Received: 15 May 2024 | Revised: 31 July 2024 | Accepted: 15 April 2025 | Published online: 8 May 2025

RESEARCH ARTICLE

Artificial Intelligence and Applications
2025, Vol. 00(00) 1–9
DOI: 10.47852/bonviewAIA52023432

BON VIEW PUBLISHING

Enhancing Big Data Classification Accuracy Through Deep Learning Techniques

Renuka Devi D.[1,*] **and Swetha Margaret T. A.**[1]

[1]*Department of Computer Science, Stella Maris College (Autonomous), India*

Abstract: Classifying data stands as a pivotal stage within the machine learning process, wherein extracting insights from vast datasets poses a formidable challenge. Within the realm of big data research, numerous methodologies have been employed to tackle these obstacles. Machine learning methodologies must evolve to effectively address the burgeoning challenges and complexities inherent in research. Deep learning methodologies have emerged as a solution for big data classification, effectively managing the rapid influx of data through deep neural networks. These networks' multi-layered architectures excel in discerning patterns within extensive datasets. Real-world applications, such as speech recognition, sentiment analysis, prediction, and recommender systems, prominently feature the utilization of deep learning algorithms. This study integrates incremental learning with the Deep Multiple Layer Perceptron utilized as a classifier. Experimental results encompassing six datasets showcase notable enhancements in classification accuracy. The proposed approach considerably contributed to reduce the processing time; at the same time, incremental deep learning classification has contributed for enhanced accuracy percentage. From the results' observation, the proposed model achieves higher accuracy and less processing time.

Keywords: machine learning, big data, classification, deep multiple layer perceptron, deep learning, incremental learning

1. Introduction

Big data analytics (BDA) [1] is a buzz term for its widespread use in business, education, healthcare, and many more. The advantage of such analytics is that it provides interpretations that are advantageous to society and humankind. The benefits include marketing strategies, scientific developments, prediction systems, customer segmentation, disease prediction, and so on. Many companies have started deploying BDA on a large scale (immense quantity of data), and research interests have boomed up across the world. The aim of the research endeavor is to provide society with valuable insights gleaned from thorough analysis.

The BDA research [2] developments have been categorized into classification, feature selection (FS), storage management, and security aspects. The advantages of BDA are faster analytics of immense data of varied types (multiple sources), effective decision-making at the right time, cost benefits, early prediction of diseases/ weather prediction that may alarm people, optimization of operations across the verticals, a better understanding of social network data and aid in developing better plans, enhanced efficiency, societal development, health care, and medical advancements.

Data management [3] is indeed the greatest challenge because voluminous data have to be stored, maintained, and later processed. In real time, the storage space complexity is higher and the cost incurred for such storage is also higher. The data are gathered from varied sources with a diverse format. The quality of the data is maintained using sophisticated big data tools and techniques. Resources are needed to accommodate storage requirements. The data security system is enhanced to protect the data. The intricate big data ecosystem must address all the security concerns and issues. Numerous big data tools have been developed for storage, processing, and deep analytics. The appropriate tools are selected based on the requirement and type of analytics required. The basics of analytics are classified into main categories based on the mode the analytics is being carried out. Analytics has been used as a new trend in most organizations to support decision-making through the implications.

Descriptive and exploratory analytics [4] are implemented by a set of mathematical models and a set of visualization techniques. Predictive analytics are closely associated with machine learning techniques such as classification, regression, text analytics, and clustering. Prescriptive methods are intended for optimal solutions. Mathematical optimization and stochastic models are developed for this kind of analytics. Classification is the substantial phase in the machine learning process, where deriving implications from huge data is a challenging task. In big data mining research, many approaches have already been in use to address the problems. The machine learning algorithm for big data mining should be adapted in such a way as to accommodate all the intensifying research challenges and issues.

The novel classification model has been projected and assessed across high-dimensional datasets [5]. The objective of such analytics, encompassing medical, sentiment, emotion, psychology, and drug development domains, is to extract insights and identify patterns aimed at enhancing people's lifestyles. A range of dataset types was employed to showcase the adaptability of the proposed model. This analytical approach is vital for societal system enhancement. The research findings indicate enhanced classification accuracy and reduced computation time.

[*]**Corresponding author:** Renuka Devi D., Department of Computer Science, Stella Maris College (Autonomous), India. Email: drenukadevi@stellamariscollege.edu.in

The major challenge in data mining is the classification of big datasets. The big data mining process is classified into either classification or prediction models. Any classification algorithm aims to reduce the error between predicted and actual values. In the case of BDA, the challenge is to achieve enhanced accuracy with reduced time complexity. The selection of learning algorithms is a research challenge based on the nature of the application and problem specificity. Many research aspects have been proposed for big data classification.

The general classification of learning algorithms falls into classes such as supervised, unsupervised, semi-supervised, and reinforcement. Supervised learning methods are employed to identify which category or class labels the given dataset samples belong to. The learning is guided by known output variables. Classification and regression are types of guided learning approaches. Unsupervised learning approaches are used when the classification label is unknown, and the learning algorithm groups similar patterns together. Semi-supervised learning combines both approaches. The reinforcement learning method is based on real-time learning and rewards or penalties.

1.1. Steps in the classification process

The big data mining [6] process is supported by the techniques used for data collection, data cleaning, pre-processing, model building, evaluation, and testing. The significant aspect of this part of the section is to elaborate on the steps and base methods involved in the flow of BDA instigated in this research.

1.1.1. Data collection

The foremost task in data analytics is data collection. It is the process of investigating and exploring the characteristics of the data under study. Data can be accumulated from the repositories or it can be acquired from data warehouses. The proposed model is implemented and tested with datasets from the UCI repository and Twitter datasets from Kaggle.

1.1.2. Pre-processing of data

The data pre-processing [7] is aimed at treating the collected data and removing any anomalies. This step is considered significant because the quality of the data is directly correlated to the outcome of the results. “Data Cleaning” involves treating null values and checking for data integrity. “Data Integration” aims to combine data in different formats into a common form to resolve representation conflicts. “Data Transformation” is applied when the data are on a different range or scale, and it involves generalizing datasets. “Data reduction” is a technique used to convert high-dimensional space into a lower dimension.

1.1.3. Big data classification

The major challenge in data mining is the classification of large datasets [8]. The big data mining process is categorized into either classification or prediction models. Any classification algorithm aims to minimize the error between predicted and actual values. In the case of BDA, the challenge is to achieve enhanced accuracy with reduced time complexity. The selection of learning algorithms is a research challenge based on the nature of the application and problem specifics. The general classification of learning algorithms falls into classes such as supervised, unsupervised, semi-supervised, and reinforcement learning. Supervised learning methods are employed to identify to which category or class label the given dataset samples belong. The learning process is guided by known output variables. Classification [9] and regression are types of guided learning approaches.

2. Background Study

Machine learning algorithms are being considered for classification problems because they can effectively analyze large volumes, speeds, and a variety of data. The recent research summary provides an overview of a bibliometric study investigating research trends in Artificial Intelligence (AI) and BDA across diverse academic domains. It delineates the rising interest in AI and BDA among scholars and practitioners and delineates the methodology employed to gather and analyze data from 711 articles published between 2012 and 2022. The study delineates significant contributors to AI and BDA research, encompassing countries, institutions, and research clusters. Notably, the USA emerges as the most influential nation in terms of citations, while China stands out as the most prolific nation in terms of publications. Additionally, the study identifies five principal research clusters where AI and BDA are extensively explored and forecasts key areas of future research focus. Overall, the summary furnishes valuable insights into the global landscape of AI and BDA research, serving as a resource for both novice and experienced researchers seeking to comprehend current trends and anticipate future directions in the field [10].

An optimized multi-kernel support vector machine (SVM) classifier customized with hyper-heuristic salp swarm optimization (HHSSO) is used to solve the big data classification challenge. By choosing the best feature subsets and fine-tuning SVM kernels, the method attempts to solve the computational time problems with SVM for large data. In order to determine the optimal kernel function and improve SVM parameters, it combines the HHSSO technique with conventional salp swarm optimization (SSO). The MATLAB-implemented approach, assessed on reference datasets, shows enhanced precision and decreased processing time for large-scale data classification [11].

The credit scoring is to financial institutions and how huge data present obstacles for precisely determining the financial trustworthiness of loan applicants. The use of machine learning algorithms that can handle massive volumes of data from sources like social networks is a result of traditional data mining approaches’ difficulties with this task. However, static credit scoring models become outdated over time due to the dynamic nature of customer behavior and variables. The Incremental Adaptive and Heterogeneous Ensemble (IAHE) credit scoring model is a solution, which is made to identify changes in consumer behavior, learn incrementally, and adapt to variable drift. Empirical tests show that IAHE outperforms nine other credit scoring models in identifying default samples and generalization capacity across several datasets [12].

The difficulties in utilizing artificial neural networks (ANN) for Big Data are discussed in the paragraph, which stems from the sluggish convergence rates of conventional learning methods like backpropagation. An ANN learning algorithm based on distributed genetic algorithms is suggested as a solution to this problem. For distributed learning, Genetic Algorithm provides effective parallelization that improves accuracy and convergence. As compared to conventional methods, experimental findings show a considerable improvement in computing time and accuracy, demonstrating the efficacy of the suggested approach for ANN learning with Big Data [13].

Deep learning (DL) plays a crucial role in learning data representations and is characterized by its multilayered nonlinear structure in machine learning. It talks about how model transfer and other techniques like generative and discriminative models have allowed DL to completely transform information processing. A thorough analysis of DL algorithms, such as Multilayer

Perceptrons, Self-Organizing Maps, and Deep Belief Networks, is suggested in this article. It focuses on the latest and historical developments in DL implementation techniques and architectures. Moreover, it classifies different uses of these methods in a range of industries, including natural language processing, wireless networks, speech recognition, and medical applications [14].

The ability of deep learning (DL) algorithms to analyze unstructured data makes them valuable for applications such as natural language processing, image classification, speech recognition, and more. The survey study offers a thorough analysis of DL techniques in BDA, including information on their taxonomy, fundamental methods, intricacies, difficulties, and practical applications. In addition, it examines benchmarked frameworks and datasets, compares and contrasts current methods, and talks about obstacles and future possibilities in DL modeling [15]. The review study offers insights into a range of research projects targeted at BDA and machine learning opportunities and difficulties. In addition to providing significant advances in knowledge and methodologies in these fields, they include studies on bibliometric analysis of trends in AI and BDA, optimization of SVM classifiers for big data classification, development of adaptive credit scoring models, proposal of distributed genetic algorithms for ANN learning, exploration of DL applications across various domains, and thorough surveys on DL techniques in BDA.

3. Proposed Methodology: Experimental Design

The proposed model is illustrated in Figure 1. In the initial phase, input samples are normalized to a common scale. During the FS phase, the MapReduce approach is utilized to enhance accuracy and reduce execution time, addressing scalability issues. FS is carried out in a distributed (parallel) manner using the Accelerated BAT algorithm. This is followed by the classification phase, where the Enhanced Incremental Deep Multi-Layer Perceptron (EIDMLP) classifier is proposed.

3.1. Pre-processing

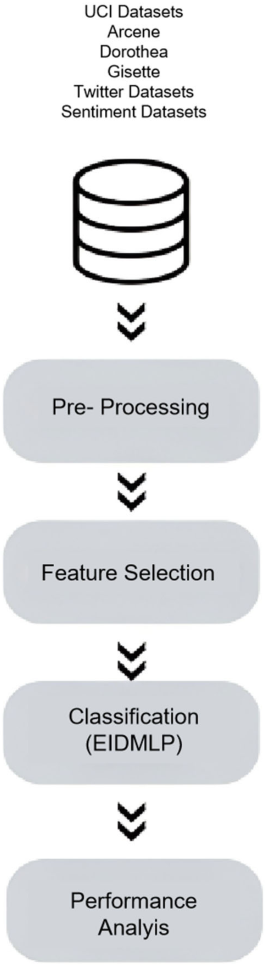
In the initial phase, input samples undergo Min-Max normalization to scale all data to a common range, typically [0,1]. This step ensures that each feature contributes equally to the analysis, preventing features with larger ranges from dominating the learning algorithms.

3.2. Feature selection

During the FS phase, the MapReduce approach is employed to enhance accuracy and decrease execution time, effectively addressing scalability challenges associated with large datasets. MapReduce enables efficient processing by distributing the data across multiple nodes, thereby parallelizing the computation. FS is then performed using the Accelerated BAT algorithm in a distributed (parallel) manner. The Accelerated BAT algorithm optimizes FS by mimicking the echolocation behavior of bats, ensuring that the most relevant features are chosen quickly and efficiently. This distributed approach leverages parallel processing to handle large volumes of data more effectively.

In the MapReduce model, the dataset Ds is divided into subsets and distributed across different nodes. The FS algorithm is applied to each partition, allowing the algorithm to run in parallel. This parallel execution ensures that the subsets are processed while preserving class balance. For a subset *Dsi*, a mapping function called map$_i$ is applied. During the mapping phase, Ds_i is processed using Accelerated Bat Algorithm. The FS algorithm is applied to each partition, and the map$_i$ function generates an output $fe_i = (fe_{i1}, \dots, fe_{iD})$, where D represents the number of selected features. In the reduce phase, the individual outputs from each node are merged to produce the final result. The Accelerated BAT Algorithm is given in Algorithm 1.

Figure 1
Proposed methodology: Classification architecture

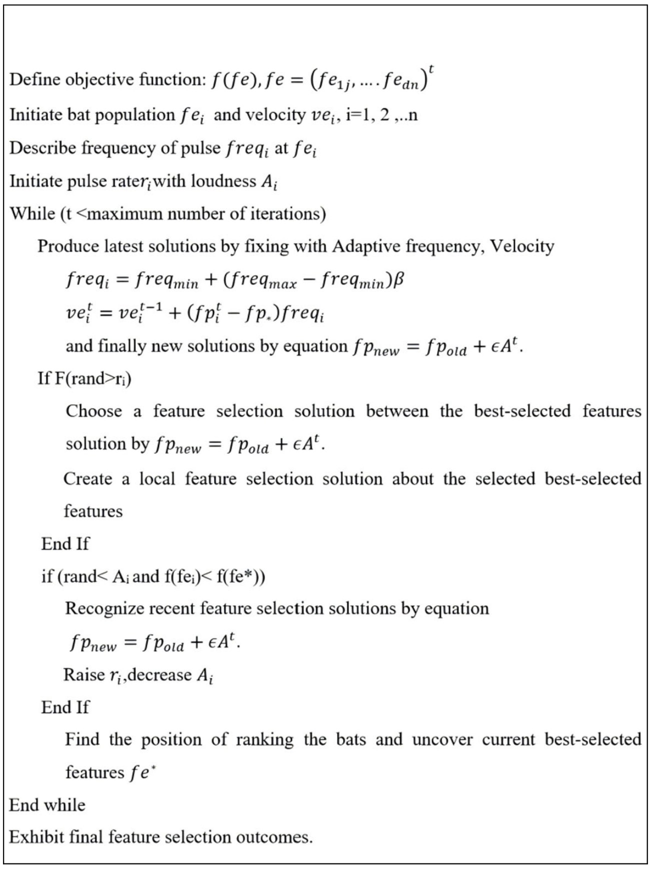


3.3. Classification

Finally, the classification phase employs the EIDMLP classifier. Big data classification is expected to adapt through incremental learning (IL). Many advancements in online and IL techniques have been achieved recently, with a focus on real-time massive data stream classification. The idea behind this approach is to update the induction model gradually rather than rebuilding it every time new instances appear. This kind of situation is frequently encountered in real-time systems where instances are streaming rather than fixed.

To augment big data stream classification, DMLP is employed. An ensemble method combines the classification results of individual classifiers, thus benefiting from classification accuracy. The classification decision is made by a majority of the votes [16] proposed by the classifier. DMLP is a feedforward network [17] that maps the input to corresponding output. It is connected across the input, hidden, and output layers. Layer-to-layer connectivity is established for learning and training. Except for the input layers, the activation function (sigmoid) is applied to the neurons. This network is intended for nonlinear data with three ensembles.

Algorithm 1



```
Define objective function: $f(fe), fe = \left(fe_{1j}, \ldots . fe_{dn}\right)^t$
Initiate bat population $fe_i$ and velocity $ve_i$, i=1, 2 ,..n
Describe frequency of pulse $freq_i$ at $fe_i$
Initiate pulse rate$r_i$with loudness $A_i$
While (t <maximum number of iterations)
    Produce latest solutions by fixing with Adaptive frequency, Velocity
        $freq_i = freq_{min} + (freq_{max} - freq_{min})\beta$
        $ve_i^t = ve_i^{t-1} + (fp_i^t - fp_*)freq_i$
        and finally new solutions by equation $fp_{new} = fp_{old} + \epsilon A^t$.
    If F(rand>$r_i$)
        Choose a feature selection solution between the best-selected features solution by $fp_{new} = fp_{old} + \epsilon A^t$.
        Create a local feature selection solution about the selected best-selected features
    End If
    if (rand< $A_i$ and f($fe_i$)< f(fe*))
        Recognize recent feature selection solutions by equation
        $fp_{new} = fp_{old} + \epsilon A^t$.
        Raise $r_i$,decrease $A_i$
    End If
        Find the position of ranking the bats and uncover current best-selected features $fe^*$
End while
Exhibit final feature selection outcomes.
```

In DMLP, more than 5 hidden layers are constructed compared to the conventional MLP network. In general, sigmoid and tanh functions have shown elevated performance for large-sized networks. The connection is linked between layer (i) and layer (j), and the weight measure we_{ji} is applied to each neuron.

The input (net_i) is given to the layer, and ac_i is the activation output. Even though the classification task is very well handled by this network, big data streams still pose a challenge of updating the streams that may arrive at different intervals. To overcome this glitch, an incremental approach is coupled with this procedure. IL is designed to adapt to big data stream classification by updating the induction model incrementally rather than reconstructing it with each new instance. This approach is particularly useful in real-time applications where data instances continuously stream. The Hoeffding tree is an incremental classifier tailored for big data streams, assuming a stable data distribution. It constructs a tree-like structure by evaluating node conditions using the Hoeffding bound for node splitting decisions. The algorithm begins with a root node and updates each node with sufficient statistics to make further split decisions. The fitness function is calculated for each leaf node, and the top leaf nodes with the highest bound values are identified. The Hoeffding value (B) is generated and compared with information gain, and this process is repeated to enhance the algorithm's efficacy, ultimately splitting nodes and growing the tree accordingly.

Thus, the reconstruction of the learning model is greatly minimized. The model updation is updated incrementally. To combat the aforementioned issues, an incremental DMLP classification model is projected. The frequency of (at_{ij}), for a class (y_k), Hoeffding Bound (HB) [18] is found using Equation (1).

$$HB = \sqrt{\frac{R^2 \ln\left(\frac{1}{\delta}\right)}{2n}} \tag{1}$$

where R is the distribution of class and number of instances(n). At any specific time, the upper value of H(.) is found by Equation (2),

$$at_{ia} = \; argmax \; \mathrm{H}\left(at_{ij}\right) \tag{2}$$

Similarly, the second highest value is at_{ib} found by Equation (3),

$$at_{ib} = argmax \, H(at_{ij}), \; \forall j \neq a \tag{3}$$

The two highest values are incrementally taken as induction leaves. The $\Delta H(at_i)$ is figured by the Equation (4),

$$\Delta H(at_i) = \Delta H(at_{ia}) - \Delta H(at_{ib}) \tag{4}$$

$\Delta H \; (at_i)$ is applied to each attribute (at_{ib}) and $i \in I$ gives the difference between the two highest values. The confidence interval (r_{true}) is calculated for "n" number of instances. This is achieved to correlate the attribute (at_{ij}) to class (y_k). The confidence intervals are updated incrementally for at_i, $r - HB \leq r_{true} < r + HB$, where $r = (1/n)\sum r_i$ is maintained. When $r_{true} < 1$ is true for samples, then at_i is considered as the best statistical candidate with enhanced accuracy.

The classification output of each ensemble is collated with a majority voting rule.

Consider the following notations.

For each $Tr \in r_i$, the prediction is done for the classifier (Q) that have all the predicted classes. The majority of the classes are identified by the maximum votes as described below.

Let $cl_l \in CL$ describe the class and Tr is predicted by a classifier Al, and a counting function (F_k) is given in Equation (5),

$$F_k(cl_l) = \begin{cases} 1 \; cl_l = cl_k \\ 0 \; cl_l \neq cl_k \end{cases} \tag{5}$$

where cl_l and cl_k = Classes of CL.

The accumulation of votes for cl_k by a voting function is given in Equation (6),

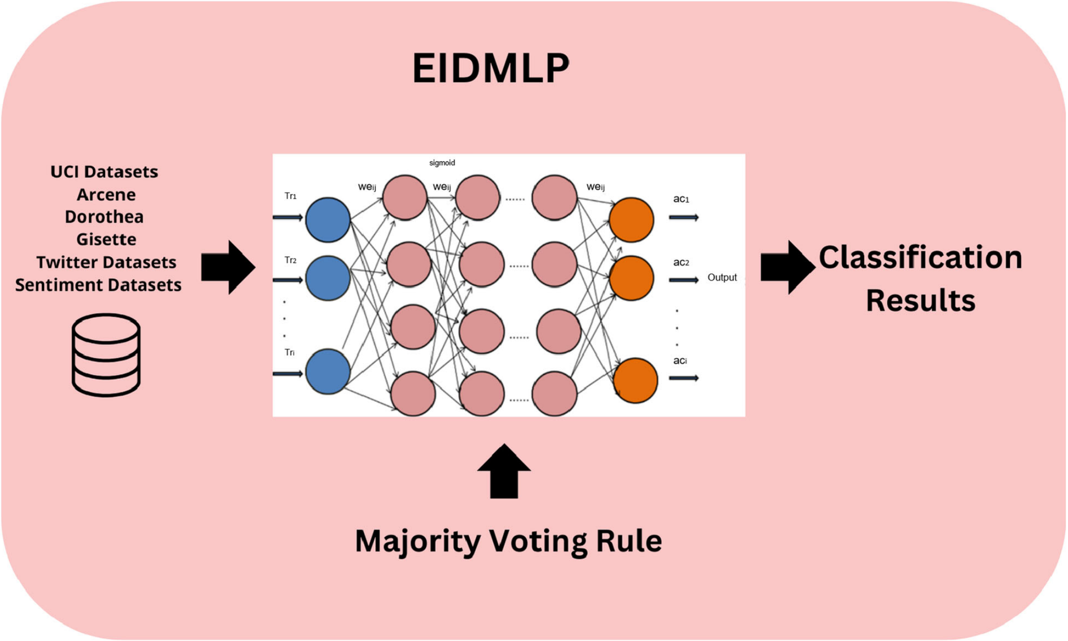
$$mv_{Mk} = Tc_k = \sum\nolimits_{l=1}^{M} F_k(cl_l) \tag{6}$$

The collection of classes set is given by Equation (7),

$$S(tr) = \underset{k \in \{1, \ldots Q\} =}{\operatorname{argmax}} \; Tc_k \tag{7}$$

The inclusive steps of proposed EIDMLP classifier [19] are presented in Algorithm 2 and Figure 2. For classification problems, the standard evaluation metrics used are Precision, Recall, F-measure, and Accuracy. Apart from these metrics, time factor is also vital for BDA classification problem; thus, execution time is taken for one of the performance metrics. Conventional approaches are time-consuming and require reconstructing the learning model from scratch every time. The conventional training model assumes that all instances are fully available before training, which is in contrast to IL. Algorithms that facilitate this type of learning are incremental classifiers. These classifiers retain previous knowledge and update future learning as well. This is a

Figure 2
EIDMLP classifier



dynamic technique to process data that arrives gradually over time. They maintain and update the learning parameters relevant to past values (history). The proposed EIDMLP is a robust DL model that consists of multiple layers of neurons, capable of capturing complex patterns in the data.

Algorithm 2

```
INPUT: Training samples $Tr_i$, count all nodes in topological sequence, $S = \{A_1, A_2, A_M\}$
OUTPUT: Classified Results
ENSURE:
    1. Number of ensembles M=3, $CL$ be a set of $Q$ classes
        2. For each (M=1,2, 3) ($S = \{A_1, A_2, A_M\}$)
        3. Compute results of EIDMLP
        4. For each ($Tr_i$) do
        Compute activation of input nodes $ac_i \leftarrow Tr_i$
        5. Apply to the hidden and output nodes by equations,
                $net_i \leftarrow we_{i0} + we_{ij}ac_{ij}$
                $ac_i = f_l net_i * HB$
        6. For every output nodes i do
        7. Compute $mv_{Mk}\&S(Tr_i)$
        8. Gather $ac_i$ in output vector ‘Y’
        9. Return ensemble results via majority voting
END
```

4. Results and Discussion

The datasets (UCI & Kaggle) used are given in Figures 3 and 4. Arcene comprises of 900 instances with 10,000 attributes in which the numerous feature characteristics have been categorized into either cancerous or non-cancerous. The notion of classification is to identify the cancerous patient from non-cancerous. Dorothea is the drug discovery dataset, where the classification task is identifying the given molecular structure combination that makes the drug active or not. This decision is vital for discovery of new drugs. The Madelon is an artificial dataset that is known for classification of instances into positive and negative labels, where the redundant features are added purposely. Gisette dataset is to categorize the handwritten digit. The process is to categorize the digit into four or nine.

The proposed methodology is also experimented with streaming nature of three representative Twitter datasets, such as Sentiment140 dataset, Apple Twitter dataset, and US Airline Twitter dataset. The dataset contains the processed tweets that are categorized under different class labels. The notion of analysis is to classify the same. Analysis of human sentiment is the need of the hour, where finding negative emotions is crucial to help and support them in a more psychological way and quick action can be taken when analysis is done with the lesser time with more accuracy of classification. The EIDMLP algorithm is compared with other classifiers such as Naïve Bayes(NB) [20], SVM [21]) Hoeffiding Tree [22], and Fuzzy Minimal Consistent Class Subset Coverage KNN (FMCCSC-KNN) [22].

This classifier is designed to provide high accuracy and strong generalization capabilities, making it well-suited for handling diverse and complex datasets. These metrics are assessed by means of the confusion matrix (Figure 5).

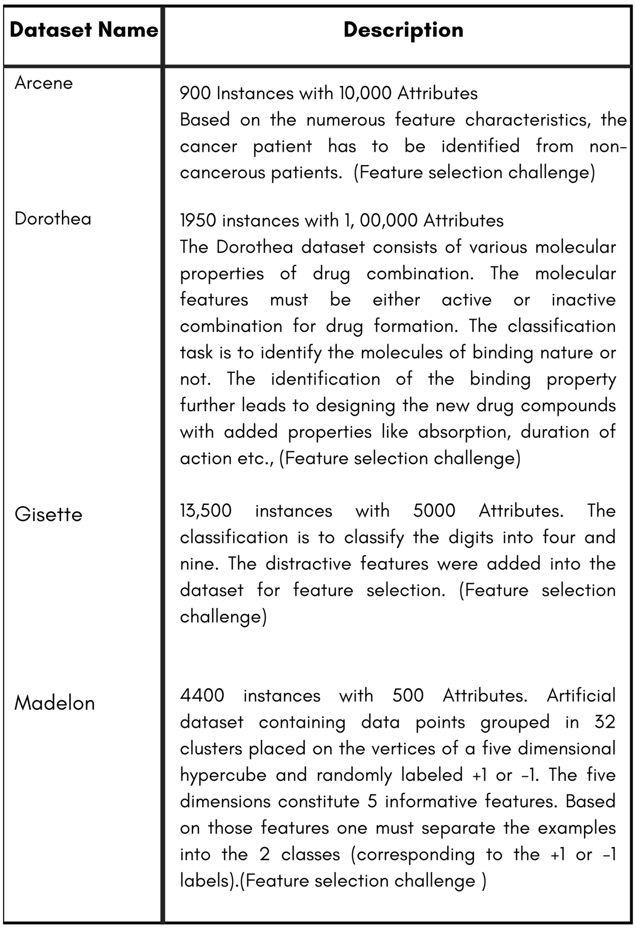
4.1. Analysis: Dorothea

Table 1 displays the performance assessment of varied classifiers in terms of metrics assessed for Dorothea dataset.

4.2. Analysis: Arcene

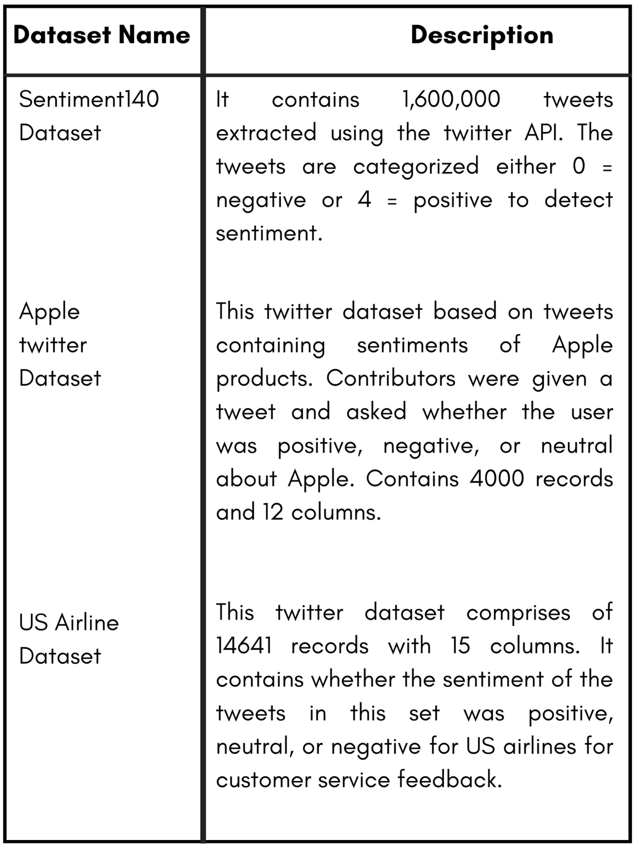
Table 2 displays the performance assessment of varied classifiers in terms of metrics assessed for Arcene dataset.

Figure 3
UCI datasets



Dataset Name	Description
Arcene	900 Instances with 10,000 Attributes Based on the numerous feature characteristics, the cancer patient has to be identified from non-cancerous patients. (Feature selection challenge)
Dorothea	1950 instances with 1, 00,000 Attributes The Dorothea dataset consists of various molecular properties of drug combination. The molecular features must be either active or inactive combination for drug formation. The classification task is to identify the molecules of binding nature or not. The identification of the binding property further leads to designing the new drug compounds with added properties like absorption, duration of action etc., (Feature selection challenge)
Gisette	13,500 instances with 5000 Attributes. The classification is to classify the digits into four and nine. The distractive features were added into the dataset for feature selection. (Feature selection challenge)
Madelon	4400 instances with 500 Attributes. Artificial dataset containing data points grouped in 32 clusters placed on the vertices of a five dimensional hypercube and randomly labeled +1 or -1. The five dimensions constitute 5 informative features. Based on those features one must separate the examples into the 2 classes (corresponding to the +1 or -1 labels).(Feature selection challenge)

Figure 4
Kaggle datasets



Dataset Name	Description
Sentiment140 Dataset	It contains 1,600,000 tweets extracted using the twitter API. The tweets are categorized either 0 = negative or 4 = positive to detect sentiment.
Apple twitter Dataset	This twitter dataset based on tweets containing sentiments of Apple products. Contributors were given a tweet and asked whether the user was positive, negative, or neutral about Apple. Contains 4000 records and 12 columns.
US Airline Dataset	This twitter dataset comprises of 14641 records with 15 columns. It contains whether the sentiment of the tweets in this set was positive, neutral, or negative for US airlines for customer service feedback.

4.3. Analysis: Gisette

Table 3 displays the performance assessment of varied classifiers in terms of metrics assessed for Gisette dataset.

4.4. Analysis: Sentiment 140

Table 4 displays the performance assessment of varied classifiers in terms of metrics assessed for Sentiment 140 dataset.

4.5. Analysis: Apple Sentiment dataset

Table 5 displays the performance assessment of varied classifiers in terms of metrics assessed for Apple Sentiment dataset.

4.6. Analysis: US Airline Twitter dataset

Table 6 displays the performance assessment of varied classifiers in terms of metrics assessed for US Airline Twitter dataset.

The consolidated classification results are tabulated in Table 7 and Figure 6.

The proposed model has achieved accuracy of 98.6%, 98.4%, 99.48%, 98.9%, and the processing time (refers specifically to both training and inference time) for classification has achieved 0.056 s, 0.053 s, 0.044 s, 0.0034 s, 0.0024, 0.0053 s for Dorothea, Arcene, Gisette, Sentiment 140, Apple sentiment, and US Airline Twitter, respectively. The proposed approach considerably contributed to reduce the processing time; at the same time, incremental DL classification has contributed for enhanced accuracy percentage. From the results' observation, the proposed model achieves higher accuracy and less processing time, which is evident that this model is effective in handling the big datasets in a component way.

Figure 5
Confusion matrix

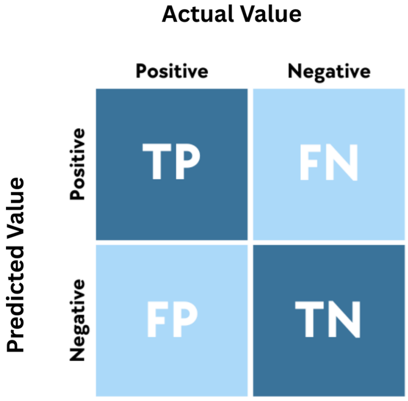


Table 1
Classification of Dorothea

	RESULTS (%)				PROCESSING
CLASSIFIERS	PRECISION	RECALL	F-MEASURE	ACCURACY	TIME (SECONDS)
NB	66.38	82.90	74.05	83.00	0.068
SVM	68.29	86.76	76.28	85.62	0.180
HT	75.51	93.45	83.12	91.00	0.040
FMCCSC-KNN	85.93	94.80	89.91	95.87	0.038
EIDMLP	94.18	98.66	93.37	98.60	0.056

Table 2
Classification of Arcene

	RESULTS (%)				PROCESSING
CLASSIFIERS	PRECISION	RECALL	F-MEASURE	ACCURACY	TIME (SECONDS)
NB	82.98	82.38	82.68	83.00	0.07
SVM	86.77	87.17	86.97	87.00	0.14
HT	90.80	90.99	90.89	91.00	0.048
FMCCSC-KNN	96.22	96.42	96.13	96.00	0.045
EIDMLP	99.12	99.10	98.90	99.00	0.053

Table 3
Classification of Gisette

	RESULTS (%)				PROCESSING
CLASSIFIERS	PRECISION	RECALL	F-MEASURE	ACCURACY	TIME (SECONDS)
NB	82.99	82.90	82.99	83.00	0.05
SVM	86.71	86.68	86.69	86.70	0.181
HT	90.80	90.90	90.89	90.90	0.036
FMCCSC-KNN	95.78	95.66	95.72	95.70	0.042
EIDMLP	98.60	98.59	98.59	98.60	0.044

Table 4
Classification of Sentiment 140

	RESULTS (%)				PROCESSING
CLASSIFIERS	PRECISION	RECALL	F-MEASURE	ACCURACY	TIME (SECONDS)
NB	81.9461	81.90	81.9231	81.90	0.1844
SVM	86.9427	86.90	86.9213	86.90	0.2272
HT	90.6234	90.60	90.6117	90.60	0.1781
FMCCSC-KNN	95.5089	95.50	95.5045	95.50	0.0041
EIDMLP	98.4008	98.40	98.4004	98.40	0.0034

Table 5
Classification of Apple Sentiment dataset

	RESULTS (%)				PROCESSING
CLASSIFIERS	PRECISION	RECALL	F-MEASURE	ACCURACY	TIME (SECONDS)
NB	87.4298	87.6543	87.5419	87.5862	0.0359
SVM	87.9002	87.9415	87.8751	87.9218	0.1006
HT	88.3711	88.6321	88.5014	88.5310	0.0037
FMCCSC-KNN	99.4355	99.4619	99.4533	99.4681	0.0063
EIDMLP	99.4556	99.4712	99.4723	99.4824	0.0024

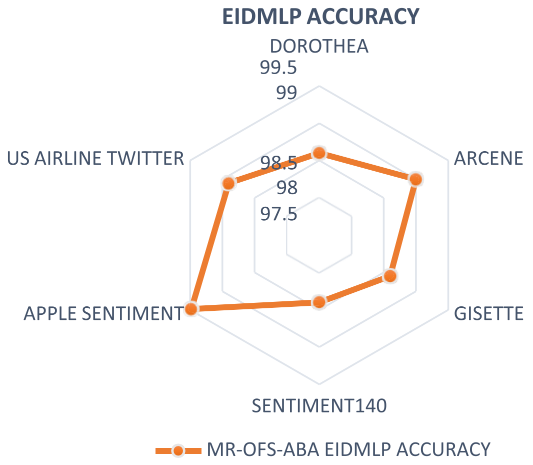
Table 6
Classification of US Airline Twitter dataset

	RESULTS (%)				
CLASSIFIERS	PRECISION	RECALL	F-MEASURE	ACCURACY	PROCESSING TIME (SECONDS)
NB	78.0450	83.1741	80.5280	83.0000	0.0107
SVM	80.5313	85.7494	83.0585	85.3000	0.0869
HT	87.2397	90.6067	88.8913	91.0000	0.0440
FMCCSC-KNN	93.6010	95.5047	94.6232	95.8000	0.0058
EIDMLP	98.1469	99.0038	98.5735	98.9000	0.0053

Table 7
Classification results

S.NO	DATASET NAME	ACCURACY %	PROCESSING TIME (SECONDS)
1	DOROTHEA	98.6	0.056
2	ARCENE	99	0.053
3	GISETTE	98.6	0.044
4	SENTIMENT140	98.4	0.0034
5	APPLE SENTIMENT	99.48	0.0024
6	US AIRLINE TWITTER	98.9	0.0053

Figure 6
Classification results



5. Conclusion

The study's result underlines the importance of big data classification in particular and underscores how well the EIDMLP performs in huge dataset applications. Using high-dimensional datasets from the UCI repository and Twitter, the study illustrates the effectiveness of the suggested method. The outcomes show better performance in terms of faster processing times and more accurate classification. The overall goal of the research is to minimize time complexity while producing accurate categorization findings that yield valuable insights. By enabling well-informed decisions based on deduced insights, these insights and decisions produced from the research findings are anticipated to help society. Furthermore, the effectiveness of EIDMLP across diverse domains and data types was not explored, limiting the generalizability of the findings. The model's performance may also be sensitive to the choice of hyperparameter, which were optimized for the specific datasets used, requiring extensive parameter tuning in different contexts. One limitation is that the enhanced model may demand substantial computational resources, which can restrict its use in environments with limited resources. These limitations highlight areas for future research to enhance the robustness and applicability of EIDMLP in various big data contexts.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in UCI at https://archive.ics.uci.edu/dataset/169/dorothea, https://archive.ics.uci.edu/dataset/167/arcene, https://archive.ics.uci.edu/dataset/170/gisette; in Kaggle at https://www.kaggle.com/datasets/anishdabhane/apple-tweets-sentiment-dataset, https://www.kaggle.com/datasets/kazanova/sentiment140, https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment.

Author Contribution Statement

Renuka Devi D.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Swetha Margaret T. A.:** Conceptualization, Methodology, Software, Validation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Xia, S., Song, J., Ameen, N., Vrontis, D., Yan, J., & Chen, F. (2023). What changes and opportunities does big data analytics capability bring to strategic alliance research? A systematic literature review. *International Journal of Management Reviews*, *26*(1), 34–53. https://doi.org/10.1111/ijmr.12350

[2] Gupta, S. K., Hrybiuk, O., Cherukupalli, N. L. S., & Shukla, A. K. (2023). Big data analytics tools, challenges and its applications. In A. Khang, S. K. Gupta, S. Rani, & D. A. Karras (Eds.), *Smart cities: IoT technologies, big data solutions, cloud platforms, and cybersecurity techniques* (pp. 307–320). CRC Press.

[3] Theodorakopoulos, L., Theodoropoulou, A., & Stamatiou, Y. (2024). A state-of-the-art review in big data management engineering: Real-life case studies, challenges, and future research directions. *Eng*, *5*(3), 1266–1297. https://doi.org/10.3390/eng5030068

[4] Igamberdieva, Z. (2024). Data analytics and its methods of development. *Web of Discoveries: Journal of Analysis and Inventions*, *2*(2), 55–58.

[5] Ikegwu, A. C., Nweke, H. F., & Anikwe, C. V. (2024). Recent trends in computational intelligence for educational big data analysis. *Iran Journal of Computer Science*, *7*(1), 103–129. https://doi.org/10.1007/s42044-023-00158-5

[6] Huang, K. (2024). *Estudio bibliográfico sobre la aplicación en inteligencia artificial y análisis de big data a gestión de calidad de proyectos de ingeniería civil [Bibliographic study on the application of artificial intelligence and big data analysis to quality management of civil engineering projects]*. Master's Thesis, Universitat Politècnica de València.

[7] Dhawas, P., Ramteke, M. A., Thakur, A., Polshetwar, P. V., Salunkhe, R. V., & Bhagat, D. (2024). Big data analysis techniques: Data preprocessing techniques, data mining techniques, machine learning algorithm, visualization. In D. Darwish (Ed.), *Big data analytics techniques for market intelligence* (pp. 183–208). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-0413-6.ch007

[8] Hernández, G., Zamora, E., Sossa, H., Téllez, G., & Furlán, F. (2020). Hybrid neural networks for big data classification. *Neurocomputing*, *390*, 327–340. https://doi.org/10.1016/j.neucom.2019.08.095

[9] Thayyib, P. V., Mamilla, R., Khan, M., Fatima, H., Asim, M., Anwar, I., . . . , & Khan, M. A. (2023). State-of-the-art of artificial intelligence and big data analytics reviews in five different domains: A bibliometric summary. *Sustainability*, *15*(5), 4026. https://doi.org/10.3390/su15054026

[10] Ali, I. M. S., & Hariprasad, D. (2023). Hyper-heuristic salp swarm optimization of multi-kernel support vector machines for big data classification. *International Journal of Information Technology*, *15*(2), 651–663. https://doi.org/10.1007/s41870-022-01141-2

[11] Museba, T. (2024). Incremental machine learning-based approach for credit scoring in the age of big data. In T. Moloi, & B. George (Eds.), *Towards digitally transforming accounting and business processes: Proceedings of the international conference of accounting and business iCAB* (pp. 547–565). Springer Cham. https://doi.org/10.1007/978-3-031-46177-4_29

[12] Haritha, K., Shailesh, S., Judy, M. V., Ravichandran, K. S., Krishankumar, R., & Gandomi, A. H. (2023). A novel neural network model with distributed evolutionary approach for big data classification. *Scientific Reports*, *13*(1), 11052. https://doi.org/10.1038/s41598-023-37540-z

[13] Naskath, J., Sivakamasundari, G., & Begum, A. A. S. (2023). A study on different deep learning algorithms used in deep neural nets: MLP SOM and DBN. *Wireless Personal Communications*, *128*(4), 2913–2936. https://doi.org/10.1007/s11277-022-10079-4

[14] Selmy, H. A., Mohamed, H. K., & Medhat, W. (2024). Big data analytics deep learning techniques and applications: A survey. *Information Systems*, *120*, 102318. https://doi.org/10.1016/j.is.2023.102318

[15] Solomon, D. D., Khan, S., Garg, S., Gupta, G., Almjally, A., Alabduallah, B. I., . . . , & Abdallah, A. M. A. (2023). Hybrid majority voting: Prediction and classification model for obesity. *Diagnostics*, *13*(15), 2610. https://doi.org/10.3390/diagnostics13152610

[16] Liu, W., Yang, C., Nan, J., Gao, M., & Niu, H. (2023). Antenna notch structure optimization using deep neural networks. *Progress in Electromagnetics Research Letters*, *114*, 37–44. https://doi.org/10.2528/PIERL23071902

[17] Fong, S., Wong, R., & Vasilakos, A. V. (2016). Accelerated PSO swarm search feature selection for data stream mining big data. *IEEE Transactions on Services Computing*, *9*(1), 33–45. https://doi.org/10.1109/TSC.2015.2439695

[18] Renuka Devi, D., & Sasikala, S. (2019). Online feature selection (OFS) with accelerated bat algorithm (ABA) and ensemble incremental deep multiple layer perceptron (EIDMLP) for big data streams. *Journal of Big Data*, *6*, 1–20. https://doi.org/10.1186/s40537-019-0267-3

[19] Verma, G., & Sahu, T. P. (2024). A correlation-based feature weighting filter for multi-label Naive Bayes. *International Journal of Information Technology*, *16*(1), 611–619. https://doi.org/10.1007/s41870-023-01555-6

[20] Prabhavathy, T., Elumalai, V. K., & Balaji, E. (2024). Hand gesture classification framework leveraging the entropy features from sEMG signals and VMD augmented multi-class SVM. *Expert Systems with Applications*, *238*, 121972. https://doi.org/10.1016/j.eswa.2023.121972

[21] Bartz-Beielstein, T. (2024). Supervised learning: Classification and regression. In E. Bartz, & T. Bartz-Beielstein (Eds.), *Online machine learning: A practical guide with examples in Python* (pp. 13–22). Springer Singapore. https://doi.org/10.1007/978-981-99-7007-0_2

[22] Renuka Devi, D., & Sasikala, S. (2019). Accelerated simulated annealing and mutation operator feature selection method for big data. *International Journal of Recent Technology and Engineering*, *8*(2), 910–916. https://doi.org/10.35940/ijrte.B1712.078219

How to Cite: D., R. D., & Margaret T. A., S. (2025). Enhancing Big Data Classification Accuracy Through Deep Learning Techniques. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA52023432