IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 February 2024, accepted 22 February 2024, date of publication 1 March 2024, date of current version 8 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3372082

RESEARCH ARTICLE

Comparative Analysis of Predictive Algorithms for Performance Measurement

SWATI GUPTA, BAL KISHAN, AND PREETI GULIA
Department of Computer Science and Applications, Maharshi Dayanand University (MDU), Rohtak, Haryana 124001, India

Corresponding authors: Swati Gupta (Swati.rs20.dcsa@mdurohtak.ac.in), Bal Kishan (balkishan@mdurohtak.ac.in), and Preeti Gulia (preeti@mdurohtak.ac.in)

ABSTRACT Predictive algorithms, also known as mathematical models, utilize historical data to accurately predict future outcomes. These algorithms identify patterns and relationships within the data, resulting in precise predictions. The growing importance of predictive algorithms in various domains, such as finance, healthcare, marketing, weather forecasting, E-commerce, etc., has led to an increasing need for robust and accurate models. Machine learning (ML) and deep learning (DL) algorithms, including supervised, unsupervised, & reinforcement learning, play a crucial role in prediction. Supervised algorithms include classification and regression, while unsupervised algorithms primarily focus on clustering. In this study, a detailed comparative analysis of eight classification algorithms, six regression algorithms, and five clustering algorithms is performed using diverse datasets and performance metrics. ROBERTA, ResNet, Random Forest Regression, and K-means clustering algorithms outperformed traditional algorithms in textual classification, image classification, regression, and clustering. This study enables data scientists and practitioners to make informed decisions when selecting appropriate models for their specific applications.

INDEX TERMS Predictive algorithms, supervised algorithm, unsupervised algorithm, classification, regression, clustering, performance metrics.

I. INTRODUCTION

Predictive algorithms are advanced mathematical models that leverage historical data to predict future outcomes. By analyzing patterns, correlations, and trends within the data, these algorithms provide accurate forecasts [1]. Predictive algorithms are widely employed to extract meaningful insights, forecast trends, and make data-driven decisions. As the accuracy and reliability of these algorithms significantly impact real-world applications, researchers continuously seek to improve and assess their performance across various scenarios. This study aims to address the lack of comprehensive evaluations by conducting a comparative analysis of predictive algorithms across different data sets using a range of performance metrics [2]. ML and DL algorithms, like supervised, unsupervised, and reinforcement, are instrumental in facilitating prediction tasks across various domains. Supervised algorithms are very important because they use labeled training data to discover associations between inputs and desired outcomes. This enables them to classify new instances into predefined categories (classification) or predict numerical values (regression) [3]. Supervised algorithms are widely used in applications like email spam detection, image recognition, and stock price prediction. Classification algorithms are employed to categorize and organize data into classes and categories. They can be applied to both textual and graphical data, and there are three approaches to classification: binary, multiclass, and multi-label classification [4]. The current study considers classification algorithms such as Support Vector Machine (SVC), Naïve Bayes (NBC), Long Short-Term Memory (LSTM), ADA boost (ABC), ROBERTA, ANN, CNN, and RNN. The efficacy of these algorithms is evaluated using performance metrics such as Support, Recall, Accuracy, F1-score, and Precision. CNN algorithms like RESNET, DENSENET, and INCEPTION are used to categorize visual information with the help of these performance metrics. Regression is an analytical method used to examine the association between many sets of independent variables or features and a single set of outcomes. Predictive modeling is a method that employs this approach in

The associate editor coordinating the review of this manuscript and approving it for publication was Ikramullah Lali.

machine learning. Performance indicators including MSE, MAR, R Squared, RMSE, and MAPE are used to evaluate six different regression algorithms: Logistic, Linear, Polynomials Regression, Random Forest, decision tree, and LASSO.

On the other hand, unsupervised algorithms operate without labeled data and focus on uncovering hidden patterns or structures within the data. Clustering algorithms are a prominent example of unsupervised learning, where datasets are grouped together based on their similarity or proximity. This helps in identifying natural groupings or clusters in the data without prior knowledge of the categories. In cluster analysis, items are grouped into clusters where members of the same cluster have more similarities than differences with members of other clusters. It is widely utilized in various disciplines. Five clustering algorithms, namely Density-Based Clustering, K Mean Cluster, Birch Clustering, Agglomerative Clustering, and Spectral Clustering, are analyzed with performance metrics such as Adjusted rand index (ARI), Silhouette coefficient (SC), Calinski-Harabasz index (CHI), Dunn Index (DI), and Davies – Bouldin index (DBI). Five textual datasets of Health care, stock market prediction, anomaly detection, criminal detection, and sentiment analysis are analyzed with classification, regression, and clustering algorithms. CNN models like RESNET, DENSENET, and INCEPTION are compared using three graphical datasets of CT scans, Brain MRI Scans, and Sentiment analysis.

The goal of comparing predictive algorithms is to find the best and most efficient solutions for solving data-related problems. Researchers and professionals evaluate different algorithms carefully to find the one that consistently offers high accuracy, flexibility, and performance. This effort helps make informed decisions and drives progress in various fields like machine learning, finance, and healthcare and improved efficiency in manufacturing.

This study demonstrates a comprehensive comparative evaluation of a variety of algorithms by assessing their performance across diverse datasets, that provides valuable insights into how these algorithms function under different conditions and with varying types of data. It employs a wide range of performance metrics, ensuring a comprehensive assessment of the algorithms. Ultimately, the intention is to utilize the knowledge gained from these comparative evaluations to develop a new hybrid algorithm, which will demonstrate superior performance compared to conventional algorithms, thereby making a substantial contribution to the advancement of algorithmic solutions in the field. This study is organized into five sections. Section II reviews the relevant literature. Section III describes the methodology used in this study. Section IV presents the comparative study of the performance of different algorithms using various metrics. Section V concludes this study.

II. LITERATURE REVIEW

Numerous studies comparing various ML and DL algorithms are available in the literature. These studies focus on both supervised and unsupervised algorithms. Specifically, they discuss classification and regression algorithms in supervised learning, as well as clustering algorithms in unsupervised learning. A review of classification, regression, and clustering is presented in sections A, B, and C respectively.

A. CLASSIFICATION ALGORITHMS

Hashem et al., compared five different classifications to predict attack anomalies based on performance measures. Test accuracy results are interpreted as 99.4%. Although the accuracy of these methods is equal, proxy scores show that RFC ultimately performs better [5]. Priyadarsini and Titus compared NBC, SVC, DTC, and RFC's four classification algorithms to predict diabetes based on performance measures. SVC showed the highest precision & accuracy, while NBC & SVC performed similarly [6]. Saranya et al. compared the performance of 12 different classifications with accuracy. RFC is the top-performing algorithm, with an accuracy of 99.81% [7]. Susan Cheragi et al., the ability of radiation signatures derived from computed tomography to foretell the development of chronic kidney disease in individuals receiving radiotherapy for abdominal cancer was assessed. Fifty people who were treated with radiation treatment for a full year were included in the research. There are six different classifiers utilized for CKD prediction, with RFC having the best accuracy at 94%. Most patients (58%) suffer from CKD [8]. Dey et al., a machine learning approach to accurately diagnose stroke using imbalanced data is presented. They used the ROS tapping technique to balance the data and analyzed 11 classifiers. SVC and RFC performed best [9]. Liu et al. analyzed student learning behavior using ML in an interactive learning data environment. They compare performance metrics such as F1 scores and accuracy to predict student learning outcomes. NN gave the best results among these algorithms [10].

B. REGRESSION ALGORITHMS

Cemal Hanilçi et al. compared the performance of different regression methods in remote tracking of Parkinson's disease progression. Results indicate that LS-SVM exhibits the best performance among 3 alternative methods [11]. Tehseen Zia et al., a comparison was made among 4 different regression algorithms. The evaluation was based on performance metrics including MAE, RMSE, and RRSE. Results of the research revealed that the approach utilizing the "IBK with No Distance Weighting" algorithm showed effective utilization for modeling reusability evaluation based on metrics [12]. Rui Abreu et al., conducted research comparing seven regression algorithms, namely SVM, RF, AdaBoost, KNN, Baseline, OLS, and CART, based on the RMSE metric. The results indicated that SVM performed the best for Students' Academic Performance datasets [13]. Christiaan M. van der Walt et al., performed predictions on a wind speed time series using three machine learning. SVR performed marginally better than regular least squares and Bayesian ridge regression, demonstrating

TABLE 1. Description of datasets.

Dataset Name	Dataset	Type of Dataset	Size of Datasets
D1	Stock Price Prediction [22]	Textual	100 scripts
D2	fraud detection [23]	Textual	492 frauds out of 284,807 transactions
D3	Sentiment Analysis (IMDB review) [24]	Textual	50,000 IMDB movie reviews for sentiment analysis
D4	Anomaly Detection (Elliptic Dataset) [25]	Textual	There are 203,769 nodes 234,355 edges and 166 features associated with each node.
D5	Health Care [26]	Textual	9 features related to heart diseases in 3,000 persons
D6	Brain: MRI scan [27]	Graphical	2501 MRI Images.
D7	Brain: CT scan [27]	Graphical	2503 brain CT scans
D8	Flickr 30k Dataset [28]	Graphical	Flickr30k Entities, which augments the 158k captions from Flickr30k with 244k coreference chains with 158915 images.

a considerable improvement over the persistence prediction [14]. Aneeta S Antony et al. conducted a comparison of different regression algorithms. The comparison was based on performance metrics MSE and R2, using the profile of the student as input. The results indicated that Linear Regression performed the best on the student dataset, exhibiting a low MSE and a high R2 score. Random Forest closely followed, demonstrating competitive performance [15].

C. CLUSTERING ALGORITHMS

J. Yang et al. conducted a comparison of three algorithms for tag clustering including MMSK-means, Latent Semantic Analysis (LSA-based algorithm), and LMMSK. The comparison utilized a real tag-resource dataset obtained from the Delicious Social Bookmarking System spanning from 2004 to 2009. Using MATLAB, it was found that LMMSK produced the most effective and accurate clustering results among the three algorithms [16]. X. Gong et al., conducted a comparative research study on Particle Swarm Optimization. The research focused on both synthetic and real medical datasets, specifically exploring the application of swarm intelligence clustering for analyzing medical data. The study revealed that swarm intelligence clustering algorithms played a significant role in the analysis of medical data, showcasing their effectiveness in this domain [17]. A. E. S. Ezugwu et al., Automatic data clustering was evaluated. 12 standard ground truth clustering datasets were used in this research, all obtained from the UCI ML Repository. Extensive experimental evidence showed that the FAPSO method beat other hybrids in terms of solution quality and convergence speed, including FAABC, FAIWO, and FATLBO [18]. A. A. Bushra et al., conducted an analysis of the density-based clustering algorithm using datasets such as Iris, travel review, wine, Ecoli, Yeast, Glass, TAE, and WDBC. The study examined the performance of the original DBSCAN algorithm and compared it to improved algorithms. The results obtained from the FCPS collection showed that while the improved algorithms provided better quality results in specific cases, the original DBSCAN algorithm generally demonstrated proficiency in handling clustering operations [19]. F. Malik et al. conducted an evaluation of distance metrics for datasets with uneven clusters. The study focused on unsupervised learning and k-means clustering, specifically using the Hybrid BH algorithm to handle the optimal value of k efficiently. After examining many methods for doing cluster analysis, the researchers settled on the Hybrid Black Hole algorithm [20]. M. Raeisi et al., implemented unsupervised learning and k-means clustering. Focusing on clustering datasets with unequal-size clusters in wireless and autonomous network applications, the research assessed the efficacy of a suggested metric in applications with non-linear distance requirements. The suggested measure was shown to be effective in such simulations [21]. The study reviewed in this section provided insight into the work of various researchers along with the algorithms and their results. These studies are performed on different datasets. To compare various models further on the same datasets, the methodology is designed so that a better selection of algorithms can be made.

III. RESEARCH METHODOLOGY

The methodology employed in this study encompasses several key steps including the selection of the data set, data preprocessing, applying algorithms, and performance metrics used, and result description shown in Figure 1.

A. SELECTION OF DATASET

This study focuses on two types of datasets. Textual datasets are used in the case of classification, Regression, and clustering algorithms. The graphical dataset is used for CNN-based models during accuracy prediction. A total of eight datasets were selected, and their details are given in Table 1.

B. DATA PREPROCESSING

The concept of "data preprocessing" is used to describe any action taken on unprocessed data before it is used. It's the first and most important step in every data mining process. Preparing data for use in data science processes like data mining and machine learning requires data preprocessing. During textual data preprocessing less significant attributes and records are eliminated. The record with less significance creates confusion in decision-making. Thus, such records are eliminated from the textual dataset. In the case of the graphical dataset, images are compressed and resized then

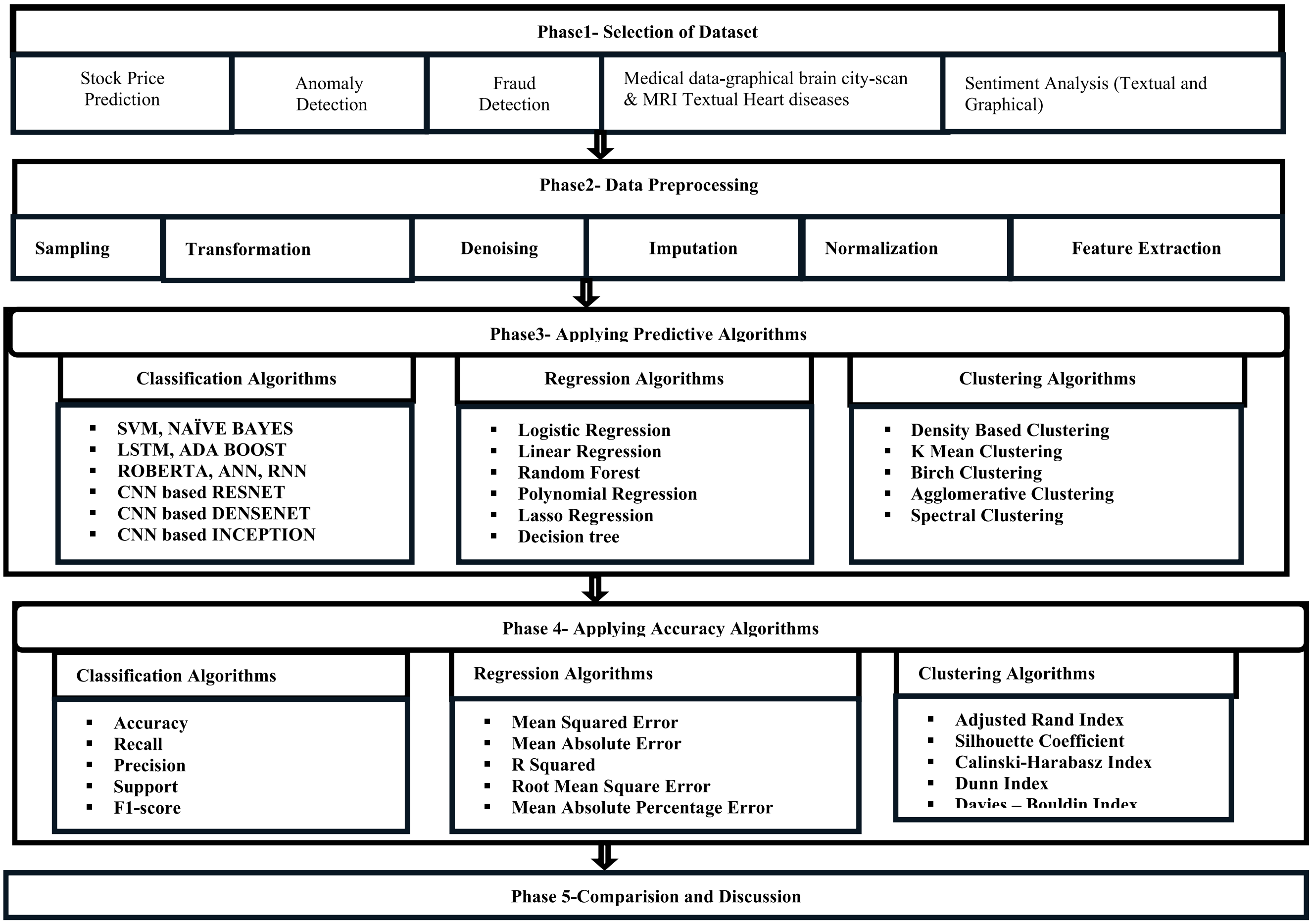

FIGURE 1. Research methodology.

a filtering mechanism is applied to those images to improve their quality so that accuracy improves during image detection. The following data preparation tools and techniques are used in this study:

- From a big pool of information, a representative sample is produced using sampling techniques.
- Raw data is transformed into a unified form using transformation techniques.
- Noise in data is eliminated by denoising.
- For missing values, imputation synthesizes statistically meaningful data from raw data.
- The data is easier to find and use after being normalized.

After data preprocessing, parameters of classification, regression, and clustering are compared for different techniques considering various datasets. Deep learning mechanisms such as ANN, CNN, and RNN are implemented to evaluate their accuracy in graphical data.

C. APPLYING PREDICTIVE ALGORITHMS

The present study considers different classification, regression, and clustering algorithms as predictive algorithms. Classification algorithms are SVM, Naive, LSTM, ADA Boost, Roberta for text, and CNN for graphics. Regression algorithms are logistic, linear, random forest, polynomial regression, and decision trees. Clustering algorithms are density-based clustering, k-mean clustering, Birch clustering, spectral clustering, and agglomerative clustering.

D. PERFORMANCE EVALUATION

Performances parameters considered during classification are Support, Accuracy, Recall, Precision, and F1-score. During regression, performance metrics are MSE, MAR, R Squared, RMSE, and MAPR. On the other hand, adjusted rand index, Silhouette coefficient, CHI, DI, and DBI are performance parameters for clustering.

IV. COMPARISON AND DISCUSSION

Considering the algorithms described in section C, five datasets are processed for classification, regression, and clustering. After comparing classification algorithms based on textual data, a comparative analysis of accuracy is performed for three graphical datasets. Finally, regression and clustering parameters are compared using five textual datasets. The Python Scikit-Learn library is used for managing textual datasets (D1 to D5) by conducting preprocessing, feature extraction, and model training. Additionally, when dealing with graphical datasets (D6 to D8) and coupled with Scikit-Network, the library is instrumental in creating, and

TABLE 2. Comparison of performance parameters for different classifiers.

Classification Algorithm VS Performance Metrics	Data Set	Support	Accuracy	Recall	Precision	F1-score
SVM	D1	323	91.92	86.11	84.82	80.46
	D2	379	91.12	84.15	81.58	80.86
	D3	355	91.21	85.23	87.69	85.92
	D4	365	91.02	82.02	83.85	88.96
	D5	329	91.22	83.05	82.24	82.00
NAÏVE BAYES	D1	370	92.06	86.40	84.59	86.64
	D2	320	92.35	84.40	89.31	82.19
	D3	398	92.55	84.08	86.90	87.71
	D4	315	92.00	83.25	87.36	84.38
	D5	305	92.99	86.64	85.90	88.84
LSTM	D1	394	94.84	89.55	89.20	89.60
	D2	348	94.27	89.59	89.68	89.87
	D3	368	94.18	89.53	89.50	89.72
	D4	357	94.14	89.10	89.07	89.62
	D5	341	94.13	89.19	89.55	89.39
ADA BOOST	D1	379	97.47	89.67	89.81	89.39
	D2	378	97.39	89.79	89.73	89.25
	D3	325	97.35	89.24	89.93	89.31
	D4	372	97.21	89.81	90.00	89.55
	D5	374	97.03	89.18	89.87	89.09
ROBERTA	D1	385	98.57	92.27	92.98	92.81
	D2	380	98.03	92.90	92.81	92.77
	D3	399	98.77	92.10	92.25	92.41
	D4	387	98.13	92.03	92.92	92.61
	D5	378	98.86	92.41	92.32	92.44
ANN	D1	432	95.56	92.00	89.00	91.00
	D2	456	94.61	91.00	90.00	90.00
	D3	412	95.63	89.00	88.00	90.00
	D4	431	95.12	92.00	89.00	91.00
	D5	456	95.54	90.00	88.00	90.00
RNN	D1	478	97.60	95.00	90.00	92.00
	D2	462	96.14	92.00	93.00	92.00
	D3	417	97.30	91.00	91.00	92.00
	D4	423	97.27	92.00	93.00	91.00
	D5	490	97.49	92.00	93.00	92.00

analyzing graphs, extracting features, and seamlessly integrating them with Scikit-Learn for tasks like classification and visualization.

A. COMPARISON OF CLASSIFICATION ALGORITHMS BASED ON PERFORMANCE PARAMETERS ON TEXTUAL DATASETS

Classification metrics, also known as confusion matrices, are often used to assess the efficacy of statistical models. Classification algorithms are used to determine a variety of performance metrics. Performance metrics for different classifiers are shown in Tables 2 & 3, where support, accuracy, recall precision, and f1-score are considered.

Figure 2 shows classification performance metrics for stock price datasets.

Fig. 3 shows classification performance metrics for fraud detection.

Fig. 4 shows classification performance metrics for sentiment analysis.

Fig. 5 shows classification performance metrics for anomaly detection.

Fig. 6 shows the classification performance metrics of performance metrics for healthcare detection.

Figure 7 shows the classification performance metrics of performance metrics for anomaly detection.

B. COMPARISON OF CLASSIFICATION ALGORITHMS BASED ON PERFORMANCE PARAMETERS ON GRAPHICAL DATASETS

A trained CNN model is tested for anomaly detection using datasets of MRI scans, CT scans, and sentiment analysis considering ResNet, DenseNet, and Inception. Throughout the training, many hyperparameters are adjusted, such as the batch size, epochs, training dataset, and testing dataset. In training operation, the ResNet, DenseNet, and Inception models are trained, and both the validation accuracy and the training accuracy are taken into consideration. Following the completion of the testing procedure, a confusion matrix is

TABLE 3. Average comparison of performance parameters for different classifiers.

Classification Algorithm VS Performance Metrics	Support	Accuracy	Recall	Precision	F1-score
SVM	350.20	73.26	84.11	84.04	83.64
NAÏVE BAYES	341.60	92.39	84.96	86.81	85.95
LSTM	361.60	94.31	89.39	89.40	89.64
ADA BOOST	365.60	97.29	89.54	89.87	89.32
ROBERTA	385.80	98.47	92.34	92.66	92.61
ANN	437.40	95.29	90.80	88.80	90.40
RNN	454.00	97.16	92.40	92.00	91.80

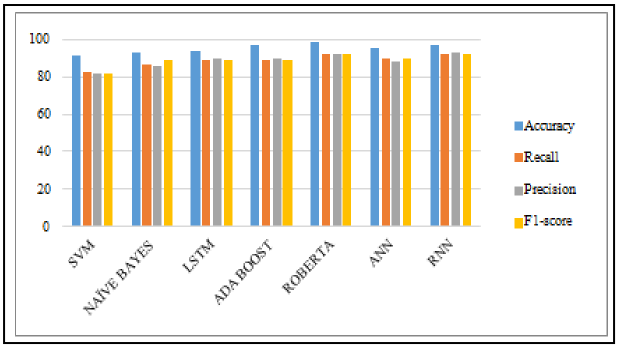

FIGURE 2. Comparison of classification parameters for stock price prediction.

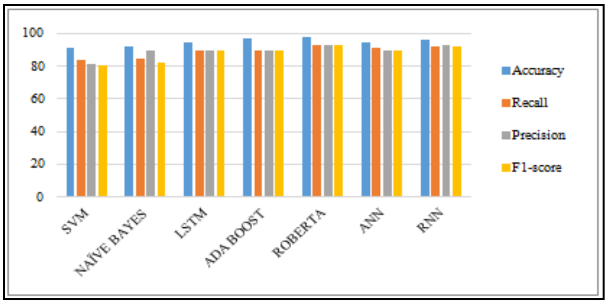

FIGURE 3. Comparison of classification parameters for Fraud detection.

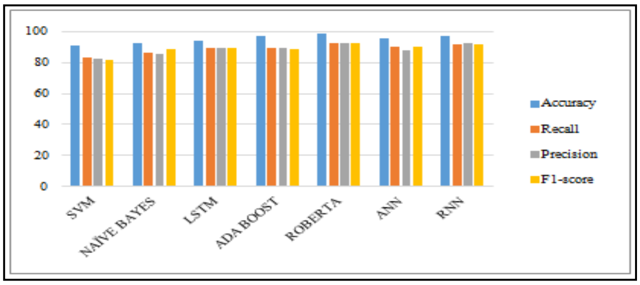

FIGURE 4. Comparison of classification parameters for sentiment detection.

produced. Table 4 displays the results of the calculations done, to determine the accuracy of each model and RESNET is best among them. CNN models ResNet, DenseNet, and Inception are trained and tested on datasets of MRI scan, CT scan, and sentiment analysis, accuracy is calculated.

Figure 8 shows accuracy in the case of RESNET, DenseNet, and Inception networks for MRI scans, CT scans, and sentiment analysis.

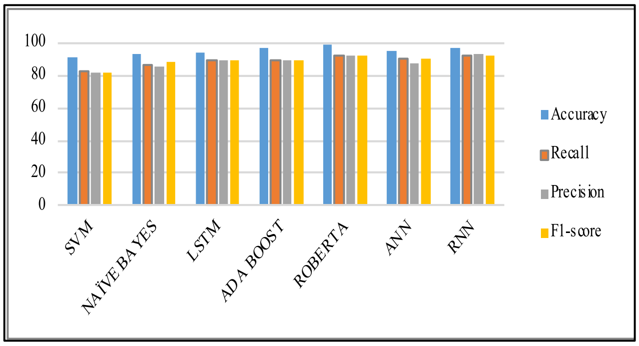

FIGURE 5. Comparison of classification parameters for anomaly detection.

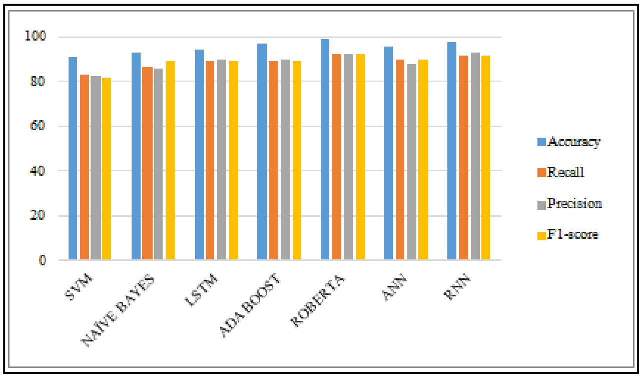

FIGURE 6. Comparison of classification parameters for healthcare detection.

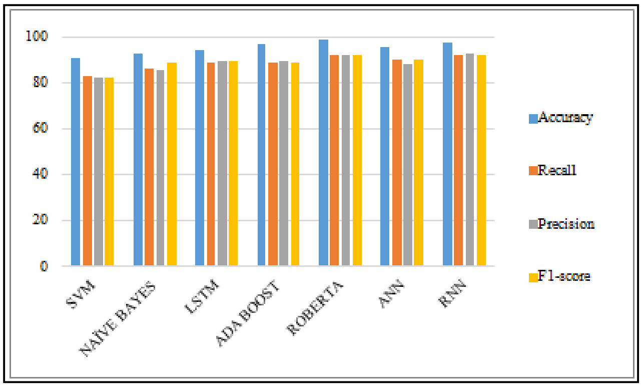

FIGURE 7. Comparison of classification parameters for five datasets altogether.

TABLE 4. Comparison accuracy of cnn models for different dataset.

CNN	RESNET	DENSENET	INCEPTION
MRI Scan Healthcare	96.54	94.53	93.54
CT scans	96.37	94.93	93.98
Sentiment analysis (flikr 30k)	97.43	95.23	93.25

C. COMPARISON OF REGRESSION ALGORITHMS BASED ON PERFORMANCE PARAMETERS

MSE, MAE, R Squared, RMSE, and MAPE are calculated for logistic, linear, random forest, polynomial, lasso, and decision tree-based regression. Performance metrics for different regressors are shown in Table 5 and an average of dataset performance is shown in Table 6.

TABLE 5. Regression algorithms performance comparison using various performance metrics.

Regression Algorithm VS Performance Metrics	Dataset	MSE	MAE	R Squared	RMSE	MAPE
Logistic Regression	D1	2.80	1.40	0.31	0.44	0.30
	D2	2.70	1.40	0.32	0.44	0.30
	D3	2.80	1.40	0.40	0.44	0.30
	D4	2.60	1.40	0.35	0.44	0.30
	D5	2.86	1.46	0.32	0.44	0.30
Linear Regression	D1	2.83	1.42	0.26	0.44	0.33
	D2	2.79	1.42	0.24	0.44	0.32
	D3	2.87	1.43	0.28	0.44	0.32
	D4	2.75	1.42	0.23	0.44	0.32
	D5	2.86	1.46	0.28	0.44	0.33
Random Forest Regression	D1	0.51	0.41	0.87	0.44	0.09
	D2	0.49	0.42	0.85	0.44	0.09
	D3	0.51	0.44	0.84	0.44	0.09
	D4	0.52	0.42	0.85	0.44	0.10
	D5	0.52	0.41	0.81	0.44	0.09
Polynomial Regression	D1	1.78	1.34	0.41	0.44	0.29
	D2	1.77	1.34	0.42	0.44	0.26
	D3	1.78	1.34	0.40	0.44	0.25
	D4	1.76	1.34	0.45	0.44	0.28
	D5	1.76	1.36	0.42	0.44	0.25
Lasso Regression	D1	1.68	1.33	0.44	0.44	0.22
	D2	1.67	1.33	0.44	0.44	0.22
	D3	1.68	1.33	0.44	0.44	0.22
	D4	1.66	1.33	0.45	0.44	0.22
	D5	1.56	1.34	0.46	0.44	0.22
Decision tree	D1	2.17	1.21	0.43	0.45	0.29
	D2	2.17	1.22	0.41	0.45	0.28
	D3	2.17	1.23	0.42	0.45	0.28
	D4	2.17	1.26	0.41	0.45	0.28
	D5	2.17	1.21	0.43	0.45	0.29

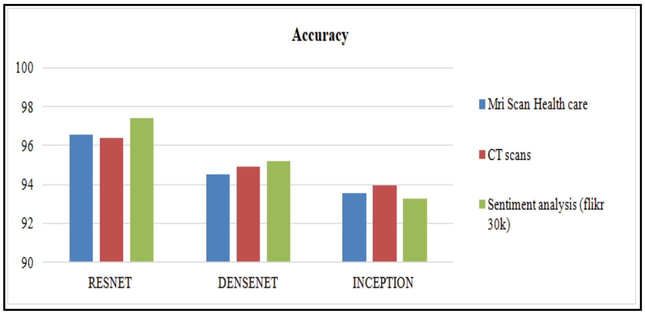

FIGURE 8. Comparison of CNN algorithms for different datasets.

TABLE 6. Average of Regression Algorithms Performance comparison using various performance metrics.

Regression Algorithm VS Performance Metrics	MSE	MAE	R Squared	RMSE	MAPE
Logistic Regression	2.75	1.41	0.34	0.44	0.30
Linear Regression	2.82	1.43	0.26	0.44	0.32
Random Forest Regression	0.51	0.42	0.84	0.44	0.09
Polynomial Regression	1.77	1.34	0.42	0.44	0.27
Lasso Regression	1.65	1.33	0.45	0.44	0.22
Decision tree	2.17	1.23	0.42	0.45	0.28

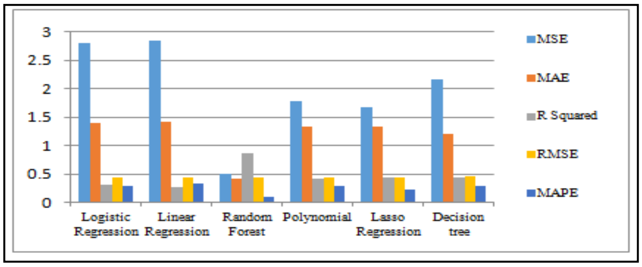

FIGURE 9. Comparative analysis of regression parameters for Stock price prediction.

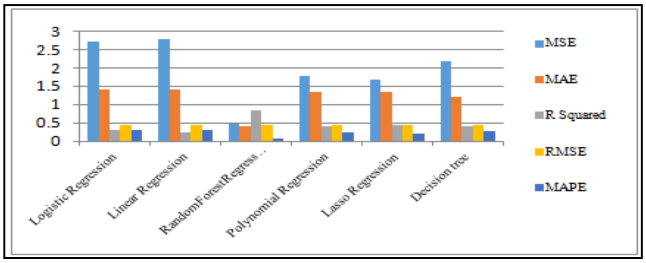

FIGURE 10. Comparative analysis of regression parameters for fraud detection.

This study shows that random Forest performed best in terms of performance metrics compared to other regressors. Figure 9 presents MSE, MAE, R-squared, RMSE, and MAPE in the case of stock price prediction.

Figure 10 presents MSE, MAE, R-Squared, RMSE, and MAPE in case of fraud detection.

Figure 11 presents MSE, MAE, R-squared, RMSE, and MAPE in the case of sentiment analysis.

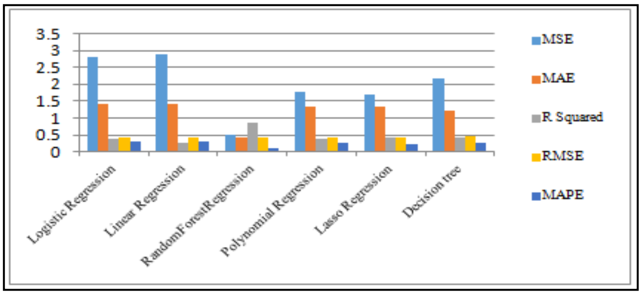

FIGURE 11. Comparative analysis of regression parameters for sentiment analysis.

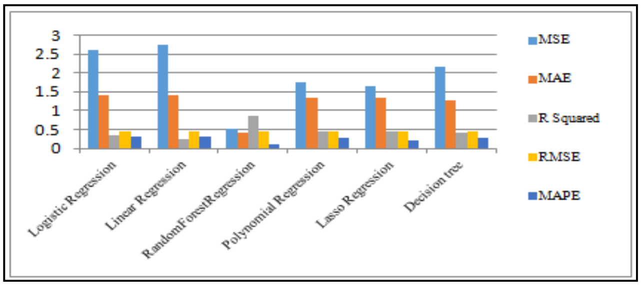

FIGURE 12. Comparative analysis of regression parameters for anomaly detection.

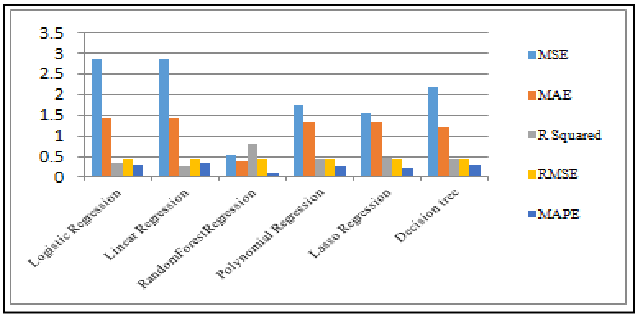

FIGURE 13. Comparative analysis of regression parameters for Healthcare prediction.

Figure 12 presents MSE, MAE, R-squared, RMSE, and MAPE in the case of anomaly detection for logistic regression, linear regression, random forest, polynomial, lasso regression, and decision tree.

Figure 13 presents MSE, MAE, R-squared, RMSE, and MAPE in the case of Health care prediction for logistic regression, linear regression, random forest, polynomial, lasso regression, and decision tree.

Figure 14 presents MSE, MAE, R-Squared, RMSE, and MAPE considering the average of 5 datasets of stock price prediction, fraud detection, sentiment analysis, anomaly detection, healthcare detection for logistic regression, linear regression, random forest, polynomial, lasso regression, decision tree.

D. COMPARISON OF CLUSTERING ALGORITHMS BASED ON PERFORMANCE PARAMETERS

ARI, SC, CHI, DI, and DBI are performance metrics for clustering algorithms i.e., density-based clustering, k-mean clustering, Birch clustering, spectral clustering, and agglomerative clustering. Performance metrics for different clusters are shown in Table 7 and an average of dataset performance is shown in Table 8.

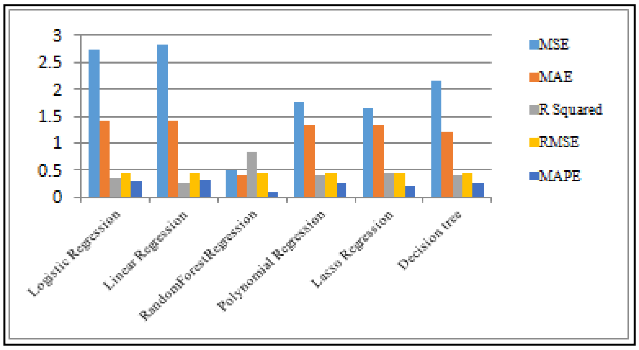

FIGURE 14. Comparison of regression parameters for the Average of all the datasets together.

TABLE 7. Clustering algorithms performance comparison using various performance metrics.

Regression Algorithm VS Performance Metrics	Data Set	ARI	SC	CHI	DI	DBI
DENSITY-BASED CLUSTERING	D1	0.53	0.40	2.93	0.20	0.76
	D2	0.55	0.40	2.73	0.20	0.70
	D3	0.53	0.40	2.30	0.20	0.76
	D4	0.51	0.39	2.11	0.20	0.70
	D5	0.56	0.40	2.92	0.15	0.76
K MEAN CLUSTER	D1	0.58	0.41	3.19	0.22	0.76
	D2	0.58	0.42	3.27	0.21	0.73
	D3	0.58	0.42	3.33	0.24	0.76
	D4	0.58	0.43	3.51	0.29	0.73
	D5	0.58	0.42	3.72	0.19	0.76
BIRCH CLUSTERING	D1	0.53	0.41	2.94	0.20	0.77
	D2	0.55	0.40	2.73	0.20	0.71
	D3	0.54	0.40	2.31	0.20	0.77
	D4	0.51	0.39	2.12	0.20	0.71
	D5	0.57	0.41	2.92	0.16	0.77
AGGLOMERATIVE CLUSTERING	D1	0.59	0.42	3.19	0.22	0.76
	D2	0.58	0.42	3.28	0.21	0.74
	D3	0.59	0.42	3.33	0.24	0.76
	D4	0.59	0.43	3.52	0.30	0.74
	D5	0.58	0.42	3.73	0.20	0.76
SPECTRAL CLUSTERING	D1	0.54	0.42	2.94	0.21	0.77
	D2	0.56	0.41	2.74	0.21	0.71
	D3	0.55	0.41	2.32	0.21	0.77
	D4	0.52	0.39	2.12	0.21	0.71
	D5	0.57	0.42	2.93	0.16	0.77

TABLE 8. Average of Clustering Algorithms Performance comparison using various performance metrics.

Regression Algorithm VS Performance Metrics	ARI	SC	CHI	DI	DBI
Density Based Clustering	0.54	0.40	2.60	0.19	0.74
K Mean Clustering	0.58	0.42	3.41	0.23	0.75
Birch Clustering	0.54	0.40	2.60	0.20	0.74
Agglomerative Clustering	0.59	0.43	3.41	0.24	0.75
Spectral Clustering	0.54	0.41	2.61	0.20	0.74

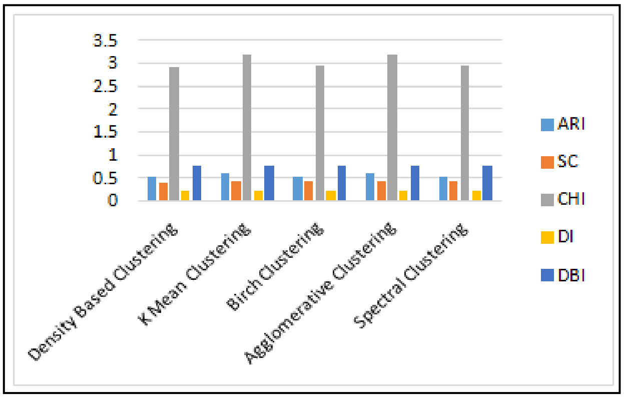

FIGURE 15. Clustering performance metrics for stock price prediction.

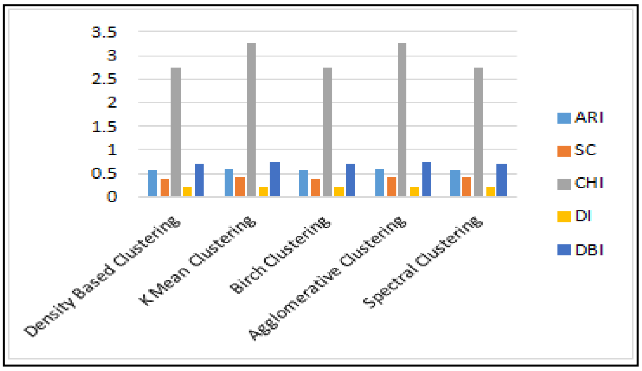

FIGURE 16. Clustering performance metrics for Fraud detection.

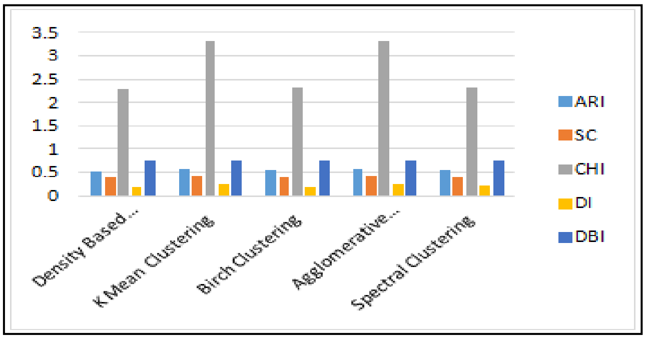

FIGURE 17. Clustering performance metrics for sentiment analysis.

The graphical simulation table presents clustering performance metrics and the average of performance metrics for five datasets. The study shows that K-Mean clustering performed best in terms of performance metrics compared to other clusters.

Figure 15 presents clustering performance metrics for the stock price prediction dataset.

Figure 16 presents Clustering performance metrics for fraud detection.

Figure 17 presents Clustering performance metrics for sentiment analysis.

Figure 18 shows clustering performance metrics for Anomaly analysis.

Figure 19 shows clustering performance metrics for healthcare detection.

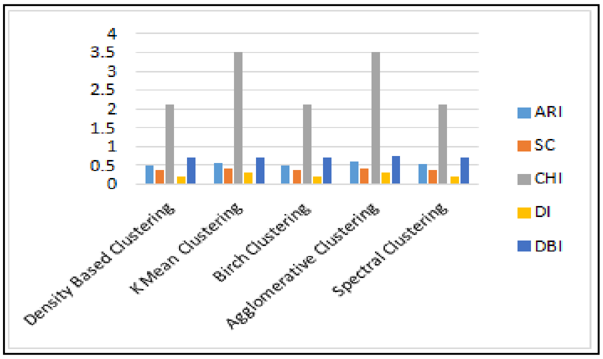

FIGURE 18. Clustering performance metrics for Anomaly detection.

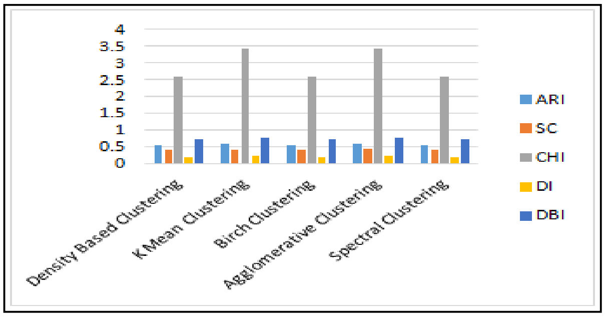

FIGURE 19. Comparison of clustering parameters for healthcare dataset.

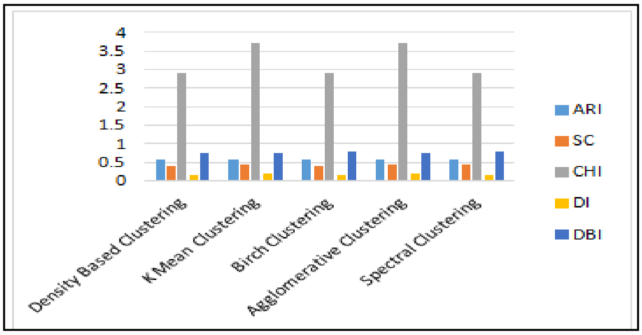

FIGURE 20. Comparison of clustering parameters for five datasets and Average of all the datasets together.

Figure 20 compares clustering performance metrics for an average of five datasets.

V. CONCLUSION

This study compared the performance of different classification, regression, and clustering algorithms. The results showed that ROBERTA is the most accurate classification algorithm for text data, outperforming SVM, Naïve Bayes, LSTM, and ADA Boost. RESNET is the most accurate classification algorithm for image data, outperforming DENSENET and INCEPTION. Random forest regression is the most accurate regression algorithm for textual data, outperforming logistic regression, linear regression, polynomial regression, lasso regression, and decision tree regression. K-means clustering is the most effective clustering algorithm for textual data, outperforming density-based clustering, Birch clustering, spectral clustering, and agglomerative clustering. This study provides valuable insights for the development of hybrid methodologies, which can be used to make well-informed decisions in complex situations.

REFERENCES

[1] S. Gupta and B. Kisan, ''Assessing the effectiveness of predictive machine learning algorithms based on classification techniques,'' *Eur. Chem. Bull.*, vol. 12, no. 4, pp. 13882–13895, 2023.

[2] S. Gupta and B. Kisan, ''A review study of prediction-based models,'' *J. Oriental Res. Madras*, vol. 202, pp. 62–72, Jan. 2021.

[3] S. Gupta and B. Kisan, ''A literature study of predictive machine learning algorithms,'' *Int. J. Interdiscipl. Organizational Stud.*, vol. 16, no. 4, pp. 59–71, Dec. 2021.

[4] J. Goyal and B. Kishan, ''TLHEL: Two layer heterogeneous ensemble learning for prediction of software faults,'' *Int. J. Eng. Trends Technol.*, vol. 69, no. 4, pp. 16–20, Apr. 2021, doi: 10.14445/22315381/ijett-v69i4p203.

[5] M. Hasan, M. M. Islam, M. I. I. Zarif, and M. Hashem, ''Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches,'' Dept. Comput. Sci. Eng., Khulna Univ. Eng. Technol., Khulna, Bangladesh, Tech. Rep., 2019, doi: 10.1016/j.iot.2019.10.

[6] D. R. J. Priyadarsini and D. S. Titus, ''Survey on predictive analysis of diabetes disease using machine learning algorithms,'' *Int. J. Comput. Sci. Mobile Comput.*, vol. 9, no. 10, pp. 19–27, Oct. 2020, doi: 10.47760/ijcsmc.2020.v09i10.003.

[7] T. Saranya, S. Sridevi, C. Deisy, T. D. Chung, and M. K. A. A. Khan, ''Performance analysis of machine learning algorithms in intrusion detection system: A review,'' *Proc. Comput. Sci.*, vol. 171, pp. 1251–1260, Jan. 2020, doi: 10.1016/j.procs.2020.04.133.

[8] S. Amiri, M. Akbarabadi, F. Abdolali, A. Nikoofar, A. J. Esfahani, and S. Cheraghi, ''Radiomics analysis on CT images for prediction of radiation-induced kidney damage by machine learning models,'' *Comput. Biol. Med.*, vol. 133, Jun. 2021, Art. no. 104409, doi: 10.1016/j.compbiomed.2021.104409.

[9] N. Biswas, K. M. M. Uddin, S. T. Rikta, and S. K. Dey, ''A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach,'' *Healthcare Anal.*, vol. 2, Nov. 2022, Art. no. 100116, doi: 10.1016/j.health.2022.100116.

[10] Y.-S. Su, Y.-D. Lin, and T.-Q. Liu, ''Applying machine learning technologies to explore students' learning features and performance prediction,'' *Frontiers Neurosci.*, vol. 16, Dec. 2022, Art. no. 1018005, doi: 10.3389/fnins.2022.1018005.

[11] Ö. Eskidere, F. Ertas, and C. Hanilçi, ''A comparison of regression methods for remote tracking of Parkinson's disease progression,'' *Expert Syst. Appl.*, vol. 39, no. 5, pp. 5523–5528, Apr. 2012, doi: 10.1016/j.eswa.2011.11.067.

[12] S. I. Zahara, M. Ilyas, and T. Zia, ''A study of comparative analysis of regression algorithms for reusability evaluation of object oriented based software components,'' in *Proc. Int. Conf. Open Source Syst. Technol.*, Dec. 2013, pp. 75–80.

[13] P. Strecht, L. Cruz, C. Soares, J. Mendes-Moreira, and R. Abreu, ''A comparative study of classification and regression algorithms for modelling students' academic performance,'' in *Proc. 8th Int. Conf. Educ. Data Mining*, Jun. 2015, pp. 1–4.

[14] C. M. van der Walt and N. Botha, ''A comparison of regression algorithms for wind speed forecasting at Alexander bay,'' in *Proc. Pattern Recognit. Assoc. South Afr. Robot. Mechatronics Int. Conf. (PRASA-RobMech)*, Stellenbosch, South Africa, Nov. 2016, pp. 1–5.

[15] M. S. Acharya, A. Armaan, and A. S. Antony, ''A comparison of regression models for prediction of graduate admissions,'' in *Proc. Int. Conf. Comput. Intell. Data Sci. (ICCIDS)*, Feb. 2019, pp. 1–5.

[16] J. Yang and J. Wang, ''Tag clustering algorithm LMMSK: Improved K-means algorithm based on latent semantic analysis,'' *J. Syst. Eng. Electron.*, vol. 28, no. 2, pp. 374–384, Apr. 2017, doi: 10.21629/JSEE.2017.02.18.

[17] X. Gong, L. Liu, S. Fong, Q. Xu, T. Wen, and Z. Liu, ''Comparative research of swarm intelligence clustering algorithms for analyzing medical data,'' *IEEE Access*, vol. 7, pp. 137560–137569, 2019, doi: 10.1109/ACCESS.2018.2881020.

[18] A. E. Ezugwu, M. B. Agbaje, N. Aljojo, R. Els, H. Chiroma, and M. A. Elaziz, ''A comparative performance study of hybrid firefly algorithms for automatic data clustering,'' *IEEE Access*, vol. 8, pp. 121089–121118, 2020, doi: 10.1109/ACCESS.2020.3006173.

[19] A. A. Bushra and G. Yi, ''Comparative analysis review of pioneering DBSCAN and successive density-based clustering algorithms,'' *IEEE Access*, vol. 9, pp. 87918–87935, 2021, doi: 10.1109/ACCESS.2021.3089036.

[20] F. Malik, S. Khan, A. Rizwan, G. Atteia, and N. A. Samee, ''A novel hybrid clustering approach based on black hole algorithm for document clustering,'' *IEEE Access*, vol. 10, pp. 97310–97326, 2022, doi: 10.1109/ACCESS.2022.3202017.

[21] M. Raeisi and A. B. Sesay, ''A distance metric for uneven clusters of unsupervised K-means clustering algorithm,'' *IEEE Access*, vol. 10, pp. 86286–86297, 2022, doi: 10.1109/ACCESS.2022.3198992.

[22] *Stoke Price Dataset*, Copyright e-Eighteen.com Ltd. All Rights Reserved. Reproduction of News Articles, Photos, Videos, or any Other Content in Whole or in Part in Any Form or Medium Without Express Written Permission of moneycontrol.com is Prohibited, Money Control, New Delhi, India, 1999

[23] A. Chauhan. (2023). *Fraud Detection*. Anonymized Credit Card Transactions Labeled as Fraudulent or Genuine. [Online]. Available: https://www.kaggle.com/datasets/whenamancodes/fraud-detection

[24] U. GHosh. (2018). *IMDB Review Dataset*. Database: Open Database, Contents: Database Contents. https://www.kaggle.com/datasets/utathya/imdb-review-dataset

[25] Elliptic and 1 Collaborator. (2019). *Elliptic Dataset: Bitcoin Transaction Graph*. Attribution-Non Commercial-No Derivatives 4.0 International (CC BY-NC-ND 4.0). [Online]. Available: https://www.kaggle.com/datasets/ellipticco/elliptic-data-set

[26] ROB HARRAND. (2019). *What Causes Heart Disease? Explaining the Model*. [Online]. Available: https://www.kaggle.com/datasets/ellipticco/elliptic-data-set

[27] DARREN2020. (2020). *CT and MRI Brain Scans: Cross-Sectional Scans for the Unpaired Image to Image Translation*. CC BY-NC-SA 4.0. [Online]. Available: https://www.kaggle.com/datasets/darren2020/ct-to-mri-cgan

[28] HSANKESARA. (2018). *Flickr Image Dataset: Flickr Image Captioning Dataset*. Flickr30k, CC0: Public Domain. [Online]. Available: https://www.kaggle.com/datasets/darren2020/ct-to-mri-cgan

SWATI GUPTA received the B.E. degree in computer science from BRCM, Bahal, affiliated with Maharshi Dayanand University, Rohtak, Haryana, India, in 2009, and the M.Tech. degree in computer science and engineering from TITS, Bhiwani, affiliated to Maharshi Dayanand University, in 2011. She is currently pursuing the Ph.D. degree in data science with the Department of Computer Science and Application, Maharshi Dayanand University. Her research interests include artificial intelligence, prediction algorithms, machine learning, deep learning, python, R, and Julia.

BAL KISHAN is currently an Assistant Professor with the Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak, Haryana. He has teaching experience of 12 years at the postgraduate level. He has published more than 30 research articles in national and international journals.

PREETI GULIA is currently an Associate Professor with the Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak, Haryana. She has teaching experience of 15 years at the postgraduate level. She has published more than 84 research articles in national and international journals.