by $\boldsymbol{u_i^{(j)}(.)}$by using Gaussian membership function.

## 4. RESULTS AND DISCUSSION

Thesuggested method experiment outcomesis scrutinized based on the simulation performed under MATLAB. By considering the parameters, such as Precision, Recall, Accuracy, F-Measure as well as Error Rate, suggested WCNN-SMT (Weighted Convolutional Neural Network with SMOTE) and the existingvariable HMM, FKNN and WCNN algorithm performances are compared as regards the implementation done for the COV

(https://archive.ics.uci.edu/ml/datasets/covertype), ECBDL14s (https://archive.ics.uci.edu/ml/datasets/Dermatology) and poker databases (https://archive.ics.uci.edu/ml/datasets/Poker+Hand).

In COV database, only forest cover type is predicted from cartographic variables. From US Forest Service (USFS) Region 2 Resource Information System (RIS) data, the actual forest cover type was determined for a specified observation (i.e. 30 x 30 meter cell). From the data that is originally acquired from US Geological Survey (USGS) and USFS data, the independent variables were derived. For qualitative independent variables (wilderness areas and soil types), data is in raw form (not scaled) that consist of binary (0 or 1) datacolumns . There are four wilderness areas considered in this work, which is situated in Roosevelt National Forest of northern Colorado. Instead of forest management practices, a result of ecological processes is focused by the existing forest cover types, since these locations are the forests, where the interruptions caused by human are very low.

In the ECBDL14s database, there are 34 attributes involved, out of which 33 are linear valued, whereas the remaining one is nominal. In the area of dermatology, it is highly challenging to diagnose the type of erythemato-squamous diseases, since all of them are mutually associated with regard to erythemaclinical features and scaling accompanied by extremely minor differences. Pityriasisrosea, pityriasisrubrapilaris, seboreic dermatitis, psoriasis, cronic dermatitis, besides lichen planus are the few examples of the diseases involved in this group. Since all these diseases have several common histopathological features, performing the diagnosis using biopsy becomes inadequate. Moreover, during the initial stage differential diagnosis, the features of one disease may resemble another, besides it may show the specific features at the subsequent stages. Based on the 12 features, the patients are clinically assessed, for which the skin samples are collected, concerning the evaluation of 22 histopathological features. Further, the samples are analysed under a microscope for determining the values of histopathological features. For this domain, the dataset is explicitly developed,whereif any of these diseases has been perceived in family, family history feature has value 1, if not it will be 0. The patient's age simply represented by the age feature.Both the clinical and histopathological features are set within 0 to 3 range, where the presence of feature is signified by 0; the relative intermediate values are denoted by 1, and 2; and the highest possible value is signified by 3.

**TABLE: 1.PERFORMANCE COMPARISON RESULTS**

| **METRICS** | **METHODS** | **DATABASES** | | |
|---|---|---|---|---|
| Runtime (S) | HMM | COV | ECBDL14S | poker |
| | FKNN | 800 | 825 | 820 |
| | WCNN | 750 | 780 | 790 |
| | WCNN-SMT | 600 | 700 | 680 |
| Accuracy (%) | HMM | 70 | 72 | 73 |
| | FKNN | 75 | 75 | 75 |
| | WCNN | 78 | 79 | 78 |
| | WCNN-SMT | 82 | 85 | 84 |
| | HMM | 75 | 74 | 75 |
| | | | | |
| | FKNN | 76 | 75 | 77 |

| | | | | |
|---|---|---|---|---|
| Precision (%) | WCNN | 79 | 78 | 79 |
| | WCNN -SMT | 81 | 82 | 81 |
| Recall (%) | HMM | 80 | 85 | 82 |
| | FKNN | 82 | 85.5 | 83.3 |
| | WCNN | 85 | 87 | 85 |
| | WCNN-SMT | 87 | 88 | 87 |
| Error rate (%) | HMM | 30 | 28 | 27 |
| | FKNN | 25 | 25 | 25 |
| | WCNN | 22 | 21 | 22 |
| | WCNN-SMT | 18 | 15 | 16 |

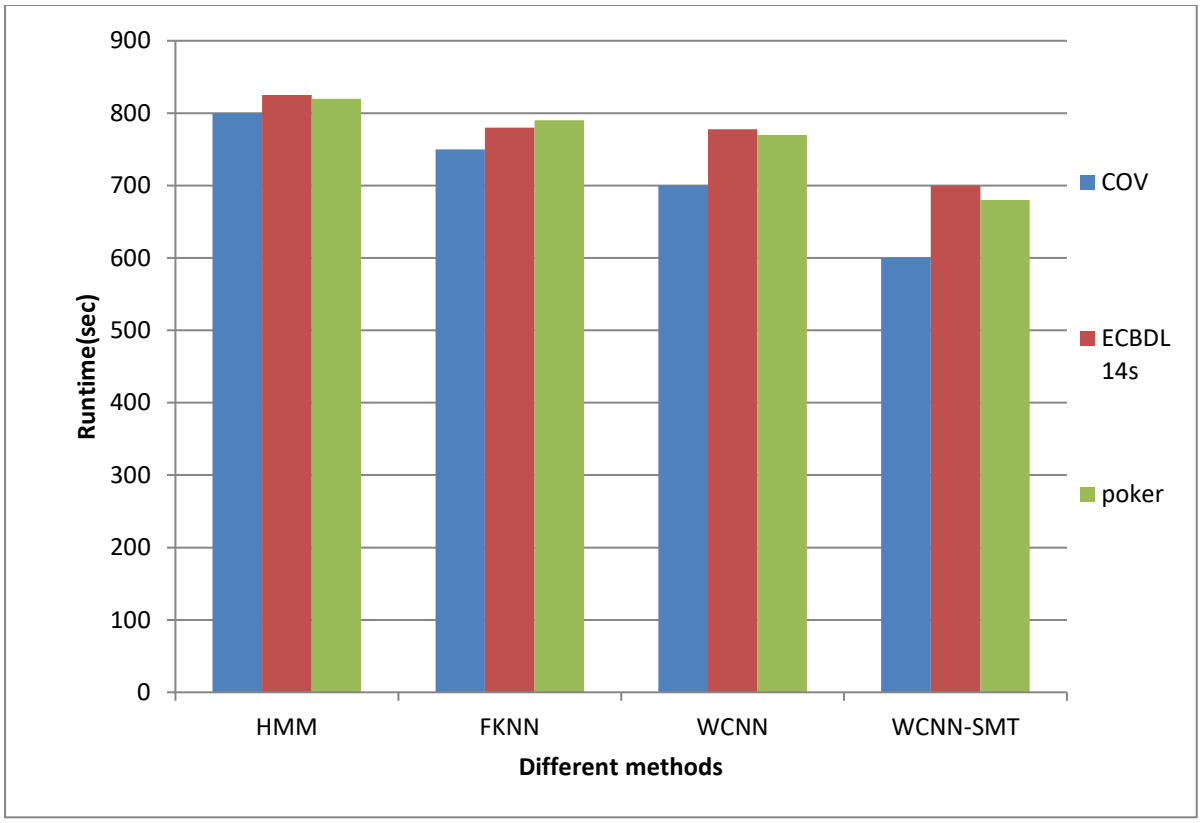



**Figure: 6.Runtime results vs. classification methods**

Figure 6 compares the Runtime values obtained by the proposed WCNN-SMT, and existing HMM, FKNN, WCNN classifiers. In the figure, the implemented methods lie on X-axis, and Y-axis stands for Runtime values. For COV dataset, the graphs depict that the proposed method is capable of providing 600(s) of Runtime, which is comparatively lesser than the existing HMM, FKNN,and WCNN since they solely yields 800(s), 750(s), and 700(s), correspondingly.

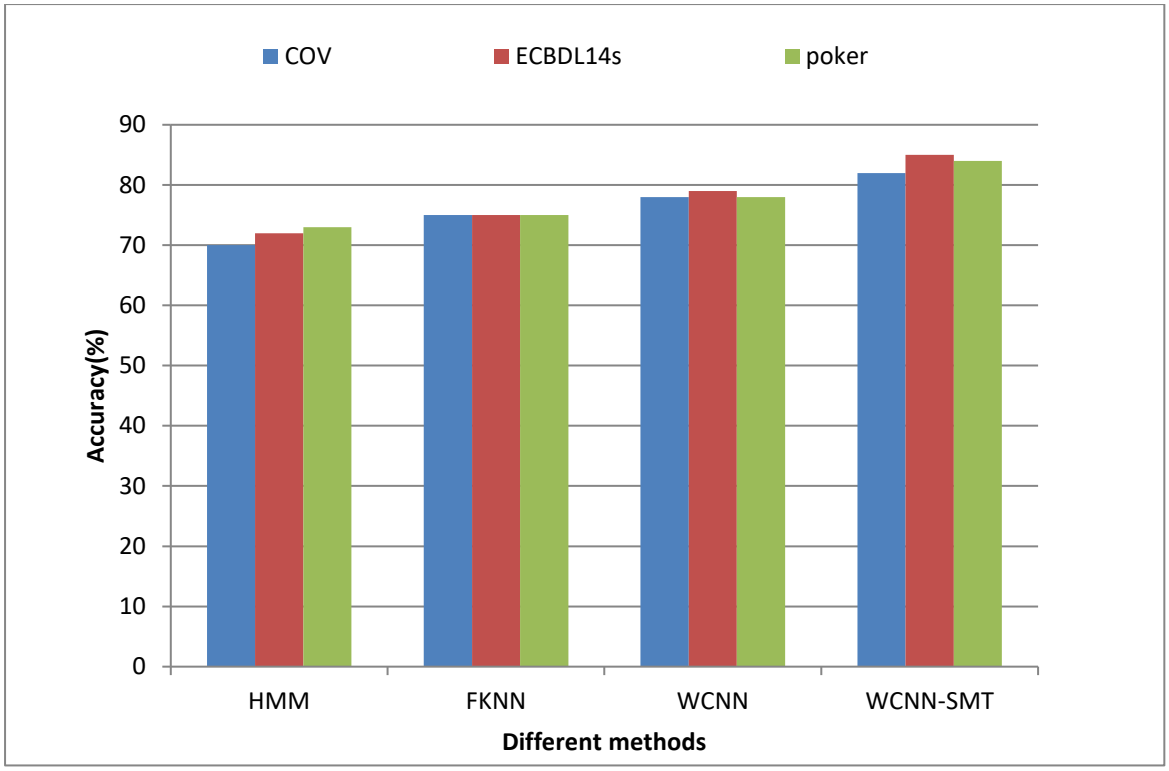



**Figure: 7. Accuracy results vs. classification methods**

The suggested WCNN-SMT Accuracy rate is compared with the existing HMM, FKNN, WCNN classifiers' accuracy rates is shown in figure 7. In the figure, the implemented methods lie on X-axis, and Y-axis stands for Accuracy rates. For COV dataset, the graphs depict the efficiency of the proposed approach toobtain82%Accuracy rate, which is comparatively greater than the existing HMM, FKNN, and WCNN since they obtain only70%, 75%, and 78%, correspondingly. Because, the proposed work employs the grey wolf optimization as a fitness function during the process of significant features selection that improves the accuracy.

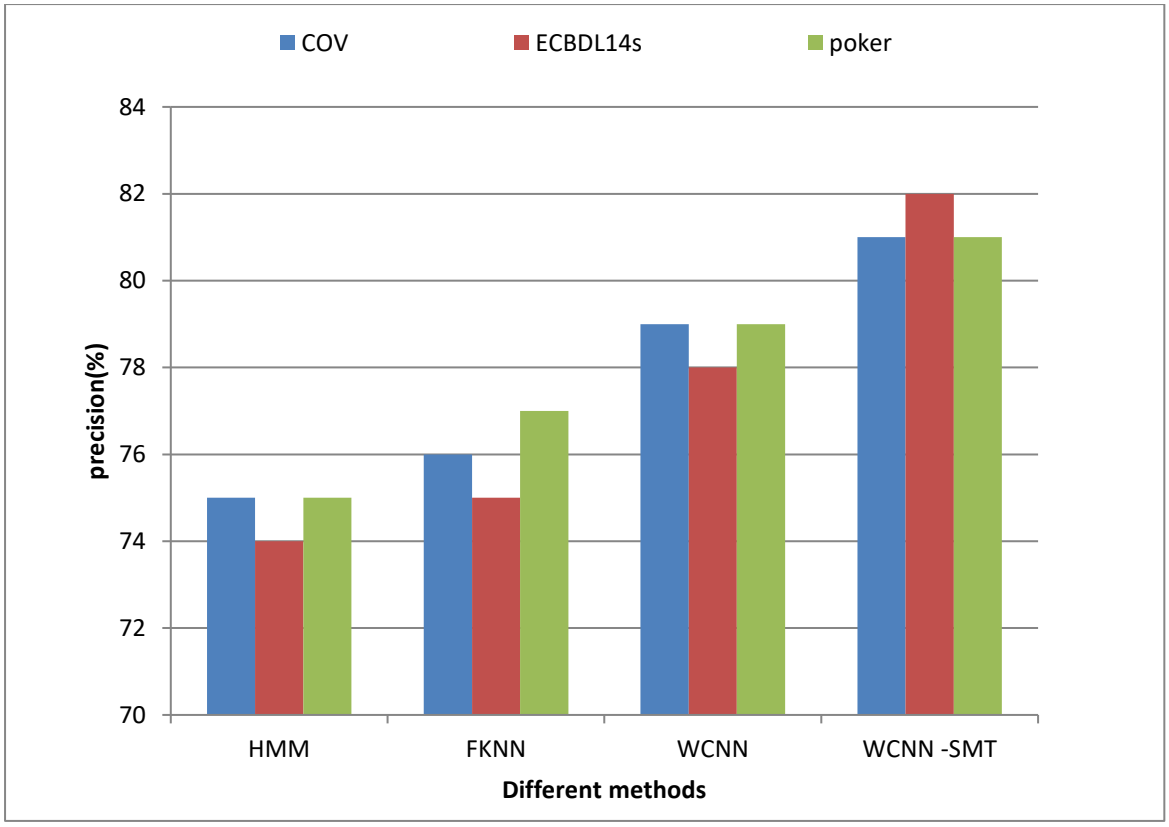



**Figure .8: Precision results Comparison of Various Classifiers**

Figure 8 compares the Precision rates obtained by the proposed WCNN-SMT, and existing HMM, FKNN, WCNN classifiers. In the figure, X-axis represents the implemented methods, and Y-axis stands for precision values. For COV dataset, the graphs prove that the proposed method has the efficiency to deliver 81% precision rate, which is considerably higher than the existing HMM, FKNN, and WCNN as they solely obtain75%, 76% and 79%, correspondingly. The reason is that the proposed method exploits the feature selection as a pre-processing step, which improves the precision rate.

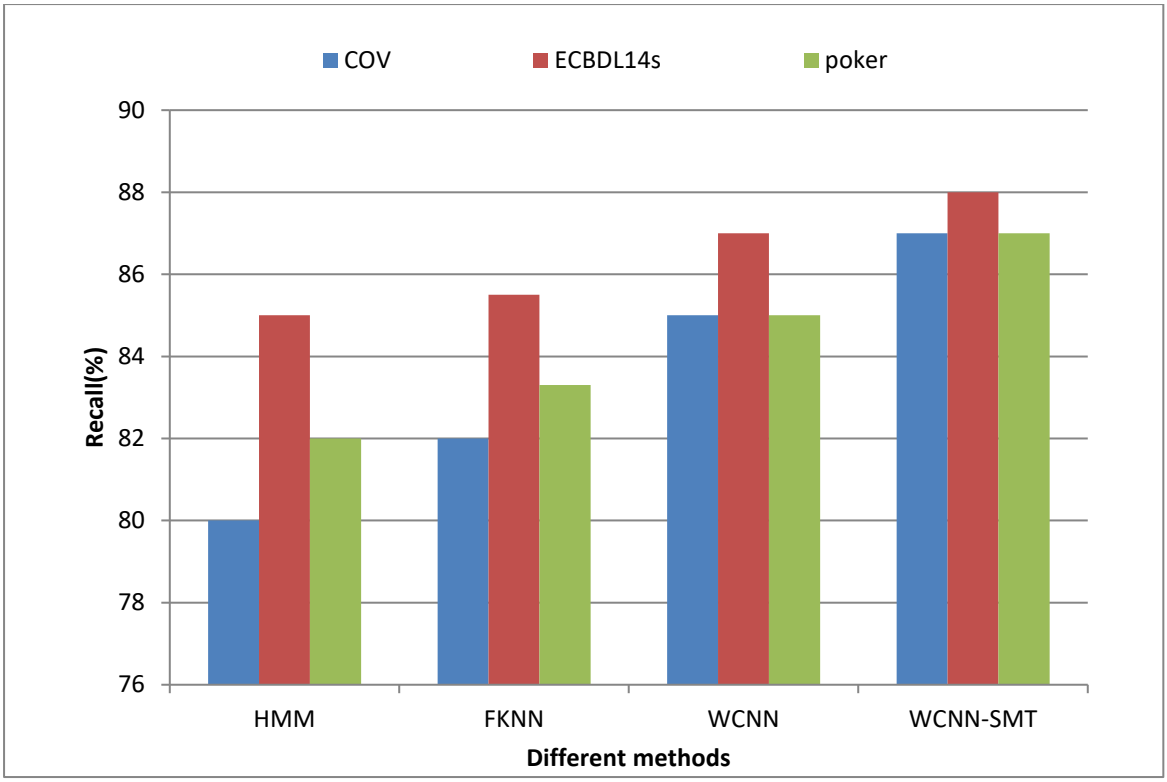



**Figure: 9. Recall results vs. classification methods**

In Figure 9, the individual Recall rates of the proposed WCNN-SMT and the existing HMM, FKNN, WCNN classifiers are compared. In the figure, the various methods lie on X-axis, and Y-axis representsRecallvalues. For COV dataset, the graphs proves thatthe proposed approach has the capability to outperform the existing methods by obtaining 87% recall rate, which is comparatively greater than the others, since HMM, FKNN, and WCNN methods obtain only 80%, 82% and 85%, correspondingly. Because, the proposed work employs fuzzy function for weight value calculation in CNN, through which the recall rate is improved.

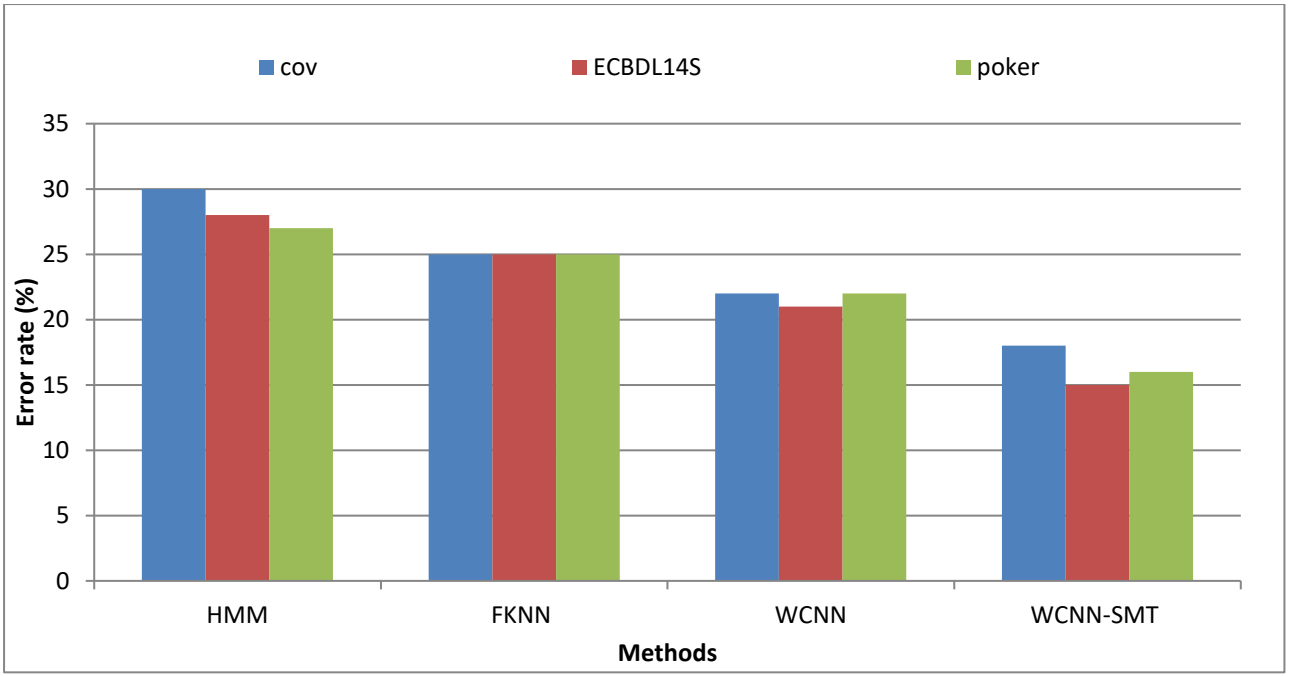



**Figure:10.Error rate result vs. classification methods**

Figure 10 compares the error rate of the proposed WCNN-SMT, and existing HMM, FKNN, WCNN classifiers. In this, methods implemented are represented in X-axis, and Y-axis stands for error rates. For COV dataset, it is substantiated thatsuggestedscheme has the efficiency for error rate reductionup to18%, which is leastwhen compared to the existing HMM, FKNN, and WCNN methods as they solely reduce30%, 25% and 22%, correspondingly.

## 5.CONCLUSION AND FUTURE WORK

In recent days, the data associated to patient, compliance, and record keeping are utilized by the healthcare industry, in which huge quantity of data is included hence it is termed as big data. In emerging digital world, digitization of these data is highly necessitated. Besides, massive data generated desires to be scrutinized in an efficient manner to reduce the costs, as regards healthcare quality improvement and resolving further challenges.In this study, providing an advanced method for the classification of big data is predominantly focused. Accordingly, the oversampling of minority class data is carried out through Synthetic Minority Oversampling Technique (SMOTE) by generating synthetic data. Further, levy flight grey wolf optimization is employed to perform the feature selection, since it effectively diminishes the computation time and improves the accuracy of classification. In addition, big data classification is executed by weighted convolutional neural network. Empirical findings depict suggestedapproach efficiency with regard to Precision, Recall, Accuracy and Reduced Error Rate based on the implementation done for the Covtype, ECBDL14-S and Poker database. Nevertheless, the data considered in this work possesses variance in range, and there exists the possibility of low accuracy in classification.

## REFERENCES:

1. Palanisamy, V. and Thirunavukarasu, R., 2019. Implications of big data analytics in developing healthcare frameworks–A review. Journal of King Saud University-Computer and Information Sciences, 31(4), pp.415-425.
2. Chen, M., Yang, J., Hu, L., Hossain, M.S. and Muhammad, G., 2018. Urban healthcare big data system based on crowd sourced and cloud-based air quality indicators. IEEE Communications Magazine, 56(11), pp.14-20.
3. Ke, H., Chen, D., Li, X., Tang, Y., Shah, T. and Ranjan, R., 2018. Towards brain big data classification: Epileptic EEG identification with a lightweight VGGNet on global MIC. IEEE Access, 6, pp.14722-14733.
4. Alhussein, M. and Muhammad, G., 2018. Voice pathology detection using deep learning on mobile healthcare framework. IEEE Access, 6, pp.41034-41041.
5. Mehta, N. and Pandit, A., 2018. Concurrence of big data analytics and healthcare: A systematic review. International journal of medical informatics, 114, pp.57-65.
6. Pramanik, M.I., Lau, R.Y., Demirkan, H. and Azad, M.A.K., 2017. Smart health: Big data enabled health paradigm within smart cities. Expert Systems with Applications, 87, pp.370-383.
7. García-Gil, D., Luengo, J., García, S. and Herrera, F., 2019. Enabling smart data: noise filtering in big data classification. Information Sciences, 479, pp.135-152.
8. Duan, M., Li, K., Liao, X. and Li, K., 2017. A parallel multiclassification algorithm for big data using an extreme learning machine. IEEE transactions on neural networks and learning systems, 29(6), pp.2337-2351.
9. Nair, L.R., Shetty, S.D. and Shetty, S.D., 2018. Applying spark based machine learning model on streaming big data for health status prediction. Computers & Electrical Engineering, 65, pp.393-399.
10. Hadi, M.S., Lawey, A.Q., El-Gorashi, T.E. and Elmirghani, J.M., 2019. Patient-centric cellular networks optimization using big data analytics. IEEE Access, 7, pp.49279-49296.
11. Galicia, A., Talavera-Llames, R., Troncoso, A., Koprinska, I. and Martínez-Álvarez, F., 2019. Multi-step forecasting for big data time series based on ensemble learning. Knowledge-Based Systems,163, pp.830-841.
12. Wan, J., Tang, S., Li, D., Wang, S., Liu, C., Abbas, H. and Vasilakos, A.V., 2017. A manufacturing big data solution for active preventive maintenance. IEEE Transactions on Industrial Informatics, 13(4), pp.2039-2047.
13. Yassine, A., Singh, S. and Alamri, A., 2017. Mining human activity patterns from smart home big data for health care applications. IEEE Access, 5, pp.13131-13141.
14. Kumar, A., Shankar, R. and Thakur, L.S., 2018. A big data driven sustainable manufacturing framework for condition-based maintenance prediction. Journal of Computational Science, 27, pp.428-439.
15. Feng, W., Huang, W., Ye, H. and Zhao, L., 2018, July. Synthetic Minority Over-Sampling Technique Based Rotation Forest for the Classification of Unbalanced Hyperspectral Data. In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium (pp. 2651-2654). IEEE.
16. Elreedy, D. and Atiya, A.F., 2019. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Information Sciences, 505, pp.32-64.

17. Hu, D., Cao, J., Lai, X., Liu, J., Wang, S. and Ding, Y., 2020. Epileptic Signal Classification based on Synthetic Minority Oversampling and Blending Algorithm. IEEE Transactions on Cognitive and Developmental Systems.
18. Tarawneh, A.S., Hassanat, A.B., Almohammadi, K., Chetverikov, D. and Bellinger, C., 2020. Smotefuna: Synthetic minority over-sampling technique based on furthest neighbour algorithm. IEEE Access, 8, pp.59069-59082.
19. Feng, W., Dauphin, G., Huang, W., Quan, Y., Bao, W., Wu, M. and Li, Q., 2019. Dynamic synthetic minority over-sampling technique-based rotation forest for the classification of imbalanced hyperspectral data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(7), pp.2159-2169.
20. Asha, C.S., Lal, S., Gurupur, V.P. and Saxena, P.P., 2019. Multi-modal medical image fusion with adaptive weighted combination of NSST bands using chaotic grey wolf optimization. IEEE Access, 7, pp.40782-40796.
21. Khandelwal, A., Bhargava, A., Sharma, A. and Sharma, H., 2019. ACOPF-Based Transmission Network Expansion Planning Using Grey Wolf Optimization Algorithm. In Soft Computing for Problem Solving (pp. 177-184). Springer, Singapore.
22. Muniraj, M., Arulmozhiyal, R. and Kesavan, D., 2020. An Improved Self-tuning Control Mechanism for BLDC Motor Using Grey Wolf Optimization Algorithm. In International Conference on Communication, Computing and Electronics Systems (pp. 315-323). Springer, Singapore.
23. Shankar, K., Lakshmanaprabu, S.K., Khanna, A., Tanwar, S., Rodrigues, J.J. and Roy, N.R., 2019. Alzheimer detection using Group Grey Wolf Optimization based features with convolutional classifier. Computers & Electrical Engineering, 77, pp.230-243.
24. Dong, H. and Dong, Z., 2020. Surrogate-assisted grey wolf optimization for high-dimensional, computationally expensive black-box problems. Swarm and Evolutionary Computation, 57, p.100713.