OPEN ACCESS
Saudi Journal of Engineering and Technology
Abbreviated Key Title: Saudi J Eng Technol
ISSN 2415-6272 (Print) |ISSN 2415-6264 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: https://saudijournals.com

Original Research Article

Performance analysis of feature selection and classification in Big Data Information extraction

Manjunatha Swamy, C[1*], Dr. S. Meenakshi Sundaram[2], Dr. Lokesh, M. R[3]

[1]Research Scholar, Department of CSE, GSSS Institute of Engineering & Technology for Women, Affiliated to Visvesvaraya Technological University, Belagavi, Mysuru, Karnataka, 570016, India
[2]Professor & Head, Department of CSE, GSSS Institute of Engineering & Technology for Women, Affiliated to Visvesvaraya Technological University, Belagavi, Mysuru, Karnataka, 570016, India
[3]Professor, Department of CSE, Vivekananda College of Engineering & Technology, Affiliated to Visvesvaraya Technological University, Belagavi, Puttur, Karnataka, 574203, India

DOI: 10.36348/sjet.2023.v08i03.002 | **Received:** 16.02.2023 | **Accepted:** 12.03.2023 | **Published:** 19.03.2023

***Corresponding author:** Manjunatha Swamy, C
Research Scholar, Department of CSE, GSSS Institute of Engineering & Technology for Women, Affiliated to Visvesvaraya Technological University, Belagavi, Mysuru, Karnataka, 570016, India

Abstract

Purpose*:** Information extraction from big data is improved by either reducing the number of features in a data set or selecting features using intelligent data analysis. Generally, big data sets are complex to process using traditional approaches. Feature selection is highly essential in big data information extraction because it chooses the subset of features that influence the final classification. Reducing the number of selected features in the data leads to enhanced accuracy and efficiency of data extraction with other attributes used in the mathematical model. This work aims to improve the performance of the classifier using an enhanced binary bat algorithm-based effective feature selection model. formulated to enhance accuracy, efficiency of data extraction with other attributes. An enhanced binary bat algorithm (EBBA) proposed to solve the mentioned problem using local optimization and global optimization factor which improves the performance of optimization. Experiment carried out with different datasets selected to test effective performance of proposed algorithm and demonstrated performance is better with other algorithms. ***Design*:** The purpose of this paper is to provide, an effective feature selection model for big data information extraction. An enhanced binary bat algorithm has been proposed to improve attribute selection using local optimization and global optimization methods. Classification of multisource data using selected features using labeled approach. Particular Information extraction for multi view multi label (PIMM) approach is compared with EBBA algorithm. Further to enhance effectiveness of shared and specific information in big data [3] by setting the delta and omega factors in order to fuse different information from different view point, Online analysis of relevance with any redundancy analysis also been incorporated. ***Findings*:** All the experiments were carried out with different datasets on the number of iterations and fitness of the attributes to validate the effective performance of the proposed algorithm. Experimental results and graphs show that the proposed methodology improves the overall performance of optimization using PIMM models. ***Originality: A feature selection model based on the binary bat algorithm has been the focus of this paper. Subset selection and feature ranking are the two important methods used in this approach. Experiments were conducted on datasets to analyze the patterns in the number of iterations and fitness of the attributes over selection. The improvement in feature selection leads to better classification accuracy of the proposed model compared to other nature inspired techniques.

Keywords: Information Extraction, Feature Selection Classification, Bio inspired Concept, EBBA model, Multi view multi object model, PIMM model.

INTRODUCTION

Accuracy of classification in high dimensional data using feature selection model improves the information extraction in binary bat algorithm, selection of optimal subset in available dataset, selected data enhanced to improve classification [1] performance of actual dataset, further noise is eliminated which help the algorithm to execute faster and more efficient. Feature selection involves filters and wrappers approaches to measure the relevance score of each feature, wrapper use classifier to evaluate selected subsets obtained by the algorithm later any modification required also considered. To optimize [11] selected features heuristic

approaches is proposed in genetic algorithm, Particle swarm algorithm and Ant colony algorithm. Here proposed bat algorithm as more conventional to select features in bio inspired algorithms, these algorithms used to develop new techniques to increase the robustness of algorithm. Many intelligent algorithms [10] available such as ant colony algorithm, Particle swarm, Genetic algorithms, Grey wolf algorithm and Elephant search algorithm, among these algorithms Bat algorithm is demonstrated as effective algorithm to extract feature information [6] as it is flexible to enhanced. Data mining application supports different phases of application as data mining these includes preprocessing, discovery of pattern and evaluation.

To add more flexibility in handling multiple data sources and to use them to discover models in automatic personalization system. In order to match user tendency association rule is added with clustering to deliver a framework to provide effective solution Classification [7] used extensively in data mining to perform preprocessing and feature selection. Here using discretization in preprocessing to transform the continuous value to integer value known as discrete which may be supervised and unsupervised. This is to find best set of break point in the group of data as a limit to yield to best accuracy in discrete data. Many other algorithms such as Chi-square, K-Means classifier used to transform continuous data, K Means technique used to generate better decision rate and accuracy. Response time of proposed algorithm is just outperformed with precision, recall and other parameters.

Many real time applications objects will have different representations in different views, the communication between different views as each view is contributing uniquely and multi label prediction is also considered in the proposed model to calculate the gradients as to update weights of network given below using PIMM approach, Later in order to extract specific information of a particular view but we do extract from base information by excluding all shared information, the whole framework is to minimize the loss with respect to parameters of PIMM model. Every iteration verifying with multi label sets if the condition is holds good then fitness of data will be measured using standard function. As observation illustrates that each label is associated with unique feature with data, then label is added with function add() to verify the adaptability, then combination suitability is constructed if not associated then standard function used to generate data with random() function

The flow of paper starts with defining feature selection [5], information exaction, followed by related work defines an analytical model, EBBA modeling with comparative discussion. Defining multi label classification [2] approach in big data information exaction, followed by related work which defines detailed about methods and how pattern recognition works followed with architecture and algorithm to define process of multilabel approach. Mathematical model to support implementation part, later comparison of various attributes like Accuracy, HLoss, HLoss(D,L), Mean and standard deviation to give how evaluation metrics help to come back with better improvement in performance.

RELATED WORK

A. Bigdata approach in Information Extraction

Midhun Mohan used Big data approach for classification [1] used as amount of data is generated and stored is very random, information is hidden in database in order to improve performance, using predictive model approach, where data is modeled to make information is available.

Bigdata analytics with multi classifier approach addressed by priyanka Hiranandani [2] to massive data generated with high speed which is very difficult to analyze with traditional tools, Hadoop is used to extract information from large massive dataset as well, Integrating Hadoop with R can be achieved great scalability and to enhance visualization and data transformation capabilities.

Multilabel is classified into mainly two groups one is problem transformation method is used to track multi label scenarios with other problems and binary relevance method is to transform multi label learning to binary classification to obtain ranking along with adaptation algorithm takes multiple algorithm[8] to handle multi label data directly. Multi view collects the view of other method to overcome drawbacks or weakness to improve performance factor. In analysis of duplicates labels on Extracted Objects by Stephan ortona proposed three step algorithms to perform validation blocking and scoring which further focus on ontology constraints and entity extraction system to boost extraction process by using wrapper function over data.

B. Information Exaction using Bat algorithm

Suganthi and malathi [6] discussed about Bat algorithm for feature selection in dataset, which eliminates inappropriate, repeated data from original data set. Random sampling is used to select instances in dataset. Random forest is used to get train from selected features to enhance classification accuracy. Bat algorithm works with iterations and advanced with capability of echolocation behavior and two attributes such as loud, pulse of sound to select prey an echo is generated that will return back to bat to find an object, also finds distance thus it calculates distance in between. It also differentiates between objects and prey allows them to hunt in darkness.

Niraj kumar proposed Text classification [1] and topic modeling of web extracted data in 2021 which

focus on topic models with LSA algorithm, probabilistic semantic analysis and machine learning classifier approach which improves the performance of classification with bag of words model to improve accuracy and to improve dimensionality.

Adaptive systems built using convolution approach is based on image classification of data aligned with localization. Another paper proposed technique by Yanwei with unique concept of structured web data extraction[15] using forest concept and able to get better performance of system over three attributes as mentioned earlier. With all these methods summarized and proposed model achieved best in performance with attributes precision, recall and Fscore.

Information extraction using pattern recognition with multilabel concept solves major problems of existing approaches. In web extraction[13] proposed by super string algorithm extended with pattern matching algorithm to extract data from the web pages without any computational impacts on the system. Here they used crawling approach, rule based method, learning based method to fetching information efficiency and provides cost comparison analysis, noise and redundancy, mean and standard deviation with loss average analysis by which performance is analyzed.

Data can be taken as instances into labeling process where each data can recognized throughout the system, variety of data can be identified which helps extraction process and avoids the time taken to complete effectively, data set which is accessible to Central full-text. Web page information extraction using deep learning also introduced which involves data scraping application which is not affected by structural changes in web pages was developed.

Mihai Surdenau and ramesh, focused on multi instance and multilabel learning for extracting relations in web, using approaches like deterministic model, distant supervision model using attributes relation level classifier with not much effectively incorporating data sets as resources are distributed, moderately effective in nature. Wook shin Han discussed on supportive effective when extraction mainly focus on spatial relationship using elements of DOM tree when web page rendered in browser using approach robust tuple extraction system with spatial relationship.

II. BIGDATA ANALYTICAL MODEL TO EXTRACT INFORMATION

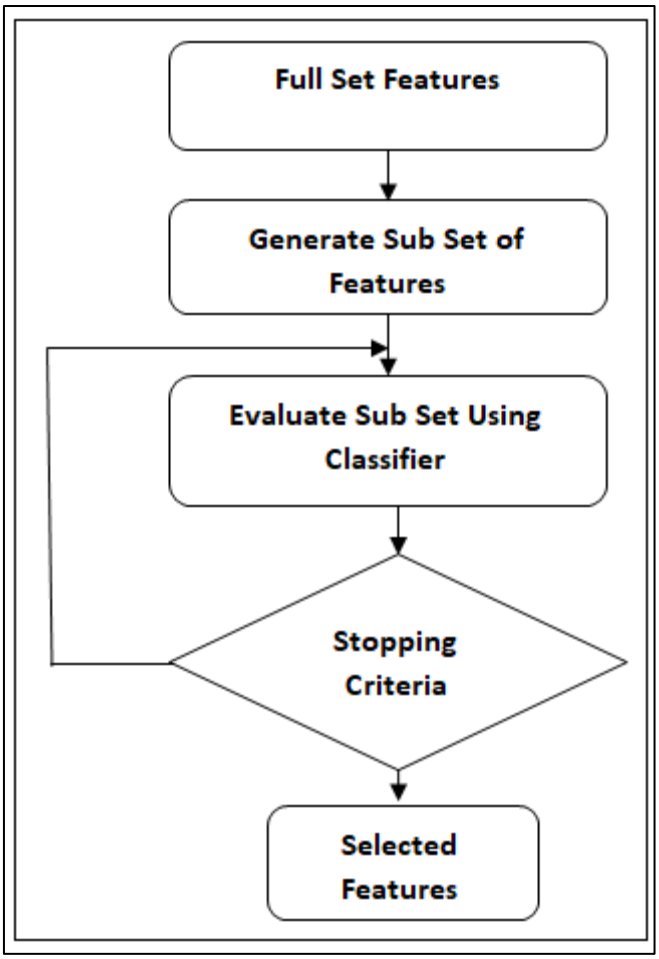


Figure 1: Architecture for Big Data Information Extraction Evaluation

The architecture of web information extraction is shown above which represents how data is received as sample data and collected from multiple sources used to manipulate and analyze representative subset of data points and to find data patterns, which classified into trained data of data set to grow single tree and data to estimate the errors. Data is trained to understand how to apply technologies and to make decisions.

Then feature selection is used to transform raw data collected as inputs to the model then that data is required to some algorithms. This process reduce the number of features by creating new features from existing data which helps to reduce set of features, data applied to select predictors to split data using best predictors so that estimate errors by applying tree to data if not true repeat until stop tree growth is fulfilled. data scraping performed to make data feasible, After estimation with data errors if it as expected go to next level that is random forest by collecting all trees else repeat until specified number of trees to obtained. Sample data and estimation process is correlated with each other, grow tree and feature selection of predictors is related to each other hence random forest algorithm will work as described. Data estimation makes data consistency to avoid failures.

4. Proposed EBBA Algorithm Based Information Extraction

In the proposed approach, information extraction using the Bat algorithm with K-Mean classification is described below:

4.1 Enhanced Binary Bat Algorithm for Information Extraction

An algorithm shown below demonstrates how a random tree classifier which classifies data to make decisions, training with replacement of data to produce a new data set. Later, a new tree is constructed with instances. If only one instance belongs to that tree, it returns that itself; otherwise, it selects randomly by splitting available features in the set. The tree with nodes N, Frequency F is modified into possible value sub child nodes and the procedure is repeated until trained error-free content is obtained. The bat algorithm is provided below, with input, output, and various steps involved.

Extraction process is improved based on the factors used as frequency, loudness factors to generate new solutions to demonstrate how random tree classify data to make decisions by selecting suitable classifier then training of data with replacement to produce new data set.

Input: A trained dataset with the attributes frequency and loudness

Output: Preexistence Dataset, Optimized Data Set

Step 1: Data set in the population with specific parameters for frequency, velocity, and loudness. Set the pulse rate. Pi and loudness Ai

Step 2: Generate new solutions by updating frequency, velocity and positions, while (t< maximum iteration)

Step 3: Apply random function concept.if(rd>ri)Select solution among best solutions randomly obtain Ls solution around BsEndif

Step 4: Generate Rs using Randfly() Randfly( )if(Rd<Ai)& f(xi)<f(B)Get Solution increase Ri and Ai

Repeat Endif Obtain Rank and Get Final best Fb.

The algorithm clearly states that the data is a randomly trained data set that is then checked for reproducibility with a preexisting data set. Possible features of the data are divided to create an N number of child nodes as an instance with relevance to the dataset taken from the build tree to optimize the information. The proposed algorithm avoids limitations in data lists with repetitive occurrences.

A. Adopted multilabel based classification Algorithm for Bigdata Extraction

An algorithm shown below, demonstrate how ensemble technique which ensure better performance obtained from any of the dataset and to compare two or more different analytical model and to synchronize results too increase accuracy of data retrieval methods with respect to boost random forest model is an best approach, also to increase classification [15, 17] performance of a model. Every iteration verifying with multi label if the condition is holds good then fitness of data will be formulated using function fitness(). Each label is associated with unique feature [21] with data then label is added with function add(), then combination suitability is constructed if not associated then sub_child=cross_over(p1,p2) mutation=TRUE

mutation mut_child=mutation(p1, p2) To get new possibilities.

Generate sub nodes of set, as p1,p2….pn if, here F is associated with (F1…,Ff)

do for I range from i=1 to f

Recall to function Cross_over(p1,p2) if ends

for ends

In the algorithm [30] it is very clearly specified that data is randomly trained and rechecked for reproduce with preexisted data set. Using label approach data is combined with different patterns and possible features of data set create N number of child nodes as instance with relevance to the dataset taken build tree to resolve the efficiency and to optimize information. Algorithm proposed avoids limitations in with data lists with repetitive occurrences.

4.3 MATHEMATICAL MODEL

In this section, equations related to variance,

sample variance with standard deviation and sample data sets are presented. In this approach, *n* indicates data entries for the mean when working with population data sets, population data contains all members of the data set, such as a part, subset, or solution by deciding whether to proceed with a sample or the entire population.

In the below equation (1), K-mean clustering is used to partition many objects into k-clusters. Every object belongs to the nearest mean of a cluster. This method produces different clusters with distinction. Using the priori concept, computation is performed to minimize total variance and the squad error function.

$$J = \sum_{i=1}^{n} xi \sum_{j=1}^{k} (x_{ji-cj)2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

In equation (2), the data set has all other relevant data with a specific feature set with all possible values. Sampling is always part or subset of available data. n is number of samples and x1,x2,x3,…xn are all the subsets of data taken with their mean.

$$Sv = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

In the below expression (3), chi-square is used to find a relationship between independent feature categories is degree of freedom with observed attributes such as O. E is expected feature values that meet intended values.

$$X_c^2 = \sum \frac{(S^i - F^i)}{(F^i)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

A. MATHEMATICAL MODEL

In accuracy set of labels are predicted for a sample must match exactly corresponding set of labels in less ambiguity with exact match ratio 1 where number of correct predictions divided by total number of examples if considered that prediction is correct if and only if predicted binary vector equal to binary vector. In expression n is the number of training examples, Yi is the true labels for ith training example and jth class, predicted labels for the ith training example, also Yi is the target and Zi is the prediction. Evaluation metrics for multi label classification that we used when comparing performance of four models, D is the multi label data set with |D| instances each has set of labels Yi and H be the classifier and Zi be the predicted labels of an instance Xi then, common labels and intersection between two data set is analyzed.

Hamming loss can be used in multi label classification helps in identifying fraction of wrong labels in total number of labels, in multiclass classification hamming loss can be calculated as the hamming distance y_true and y_pred here in multi label classification hamming loss focus on individual labels. When compared to three techniques in terms of accuracy score, binary relevance and label powerset techniques will be suited for multilabel classification due to their higher accuracy score as given in equation below.

In the above equation αβϒ are the factors which control loss term, interaction which then control the model to have output labels. Pi, Qi and Ri are the three different labels and Q j is label of i on j and denotes relevancy of labels created. mlmv can be calculated as in independent mode communication between all views is conducted.

5. RESULTS AND DISCUSSION

5.1 Implementation

In the proposed work, selection of data is done through a CSV (Comma separated values) file and is integrated and analyzed with other algorithms like the particle swarm optimization algorithm. Furthermore, the associated page information in the bat algorithm with the wrapper approach is used to practically measure the performance. In equation (12), the data sampling size is obtained with the original data set, where N is the size of the population.

$$n = \frac{n_0 N}{n0+(N-1)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

In the below equation (13), the data set is split into different attributes and entropy is calculated from the split data. Some attributes T, X are calculated to find the probability of entropies with extension. The resultant set is obtained with product to features.

$$E(T,X) = \sum_{c\in X} P(c)E(c) \quad \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

The efficiency of various wrapper functions and associated time is very high. Each time, according to the training mode as wrapper functions change. The time taken to validate the information using random functions in many models does not allow for the desired selection. It is dynamic in nature, as observed in the above equation. So for each model, the wrapper function is different.

Criterion on data set is considered as below equation

$$CFS=\max(Sk)\left[\frac{rcf1+rcf2+rcf3+\cdots.rcfk}{\sqrt{k+2(rf1f2+\cdots.+\cdots.rfifi+\cdots.+rfk,fk-1)}}\right] \quad \ldots\ldots\ldots (6)$$

Here correlation of feature selection is represented as CFS, rcfi, rfi are the correlation between data uses different measures which helps in selection of variables, attributes and variable subset selection with features. To make validation simple data is divides as fraction of data like 80% and 20% means 80 % of date is chosen over 20% of data ad will be used for testing, mapping between inputs and outputs is always in set. Data is to be tested for errors,

$$f(x) = x2 + 4 \quad \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

Above expression is to compare and to substitute the defined data in sample,

$$f(x) = \begin{cases} \sin(x) \; for \; x < -2 \\ 2 - \frac{x}{2} \; for \; -2 \leq x < 2 \\ x2 - 8x + 10 \; for \; x \geq 2 \end{cases} \quad \ldots\ldots\ldots (8)$$

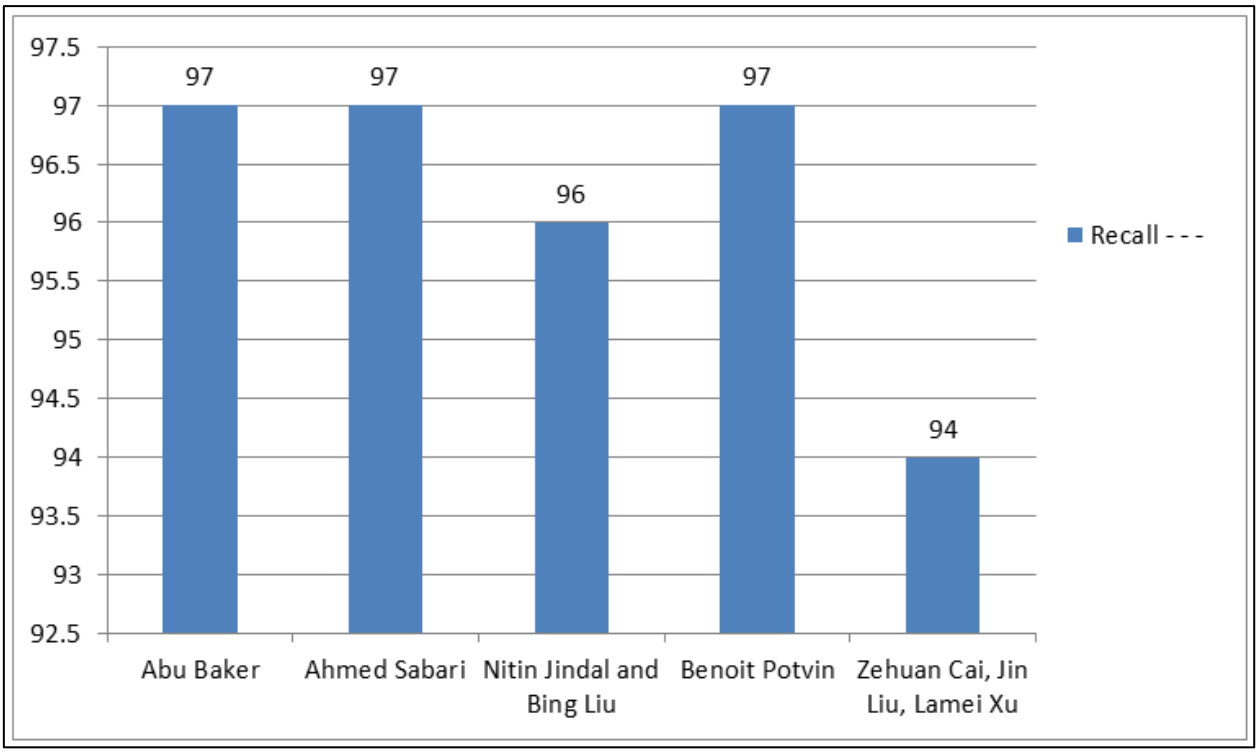
Here in this equation first case is defined as sin(x), second case as straight line. To estimate errors in data set use of accepted value with experimental values.

5.2 Software Requirements:

To demonstrate the significance improvement in the proposed algorithm is tested using data set, comparative methods and evaluation metrics. The data set considered is ionosphere.csv file with 34 features set, 352 rows of data and validation factor considered as 70 to 30 over the split up data using metrics as xtrain, xtest, ytrain and ytest. Other parameters as K value in KNN, number of variables, maximum number of iterations to perform feature selection. The data is modeled with selected features as number of train, number of validation to increase the accuracy and to obtain data convergence and used python libraries.

In general, three approaches are usable: web mining usage, content mining on the web, and structure mining. In addition to the combined tags and value similarity, DOM (Document object model) tree structures can also be used. Redundant data records RDR rule, QRR query-related record extraction, operator used The DOM model, the Machine learning method, and the successive steps of the proposed method, the iForest anomaly detection algorithm are listed.

Random Forest and Multi-Layer Perceptron (MLP) classifiers are also used in order to address web information extraction and make it more efficient. Fitness of attributes has been obtained for a number of iterations, as shown in section 4, which describes how algorithm yields better fitness with number of iterations, the algorithm affects fitness attributes over n number of iterations using -1 iteration values and in Figure 6 elaborated to describe KNN attributes using N iterations in each case it significantly shows the improvement in obtaining information extraction.



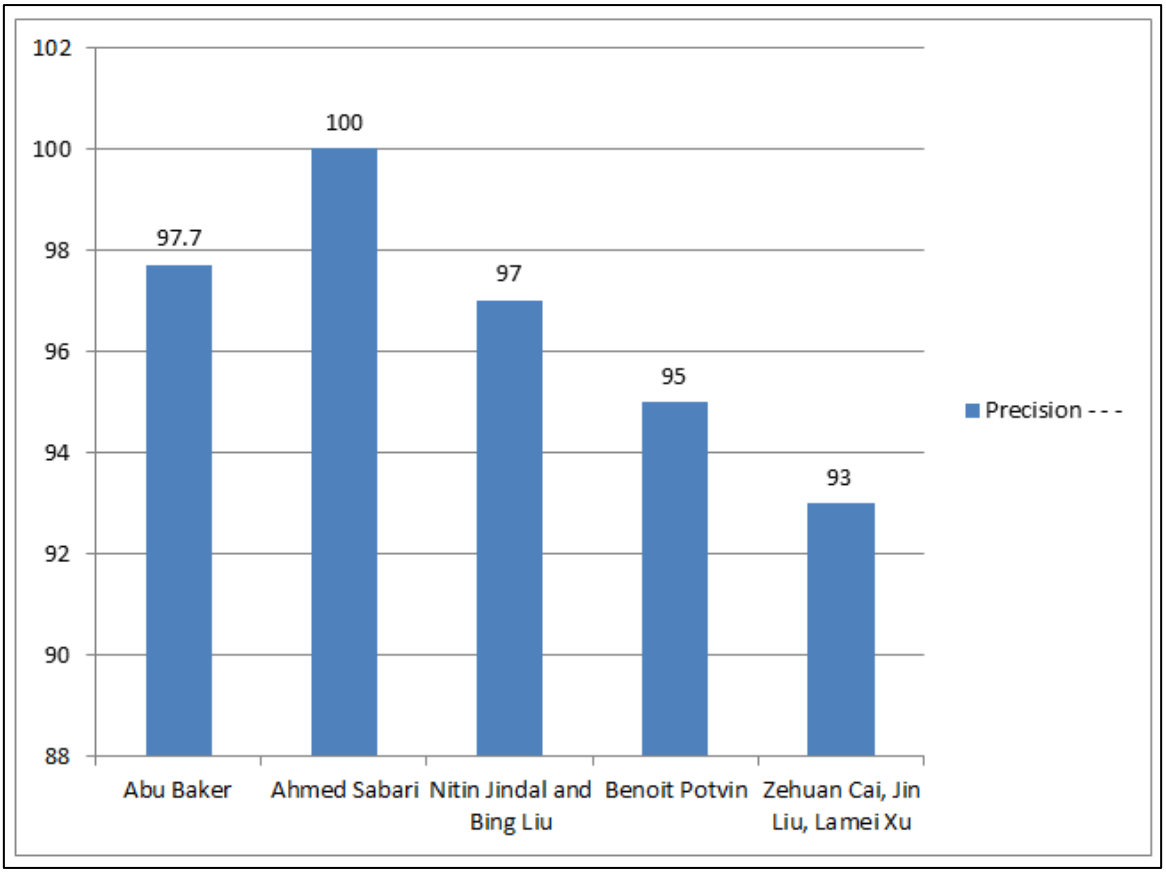


Figure 2: Performance comparison precision and recall attributes

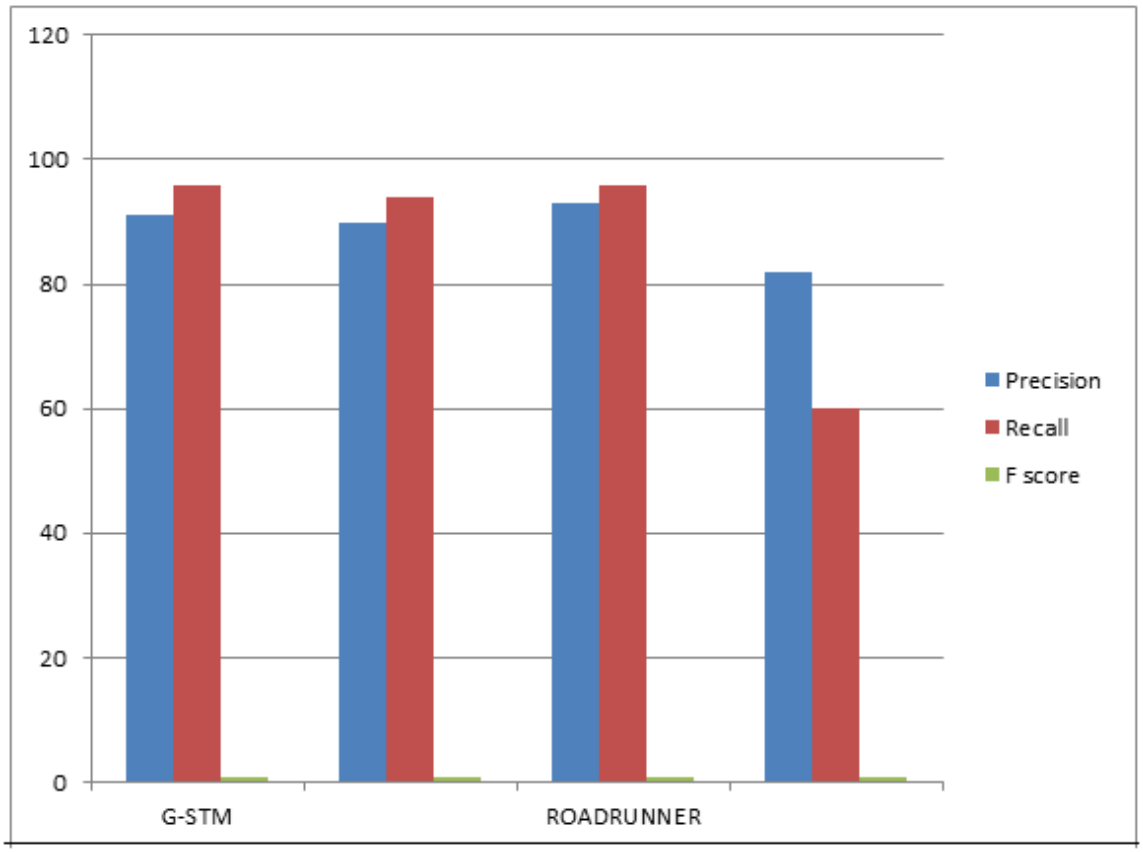


Figure 3: Attributes comparison

A. Evaluation matrices for multi label information Extraction

In below diagram noticed that analysis clearly shows the significance of the multi view and multilabel approach to make more accurate and improve the performance each and every method existing shows discrepancy with HLoss parameter where its maximum, so PIMM method will minimize that loss to help web information extraction more effective, with respect to data set D instances and set of labels our approach improves. Mean and standard deviation is also important parameter to ensure effective information extraction.

5.2 Comparative analysis and discussion

Traditional systems and advanced system models use automated methods to extract information. The contents and model structure play a vital role in establishing relations between a page and page level attributes. If there are any changes in the model, then the wrapper function will enable the deployment to make the function work. Compared to our model, it's not possible to ensure the efficiency and here more complication and achieve better extraction.

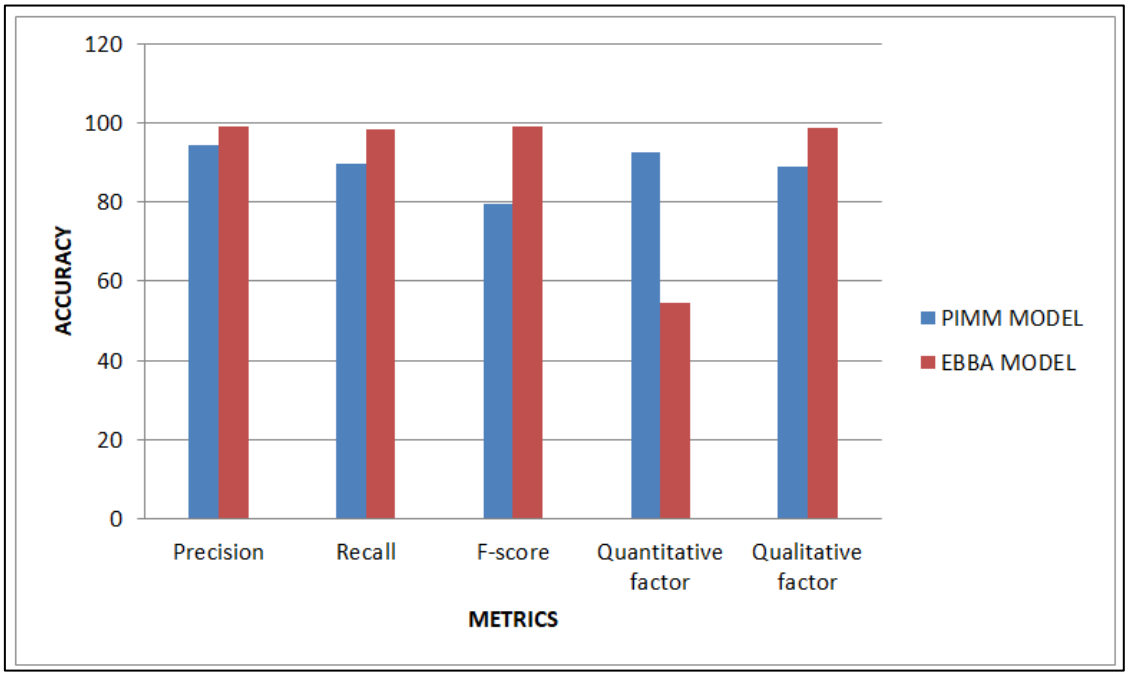


Figure 4: PIMM Model comparison with EBBA Model to compare accuracy

In Figure 4 comparison of two models used in data extraction with bigdata analytics concept. Different metrics where accuracy is analyzed to show performance of two methods. Dom Tree Model, the Mining data records MDR Algorithm, the improved HMM (Hidden Markov Model), and the Long short text classification method LSTM used in neural networks are compared for attribute accuracy and flexibility in selecting designs when changes occur in websites. Clustering algorithms are applied to formatted and unformatted records in data slots. If any modifications in layout design are not accommodated, these shortfalls are used to generate layout designs with different patterns. The Random Forest Algorithm (RFA) shows much better results of personalization with 4-parameter recall, an F score, and qualitative and quantitative attributes with various data sets from performed experiments to tackle variations and increase efficiency.

Accuracy is compared with various methodologies as given proposed method shows standard accuracy factor compared to other methods where one more method discussed by kiran adnan achieved accuracy with 96.3 other models is not commanding over the factors. Syed usama again discussed the effect of multi label approach to obtain accuracy much better way having drawbacks with n number of labels that will be improved by 90.1 shows elementary improvement in accuracy. Proposed methods EBBA and PIMM model demonstrated better improvement in the accuracy.

Table-I: Comparison of two algorithms for performance evaluation

Authors	Precision	Recall	F-score	Quantitative factor	Qualitative factor
Proposed EBBA	99.14	98.5	98.9	54.5	98.8%
Proposed PIMM model	97.5	90.2	70.5	95.3	98

6. CONCLUSION

Feature selection has been utilized in most of the data preprocessing techniques, but it will have a major impact in the context of big data information extraction. An EBBA algorithm-based feature selection model is proposed in this paper for improving information extraction from big data. Feature subset selection and feature ranking are the two important methods used in this approach. Experiments were conducted on datasets to analyze the patterns in the number of iterations and fitness of the attributes over selection. The proposed model EBBA shows a significant improvement of 3% in feature selection compared with PSO algorithm.

Classification task with classes and each label is mutual exclusive and each sample of data is assigned to only one label generally, Information extraction collects different documents as input and produce different representations of relevant information with different criteria. EBBA method compared with PIMM model is put forward to handle many cases with accuracy factor. Multi view representation and integrating view specific discriminating modeling is not much considered. Further to enhance effectiveness of shared and specific information by setting the delta and omega factors in order to fuse different information from different view point, Online analysis of relevance with any redundancy analysis also been incorporated.

REFERENCES

1. Al-Thanoon, N. A., Algamal, Z. Y., & Qasim, O. S. (2021). Feature selection based on a crow search algorithm for big data classification. *Chemometrics and Intelligent Laboratory Systems*, *212*, 104288.
2. Al-Thanoon, N. A., Algamal, Z. Y., & Qasim, O. S. (2021). Feature selection based on a crow search algorithm for big data classification. *Chemometrics and Intelligent Laboratory Systems*, *212*, 104288.
3. Subasi, A., Molah, E., & Almkallawi, F. (2019). "Intelligent website detection using random forest classifier", ICCIS, 2019.
4. Lin, B. Y., Sheng, Y., Vo, N., & Tata, S. (2020, August). Freedom: A transferable neural architecture for structured information extraction on web documents. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1092-1102).
5. Potvin, B., & Villemaire, R. (2019). Robust web data extraction based on unsupervised visual validation. In *Intelligent Information and Database Systems: 11th Asian Conference, ACIIDS 2019, Yogyakarta, Indonesia, April 8–11, 2019, Proceedings, Part I 11* (pp. 77-89). Springer International Publishing.
6. Jeon, D., & Kim, W. (2015). Random forest algorithm for linked data using a parallel processing environment. *IEICE Transactions on Information and Systems*, *98*(2), 372-380.
7. Gill, S. S., & Buyya, R. (2019). Bio-inspired algorithms for big data analytics: a survey, taxonomy, and open challenges. In Big data analytics for intelligent healthcare management (pp. 1-17). Academic Press.
8. Gutierrez, F., Dou, D., Fickas, S., Wimalasuriya, D., & Zong, H. (2016). A hybrid ontology-based information extraction system. *Journal of Information Science*, *42*(6), 798-820.
9. Hiranandani, P., Pilli, E. S., Chand, N., Ramakrishna, C., & Gupta, M. (2018, January). Big Data Analytics Using Multi-Classifier Approach with Rhadoop. In 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 478-484). IEEE.
10. Ji, B., Lu, X., Sun, G., Zhang, W., Li, J., & Xiao, Y. (2020). Bio-inspired feature selection: An improved binary particle swarm optimization approach. *IEEE Access*, 8, 85989-86002.
11. Jeyasingh, S., & Veluchamy, M. (2017). Modified bat algorithm for feature selection with the wisconsin diagnosis breast cancer (WDBC) dataset. *Asian Pacific journal of cancer prevention: APJCP*, 18(5), 1257.
12. Mohan, M. M., Augustin, S. K., & Roshni, V. K. (2015, December). A BigData approach for classification and prediction of student result using MapReduce. In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) (pp. 145-150). IEEE.
13. Mlakar, U., Zorman, M., FisterJr, I., & Fister, I. (2019). Modified binary cuckoo search for association rule mining. *Journal of Intelligent & Fuzzy Systems*, 32(6), 4319-4330.
14. Qi, C., Zhou, Z., Sun, Y., Song, H., Hu, L., & Wang, Q. (2018). Feature selection and multiple kernel boosting framework based on PSO with mutation mechanism for hyperspectral classification. *Neurocomputing*, 220, 181-190.