Nat. Volatiles & Essent. Oils, 2021; 8(4): 10993-11019

Bininng And Hybrid Feature Selection Based Big Data Classification

S. Saravanabavanandam[1] , Dr. S. Duraisamy[2]

[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

ABSTRACT

The improved information technology resources, genomics, health records ear creates opportunity for leveraging these developments. A learning health system is created using these developments for delivering informative clinical evidence. An information and data set which is highly complex and large is termed as big data. It is highly complicate to process this big data using conventional database management tools.To overcome those issues in recent work first introduces the pre-processing step using min-max normalization.And then synthetic minority oversampling is used to balancing the data set by generating synthetic data. And Features selection is computed based on levy flight grey wolf optimization additionally introduces hybrid model using CNN with SVM for big data classification. However Data set taken for this work may have noisy data values and it may affect thebig data classification performance and it does not focused in existing work. Performing classification task is based on the single classifiers which not improve the accuracy of the classifier. To avoid those issues this work first introducesthe binning for smoothingand removing unwantedpixels. And then introduces the pre-processing step using min-max normalization. It will normalize the input data into same scale. And then synthetic minority oversampling is used to balancing the data set by generating synthetic data.And then feature selection will be performed based on hybrid Chicken Swarm Optimization andWhale Optimization algorithm. Finally classification doneby usingensemble CNN-SVM. In which Ensembling of the classifier will be done by majority voting functionfor the outputs.Experimental results demonstrates proposed model's effectiveness using Covtype, ECBDL14-S and Poker database interms of precision, recall, error rate and accuracy metrics by comparing with existing HMM, FKNN, WCNN, WCNN-SMT and CNN-SVM using MATLAB.

Keywords: Min-max normalization,synthetic data,binning,minority oversampling, hybrid grey wolf optimization andwhaleoptimization, Ensembling and Big Data.

1. INTRODUCTION

For patients, different healthcare information system models are proposed in various countries for obtaining best care and services by healthcare organizations. These models are participatory, predictive and personalized. Using electronic health records (EHRs), preventive medicine are prescribed. There exist a huge complex biomedical data with high-quality called omics data. In living organisms, about regulatory process and complex biochemical processes, hige raw data is produced in post genomics technologies.

Heterogeneous nature is exhibited by these omics data and in various data formats, they are stored. EHR data is also exhibits a heterogeneous nature as like omics data. This HER data may be continuous or discontinuous and it may be unstructured, semi-structured and structured [1,2].

In medicine and healthcare, different complex and large data are referred as big data. Using traditional hardware or software, it is highly difficult to manage or analyse these data. Heterogeneous data validation, interpolation, modelling, analysis and data quality control integration are covered using big data analytics.

From available huge amount of data, comprehensive knowledge can be discovered using big data application. In healthcare and medicine, large datasets which is collected from various patients are analysed using big data analytics. Clusters can be identified and between datasets, correlation can also be obtained. Using data mining techniques, predictive models are developed using this.

In healthcare and medicine, various scientific areas like health informatics, medical informatics, sensor informatics, medical imaging and bioinformatics are analysed using big data analytics. Comprehensive benefits to health policy makers, clinicians and patients should be provided by new knowledge discovered using big data analytics [3,4].

In recent work first introduces the pre-processing step using min-max normalization. And then synthetic minority oversampling is used to balancing the data set by generating synthetic data.And Features selection is computed based on levy flight grey wolf optimization additionally introduces hybrid model using CNN with SVM for big data classification.

However Data set taken for this work may have noisy data values and it may affect the big data classification performance and it does not focused in existing work. Performing classification task is based on the single classifiers which not improve the accuracy of the classifier[5,6].

To avoid those issues this work first introduces the binning for smoothing and removing unwanted pixels. And then introduces the pre-processing step using min-max normalization. It will

normalize the input data into same scale. And then synthetic minority oversampling is used to balancing the data set by generating synthetic data.

And then feature selection will be performed based on hybrid Chicken Swarm Optimization and whale optimization algorithm. Finally classification done by using ensembleCNN-SVM. In which Ensembling of the classifier will be done by taking majority voting function the outputs[7,8,9].

The paper is structured in following five sections. Introduction to big data classification in healthcare system is provided in Section I. Section II reviews the different methods for big data classification. Design methodology for proposed big data classification model is produced in Section III. Section IV describes experimental study and includes multiple results analyses. Section V concludes work and outlines future work.

2. LITRATURE REVIEW

For big data classification, various techniques are reviewed in this section.

An improved KNN algorithm is proposed by Xing, W. and Bei, Y., [10] and comparison is made between traditional KNN algorithm and enhanced KNN algorithm. In conventional KNN classifier's query instance neighbourhood, performed the classification and for every class weights are assigned. Around query instance, class distribution are considered in this algorithm and it ensures that, outliers are not effected by assigned weight.

Density cropping and cluster denoising based enhanced KNN algorithm is proposed in this paper for eliminating issues of traditional KNN algorithm in processing large data sets. Using clustering, denoising is performed in this algorithm and K-nearest neighbors search speed is enhanced and it enhances KNN algorithm's classification efficiency is enhanced. KNN algorithm's classification accuracy is maintained using this.

In processing large data sets, KNN algorithm's classification efficiency is enhanced effectively using proposed algorithm as shown in experimental results. KNN algorithm's classification accuracy is maintained while producing good performance.

Using open sources like Cassandra, NoSQL, Kafka, Apache Storm and Hadoop, for big data healthcare analytic, a genetic architecture is proposed by Ta et al [11]. With a rapid rate, a huge healthcare data can be analysed using a combination of distributed storage system, distributed real-time computing and high throughput publish-subscribe messaging for streams.

A probabilistic data collection mechanism is designed by Sahoo, et al [12] and those collected data's correlation analysis is performed using this. At last, for seeing most correlated patientsfuture health condition according to current health status, designed a stochastic prediction model. In cloud environment, through extensive simulations, proposed protocols performance is evaluated. Around 98% prediction accuracy and 90% CPU utilization are achieved using this. Analysis time is minimized using bandwidth utilization.

Using a powerful emotion detection module, an emotion-aware connected healthcare system is proposed by Hossain and Muhammad, G., [13]. In a smart home scenario, patient's image and speech signal's are captured using various devices. For emotion detection module, these signals are given as an input. Separately processed the speech and image signals. A final score is produced by fusing these signals with classification scores and decision regarding emotion is decided using this score.

Caregivers can visit patient, if detected emotion is pain. For validating proposed system, various experimentation are performed and around 99.87% accuracy is produced using this in emotion detection. Proposed framework greatly contribute personalized and seamless emotion-aware healthcare services toward 5G.

An alternative parameterization model is proposed by Mohamad et al [14]. Without high learning, usage and maintenance cost, most optimized attribute set can be generated using this.

Model is based on two integrated models which are combined with correlation-based feature selection, soft set, best-first search algorithm, and rough set theories which were compliments to each other as parameter selection technique. In big data analysis process, proposed model has significantly shown as an alternative model as shown in experimental results.

Effective processing framework termed as deep multilayer and non-linear Kernelized Lasso feature learning (DM-NKLFL) is proposed by Prakash and Sangeetha [15]. In image processing field, it cam cope with data explosion powerfully. For complex non-linear and simple linear relationships, general framework is provided by this work. Without degrading performance, increase

Two parts are included in this proposed DM-NKLFL technique. They are deep multilayer pattern learning (DMPL), stepwise regression nonlinear Kernelized Lasso (SR-NKL) feature selection. Non-linear features are processed using SR-NKL, which minimizes complexity and time consumption in feature selection process. Data driven features are learned deeply using DMPL and it is used for computing underlying patterns. With respect to results quality and time efficiency, better performance is shown by DM-NKLFL technique in big biological data.

In health care, a big data classification model (heart disease) is proposed by Game and Emmanuel [16], where certain steps or phases are included. Following are the major steps involved in this process, Map-reduce framework, support vector machine (SVM) and optimized decision tree classifier (DT). At first, to MapReduce Framework, big data is supplied as input. Using some major operations, data content are reduced in this. For reducing data dimension, principle component analysis is used in this framework.

Minimized data is given to SVM and it produces output classes. Other conventional techniques like grey wolf optimizer algorithms, genetic algorithm, particle swarm optimization algorithm, artificial bee colony algorithm and firefly algorithm are used for making comparison with proposed DGWO model.

3. PROPOSED METHODOLOGY

Big data classification based on Ensemble CNN-SVM is proposed in this section. In this, unwanted pixels are removed and smoothened using binning. Then, input data is normalized to a same scale by introducing min-max normalization. Dataset is balanced using a synthetic minority oversampling and synthetic data is generated using this. Features are selected using hybrid whale and chicken swarm optimization. Figure 1 shows the proposed work's overall architecture.

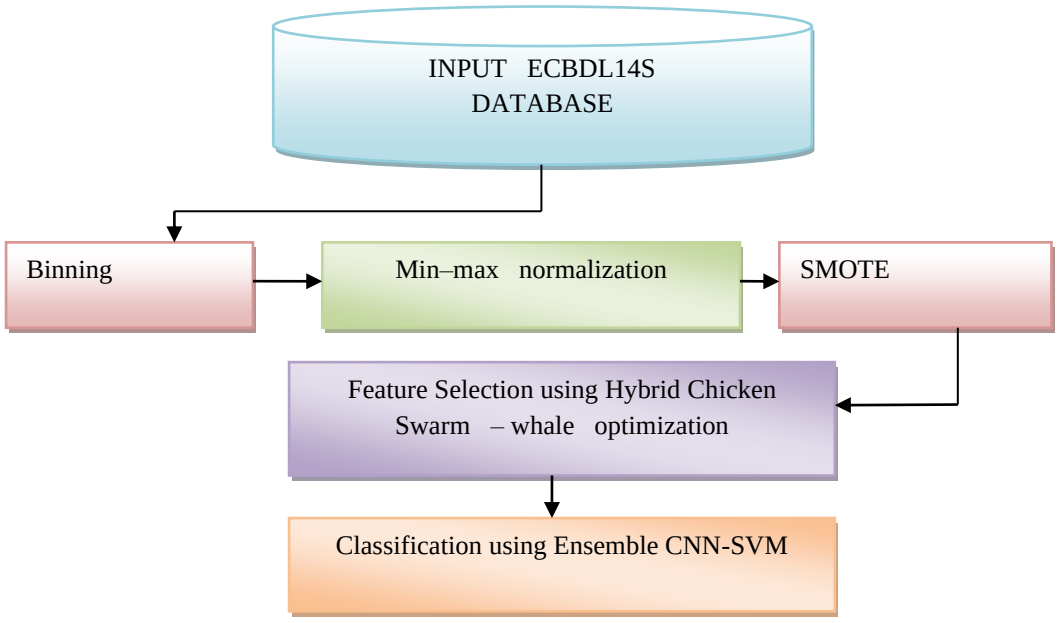

Figure: 1. Overall architecture of the proposed work

3.1. **Data Binning using bin median points**

There may be some noises in data, so it needs to be removed by performing data smoothing operation. Bin median points are used in this work. Neighbouring values or values around are referred for smoothening data in binning technique. Data is split into buckets or bins of same count in this technique for performing smoothing via bin median points [17-19].

- For crime incident, sorted data is used
- Data is split into equal groups.
- Bin points are used for data smoothing
- Bin boundaries are used for smoothing

3.2.**Min-max normalization**

Input data needs to be normalized after binning. In data, there is a chance of scale variation. Inaccurate results may be produced because of this variations. So, for eliminating this issue, data is normalized. Min-max normalization model is used in this work and using a mathematical function, new range values are formulated by converting numerical values in normalization.

Data are generally normalized using a common technique called min-max normalization. From dataset, within the specified maximum and minimum range, values are normalized and using following expression, every value is replaced.

$$v' = \frac{v - \min_A}{\max_A - \min_A}\left(\text{new\_max}_A - \text{new\_min}_A\right) + \text{new\_min}_A \quad (1)$$

Where,

A indicates Attribute data,

Min(A) is a minimum absolute value of A, Max(A) is a maximum absolute value of A

v' indicates every entry's new value in data

v indicates every entry's old value in data

new_max(A) indicted maximum value of range, new_min(A) indicated min value of range (i.e required boundary value range)

3.3. Data Balancing Using Synthetic Minority Oversampling Technique

There will be an imbalanced data in input ECBDL14S dataset. So, in pre-treatment process, applied the oversampling and normalized output is given as input. Dataset's class distribution is adjusted in

data analysis using a technique called oversampling. In this, minority samples are copied randomly for enhancing minority class samples count. This balances the majority and minority classes example size.

For solving imbalance problems, generally used oversampling technique is SMOTE (Synthetic Minority Oversampling Technique). For minority class, using linear interpolation, synthetic training examples are generated. There are two stages in SMOTE algorithm.

From every minority data, Euclidean distance is computed for finding k nearest neighbours in first stage based on all other minority data and in ascending order they are sorted. Then nearest neighbors (KNN) corresponds to k lowest distance data. From first attribute to n(maximum attributes count), between one minority data (x) and another minority data (y), Euclidian distance is computed using expression (2),

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

Between two minority data, interpolation technique is used for generating synthetic data in second stage. In synthetic data generation process, one kNN is randomized as a candidate. Therefore, for generating new synthetic data between x and y, one selected candidate (y) and original minor data (x) are used. For a-th attribute, among x and y, synthetic data is computed using expression (3),

$$\text{SyntheticData}\quad a(x, y) = x_a + r \cdot (x_a - y_a) \text{ for } 0 \le r \le 1 \quad (3)$$

Where,

R represents a random number whose value lies between 0 and 1

For n attributes, applied the above expression. Until reaching desired synthetic data, repeated this process.

3.4. **Feature Selection Using Hybrid Chicken Swarm– Whale Optimization**

Important features are selected after pre-processing as there are more features and for computation, more time is consumed by this. Significant features must be selected for minimizing computation. A Hybrid Chicken Swarm – Whale Optimization is used in this work.

3.4.1.Chicken swarm optimization (CSO)

A bio-inspire meta heuristic optimization algorithm is Chicken swarm optimization (CSO). Chicken swarm's hierarchal order and individual chickens behaviours are mimicked in this algorithm. Various

groups are formed by dividing chicken swarm's hierarchal order. Many chicks, hens and one rooster will be there in every group. Various motions law are followed by every chickens types.

In chicken's social lives, a significant role is played by hierarchal order. In a flock, weak chickens are dominated by superior chickens. More dominant hens will be there and they will be positioned near to head rooters. At groups periphery, rooters and submissive hens are positioned [20-23].

Based on following rules, proposed CSO's mathematical model. Chicken's behaviours are summarized using these rules.

1)Various groups are formed by splitting chicken swarm. There is a dominant rooster in every group and some chicks and hens will follow it.

2) Swarm's hierarchy is outlined by chickens fitness value. In every group, individuals having best fitness value is assumed as roosters and will act as a group leader. Chicks are assumed as an individuals having worst fitness values. Others are hens.

3) In a group, mother-child relationship, dominance relationship and swarm hierarchy are unchanged. At every several (G) time steps, status of these values are updated.

4) There are N virtual chickens in swarm. They are split as, roosters count RN, hens count HN, chicks count CN, mother hens count MN. In a D-dimensional space, every individuals are represented using its position.

$$x_{i,j}\ (\ i\in[i,\ldots\ldots N], j\in[1,\ldots\ldots,D]), \quad (4)$$

Rooster Movement: In a wider place range, foods are searched using roosters with better fitness values when compared with the ones in worse fitness values. Expressions (5) and (6) are used for representing this movements.

$$x_{i,j}^{t+1} = x_{i,j}^{t} * \left(1 + \text{Randn}(0,\sigma^2)\right) \quad (5)$$

$$\sigma^2 = \begin{cases} 1, & \text{if } f_i \le f_k, \\ \exp\left(\dfrac{f_k - f_i}{|f_i| + \in}\right) & \text{otherwise} \end{cases} \qquad k \in [1, N], k \neq i, \qquad (6)$$

Where, selected rooster is represented as $x_{i;j}$ and it has index i, Gaussian distribution is represented as $\text{Randn}(0,\sigma^2)$and it has standard deviation σ^2 and mean 0, a smallest constant used is represented as s and it is used for avoiding zero-division-error, from roosters group, randomly

selected roosters index is represented as k, corresponding rooster x_i's fitness value is represented as f_i.

Hen movement: For searching food, group-mate roosters are followed by hens. Moreover, good food found by other chickens are steal by hens randomly, though they are repressed using other chickens. In competing for food, more advantages are shown by dominant hens when compared with submissive ones. Expression (7) and (8)are used for formulating this mathematically.

$$x_{i,j}^{t+1} = x_{i,j}^{t} + S1 * rRand * \left(x_{r_{1,j}}^{t} - x_{i,j}^{t}\right) + S2 * Rand * \left(x_{r_{2,j}}^{t} - x_{i,j}^{t}\right) \quad (7)$$

$$S1 = \exp\left((f_i - f_{r1})/abs(f_i) + \in\right) \quad (8)$$

$$S1 = \exp\left(\left(f_{r_2} - f_i\right)\right) \quad (9)$$

Where, a random number with uniform distribution is represented as Rand and its value lies between [0,1]. Rooster's index is represented as $r_1 \in [1,\ldots..,N]$, it is a i^{th} hen's group-mate, from swarm, chicken's randomly selected index is represented as $r_2 \in [1,\ldots..,N]$.

Chick movement: For searching food, around its mother, chick moves. In expression (17), this is formulated as,

$$x_{i,j}^{t+1} = x_{i,j}^{t} + FL * \left(x_{m,j}^{t} - x_{i,j}^{t}\right) \quad (10)$$

Where, i^{th} chick's mother's position is represented as $x_{m,j}^{t}$, so that $m \in [1;N]$,speed of chicks in following its mother is represented using a parameter FL. In range [0,2], between every chick FL, differences are considered.

In an individual dimension, represented a feature space with every feature and every dimension span ranges between 0 to1. So, in search space for computing optimum point, there is a need to have an intelligent searching technique and fitness function should be maximized.

The fitness function for CSO is for maximizing classification performance over validation set given training data, as expressed in (18) while keeping minimum features count elected.

$$f_\theta = \omega * E + (1 - \omega)\frac{\sum_i \theta_i}{N} \quad (11)$$

Where, for a specified vector θ with size N, fitness function is represented as f_θ, it has elements, where unselected features are represented as 0 and selected features are represented as 1. In dataset, total features are represented as N, classifier error rate is represented E and constant value used to control classification performance is represented as ω.

In a specified dataset, features count is similar to used variable. Range of variable lies between 0 to 1. Feature of variable which is approaching to 1 is selected as a candidate for classification. For deciding threshold, features are extracted in individual fitness computation and it is expressed as,

$$f_{i,j} = \begin{cases} 1 \text{ if } X_{i,j} > 0.5 \\ 0 \text{ otherwise,} \end{cases} \quad (12)$$

Where, at dimension j, for search agent i, dimension value is represented as X_{ij} while updating firefly position called solution, limiting constraints [0,1] is violated in some dimensions during the update. So, for ensuring variable limits, simple truncation rule is used.

1. G, MN, CN,HN,RN are initialized;
2. In swarm, every chicken is initialized randomly.
3. X_i (i = 1,2,……;N).;
4. Max iteration count $T_{max,}$ is initialized;
5. while T < T_{max} do for every iteration
6. if T % G equals 0 then
7. Chickens fitness values are ranked and in swarm, a hierarchal order is established;
8. Swarm is split into various groups, and relationship between chicksand mother hens are computed in a group;
9. end
10. for every chicken Xi in swarm do
11. if Xi is a roster then
12. Using expression 5, Xi's location is updated
13.
14. end
15. if Xi is a hen then
16. Using expression 7, Xi's location is updated
17.
18. end
19. if Xi is a chick then
20. Using expression 10, Xi's location is updated

21. end
22. Using expression11, new solution is evaluated;
23. If new solution is better than its previoussolution, update it;
24. end
25. end

3.4.2. Whale Optimization

A novel stochastic optimization technique based on nature-inspired population is WOA and it is developed recently. For an optimization problem, best solution is computed using a search agents set in WOA. Humpback whale's behaviour is imitated in WOA while hunting prey through a technique called bubble-net hunting.

Three general steps are included in WOA namely, searching around best prey, bubble net attacking and encircling prey. In some benchmark testing, when compared with other meta-heuristic technique like PSO and GWO, better performance is shown by WOA [24-27].

For circling around prey and hunting, bubble-net mechanism is used by Humpback whales. Prey like fishes are enclosed in whale and for computing optimum solution, their positions are updated in this. Expression (13) and (14), expresses the WOA mathematically.

$$X(t+1) = X^*(t) - A.|C.X^*(t) - X(t)| \text{ if } p < 0.5 \qquad (13)$$

$$X(t+1)|C.X^*(t) - X(t)|.e^{bl} \cos(2\pi t) + X^*(t) \text{ if } p \geq 0.5 (14)$$

where, all whale's positions are represented using a vector X, iteration index or time is represented as t, best solution is represented as X*, A=2a. (r-a), coefficient vector is represented as a and it decreases linearly to 0 from 2 over iteration progress. Random vector is represented as r with a value between 0 to 1. Logarithmic spiral's shape is defined using a constant b and it depends on specific path and its value is set as 1.

A random number with a value between -1 to 1 is represented as l, p is also a random number with a value between 0 to1. These constants are used in whale's position update. Probabilities in expression (13) and (14) is 50%. With an equal chance, paths are randomly selected by whales in optimization process. In bubble-net phase, A value lies between -1 to 1 and in search phase, its value is greater than 1 or less than 1. Expression (15) gives the searching mechanism.

$$X(t+1) = X_{rand} - A.|C.X_{rand} - X(t)| (15)$$

Searching operation is emphasized using random search technique with |A| value greater than one and a global searching of WOA algorithm is enforced by this. At WOA searching process beginning, created solutions in random manner. Then, algorithm specified in table 1 is used for updating these solutions in every iteration. Until reaching predefined maximum iterations count, this search process is continued.

Algorithm for Whale Optimization

START

1. Data is imported
2. Whale population X's location is initialized
3. Every whale's fitness is computed
4. **r**and a are initialized and C and A are computed
5. **X*** is initialized as best hunter whale's location
6. initialize t = 1
7. **while** t ≤ max iterations **do**
8. **for** every hunting whale **do**
9. **if** p < 0.5
10. **if** |**A**| < 1
11. Using (13), current hunting whale's location is updated
12. **else if** |**A**| ≥ 1
13. Another search agent is selected randomly
14. Using (15), current hunting whale's location is updated
15. **end if**
16. **else if** p ≥ 0.5
17. Using (16), current hunting whale's location is updated
18. **end if**
19. **end for**
20. If there is a better solution, **X*** is updated
21. t = t +1
22. **end while**
23. output **X***
24. **END**

Search space is not explored properly, which is a major drawback of WOA.In rooter's position update technique, some issues are shown by chicken swarm optimization. Effective results are not shown by conventional CSO as step size is generated using a Gauss distribution. Algorithm's exploitation ability is affected due to this. Two optimization models are integrated in this work for rectifying these issues.

3.4.3. Hybrid Chicken Swarm– Whale Optimization

In both whale and chicken swarm optimization, best solution is evaluated at first in a same traditional way. At last, best solution is selected by comparing two optimization algorithm's results.

1. **Input: ECBDL14S Database**
2. **Output**: **Optimal features**
3. G, MN, CN, HN, RN are initialized;
4. In swarm, every chicken is initialized randomly
5. Maximum iteration count T_{max} is initialized
6. Chickens fitness values are ranked and hierarchal order in swarm is established
7. New solution is evaluated
8. If new solution is better than its previous one, update it
9. Whale population X's locations are initialized
10. Every whale's fitness is computed
11. r and a are initialized and C and A are computed
12. **X*** is initialized as best hunter whale's location
13. Constrains are checked
14. If there is a better solution, **X***is updated
15. Whale and chicken swarm optimization solutions are compared
16. Solution with better fitness value among others is selected
17. Solutions are updated

3.5. Classification using Ensemble CNN-SVM

At last, for classification process, selected features are given as input. An Ensemble CNN-SVM algorithm is used for performing classification. There are three steps in this proposed ensemble classifier namely incremental, ensemble and base learners. First step is base and an incremental classifier is used in this, which are run N times.

Cross-validated predictions are produced using these learners. Hybrid Convolutional Neural Network Support Vector Machine is used for producing cross validated predictions. Majority voting is used in ensemble tier. According to incremental learners probabilities, weights are assigned in MV.

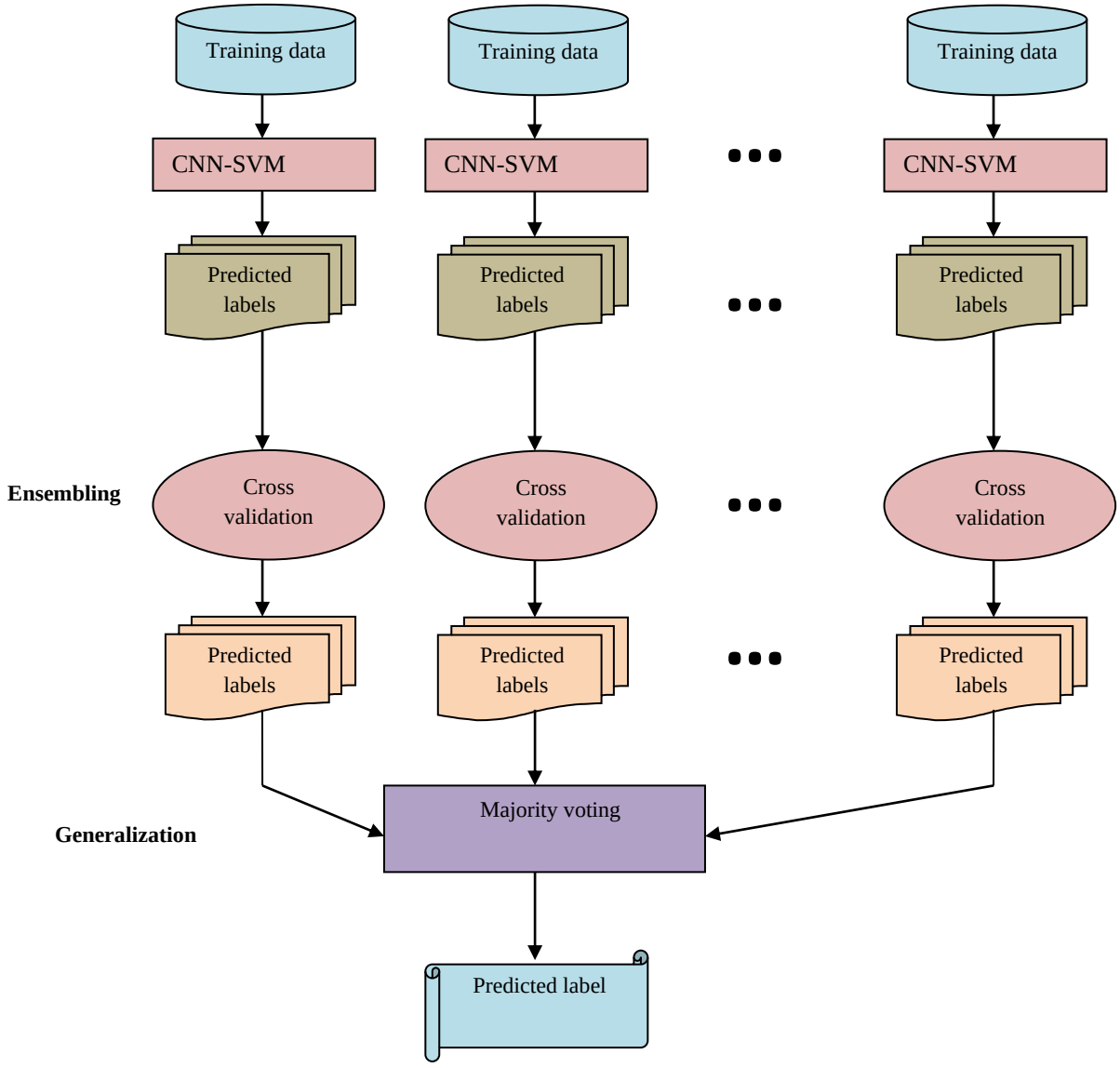

Figure: 2. CNN-SVM Classifier ensemble model

CNN

There is a single or multiple sub-sampling or convolution layers in typical CNN. After that, there exist one or more fully connected layers and output layer.

SVM

Supervised learning forms base for a binary classifier called SVM. When compared with other classifiers, better results are produced using this. In high-dimensional feature space, hyper plane is

constructed for making classification between two classes. A type of classification algorithm is SVM and various kernel techniques forms base for this.

It is the simplest one, where training patterns are separable linearly. A linear function is expressed as,

$f(x) = w^T X + b$ (16)

So that for every training sample x" function yields $f(x_i) \geq 0$ for $Y_i = +1$, and $f(x_i) < 0$ for $y_i = -1$. It can also be stated as, hyper plane is used for separating two different class's training sample.

$f(x) = w^T X + b = 0,$ (17)

Where, weight vector is represented as w and it is normal to hyper plane, threshold or bias is represented as b and data point is given by x.

There exist different hyper planes for a specified training g set. Between two classes, separating margin's are maximized using these hyper planes.

In CNN network, last layer output units gives input sample's estimated probabilities. An activation function is used for computing every output probability. Linear combination of bias term and previous hidden layer's output with trainable weights is given as an input to activation function.

However, it is meaningless to look at the hidden layer output values, but only makes sense to CNN network itself. For any other classifiers, these values can be taken as features.

Hybrid CNN-SVM is used for rectifying this. CNN model's last output layer is replaced by SVM classifier for designing hybrid CNN-SVM model architecture.

At first, input layer is given with centered and normalized input images and until training process convergence, using various epochs, trained the original CNN with output layer. Then, output layer is replaced using SVM with a Radial Basis Function (RBF) kernel. For training, from hidden layer, outputs are taken by SVM as new feature vector. Recognition task is performed by SVM classifier after its training and on testing images, new decisions are made using features which are extracted automatically.

Convolution layer

Selected features are given as an input to this proposed work. With a kernel called filter, convolved the input features in this convolution layer. The n output features maps are generated using kernel and input feature's convolution results. In general, filter corresponds to convolution matrix kernel.

Input and kernels are convolved for computing output features, which are termed as feature maps with i*isize.

Multiple convolutional layers are included in CNN. Feature vectors are given at next convolutional layers inputs and outputs. In every convolution layer, there exist n filters bunch. With input, these filters are convolved and in convolution operation, applied filters count is similar to generated feature maps (n*) depth. At certain input location, every filter map is assumed as a specific feature.

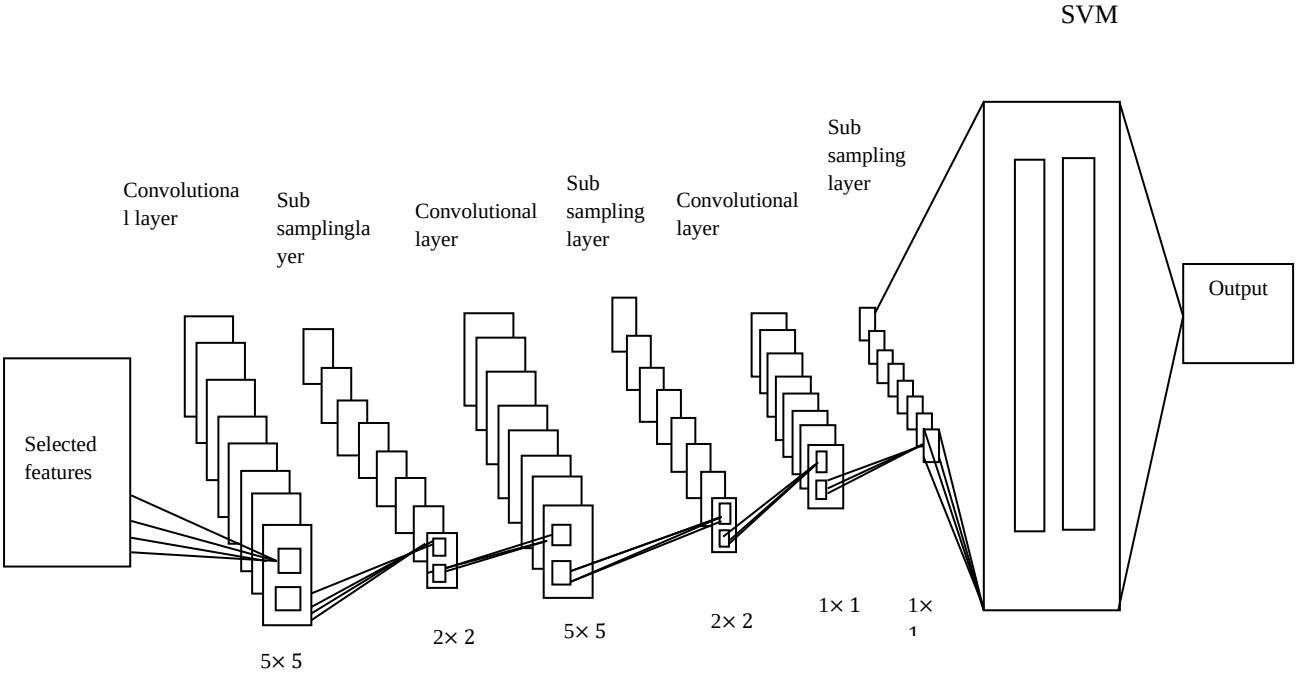

Figure.3: Convolutional Neural Network architecture

The l-th convolution layer's output is represented as $C_i^{(l)}$. This contains feature maps and is computed as

$$C_i^{(l)} = B_i^{(l)} + \sum_{j=1}^{a_i^{(l-1)}} K_{i,j}^{(l-1)} * C_j^{(l)} \quad (18)$$

Where, bias matrix is represented as $B_i^{(l)}$, convolution filter is represented as $K_{i,j}^{(l-1)}$, it has a a*a kernel size. In (l − 1) layer, j-th feature map is connected with the i-th feature map of same layer using this kernel.

Feature maps are available in output $C_i^{(l)}$ layer. In expression (10), input space is represented as a first convolutional layer $C_i^{(l-1)}$, that is, $C_i^{(0)} = X_i$. Feature maps are generated using kernel. For convolutional layer output's nonlinear transformation, activation function is applied after convolution layer.

$$Y_i^{(l)} = Y(C_i^{(l)}) \quad (19)$$

Where, activation function output is represented as $Y_i^{(l)}$, received input is represented as $C_i^{(l)}$.

Sub sampling or pooling Layer

Spatial dimension reduction of feature maps which are extracted from previous convolution layer is mainly concentrated in this layer. Between feature map and mask, sub sampling operation is performed. Proposed various sub sampling techniques like maximum pooling, sum pooling and averaging pooling.Max pooling is most commonly used technique. In this, output feature corresponds to every block's maximum value. For tolerating rotation and translating input images, convolution layer is assisted by sub sampling layer.

Fully Connected layer

A SVM classifier is used as a final CNN layer.

$$f(x) = w^T X + b \quad (20)$$

So that for every training sample x" function yields $f(x_i) \geq 0$ for $Y_i = +1$, and $f(x_i) < 0$ for $y_i = -1$.

Majority Voting (MV)

A decision making technique is Majority Voting (MV), which is retrieved from classifiers. This algorithm is run n times independently and separately to give more abilities every time. Assume N sample's set as χ and Q classes set as C. An algorithm set S = {A_1, A_2, A_M} is defined with M classifiers that are used to vote. Every example x ∈ χ is assigned for having one Q classes. For every example, predictions are give classifiers every time.

Class predicted by majority classifier (which gains majority votes) is assigned as a final class for every sample. In MV, classifier's prediction accuracy value is used for weighting every vote and it is represented as Acc. For a class c_k, total votes count is computed as,

$$T_k = \sum_{l=1}^{M} Acc\,(A_l) \times F_k(c_l) \quad (21)$$

$$F_k(c_l) = \begin{cases} 1 & c_l = c_k \\ 0 & c_l \neq c_k \end{cases} \quad (22)$$

Where, c_l and c_k are C's classes. Selected the class which is receiving highest total weight. On various independent training sets, trained all classifiers in general and assigned the weights accordingly for producing high classification rate in data classification.

4. RESULTS AND DISCUSSSION

This section analyses experimentation results carried out on proposed model. This model's implementation is carried out using MATLAB. In comparison of the already variable HMM, FKNN, WCNN ,WCNN-SMT and CNN-SVM algorithm and the proposed E-CNN-SVM are done with respect to precision, recall, accuracy, F-measure, error rate for 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).

Predicting forest cover sort from just cartographic variables in the COV database. The US Forest Service (USFS) Area 2 Resource Information System (RIS) data was used to assess the real forest cover type for a given observation (30 x 30 meter cell). Independent variables were generated using data from the United States Geological Survey (USGS) and the United States Forest Service (USFS).The data is in its natural state (not scaled) and contains conditional (0 or 1) columns for qualitative independent variables (wilderness areas and soil types).

This study area encompasses four wilderness areas in the Roosevelt National Forest in northern Colorado. In these areas, established tree cover types represent forests with little human-caused disturbances, so they are more a part of natural cycles than forest management practices. The ECBDL14s database has 34 properties, 33 of which are linear and one of which is mis nominal. In dermatology, the differential diagnosis of erythemato-squamous diseases is a great challenge. They both have erythema and scaling as therapeutic characteristics, with only minor exceptions.Psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris are all diseases of this category. A biopsy is normally expected for diagnosis, but these diseases share many histopathological features as well. Another challenge in differential diagnosis is when a disease can display symptoms of another disease in the early stages but then develop characteristic features later. Patients were first tested scientifically using a compilation of 12 criteria.Following that, skin samples were taken to determine 22 histopathological characteristics. An examination of the samples under a microscope determines the values of the histopathological features.The family history attribute in the dataset created for this domain has a value of 1 if either of these diseases has been observed in the family, and 0 otherwise. The patient's age is reflected by the age function.Any

other characteristic (clinical and histopathological) was graded on a scale of 0 to 3. Here, 0 denotes the absence of the function, 3 denotes the maximum sum possible, and 1, 2 denotes the relative intermediate values.

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
	CNN - SVM	550	625	600
	E-CNN-SVM	500	610	550
Accuracy (%)	HMM	70	72	73
	FKNN	75	75	75
	WCNN	78	79	78
	WCNN-SMT	82	85	84
	CNN-SVM	85	87	90
	E-CNN-SVM	90	91	92
Précis recision (%)	HMM	75	74	75
	FKNN	76	75	77
	WCNN	79	78	79
	WCNN -SMT	81	82	81
	CNN-SVM	86	88	85
	E-CNN-SVM	90	91	89
Recall (%)	HMM	80	85	82
	FKNN	82	85.5	83.3
	WCNN	85	87	85
	WCNN-SMT	87	88	87
	CNN-SVM	89	90	88
	E-CNN-SVM	91	92	92
Error rate (%)	HMM	30	28	27
	FKNN	25	25	25
	WCNN	22	21	22

	WCNN-SMT	18	15	16
	CNN-SVM	15	13	10
	E-CNN-SVM	10	9	8

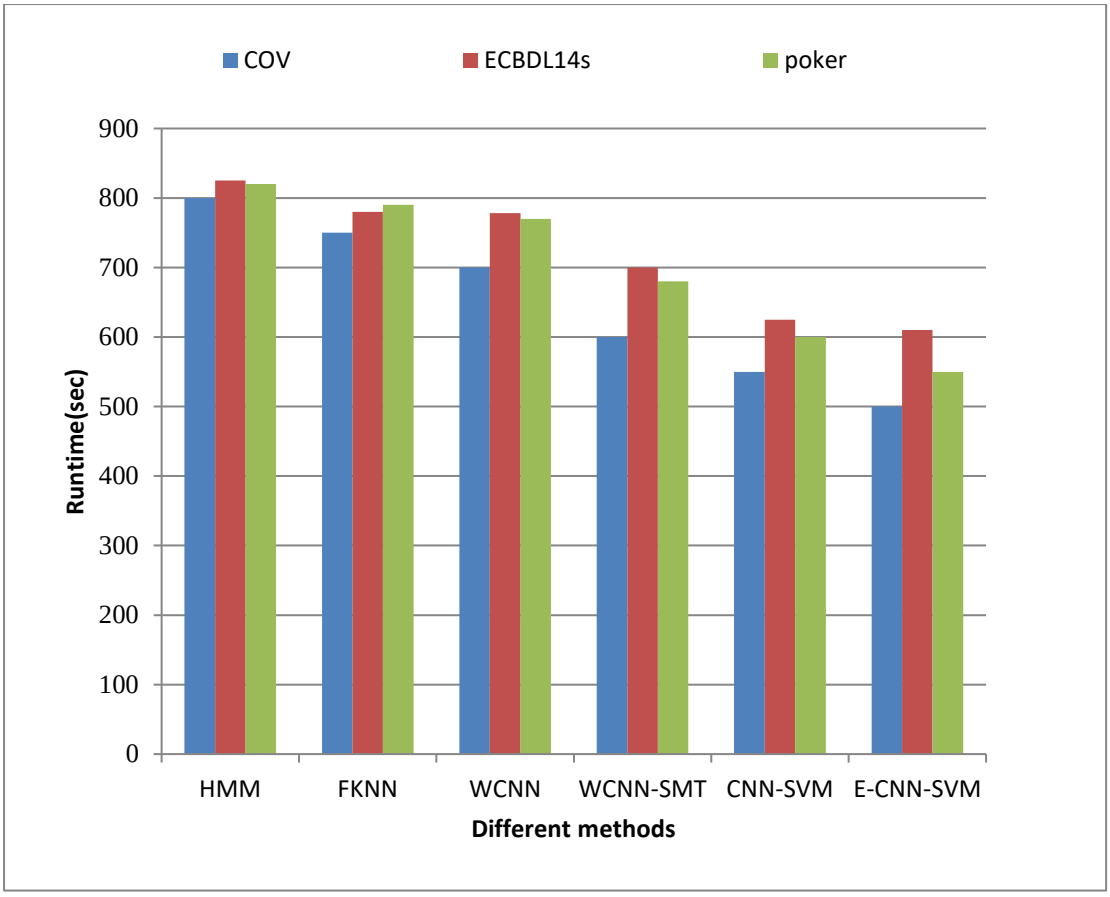

Figure:4. Runtime results vs. Classification methods

The above figure shows the Performance comparison for Runtime metrics with classifiers HMM, FKNN, WCNN, WCNN-SMT, CNN-SVM proposed E-CNN-SVM schemes. In X of above graph, represented various techniques and Runtime values are representedin Y-axis. As indicated in results, newly introduced E-CNN-SVM model produced lower Runtime results which is 500(s) for COV dataset while available, HMM , FKNN ,WCNN,WCNN-SMT and CNN-SVM technique yields only 800(s),750(s),700(s) and 600(s) ,550(S)respectively.

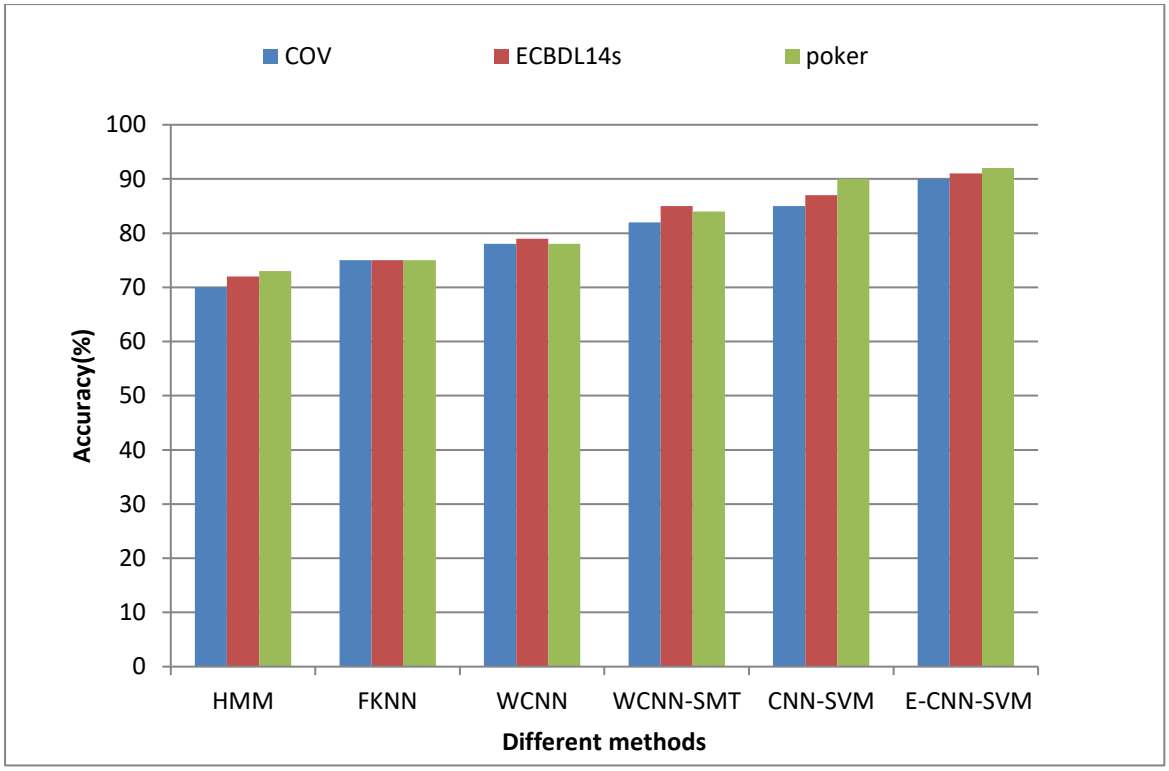

Figure: 5. Accuracy results vs. Classification methods

Accuracy metric performance comparison between existing classifier HMM, FKNN, WCNN, WCNN-SMT,CNN-SVM proposed E-CNN-SVM scheme is shown in above figure. In Proposed work, fitness function is used by hybrid features for significant features selection by which E-CNN-SVM accuracy get enhanced. In X of above graph, represented various techniques and accuracy values are representedin Y-axis. As indicated in results, it is assured that newly introduced E-CNN-SVM model produced higher Accuracy results 90% for COV dataset while available HMM, FKNN, WCNN, WCNN-SMT and CNN-SVM technique yields only 70% , 75% ,78% ,82%,85 % respectively.

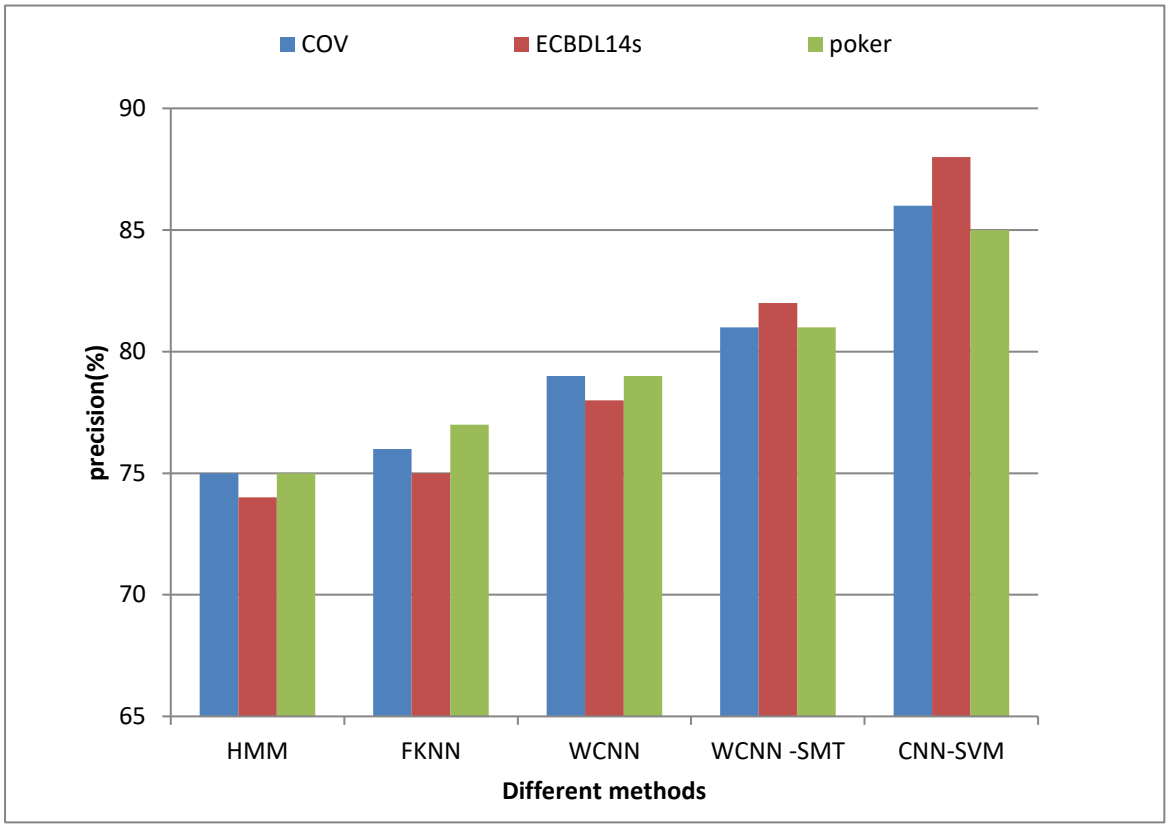

Figure . 6: Precision results Comparison of Various Classifiers

Efficiency of the proposed E-CNN-SVM is shown in the above figure by comparing this with the available HMM, FKNN, WCNN, WCNN-SMT and CNN-SVM methods in terms of precision. Proposed work uses min max normalization which transfers the input into the same scale and it increases the precision of the result. In X of above graph, represented various techniques and precision values are representedin Y-axis. As indicated in results, it is assured that newly introduced E-CNN-SVM model produced precision results of 90% for COV dataset while available HMM, FKNN, WCNN, WCNN-SMT and CNN-SVM techniques yields only 75%, 76%, 79% , 81% and 86% respectively.

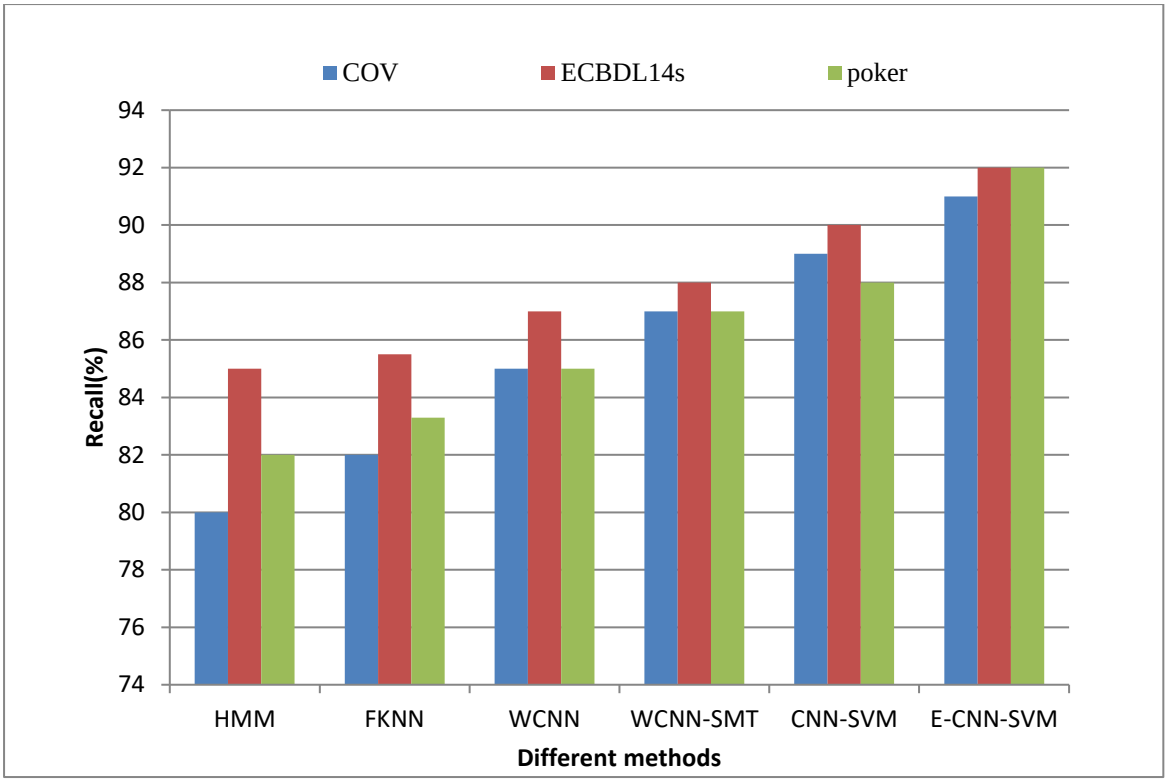

Figure: 7. Recall results vs. classification methods

Figure: 7. Shows the Performance comparison for the existing classifier HMM, FKNN, WCNN, WCNN-SMT proposed CNN-SVM scheme interms of recall. Proposed work uses SVM for classification in CNN which increases the recall rate. In X of above graph, represented various techniques and recall values are representedin Y-axis. As indicated in results,it is assured that newly introduced CNN-SVM model produces higher recall results of 91% for Cov dataset while available HMM, FKNN, WCNN, WCNN-SMT techniques yields only 80% , 82% ,85% ,87% and 89% respectively.

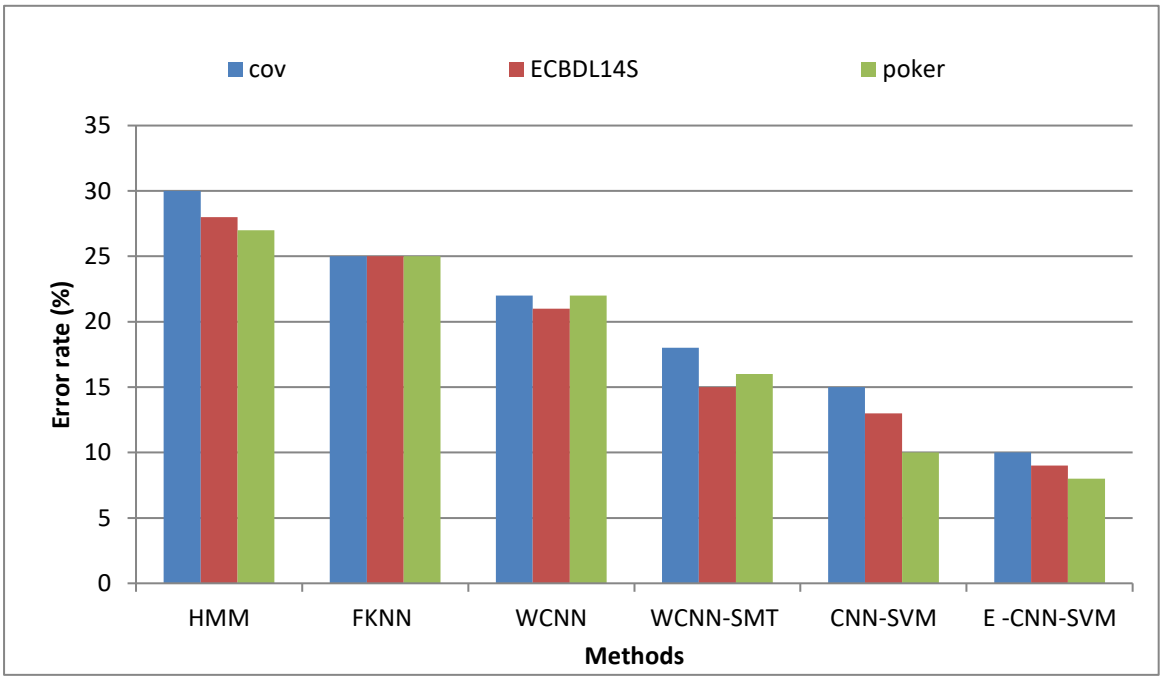

Figure: 8. Error rate result vs. classification methods

Above figure shows the Performance comparison for the existing classifier HMM, FKNN, WCNN, WCNN-SMT, CNN-SVM proposed E-CNN-SVM scheme interms of error rate. In X of above graph, represented various techniques and error rate values are representedin Y-axis. As indicated in results, it is assured that newly introduced E-CNN-SVM model produces lower error rate results of 10% for Cov dataset while available HMM, FKNN,WCNN,WCNN-SMT techniques yields only 30%, 25%, 22% ,18% and 15% respectively.

5. CONCLUSION AND FUTURE WORK

Large set of computing devices produces huge data. In big data analytics, analysis and processing of this data is highly complex. In large dataset, data scalability and consistency is a major problem. Using a novel technique, data classification, aggregation and extraction are done using this proposed algorithm. In this work binning used for smoothing and then input data will be normalized using min-max normalization. Synthetic data is generated for minority classes using Synthetic minority oversampling to balance the data set. And then hybrid Chicken Swarm Optimization and Whale Optimization algorithm is utilized for feature selection. Finally bid data classification is done by using ensemble CNN-SVM. In which Ensembling of the classifier will be done by majority voting for the outputs. Experimental results demonstrates that the effectiveness of the proposed model using Covtype, ECBDL14-S and Poker database interms of precision, recall, error

rate and accuracy metrics by comparing with existing models and shows that the proposed model provides better results. However deep learning produces more computation complexities so need to use other methods for classification in future.

REFERENCES:

1. Deng, Y., Ren, Z., Kong, Y., Bao, F. and Dai, Q., 2016. A hierarchical fused fuzzy deep neural network for data classification. IEEE Transactions on Fuzzy Systems, 25(4), pp.1006-1012.
2. Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K. and Taha, K., 2015. Efficient machine learning for big data: A review. Big Data Research, 2(3), pp.87-93.
3. Tahmassebi, A., Gandomi, A.H., McCann, I., Schulte, M.H., Schmaal, L., Goudriaan, A.E. and Meyer-Baese, A., 2017, June. An evolutionary approach for fmri big data classification. In 2017 IEEE Congress on Evolutionary Computation (CEC) (pp. 1029-1036). IEEE.
4. Suthaharan, S., 2016. Machine learning models and algorithms for big data classification. Integr. Ser. Inf. Syst, 36, pp.1-12.
5. Lin, W., Wu, Z., Lin, L., Wen, A. and Li, J., 2017. An ensemble random forest algorithm for insurance big data analysis. Ieee access, 5, pp.16568-16575.
6. Liu, B., Blasch, E., Chen, Y., Shen, D. and Chen, G., 2013, October. Scalable sentiment classification for big data analysis using naive bayes classifier. In 2013 IEEE international conference on big data (pp. 99-104). IEEE.
7. Demchenko, Y., Zhao, Z., Grosso, P., Wibisono, A. and De Laat, C., 2012, December. Addressing big data challenges for scientific data infrastructure. In 4th IEEE International Conference on Cloud Computing Technology and Science Proceedings (pp. 614-617). IEEE.
8. Hegazy, O., Safwat, S. and El Bakry, M., 2016, July. A MapReduce fuzzy techniques of big data classification. In 2016 SAI Computing Conference (SAI) (pp. 118-128). IEEE.
9. Lei, Y., Jia, F., Lin, J., Xing, S. and Ding, S.X., 2016. An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. IEEE Transactions on Industrial Electronics, 63(5), pp.3137-3147.
10. Xing, W. and Bei, Y., 2019.Medical health big data classification based on KNN classification algorithm. IEEE Access, 8, pp. 28808-28819.
11. Ta, V.D., Liu, C.M. and Nkabinde, G.W., 2016, July.Big data stream computing in healthcare real-time analytics. In 2016 IEEE international conference on cloud computing and big data analysis (ICCCBDA) (pp. 37-42). IEEE.
12. Sahoo, P.K., Mohapatra, S.K. and Wu, S.L., 2016. Analyzing healthcare big data with prediction for future health condition. IEEEAccess, 4, pp. 9786-9799.

13. Hossain, M.S. and Muhammad, G., 2017. Emotion-aware connected healthcare big data towards 5G. IEEE Internet of Things Journal, 5(4), pp.2399-2406.
14. Mohamad, M., Selamat, A., Krejcar, O., Fujita, H. and Wu, T., 2020. An analysis on new hybrid parameter selection model performance over big data set. Knowledge-Based Systems, 192, p.105441.
15. Prakash, S. and Sangeetha, K., 2020. Deep multilayer and nonlinear Kernelized Lasso feature learning for healthcare in big data environment. Journal of Ambient Intelligence and Humanized Computing, pp.1-11.
16. Game, P.S., Vaze, V. and Emmanuel, M., 2019. Optimized Decision tree rules using divergence based grey wolf optimization for big data classification in health care. Evolutionary Intelligence, pp.1-17.
17. Ren, L., Meng, Z., Wang, X., Zhang, L. and Yang, L.T., 2020. A data-driven approach of product quality prediction for complex production systems. IEEE Transactions on Industrial Informatics.
18. Bertino, E., Ooi, B.C., Yang, Y. and Deng, R.H., 2005, April. Privacy and ownership preserving of outsourced medical data. In 21st International Conference on Data Engineering (ICDE'05) (pp. 521-532). IEEE.
19. Leggas, D., Henretty, T.S., Ezick, J., Baskaran, M., von Hofe, B., Cimaszewski, G., Langston, M.H. and Lethin, R., 2020, September. Multiscale Data Analysis Using Binning, Tensor Decompositions, and Backtracking. In 2020 IEEE High Performance Extreme Computing Conference (HPEC) (pp. 1-7). IEEE.
20. Green, O., Fox, J., Watkins, A., Tripathy, A., Gabert, K., Kim, E., An, X., Aatish, K. and Bader, D.A., 2018, September. Logarithmic radix binning and vectorized triangle counting. In 2018 IEEE High Performance extreme Computing Conference (HPEC) (pp. 1-7). IEEE.
21. He, D., Lu, G. and Yang, Y., 2019. Research on optimization of train energy-saving based on improved chicken swarm optimization. IEEE Access, 7, pp.121675-121684.
22. Wang, J., Cheng, Z., Ersoy, O.K., Zhang, M., Sun, K. and Bi, Y., 2019. Improvement and application of chicken swarm optimization for constrained optimization. IEEE Access, 7, pp.58053-58072.
23. Liang, S., Fang, Z., Sun, G., Liu, Y., Qu, G. and Zhang, Y., 2020. Sidelobe reductions of antenna arrays via an improved chicken swarm optimization approach. IEEE Access, 8, pp.37664-37683.

24. Qiao, W., Huang, K., Azimi, M. and Han, S., 2019. A novel hybrid prediction model for hourly gas consumption in supply side based on improved whale optimization algorithm and relevance vector machine. IEEE Access, 7, pp.88218-88230.
25. Ahmed, M.M., Houssein, E.H., Hassanien, A.E., Taha, A. and Hassanien, E., 2017, September. Maximizing lifetime of wireless sensor networks based on whale optimization algorithm. In International conference on advanced intelligent systems and informatics (pp. 724-733). Springer, Cham.
26. Jiang, T., Zhang, C. and Sun, Q.M., 2019. Green job shop scheduling problem with discrete whale optimization algorithm. IEEE Access, 7, pp.43153-43166.
27. Chen, H., Yang, C., Heidari, A.A. and Zhao, X., 2020. An efficient double adaptive random spare reinforced whale optimization algorithm. Expert Systems with Applications, 154, p.113018.