Journal of Computational Science 69 (2023) 102011

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Impact learning: A learning method from feature's impact and competition

Nusrat Jahan Prottasha [a], Saydul Akbar Murad [b,c], Abu Jafar Md Muzahid [b], Masud Rana [c], Md Kowsher [a], Apurba Adhikary [c], Sujit Biswas [d], Anupam Kumar Bairagi [e,*]

[a] Stevens Institute of Technology, Hoboken, NJ 07030, USA
[b] Universiti Malaysia Pahang, UMP Pekan, Kuantan, 26600, Pahang, Malaysia
[c] Noakhali Science and Technology University, Sonapor, Noakhali, 3802, Chittagong, Bangladesh
[d] University of East London, University Way, London E16 2RD, UK
[e] Khulna University, Khulna, 9208, Bangladesh

ARTICLE INFO

Keywords:
Impact learning
Machine learning
Classification
Regression
Asthma prediction
Diabetes prediction
Heart disease identification

ABSTRACT

Machine learning is the study of computer algorithms that can automatically improve based on data and experience. Machine learning algorithms build a model from sample data, called training data, to make predictions or judgments without being explicitly programmed to do so. A variety of well-known machine learning algorithms have been developed for use in the field of computer science to analyze data. This paper introduced a new machine learning algorithm called impact learning. Impact learning is a supervised learning algorithm that can be consolidated in both classification and regression problems. It can furthermore manifest its superiority in analyzing competitive data. This algorithm is remarkable for learning from the competitive situation and the competition comes from the effects of autonomous features. It is prepared by the impacts of the highlights from the intrinsic rate of natural increase (RNI). We, moreover, manifest the prevalence of impact learning over the conventional machine learning algorithm.

1. Introduction

Machine learning (ML) is a state-of-the-art approach that has shown promise in the areas of categorization and prediction. To improve demand estimates, we can use a variety of methods to examine historical data, including time series analysis, machine learning techniques, and deep learning models. As needed, ML ensures program consistency and adaptability. Machine learning, while not competitive, will continue to grow in the near future due to increasing data capital and greater need for personalized applications (e.g., the development of matrix multiplications) [1]. In addition to app growth, ML is also expected to change the general perspective of computer science. ML emphasizes creating a self-monitoring, self-diagnosing, and self-repairing system by shifting the focus from "how to program a machine" to "how to make it program itself". Both statistics and computer science can contribute to the evolution of ML as they develop and apply increasingly complex ideas that change the way people think [2]. In statistics and machine learning, extracting knowledge from data is an important endeavor. Many fields, including biomedicine [3], Business Analytics [4], Computational Optimization [5], Criminal Justice [6], Cybersecurity [7], Policy Making [8], Process Monitoring [9] rely on it. Mathematical Optimization plays a vital role in building such models.

Classification and Regression Modern approaches based on recursive partitioning are trees. Those models are theoretically simple and exhibit exceptional learning performance. Nevertheless, those are very computationally expensive. There are methods and packages to instruct those models in common programming languages such as Python and R. Those models are desirable not only in terms of their interoperability but also because of their rule-based nature [10]. Those models are too attractive for a variety of applications, such as credit scoring for lending. Those models used a dataset of individuals that includes demographic and financial predictors, among others, and models use this information to predict whether consumers will be good or bad payers.

Each machine learning algorithm has its own application [11]. An approach may be optimal for a particular dataset but not for others. In real life, we work with directed data such as time periods (e.g., day, week, month, etc.), orientation, and rotation. Special algorithms are required to manage directed data. A few researchers, including [12], have developed a non-probabilistic model for directional data. Kowsher

* Corresponding author.
E-mail addresses: jahannusratprotta@gmail.com (N.J. Prottasha), saydulakbarmurad@gmail.com (S.A. Murad), mrumi98@gmail.com (A.J.M. Muzahid), masudit01404@gmail.com (M. Rana), ga.kowsher@gmail.com (M. Kowsher), apurba@nstu.edu.bd (A. Adhikary), sujitbiswas@ieee.org (S. Biswas), anupam@cse.ku.ac.bd (A.K. Bairagi).

https://doi.org/10.1016/j.jocs.2023.102011
Received 2 January 2023; Received in revised form 9 March 2023; Accepted 21 March 2023
Available online 6 April 2023

et al. [13] has proposed a set of directional Support Vector Machines: updated SVM decision function. Using cosine and triangular waves, they study the periodic parametric mapping of directional variables. Moreover, they modified the model with triangular waves to allow asymmetric circular boundaries and kernelized the different SVM variants. Nevertheless, the additional factors involved in asymmetric SVMs periodically affect the decision boundary. We generate a large amount of directional data on a daily basis. However, the standard statistical distribution is inappropriate for this type of data. Therefore, different distributions and statistics should be used to study this type of data, such as the univariate von Mises distribution and the multivariate von Mises–Fisher distribution. A unidirectional Naive Bayes classifier predictor variable was demonstrated by López-Cruz et al. [14]. Von Mises univariate must be used to apply this model. The predictor variables are modeled using the von Mises–Fisher and Gaussian distributions. The parameters of a Gaussian distribution have been found to affect the degree of complexity of the decision surfaces in a hybrid context. Von Mises Naive Bayes (vMNB) was tested on eight datasets and compared to the NB classifiers that use Gaussian distributions or discretization to model angular variables. The Selective NB classifier was the best algorithm for classifying compared to others. A generative probabilistic model, called latent Dirichlet assignment (LDA), was presented by Blei et al. [15]. They described how to estimate empirical Bayes parameters using variational approaches and an EM algorithm, as well as other useful approximate inference techniques. Document modeling, collaborative filtering, and text classification were all applications of their findings. Tree Augmented Naive Bayes (TAN) outperforms naive Bayes, which is shown by Dumitrescu [16]. Nonetheless, the TAN maintained the same computational simplicity and resilience as the Bayes model. Similar work has been completed by Yeruva et al. [17] for categorizing Cell Anemia using a Deep neural network.

Different machine learning algorithms perform better for different types of data. In this manner, there are no "best" learning algorithms to be declared [18]. We are living in an age where undertakings are information-driven, and the world is delivering more information than natural resources. There is a typical way to deal with this while building up an algorithm, which is to plan a cost function and further minimize the cost function. Practically all learning algorithms attempt to limit a cost function for a superior and streamlined outcome. Elizondo [19] presents an outline of a few of the strategies for testing linear separability between two classes. The techniques are separated into four gatherings: Those dependent on linear programming and computational math and one dependent on neural networks on quadratic programming. Shalev-Shwartz and Singer [20] depict and examine a productive boosting system that can be utilized for limiting the loss functions obtained from our group of relaxations. An et al. [21] explores how concealed layers of profound rectifier networks are equipped for change, at least two example sets to be linearly distinguishable while safeguarding the distances with an ensured degree and demonstrates the all-inclusive characterization force of such distance-saving rectifier organizations. Bzdok et al. [22] outline SVM utilizing a two-class issue and start with a case in which the classes are linearly divisible, implying that a straight line can be drawn that consummately isolates the classes, with the edge being the opposite distance between the nearest focus to the line from each category. Raissi et al. [23] uses recent developments in probabilistic machine learning to identify conditions arising from parametric linear operators. These conditions include but are not limited to, the conventional and partial differential, fractional order operators, and integrodifferential. Pavlyshenko [24] studies the utilization of logistic regression in manufacturing failure recognition. Using the generalized linear model for logistic regression makes it possible to break down the influence of the variables under study.

The core purpose of classification and regression is to predict the outcome from a single multidimensional data source. Regression attempts to fit a line with the data curve, whereas classification finds classes from the labeled data. The full functionality of an algorithm mimics a mathematical function. In [18], a robust machine learning algorithm known as Impact Learning has been proposed, where the main aim was to solve the classification and regression problems in machine learning applications. This article will briefly describe that algorithm for better understanding. If (X) represents the data and (Y) represents the outcome, then the model (F) would look like a function that converts X into Y [F(X)=Y]. This function can be called a mapping function.

In statistics and population studies, the rate of natural increase (RNI) is a well-known terminology [19]. That is collected from real-world applications, and therefore, to keep it independent and distinguished, we can utilize (RNI) for generating a machine learning model. The following expression can represent RNI:

$$\frac{dP}{dt} \approx rP, \tag{1}$$

Where r is taken as RNI and P define as population size. To extend the equation by incorporating a logistic growth model [20]. From this, we find,

$$\frac{dP}{dt} \approx rP(1 - \frac{p}{k}), \tag{2}$$

A dataset is comprised of numerous features. Almost every component of the dataset has RNI characteristics. In a competitive scenario, back forces are detrimental to the elements. Along these lines, the target variable gets affected by different highlights of the back forces, and we name that "Back Impact on Target (BIT)". Since the objective element depends on BITs, that is the reason each BIT additionally relies upon the objective factor.

In this paper, we intend to uncover the presence of impact learning. Here we have utilized three kinds of datasets to verify the impact learning. We likewise showed the graphical and factual correlations among all current machine learning techniques such as Random Forest tree, Naive Bayes, SVM, Logistic regression, etc. The main contributions of this work are presented as follows:

- We proposed an algorithm for supervised machine learning called impact learning that is used to utilize classification and regression datasets.
- We used RNI as a guide, this method makes use of the back impact of previous features.
- This algorithm is used to analyze the real competition as it learns from the competition.
- On the basis of a real dataset, we conducted a comparison between impact learning and existing algorithms like KNN, SVM, etc.

2. Introduction of impact learning

To visualize the exponential growth of population increment, we use the Malthusian logistic growth model [25], which can be expressed as follow:

$$\frac{dy}{dt} \approx ry, \tag{3}$$

where "y" indicates the population and "r" is the RNI. By incorporating carrying capacity (K) with Eq. (3) we get:

$$\frac{dy}{dt} \approx ry(1 - \frac{y}{k}). \tag{4}$$

Carrying capacity plays a crucial role in population dynamics. The population of a region cannot exceed a certain limit in ecology. Therefore, the growth rate in Eq. (4), has to be reversed after a particular time and gradually hit zero. We can visualize the phenomena through Fig. 1.

In Fig. 2, we see a logistic growth model of cellphone purchases where the RNI curve increases exponentially from 0 to day 3, and after that, the rate decreases. We observe that the purchase impacts itself

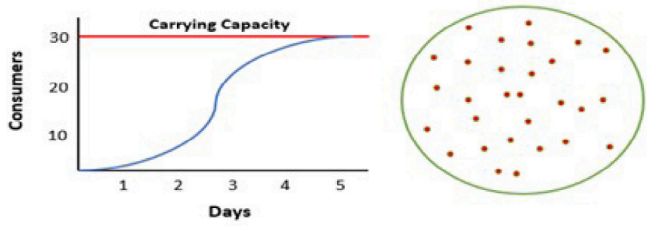

Fig. 1. Whenever we have one operator(independent variable) only, the RNI then depends on only the number of samples (consumers). The rate decreases with time passing.

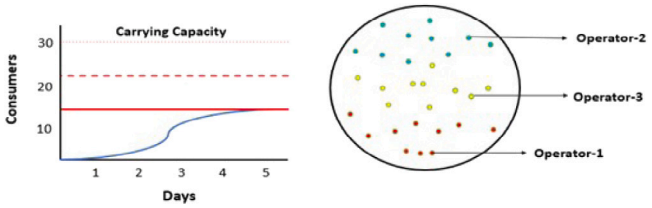

Fig. 2. In thing figure, the RNI of operator-1 depends on other operators also and it is limited by its carrying capacity because of other operators.

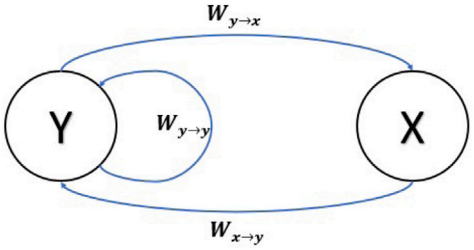

Fig. 3. Here y is a target variable that has two dependents, one dependence is for another variable and another one is self-dependence.

since a rise in the purchases automatically obstructs its growth at a certain point. Likewise, the presence of different operators can affect each other. If a particular operator makes profits, other operators face a decline simultaneously (see Fig. 3).

Back-impact on the target prevents target features from moving toward the curve of the RNI (BIT). If x is the back-impact variable, then x and y maintain their effects on each other, but y maintains its effects on itself. So, we can rewrite Eq. (4) as

$$\begin{aligned}\frac{dy}{dt} &= ry - w_{x\longrightarrow y}xy - w_{y\longrightarrow x}xy \\ &= ry - (w_{x\longrightarrow y} + w_{y\longrightarrow x})xy \\ &= ry - w * xy\end{aligned}$$

$$Or, \frac{dy}{dt} = ry - w_y y^2 - wxy, \tag{5}$$

To visually explain,

Now, combining Eqs. (3), (4), and (5) we get

$$\begin{aligned} ry(1-\frac{y}{x}) &= ry - w_y y^2 - wxy \\ Or, r(1-\frac{y}{x}) &= r - w_y y - wx \\ Or, y(\frac{r}{k} - w_y) &= -wx \end{aligned}$$

$$Or, y = \frac{kwx}{r - w_y k} + b, \tag{6}$$

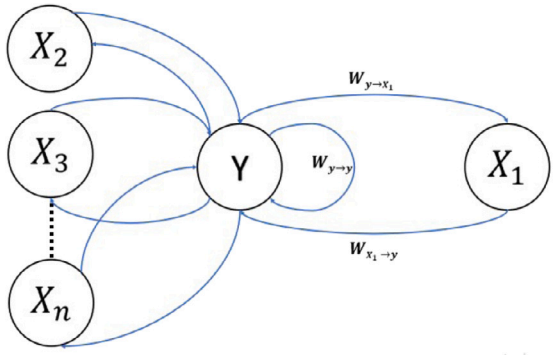

Fig. 4. For multiple variables the constructed dependence.

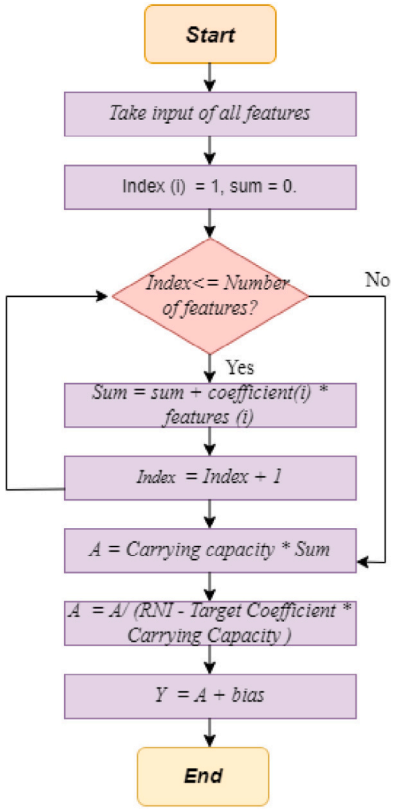

Fig. 5. Workflow of calculating impact learning step by step.

To determine the impact of x_k target feature y (trained) and y' is the target feature, then we get from the (6) as follows:

$$\text{Imp(y)} = \left(y' - \left(\frac{k\sum_{i=1}^{n} w_i x_i}{r - w_y k} + b\right)\right)^{\frac{2}{N}} \quad \text{if i} \neq \text{k}, \tag{7}$$

where, N is the size of the dataset.

If $X = [x_1, x_2, x_3, \ldots\ldots, x_n]$ and $W = [w_1, w_2, w_3, \ldots\ldots, w_n]$, then the (6) can be expressed as follows (see Fig. 4).

$$y = \frac{k(w^T.x)}{r - w_y k} + b. \tag{8}$$

In order to illustrate the polynomial structure of the impact learning, the (6) can be further expressed as follows:

$$y = \frac{k\sum_{i=1}^{n} w_i x_i^j}{r - w_y k} + b, \tag{9}$$

Table 1
Dataset attributes for diabetes identification.

Attributes	Description	Type
Age	Age of patient.	Numeral
Weight	Weight of the patient.	Numeral
Family History	If the family member has diabetes history,than 'Yes' otherwise 'No'.	Boolean
Late-night sleep habits	It defines the sleeping habit of patients.	Boolean
Heart Diseases	If the patient has heart disease, then 'Yes' otherwise 'No'.	Boolean
Sleep after eating	If the patient goes to bed after eating, then 'Yes' otherwise 'No'.	Boolean
Addiction	This field want to know about drug addiction of patient.	Boolean
Sex	Gender of patient.	Boolean
Late wake-up habit	Wake-up habit of patient.	Boolean
Exercise	Exercise habit.	Boolean

where j> 0. For the N size of data, the RNI and carrying capacity can be found from these defined functions.

$$r = \frac{\ln \frac{\max(y')}{\min(y')}}{N-1} \text{ and } k \geq \max\left(y'\right) \tag{10}$$

However, it is an excellent way to calculate the RNI (r) from the optimization techniques like gradient descent. Fig. 5 describes the flow chart of the proposed algorithm which is mathematically expressed in Eq. (1) above.

3. Experimental results and discussion

We used the proposed method to tackle various real-world situations, demonstrating the acceptability of the Impact Learning model. In addition, we compared the performance of the proposed approach to that of other standard machine learning techniques (Random Forest, KNN, SVM, Naïve Bayes, Linear Regression, and Logistic Regression). Besides, we have employed a variety of datasets. The dataset, training, and testing procedure, experimental method, outcome, and commentary will all be covered in the following sub-sections.

3.1. Datasets

All of the datasets we have used to verify the proposed approach have been gathered from the real-world environment. We constructed a Google form for each dataset to collect the information. Our first dataset is for the prediction of diabetes. We created two Google forms to collect data from participants by asking them various questions. A separate form exists for patients and another for non-patients; however, all fields are identical in both formats. The total data obtained is 469, with 232 non-diabetic patients and 237 diabetic patients included. There are 11 attributes of the first dataset, shown in Table 1.

Before collecting data for our second dataset on asthma, we consult with asthma specialties doctors. We had taken a brief note from the doctor's description of the traits that cause asthma. Finally, we extracted 23 characteristics from them. Based on such features, we created a Google Form. We collected information through social media and through visits to the doctor's office. This Google form was distributed to a public group, shared on our social media timeline, and emailed to our contacts to collect data. We spent several days physically collecting data from asthma patients and those who exhibit asthmatic symptoms in the doctor's chamber. While some previous research collected data exclusively from youngsters, we obtained data from individuals of various ages in our study. We do not have any restrictions on age. Finally, we gathered almost 500 data points from social media and the chamber of visiting doctors. Table 2 summarizes all of the characteristics of the second dataset.

Our third dataset is utilized to forecast heart disease prediction. This dataset was gathered from Kaggle and included a total of fifteen features [26]. The first fourteen variables (male, currentSmoker, age, BPMeds, cigsPerDay, prevalentStroke, prevalentHyp, totChol, diabetes, sysBP, BMI, diaBP, glucose heartrate, and region) are utilized as feature data, while the last variable (TenYearCHD) is used as level data. All of the implemented machine learning models performed well on this dataset. Table 3 contains the attributes and description of the heart disease dataset.

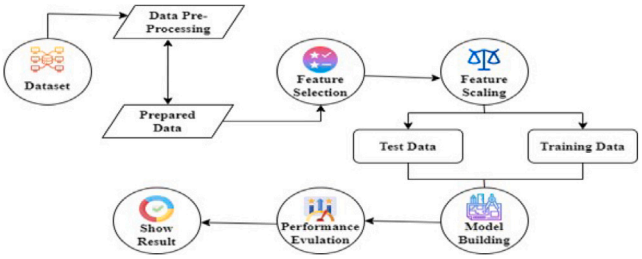

Fig. 6. Workflow of our experimental task.

3.2. Experimental setup

We use a simulation environment to present the result in this paper. First, we devised a mathematical model. To implement this model, we wrote a Python script that ran on Google Colaboratory, a cloud computing platform with a Python programming environment. For quicker and parallel computing, we employed a Google Colab-based GPU. The development of all the models took more than 200 h in total. We have also made a python module of this model available for public usage, which will make it much easier for consumers to use it. The proposed model was implemented in the following steps, as illustrated in Fig. 6.

3.2.1. Data pre-processing and preparing data

During data preprocessing, we determine whether a dataset contains null values. After the one-time data processing, we checked if the data is prepared for the following process. If we determined that the data is ready for the next process, then we stop it [27]. We repeated it numerous times. We utilized mean and median to preprocess the data. A computation is used to determine the mean, synonymous with the average value of a data collection.

3.2.2. Feature selection

Feature selection can be used to reduce the input variable model's size by selecting only relevant data and eliminating noise in the data. It is the process of automatically selecting suitable characteristics for the machine learning model utilized based on the type of problem being addressed. It is accomplished by incorporating or removing critical elements without altering their functionality. It aids in reducing noise in our data and the reduction of the size of our input data. Fig. 7 illustrates the diagram of feature selection. Feature selection models can be classified into supervised and unsupervised [28]. We use a supervised learning model to analyze the data of our datasets. The term "supervised feature selection" refers to selecting features that make use of the output label class. They use the target variables to uncover variables that can improve the model's efficiency and then incorporate these variables into the model.

Table 2
Dataset attributes for asthma identification.

Attributes	Description	Type
Gender	If gender male then 'M' or 'F' for Female	Boolean
Age	Age of the patient	Numeral
Frequent Coughing	It defines the continuous coughing. If it continues then 'YES otherwise 'NO'	Boolean
Shortness of breath	It defines Breathing problems	Boolean
Exercise or walks make feel very tired	If the person is tired, then 'YES' otherwise 'NO'	Boolean
Nausea or Coughing after walking or exercise	Walking or Running for a while if he/she is coughing then 'YES' otherwise 'NO'	Boolean
Sleep problems	If face difficulties in sleep then 'YES', otherwise 'NO'	Boolean
There is a feeling of being stuck in the chest	Usually, if you feel chest pain, answer 'YES'; otherwise, answer 'NO'	Boolean
Is anyone in family, suffers from allergies?	It defines family history. If anyone has allergies, then 'YES' otherwise 'NO'	Boolean
Does any family member has asthma?	It defines family history. If any family has asthma, then 'YES' otherwise 'NO'	Boolean
Do you have allergies?	If have allergies, then 'YES' otherwise 'NO'	Boolean
Which season does shortness of breath increase?	Here have four seasons. From there, Have to select one season, when problem is increase.	Text
Problem in Cold Water, Curd or Ice-cream, beef, ilish, prawn, brinjal, pumpkin, coconuts, Malabar night shade, duck meat	If the patient feels this kind of disease then 'Yes' otherwise 'No'	Boolean
Have asthma?	If you have asthma put 'YES', otherwise 'NO'	Boolean

Table 3
Dataset attributes for heart disease identification.

Attributes	Description	Type
Male	It defines the sex: male or female.	Nominal
Age	Patient age.	Number
CurrentSmoker	Whether the patient is a current smoker or not	Nominal
CigsPerDay	The average number of cigarettes smoked each day by the individual	Number
BPMeds	Patient's blood pressure medicine	Nominal
PrevalentStroke	Whether the patient had already experienced a stroke	Nominal
PrevalentHyp	Whether the patient was hypertensive or not	Nominal
Diabetes	Diabetes history of the patient	Nominal
TotChol	It defines cholesterol level	Number
SysBP	Systolic blood pressure of patient	Number
DiaBP	Current diastolic blood pressure	Number
BMI	Body Mass Index of patient	Number
HeartRate	Heart rate of patient	Number
Glucose	Glucose level	Number
TenYearCHD	Coronary heart disease is a 10-year risk	Boolean

3.2.3. Feature scaling

In data analysis, feature scaling is a technique used to standardize the range of independent variables or characteristics of data [29,30]. This happens in the preprocessing phase of data processing and is also called the normalization of data. Normalization and standardization are two of the most commonly used strategies for feature scaling. Normalization is utilized when we wish to confine our values to a range between two numbers, often between [0, 1] and [−1, 1], and we want to do so as efficiently as possible. The following is the general

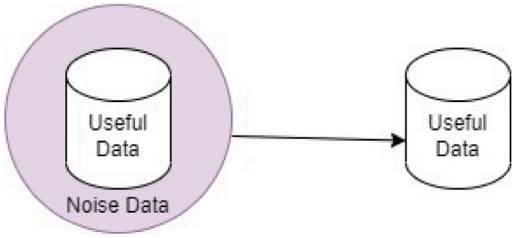

Fig. 7. Feature Selection.

normalization formula:

$$z' = \frac{z - \min(z)}{\max(z) - \min(z).} \tag{11}$$

Feature standardization reduces the data's variance and means to zero for each of the data's features. For each feature, the distribution mean and standard deviation are calculated, then the new data point is calculated using the formula below:

$$z' = \frac{z - \bar{z}}{\sigma} \tag{12}$$

We separate our dataset into two parts in the feature scaling process: training and testing, with 70 percent of the data being used for training and 30 percent for testing, respectively. For the first and second datasets, we use Normalization; however, for datasets three and four, we use standardization.

3.2.4. Model building

We employ four classification techniques (Random Forest, k-nearest neighbors, support vector machine, and Naive Bayes) and two regression models (Linear regression and Logistic Regression) to utilize the entire dataset, except the newly proposed model. We compared our proposed model to the techniques employed to analyze the whole dataset. For our proposed model, we got a good outcome. A summary of all algorithms used in this work and their parameters are represented in Table 4.

Table 4
This table contains the details of all implemented algorithms.

Name of Algorithm	Description	Parameters
Random Forest	Creates decision trees from various samples and uses the majority vote for classification.	Number of decision tress
K-Nearest Neighbors (KNN)	The new data point is classified by the majority of votes from its fixed neighbors.	Nearest Neighbors.
Support Vector Machine (SVM)	Find a hyperplane in an N-dimensional space that clearly classifies the data points (N — the number of characteristics)	Kernel function.
Naïve Bayes	Classification method based on Bayes' Theorem and the premise of predictor independence	Probabilities of different classes
Linear Regression	Undertakes the task of predicting a dependent variable (target) based on the independent variable provided (s).	Value of independent variable
Logistic Regression	It is the method of choice for binary classification issues (problems with two class values)	Value of all independent variable

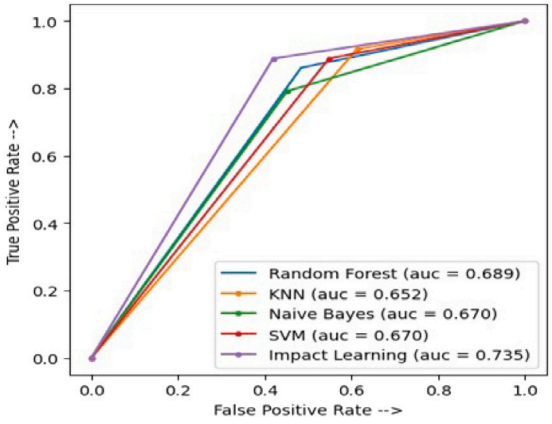

Fig. 8. Comparison of all implemented algorithms with Impact Learning using ROC curve for asthma dataset.

3.3. Performance analysis

This section is mainly divided into two sub-sections: performance analysis for classification algorithms and regression algorithms. The accuracy, precision, F1-score, recall, learning curves, and ROC curves for classification algorithms are used to examine the results of the experiments. These evaluation matrices are compared for four classification algorithms and two regression algorithms.

3.3.1. Performance analysis for classification algorithms (asthma dataset)

We collected 500 data points from the real world to evaluate the asthma dataset, 211 from asthma patients and 289 from non-asthmatic patients. As a result of the fact that this is a balanced dataset, the accuracy score will provide the right idea for the prediction result. Furthermore, the evaluation should include accuracy and recall scores, and the F1 score provides a single harmonic mean score that accounts for both recall and precision. Our experiment used four robust machine learning models (Random Forest, Support Vector Machine, k-Nearest Neighbors, and Naive Bayes) and compared those to our proposed algorithm. The accuracy of the employed models is used to evaluate their performance.

From Table 5, with an accuracy of 80.6%, the proposed algorithm (Impact learning) shows the best result compared to all other proposed algorithms. For the SVM, the accuracy is also good, which is 77.5%. The KNN achieved an accuracy of 75.7%, which is very similar to Random Forest. The lowest accuracy is found for BernoulliNB, which is 71.4%. Following a comparison of the four types of performance matrices for various methods, it is clear that our suggested models (Impact Learning) perform the best in these datasets.

The receiver operating characteristic (ROC) plot is one of the most often used methods for assessing classifier performance. The ROC plot

Table 5
Accuracy, Precision, F1-score, and Recall for all implemented classification algorithms and ranked them into ascending order. These performance metrics are for asthma Dataset.

Method Name	AC	PR	RC	F1-Score	Rank (Accuracy)
Impact learning	0.806	0.761	0.734	0.745	1
SVM	0.775	0.698	0.681	0.721	2
KNN	0.757	0.721	0.653	0.661	3
Random Forest	0.752	0.683	0.711	0.691	4
Bernoulli-NB	0.714	0.665	0.671	0.662	5

AC = Accuracy, PR = Precision, RC = Recall

is built on two key evaluation metrics: specificity and sensitivity. The negative part's performance is measured by specificity, whereas the positive part's performance is measured by sensitivity [31].

In the ideal condition, FPR = 0, indicating that the area under the ROC curve (AUC) equals 1, but in sub-optimal situations, it is less than 1. In Fig. 8, the ROC curves for the asthma dataset are shown. The higher the curve, the better the results. The area under the ROC curve (AUC) for each classifier has been determined as a quantitative measure of performance. All of the methods we tested generated satisfactory results. With a score of 73.5 percent, Impact learning is better than other algorithms.

The Learning Curve is a graphical tool that may be used to estimate the benefit of supplementing our model with extra training data. It demonstrates the relationship between training and test scores for a machine-learning model with variable sample size. Generally, the cross-validation strategy is utilized when drawing the learning curve. We visualized the learning curve using the Python Yellowbrick tool. The word "Training Score" refers to the accuracy score for the training set, whereas the term "Cross-Validation Score" refers to the accuracy score for the test set.

Fig. 9 shows the learning curves of the four classification algorithms in terms of accuracy and speed. Fig. 9(a) depicts the learning curve for the SVM classifier, which demonstrates that increasing the number of training cases leads to better generalization. This model's training and test scores are highly well-matched to the dataset. Fig. 9(b) of the KNN classifier indicates that the training and test scores have not merged, showing that this model may benefit from more training. The Random Forest classifier (Fig. 9(c)) indicates a significant degree of systematic error, which suggests that the model, regardless of the quantity of data fed into it, is incapable of reflecting the underlying relationship and has a high level of systematic error. With an increase in the training dataset, BernoulliNB (Fig. 9(d)) shows that the model's performance will continue to improve. Impact learning (Fig. 10) illustrates that the training and testing are still not working effectively. We need to enhance the dataset's size and diversity to get better results.

3.3.2. Performance analysis for classification algorithms (diabetes dataset)

For the study of diabetes evolution, we have gathered data from the actual world, both physically and through an online platform. We

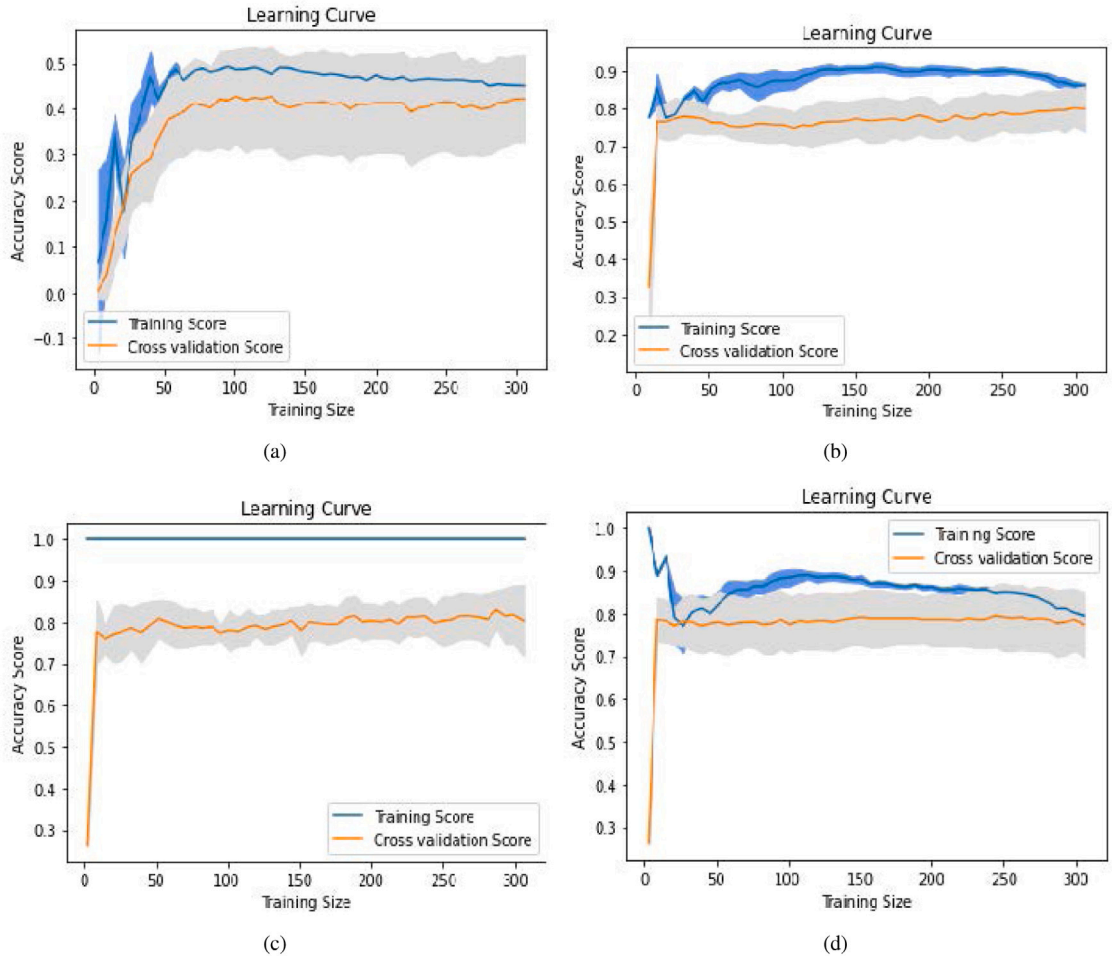

Fig. 9. Learning curve for the SVM, KNN, Random Forest, and BernoulliNb classification algorithms which are used for asthma dataset. (a) Learning curve for SVM classifier. (b) Learning curve for KNN classifier. (c) Learning curve for Random Forest classifier. (d) Learning curve for BernoulliNB classifier.

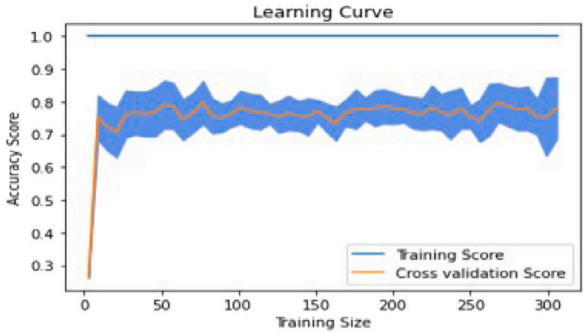

Fig. 10. Learning curve for the proposed algorithm (Impact learning) that is used for asthma dataset.

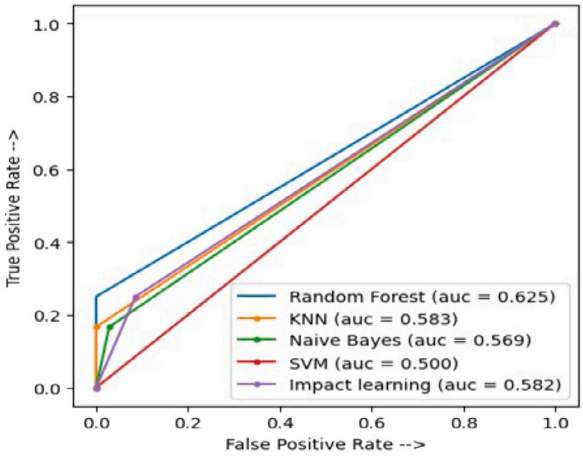

Fig. 11. Comparison of all implemented algorithms with Impact Learning using ROC curve for diabetes dataset.

have collected 469 data points, with 232 being non-diabetic and 237 being diabetic. Because the prevalence of diabetes and non-diabetes is nearly equal, this dataset is considered a balanced dataset. Accuracy is the most vital metric to consider while identifying the appropriate method for the balanced dataset. This dataset's accuracy, F1 score, precision, and recall have all evolved. We compare our proposed approach with four robust machine learning algorithms, shown in Table 6. According to the accuracy of the algorithms, we ranked them in Table 6.

Table 6 shows that we found a good accuracy for all of the implemented algorithms. Compared to other algorithms, the random forest classifier's performance is superior because of its accuracy of 89 percent. The accuracy of the KNN classifier is 87.8 percent, which is the second-highest accuracy among the classifiers. The accuracy rate for Impact learning is 86 percent, which is likewise a strong performance rating. The BernoulliNB achieves an accuracy of 85.4 percent, nearly identical to that of the SVM. For all other (Precision, Recall, F1-Score) parameters, the performance is also good for Impact learning.

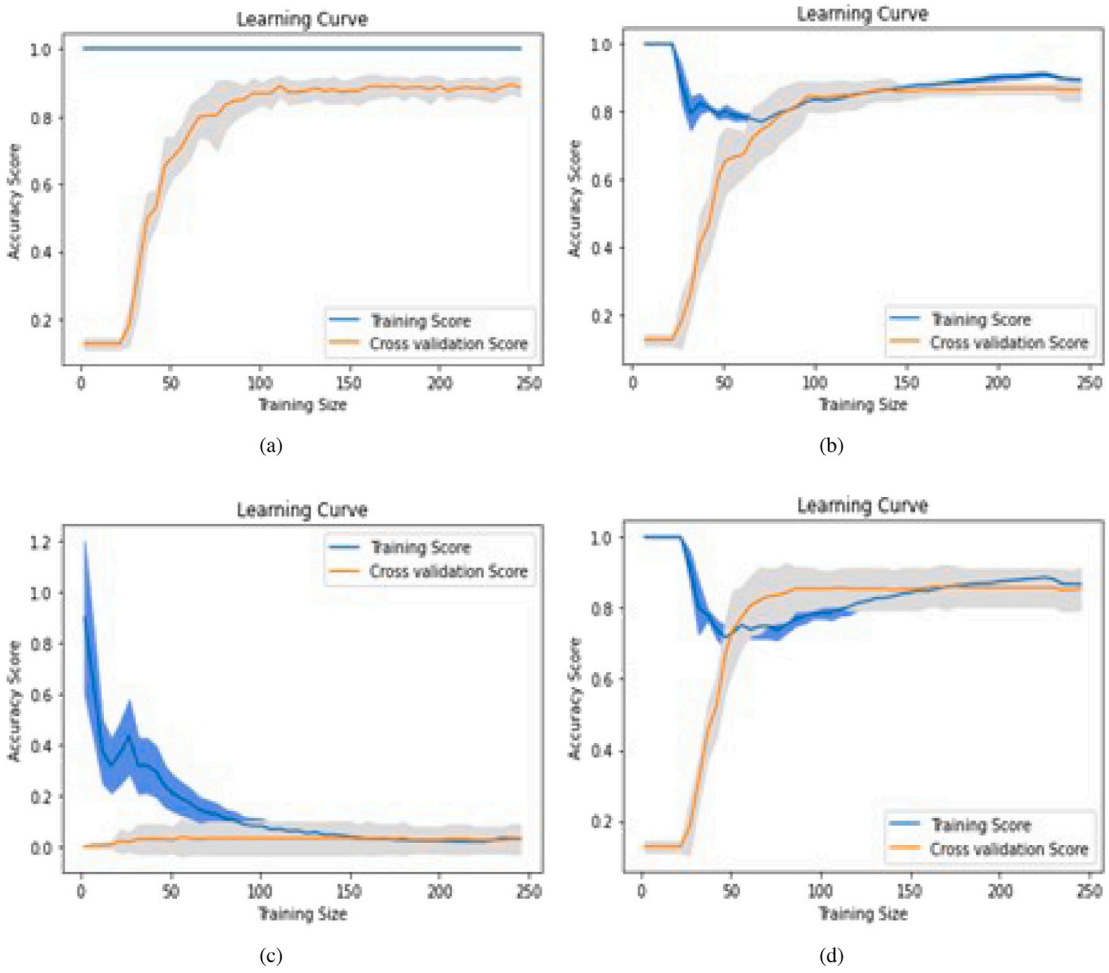

Fig. 12. Learning curve for the SVM, KNN, Random Forest, and BernoulliNb classification algorithms which are used for asthma dataset. (a) Learning curve for Random Forest classifier. (b) Learning curve for KNN classifier. (c) Learning curve for SVM. (d) Learning curve for BernoulliNB classifier.

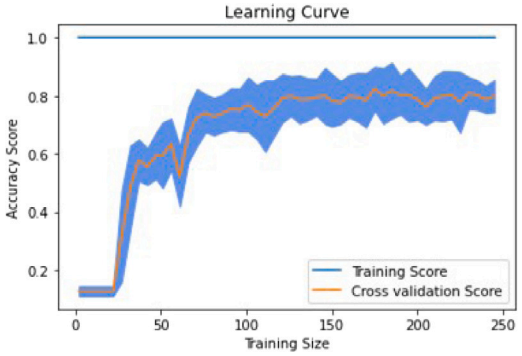

Fig. 13. Learning curve for the proposed algorithm (Impact learning) that is used for diabetes dataset.

Fig. 11 depicts the asthma dataset's receiver operating characteristic (ROC) curves. The higher the slope of the curve, the better the outcome. Each classifier's area under the receiver operating characteristic curve (AUC) has been calculated as a quantitative measure of performance. For the majority of the algorithms, we were able to achieve a minimum satisfactory outcome. Compared to other algorithms, Random Forest performs significantly better, receiving a score of 62.5 percent. This score is 58.5 percent for impact learning, comparable to 58.5 percent for KNN. SVM received a low score, indicating that it was not the best choice for this dataset.

Table 6
Accuracy, Precision, Recall, and F1-score for all implemented classification algorithms and ranked them into ascending order. These performance metrics are for Diabetes Dataset.

Method Name	AC	PR	RC	F1-Score	Rank (Accuracy)
Random Forest	0.89	0.94	0.62	0.89	1.00
KNN	0.87	0.93	0.58	0.60	2.00
Impact Learning	0.86	0.60	0.58	0.59	3.00
Bernoulli-NB	0.85	0.68	0.56	0.58	4.00
SVM	0.85	0.42	0.50	0.46	5.00

AC = Accuracy, PR = Precision, RC = Recall

No matter how much data is fed into it, the Random Forest classifier (Fig. 12(a)) cannot capture the underlying link accurately and has a significant systematic error. Fig. 8 displays the learning curves of the four classification algorithms without the impact of learning. Underfitting can be seen by looking at the KNN's training score (Fig. 12(b)), which continues to fall. In contrast, the cross-validation score (orange line) continues to rise and fall. based on the SVM's findings, Fig. 12(c) illustrates how increasing the number of training instances improves generalization. Finally, the BurnolliNB (Fig. 12(d)), increasing the number of training cases enhances generalization. Fig. 13 illustrates that the training and test scores for the Impact learning model have not yet converged, which suggests that this model could benefit from more training data being added.

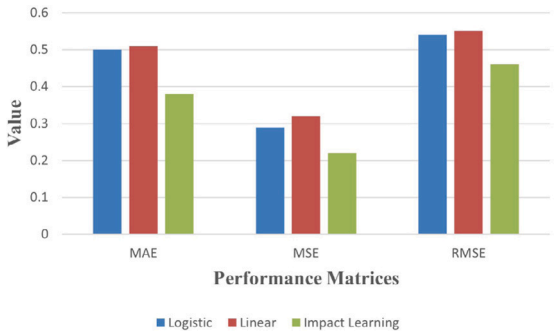

Fig. 14. Performance analysis (MSE, MAE, and RMSE) for Logistic, Linear, and Impact Learning algorithm.

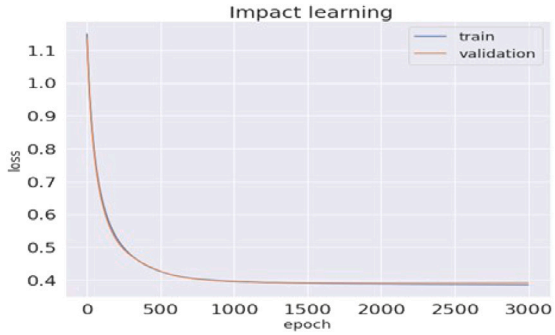

Fig. 15. Training and validation loss curve for Impact Learning.

3.3.3. Performance analysis for regression algorithms (heart disease dataset)

A heart illness dataset from Kaggle with 16 columns was used to improve the regression dataset's performance. The goal of this dataset was to forecast the likelihood of cardiovascular disease during the following ten years. We have considered three factors to analyze this dataset's performance: the MSE, MAE, and RMSE depicted in Fig. 14.

When it comes to improving the performance of our proposed algorithm (Impact Learning), we have benchmarked it against two of the most prominent algorithms available, namely linear regression and logistic regression. Fig. 14 indicates that the output of Impact Learning is superior to the output of two other algorithms for all of the performance matrices. Specifically, the effect learning value for MAE is less than 40%, the MSE value is somewhat higher than 20%, and the RMSE value is almost 50%. We obtained the lowest performance with the linear approach, although it is practically identical to that obtained with logistic regression.

Fig. 15 shows that the training and validation loss curves are shrinking as the number of epochs increases. Fig. 15 indicates that the performance of Impact learning is improving as the number of epochs increases.

4. Conclusion and future work

We have developed a novel machine learning approach for handling regression and classification issues. The system of learning from RNI and the impact of additional aspects such as competition is the primary techniques employed by this strategy. As it gains knowledge from the effect, it can be applied to other aspects of competition, such as the effect of other actors. To improve the performance of this algorithm, we used three real datasets, two of which were used for classification analysis and one of which was used for regression analysis. We compared the performance of our method to that of several other well-known algorithms. When we reached the results of our proposed algorithm to those of different algorithms, we discovered that it produced an excellent outcome. Accuracy, precision, recall, and F1-score are some matrices that we examine while looking at performance evolution. We hope to use impact learning to solve NLP challenges in the future, integrating it with existing machine learning and deep learning algorithms for improved performance. It will be trained using backpropagation and gradient descent rather than the least-squares approach. Additionally, we have developed a method for using this model to evaluate and forecast market value among many competitors in business, economics, and machine learning, among others.

CRediT authorship contribution statement

Nusrat Jahan Prottasha: Conception and design of study, Writing – review & editing. **Saydul Akbar Murad:** Conception and design of study, Acquisition of data, Writing – original draft. **Abu Jafar Md Muzahid:** Conception and design of study, Acquisition of data, Writing – original draft. **Masud Rana:** Analysis and/or interpretation of data, Writing – review & editing. **Md Kowsher:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft. **Apurba Adhikary:** Acquisition of data, Writing – review & editing. **Sujit Biswas:** Analysis and/or interpretation of data, Writing – review & editing. **Anupam Kumar Bairagi:** Conception and design of study, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

References

[1] J. Singh, G. Dhiman, A survey on machine-learning approaches: Theory and their concepts, Mater. Today Proc. (2021) http://dx.doi.org/10.1016/j.matpr.2021.05.335, S2214785321039298.

[2] J. Elith, J.R. Leathwick, T. Hastie, A working guide to boosted regression trees, J. Anim. Ecol. 77 (4) (2008) 802–813, http://dx.doi.org/10.1111/j.1365-2656.2008.01390.x.

[3] S.M.D.A.C. Jayatilake, G.U. Ganegoda, Involvement of machine learning tools in healthcare decision making, J. Healthc. Eng. 2021 (2021) 1–20, http://dx.doi.org/10.1155/2021/6679512.

[4] C. Chen, Storey, Business intelligence and analytics:From big data to big impact, MIS Quart. 36 (4) (2012) 1165, http://dx.doi.org/10.2307/41703503.

[5] H. Al-Sahaf, Y. Bi, Q. Chen, A. Lensen, Y. Mei, Y. Sun, B. Tran, B. Xue, M. Zhang, A survey on evolutionary machine learning, J. R. Soc. New Zealand 49 (2) (2019) 205–228, http://dx.doi.org/10.1080/03036758.2019.1609052.

[6] S. Hallmann, M. Moser, S. Reck, T. Eberl, K. Collaboration, Machine learning for KM3NeT/ORCA, in: Proceedings of 36th International Cosmic Ray Conference — PoS(ICRC2019), Sissa Medialab, 2019, p. 904, http://dx.doi.org/10.22323/1.358.0904.

[7] I.H. Sarker, A.S.M. Kayes, S. Badsha, H. Alqahtani, P. Watters, A. Ng, Cybersecurity data science: an overview from machine learning perspective, J. Big Data 7 (1) (2020) 41, http://dx.doi.org/10.1186/s40537-020-00318-5.

[8] T. Hendrickx, B. Cule, P. Meysman, S. Naulaerts, K. Laukens, B. Goethals, Mining association rules in graphs based on frequent cohesive itemsets, in: Advances in Knowledge Discovery and Data Mining, vol. 9078, Springer International Publishing, 2015, pp. 637–648, http://dx.doi.org/10.1007/978-3-319-18032-8_50.

[9] M.I. Jordan, T.M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science 349 (6245) (2015) 255–260, http://dx.doi.org/10.1126/science.aaa8415.

[10] Y. Ni, D. Aghamirzaie, H. Elmarakeby, E. Collakova, S. Li, R. Grene, L.S. Heath, A machine learning approach to predict gene regulatory networks in seed development in arabidopsis, Front. Plant Sci. 7 (2016) http://dx.doi.org/10.3389/fpls.2016.01936.

[11] S. Raschka, J. Patterson, C. Nolet, Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence, Information 11 (4) (2020) 193.

[12] D. Pernes, K. Fernandes, J.S. Cardoso, Directional support vector machines, Appl. Sci. 9 (4) (2019) 725.

[13] M. Kowsher, I. Hossen, A. Tahabilder, N.J. Prottasha, K. Habib, Z.R.M. Azmi, Support directional shifting vector: A direction based machine learning classifier, Emerg. Sci. J. 5 (5) (2021) 700–713.

[14] P.L. López-Cruz, C. Bielza, P. Larrañaga, Directional naive Bayes classifiers, Pattern Anal. Appl. 18 (2) (2015) 225–246.

[15] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.

[16] A.L. Dumitrescu, Chemicals in Surgical Periodontal Therapy, Springer, 2011.

[17] S. Yeruva, M.S. Varalakshmi, B.P. Gowtham, Y.H. Chandana, P.K. Prasad, Identification of sickle cell anemia using deep neural networks, Emerg. Sci. J. 5 (2) (2021) 200–210.

[18] M. Kowsher, A. Tahabilder, S.A. Murad, Impact-learning: a robust machine learning algorithm, in: Proceedings of the 8th International Conference on Computer and Communications Management, 2020, pp. 9–13.

[19] D. Elizondo, The linear separability problem: Some testing methods, IEEE Trans. Neural Netw. 17 (2) (2006) 330–344.

[20] S. Shalev-Shwartz, Y. Singer, On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms, Mach. Learn. 80 (2) (2010) 141–163.

[21] S. An, F. Boussaid, M. Bennamoun, How can deep rectifier networks achieve linear separability and preserve distances? in: International Conference on Machine Learning, PMLR, 2015, pp. 514–523.

[22] D. Bzdok, M. Krzywinski, N. Altman, Machine learning: supervised methods, Nature Methods 15 (1) (2018) 5.

[23] M. Raissi, P. Perdikaris, G.E. Karniadakis, Machine learning of linear differential equations using Gaussian processes, J. Comput. Phys. 348 (2017) 683–693.

[24] B. Pavlyshenko, Machine learning, linear and bayesian models for logistic regression in failure detection problems, in: 2016 IEEE International Conference on Big Data (Big Data), IEEE, 2016, pp. 2046–2050.

[25] R. Almeida, N.R. Bastos, M.T.T. Monteiro, A fractional Malthusian growth model with variable order using an optimization approach, 2018.

[26] N.F. Zulkiflee, M.S. Rusiman, Heart disease prediction using logistic regression, Enhanced Knowl. Sci. Technol. 1 (2) (2021) 177–184.

[27] S.A. Murad, Z.R.M. Azmi, Z.H. Hakami, N.J. Prottasha, M. Kowsher, Computer-aided system for extending the performance of diabetes analysis and prediction, in: 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), IEEE, 2021, pp. 465–470.

[28] M.S. Islam, M.M. Hasan, S. Abdullah, J.U.M. Akbar, N. Arafat, S.A. Murad, A deep spatio-temporal network for vision-based sexual harassment detection, in: 2021 Emerging Technology in Computing, Communication and Electronics, ETCCE, IEEE, 2021, pp. 1–6.

[29] A. Adhikary, S.A. Murad, M.S. Munir, C.S. Hong, Edge assisted crime prediction and evaluation framework for machine learning algorithms, in: 2022 International Conference on Information Networking, ICOIN, IEEE, 2022, pp. 417–422.

[30] N.J. Prottasha, A.A. Sami, M. Kowsher, S.A. Murad, A.K. Bairagi, M. Masud, M. Baz, Transfer learning for sentiment analysis using BERT based supervised fine-tuning, Sensors 22 (11) (2022) 4157.

[31] M. Muppalaneni, M. Ma, S. Gurumoorthy, Soft Computing and Medical Bioinformatics, Springer, 2019.

Nusrat Jahan Prottasha received the B.Sc. degree in computer science and engineering from Daffodil International University, in 2022. She is currently working with Data Science Platform as a Research Assistant at Stevens Institute of Technology. In 2020, she received the Best Paper Award from the International Conference of Cyber Security and Computer Science. Besides, in recognition of scholarly publication in the reputed indexed journal has been awarded for publishing four articles in Scopus journals from her university.

Saydul Akbar Murad is a researcher who has completed B.Sc. Engr. degree in Information and Communication Engineering (ICE) department from Noakhali Science and Technology University (NSTU), Noakhali, Bangladesh. He is Currently pursuing his master's degree in Faculty of Computing from Universiti Malaysia Pahang, Malaysia. His current research interests are Cloud Computing, Edge Computing, machine learning and Neural Network.

Abu Jafar Md Muzahid received the B.Sc. and the M.Sc. degree in Statistics from the Shahjalal University of Science and Technology, Sylhet, Bangladesh. He also completed his second master's degree in the faculty of Computing at University Malaysia Pahang, Malaysia. His main areas of research interest are Artificial Intelligence, Machine Learning, and green automotive technology.

Masud Raana has completed my B.sc engineering degree from the department of Information and communication engineering at Noakhali Science and Technology University. Currently, he is working on Machine learning, Deep learning, and Cybersecurity.

Md. Kowsher is currently pursuing the Ph.D. degree at the Stevens Institute of Technology. He is also working as an Artificial Intelligence Scientist at Hishab Ltd., and an AI Engineer at NKSoft, USA. He is also working as a Doctorate Research Assistant at the Stevens AI Laboratory, USA. He reviewed many papers in ICCIDM, ICSECS, ICOCSIM, and Visual Computing for Industry, Biomedicine, and Art. He also got the best paper awards from various international conferences, such as ICONCS, IC4ME2, ICCCM, NISS, and ICIET. Apart from that, he was the champion of Robi r-ventures 2.0 and got the National Basis ICT Award, in 2019. In 2021, he got the Scientist of the year award for his excellent research in the field of AI from IEM, India. He received the Provost Fellowship Award for Ph.D. degree.

Apurba Adhikary received his B.Sc and M.Sc Engineering degrees in Electronics and Communication Engineering from Khulna University, Khulna, Bangladesh. He is a Ph.D. Researcher in the Department of Computer Science and Engineering at Kyung Hee University (KHU), South Korea. He has been serving as an Assistant Professor in Information and Communication Engineering Department at Noakhali Science and Technology University (NSTU), Noakhali, Bangladesh since 28 January 2020. In addition, he served as a Lecturer in Information and Communication Engineering Department at Noakhali Science and Technology University (NSTU), Noakhali, Bangladesh from 28 January 2018 to 27 January 2020. His research interests are currently focused on integrated sensing and communication, target-oriented communication and computation, distributed edge intelligence, intelligent networking resource management, artificial intelligence, and machine learning.

Sujit Biswas (M'20) received his M.Sc. degree in Computer Engineering from Northwestern Polytechnical University, China in 2015 and Ph.D degree in Computer Science and Technology from Beijing Institute of Technology, China. Currently, he is working as a Lecturer in University of East London, UK. He has also served as an Assistant Professor with Computer Science and Engineering department, Faridpur Engineering College, Bangladesh. His basic research interest is in IoT, Blockchain, Mobile computing security and privacy, Data driven decision making, etc.

Anupam Kumar Bairagi (S'17-M'18-SM'22) received his Ph.D. degree in Computer Engineering from Kyung Hee University, South Korea and B.Sc. and M.Sc. degree in Computer Science and Engineering from Khulna University (KU), Bangladesh. He is an professor in Computer Science and Engineering discipline, Khulna University, Bangladesh. His research interests include wireless resource management in 5G and beyond, Healthcare, IIoT, cooperative communication, and game theory. He has authored and coauthored around 60 publications including refereed IEEE/ACM journals, and conference papers. He has served as a technical program committee member in different international conferences. He is a senior member of IEEE.