



Predicting the potential toxicity of the metal oxide nanoparticles using machine learning algorithms

Gehad Ismail Sayed^{1,2} · Heba Alshater^{2,3} · Aboul Ella Hassanien^{2,4,5}

Accepted: 20 February 2024 / Published online: 10 July 2024
© The Author(s) 2024

Abstract

Over the years, machine learning (ML) algorithms have proven their ability to make reliable predictions of the toxicity of metal oxide nanoparticles. This paper proposed a predictive ML model of the potential toxicity of metal oxide nanoparticles. A dataset consisting of 79 descriptors including 24 metal oxide nanoparticles (MexOy NPs) and their physicochemical and structural characteristics is adopted. The proposed model comprises of three main phases. The first phase is used to analyze the characteristics of nanoparticles along with their toxicity behavior. In the second phase, the problems associated with the metal oxide nanoparticles dataset are tackled. The first problem namely the class imbalance problem is handled through utilizing synthetic minority over-sampling technique (SMOTE). The second problem namely the outliers is handled through applying a novel feature selection algorithm based on the enhanced binary version of the sine tree-seed algorithm (EBSTSA). The proposed EBSTSA is used to find the relevant features affecting toxicity. The density-based spatial clustering of applications with noise (DBSCAN) is utilized as a tool for identifying outliers in the dataset and for visualizing the impact of the feature selection on the performance of the subsequent classification. Finally, in the third phase, the support vector machine (SVM) supervised machine learning algorithm and k-fold cross-validation method are applied to classify the mode of action of each instance of nanoparticle as toxic or nontoxic. The simulation results showed that the EBSTSA-based feature selection algorithm is reliable and robust across 23 benchmark datasets from the UCI machine learning repository. The results also showed that proposed EBSTSA can effectively find the relevant descriptors for nano-particles. Furthermore, the results demonstrated the efficacy of the proposed ML toxicity prediction model. It is obtained on average 1.02% of error rate, 100% of specificity, 98.87% of sensitivity, and 99.47% of f1-score.

Keywords Cytotoxicity · Metal oxide nanoparticles · Sine tree-seed algorithm · Prediction model

Abbreviations

AI	Artificial Intelligence	SMOTE	Synthetic Minority Over-sampling Technique
ML	Machine Learning	DBSCAN	Density-Based Spatial Clustering of Applications with Noise
MexOy NPs	Metal Oxide Nanoparticles	SVM	Support Vector Machine
		EBSTSA	Enhanced Binary Version of the Sine Tree-Seed Algorithm
		STSA	Sine Tree-Seed Algorithm
		XRD	X-Ray Diffraction
		FT-IR	Fourier Infrared Spectroscopy
		TEM	Transmission Electron Microscopy
		AFM	Atomic Force Microscope

✉ Gehad Ismail Sayed gehad_sayed@cic-cairo.com			
¹ School of Computer Science, Canadian International College (CIC), Cairo, Egypt			
² Scientific Research School of Egypt (SRSEG), Cairo, Egypt			
³ Forensic Medicine and Clinical Toxicology Department, Menoufia University Hospital, Al Minufiyah, Egypt			
⁴ Faculty of Computers and Artificial Intelligence, Cairo University, Giza, Egypt			
⁵ College of Business Administration, Kuwait University, Al Shadadiya, Kuwait			

SEM	Scanning Electron Microscope	mitochondrial and DNA damage, cytochrome C and inflammatory cytokine release and cell death (Khan et al. 2023; Cabral et al. 2023; Li et al. 2014). Metal oxide nanoparticles' detrimental toxicity has been demonstrated by their surface features such as electrical property (bandgap), functional groups, solubility, and so on Zhang et al. (2012). In soil studies, Al ₂ O ₃ , SiO ₂ , and SnO ₂ nanoparticles influenced bacterial and fungal species, and Al ₂ O ₃ , Mn ₃ O ₄ , and SiO ₂ nanoparticles inhibited the growth of bacteria, algae, and protozoa. Furthermore, SnO ₂ nanoparticles decreased cell viability and induced cell membrane damage in bacteria (Chavez-Calderon et al. 2016; McGee et al. 2017). Several studies on animals have shown that nanoparticles can pass through the blood–brain barrier, which protects the brain from toxins in the bloodstream due to their small size. Besides they cause air pollution such as the nano-sized soot particles in diesel exhaust, and in turn when inhaled it passes into the lungs, causing respiratory problems and heart disease (Asmatulu et al. 2022). Surface coating, doping, core-shell structure, and geometric control are currently considered safe design techniques. A variety of coating molecules on various metal oxides were tested to see which were the most successful for surface passivation. The ideal coating molecule may reduce the surface reactivity and/or dissolution of metal oxide, hence reducing their in vitro and in vivo hazard potential (Osmond-McLeod et al. 2013; Davidson et al. 2016; Gambhir et al. 2022).
ATP	Adenosine Triphosphate	
LDH	Lactate Dehydrogenase	
EG	Energy Bandgap	
EV	Energy of the Valence	
EC	Energy of Conduction	
k-FCV	K-fold Cross-Validation	
K-S	Kolmogorov-Smirnov	
LUMO	Lowest Unoccupied Molecular Orbital	
QBABC	Quantum Binary Artificial Bee Colony	
BPSO	Binary Particle Swarm Optimization	
BSSA	Binary Salp Swarm Algorithm	
BCSO	Binary Chicken Swarm Optimization	
BBOA	Binary Butterfly Optimization Algorithm	
BALO	Binary Ant Lion Optimization	
BASO	Binary Atom Search Optimization	
RBF	Radial Basis Function	
Receiver Operating Characteristic	ROC	

1 Introduction

Nanotechnology and nanoscience have gotten a lot of attention since they have so many applications in so many different sectors. The nanoparticle is a microscopic particle measured on the nanoscale. It is characterized by its large surface area and large surface energy and its reactivity so; it can be used as an excellent catalyst (Divya and Muthuvinothini 2015). Metallic oxides perform a well-known region of research owing to their distinguished electronic, optical, and physicochemical features. Metal oxide nanoparticles are also manufactured and synthesized for use in electronics, sensors, catalysis, energy storage devices, drug delivery, medicine, and magnetic resonance imaging (Hameed et al. 2019; Parashar et al. 2020). Several studies investigated the unique physicochemical properties of metal oxide nanoparticles such as chemical stability, antibacterial capability, thermal, and electrical, and mechanical properties (Cai et al. 2019). However, these numerous applications of metal oxide nanoparticles, the expanding generation as well as wide utilization of metal oxide nanoparticles have driven significant concerns about their potentially hazardous impacts on both humans and the environment. Some of them have been investigated to produce reactive oxygen species generation,

Owing to the much longer time and cost of experiments in vivo and in vitro to predict the toxicity of metal oxide nanoparticles. Furthermore, the high cost of characterizing these metal oxide nanoparticles using various techniques like X-ray diffraction (XRD), visible ultraviolet spectroscopy (UV visible), Fourier infrared spectroscopy (FT-IR), Scanning Electron Microscope (SEM), Atomic Force Microscope (AFM), and transmission electron microscopy (TEM). The features of nanoparticles, such as shape, size, surface area, and so on, must be characterized since they determine their toxicity. Therefore, the goal of this paper is to use computational methods to predict the mode of action and toxicity (cell viability) of metal oxide nanoparticles on the luminometric adenosine triphosphate (ATP) assay and the colorimetric lactate dehydrogenase (LDH) assay, employing various structural, physicochemical, and assay-related descriptors, both of which assess irreversible cell membrane damage (Banaye Yazdipour et al. 2023).

Machine learning (ML) is a subfield of artificial intelligence (AI) that focuses on developing computer systems that can accomplish certain tasks based on the data they ingest without the need for explicit instructions (Gombolay et al. 2023). It has played a key role in a variety of domains, such as pattern recognition, computer vision, data mining, and natural language processing. Machine learning algorithms are the engines that power ML. They can provide potential

solutions in all these domains and more. They have been widely adopted and proved their efficiency in many complex and massive fields such as biology, agriculture, medicine, and so on Ali et al. (2016), Gamal et al. (2017), and Sayed and Hassanien (2023). They can effectively provide possible solutions to mine the hidden information in the data. ML algorithms can be classified into main categories; supervised and unsupervised learning algorithms. In this paper, we provided a predictive ML model of the potential toxicity of metal oxide nanoparticles. Moreover, a deep analysis of the properties of nanoparticles is provided. The proposed model is comprised of three main phases. The first phase is used to analyze the characteristics of nanoparticles along with their toxicity behavior. Additionally, this phase introduces the problems associated with the used dataset. The second phase aims to solve the problems associated with the dataset, where a novel feature selection algorithm based on the sine tree-seed algorithm (STSA) is proposed, where the v-shape transfer function is used to convert the position of each individual from real value to binary values. The proposed STSA-based feature selection algorithm is applied to the metal oxide nanoparticles dataset to find the best relevant influential attributes. To the best of our knowledge is the first time to propose a binary version of STSA. It is also the first time it has been applied to the feature selection optimization problem. Finally, in the third phase, the supervised machine learning algorithm support vector machine (SVM) is utilized to classify the mechanism of action of each instance of nanoparticle as dangerous or harmless. The following are the primary contributions of this paper:

- A model for predicting the potential toxicity of metal oxide nanoparticles is proposed
- An enhanced binary version of STSA based on the V-shape transfer function is proposed.
- The proposed EBSTSA-based feature selection algorithm is applied to select the relevant descriptors (physicochemical, atomistic, and molecular) for nanoparticles.
- The proposed model can be utilized to improve risk assessments and informed decision-making in industries and biological applications that use metal oxide nanoparticles.

The rest of the paper is structured as follows. Section 2 contains an in-depth overview of the employed metal oxide nanoparticles dataset. Section 3 describes the proposed toxicity detection model for metal oxide nanoparticles. The simulation results and discussions are presented in Sect. 4. Section 5 concludes with conclusions and future work.

2 Metal oxide nanoparticles dataset characteristics

To predict the toxicity of metal oxide nanoparticles, 24 distinct metal oxide nanoparticles (17 commercial and 7 made in-house) are collected in this paper. They are described by physicochemical and structural characterizations such as (specific surface area, NP core size, hydrodynamic size, ζ potential, total surface area, metal dissolution, point of zero *zeta* potential, and crystal structure). It also comprised the standard enthalpy of formation ($E\Delta H$), the energy of the valence (EV) and conduction (EC) bands, metal electronegativity (cation), the energy bandgap (EG), MexOy absolute electronegativity (oxide). Cell lines were included because they are routinely used to assess the cytotoxicity of manufactured nanoparticles without regard to a specific method of action. Zhang et al. (2012) demonstrated the ability to employ Ec levels to determine the toxicological potential of metal oxide nanoparticles at the cellular and whole animal levels. Toxicological evaluations were carried out in vitro using single and multi-parameter toxicity assays. Because they are used to test designed nanoparticle cytotoxicity without regard to a specific mode of action, single-parameter LDH and ATP assays in murine myeloid (RAW264.7) and human bronchial epithelial (BEAS-2B) cell lines were included. A single-parameter toxicity study's findings were contrasted with those of an internal multi-parameter high-throughput screening assay that evaluated oxidative stress using five parameters (Zhang et al. 2012). The findings supported the idea that the same seven nanoparticles may be more hazardous when examined for general toxicity and when compared for how they influence oxidative stress in multi-parameter experiments. The results demonstrated a substantial link between the multi-parameter and single-parameter responses. The Vivo mouse model was then used to evaluate the emergence of acute neutrophilic inflammation and cytokine responses in the lungs of C57 BL/6 mice.

The single-parameter (LDH and ATP) studies in the aforementioned RAW 264.7 and BEAS-2B cell lines are used in this paper as a subset of the data for modeling. The kind of assay (LDH and ATP) was included as an additional parameter, and many patterns revealed a strong correlation between the type of assay and the outcomes of the cytotoxicity test (Gambhir et al. 2022; Furxhi 2022; Labouta et al. 2019). Six physicochemical, six molecular, and three assay-related descriptors (NP exposure dosage, assay type, and cell species) were employed in the paper out of the 15 descriptors (independent variables, Zhang et al. 2012). The investigation employed the physicochemical descriptors specific surface area, chemical formula, core size, hydrodynamic size, total surface area, and *zeta* potential. The percentage of cell viability was the biological endpoint (dependent variable) 24 h after treatment. The attributes of the metal oxide nanoparti-

cles employed in the paper are displayed in Table 1 along with their corresponding annotation.

3 The proposed toxicity detection of the metal oxide nanoparticles model

The proposed ML toxicity detection of the metal oxide nanoparticles model is comprised of three main phases. The first phase of the proposed model is called data analysis. Data analysis is the method of giving order, structure, and meaning to a large amount of data. In this phase, the captured metal oxide nanoparticle data from quantitative and qualitative evaluation is presented, interpreted, and described systematically. The first step in data analysis is checking for the missing data. The next step is the statistical analysis. In this step, the mean, the standard deviation, and the class distribution are used. The final step in data analysis is to figure out whether there are outliers in the data or not. After discovering the problems of the used dataset, the next phase is to introduce a solution for each problem of the provided metal oxide nanoparticles dataset. This phase includes handling the imbalanced dataset, where the class distribution is improved. Moreover, an algorithm is introduced to remove the outliers detected in the captured dataset. In this step, a novel feature selection algorithm is proposed to reduce the size of the feature set and remove the outliers. Finally, in the classification and cross-validation phase, SVM with Gaussian kernel function is used to predict the mode of action (Toxic or non-toxic). Moreover, the k-fold cross-validation algorithm is used to prove the reliability of the proposed model. Figure 1 comprises the three phases of the proposed model.

3.1 Metal oxide nanoparticles data analysis phase

Data analysis is a very essential task. It is used to present the provided dataset in an interpretable and intelligible form by identifying the relations and trends in it. The aim of this paper is to investigate if the cytotoxicity of metal oxide nanoparticles can be predicted by a combination of physicochemical, molecular, and whole nanoparticle computational descriptors. According to the dataset description section, the adopted dataset from Zhang et al. (2012) comprises 80 attributes with various physicochemical and structural characteristics. All the metal oxide nanoparticle attributes have no missing values.

3.1.1 Statistical analysis

In this part, three well-known evaluation measurements are used to characterize the data. These measurements are mean, standard deviation, and class distribution (Wan et al. 2014).

The mean is one of the useful measurements to discover the pattern of the data and discover the outliers associated with it. The mathematical definition of the mean \bar{Z} is shown in Eq. (1). Standard deviation is another statistical measurement used to measure the differences between each record of nanoparticle from the mean. It is mathematically defined in Eq. (2). Similar to the mean, outliers can significantly affect the standard deviation. The big values of standard deviation mean that lots of nanoparticles are meaningless and can remarkably affect the performance of an algorithm.

$$\bar{Z} = \sum_{j=1}^S Z_j \quad (1)$$

where S is the total number of samples in the dataset and Z_j is the j -th data point.

$$\sigma = \sqrt{\frac{\sum_{j=1}^S Z_j - \bar{Z}}{S}} \quad (2)$$

where \bar{Z} is the averages of Z .

Understanding and tackling the class imbalance problem in machine learning depends on calculating the class distribution. Insightful information about the relative frequencies of various classes within a dataset is provided by this distribution. Models tend to favor the majority class when certain classes are underrepresented, which causes skewed predictions and decreased accuracy for minority classes (Elreedy et al. 2023). The class distribution's mathematical definition is presented in Eq. (3).

$$P(Z = k) = \frac{Z_k}{S} \quad (3)$$

where $P(Z = k)$ is the probability of the data points belonging to class k and Z_k is the number of data points in the k -th class.

3.1.2 Outliers detection

Outlier detection, also known as anomaly detection, is the challenge of detecting unexpected patterns in data that do not match to predicted normal behavior. Outlier detection is considered a critical task for identifying anomalies in metal oxide nanoparticle data. Excluding these outliers from the training dataset can improve classification accuracy and consequently decision-making, according to Ijaz et al. (2018). Outlier detection strategies are divided into two types: distance-based approaches and statistical-based approaches. To determine whether the data fits within the model, statistical tools are applied. The distance between each record in the data is calculated in the distance-based

Table 1 Attributes definition

Att.	Meaning	Att.	Meaning
A1	Material type	A41	O 4rdN avg core (eV)
A2	Core size (nm)	A42	O 4rdN avg shell (eV)
A3	Hydro size (nm)	A43	NP diameter (A)
A4	Surface charge (mV)	A44	NP surface area(A ²)
A5	Surface area (m ² /g)	A45	NP volume(A ³)
A6	Method surface area	A46	Lattice energy of NP (eV)
A7	Hsf (eV)	A47	Relative lattice energy of NP to bulk material (E L bulk-E L NP) (eV)
A8	Ec (eV)	A48	Lattice energy of NP / d NP (eV/A)
A9	Ev (eV)	A49	Lattice energy of NP /S NP (eV/A ²)
A10	MeO (eV)	A50	Lattice energy of NP /V NP (eV/A ³)
A11	Assay	A51	Force vector length avg all (eV)
A12	Cell name	A52	Force vector length avg core (eV)
A13	Cell species	A53	Force vector length avg shell (eV)
A14	Cell origin	A54	Al force vector length 4rdN avg all (eV)
A15	Cell type	A55	Al force vector length 4rdN avg core (eV)
A16	Exposure dose (ug/mL)	A56	Al force vector length 4rdN avg shell (eV)
A17	log(n atoms all)	A57	O force vector length avg all (eV)
A18	log(n atoms core)	A58	O force vector length avg core (eV)
A19	log(n atoms shell)	A59	O force vector length avg shell (eV)
A20	log(n Al atoms all)	A60	Force vector surface 2 component avg all (eV)
A21	log(n Al atoms core)	A61	Force vector surface 2 component avg core (eV)
A22	log(n Al atoms shell)	A62	Force vector surface 2 component avg shell (eV)
A23	log(n O atoms all)	A63	Al force vector surface 2 component 4rdN avg all (eV)
A24	log(n O atoms core)	A64	Al force vector surface 2 component 4rdN avg core (eV)
A25	log(n O atoms shell)	A65	Al force vector surface 2 component 4rdN avg shell (eV)
A26	peng avg all (eV)	A66	O force vector surface 2 component avg all (eV)
A27	peng avg core (eV)	A67	O force vector surface 2 component avg core (eV)
A28	peng avg shell (eV)	A68	O force vector surface 2 component avg shell (eV)
A29	Al peng avg all (eV)	A69	Force vector surface tangent component avg all (eV)
A30	Al peng avg core (eV)	A70	Force vector surface tangent component avg core (eV)
A31	Al peng avg shell (eV)	A71	Force vector surface tangent component avg shell (eV)
A32	O peng avg all (eV)	A72	Al force vector surface tangent component 4rdN avg all (eV)
A33	O peng avg core (eV)	A73	Al force vector surface tangent component 4rdN avg core (eV)
A34	O peng avg shell (eV)	A74	Al force vector surface tangent component 4rdN avg shell (eV)
A35	4rdN avg all (eV)	A75	O force vector surface tangent component avg all (eV)
A36	4rdN avg core (eV)	A76	O force vector surface tangent component avg core (eV)
A37	4rdN avg shell (eV)	A77	O force vector surface tangent component avg shell (eV)
A38	Al 4rdN avg all (eV)	A78	Viability (%)
A39	Al 4rdN avg core (eV)	A79	lity
A40	Al 4rdN avg shell (eV)		

approach, and the record with a distance larger than a certain distance is referred to as an anomaly or outlier.

Clustering is a distance-based approach used to find outliers. Its basic idea is to identify dense regions as representing normal situations, whereas anomalies, or outliers, manifest as data points that do not adhere to any established cluster

and tend to form smaller, isolated groups. The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is employed in this paper. The minimal number of points and neighborhood distance epsilon are two essential DBSCAN parameters. DBSCAN's primary benefit is that it separates points into border, core, and outlier points. Borders are core

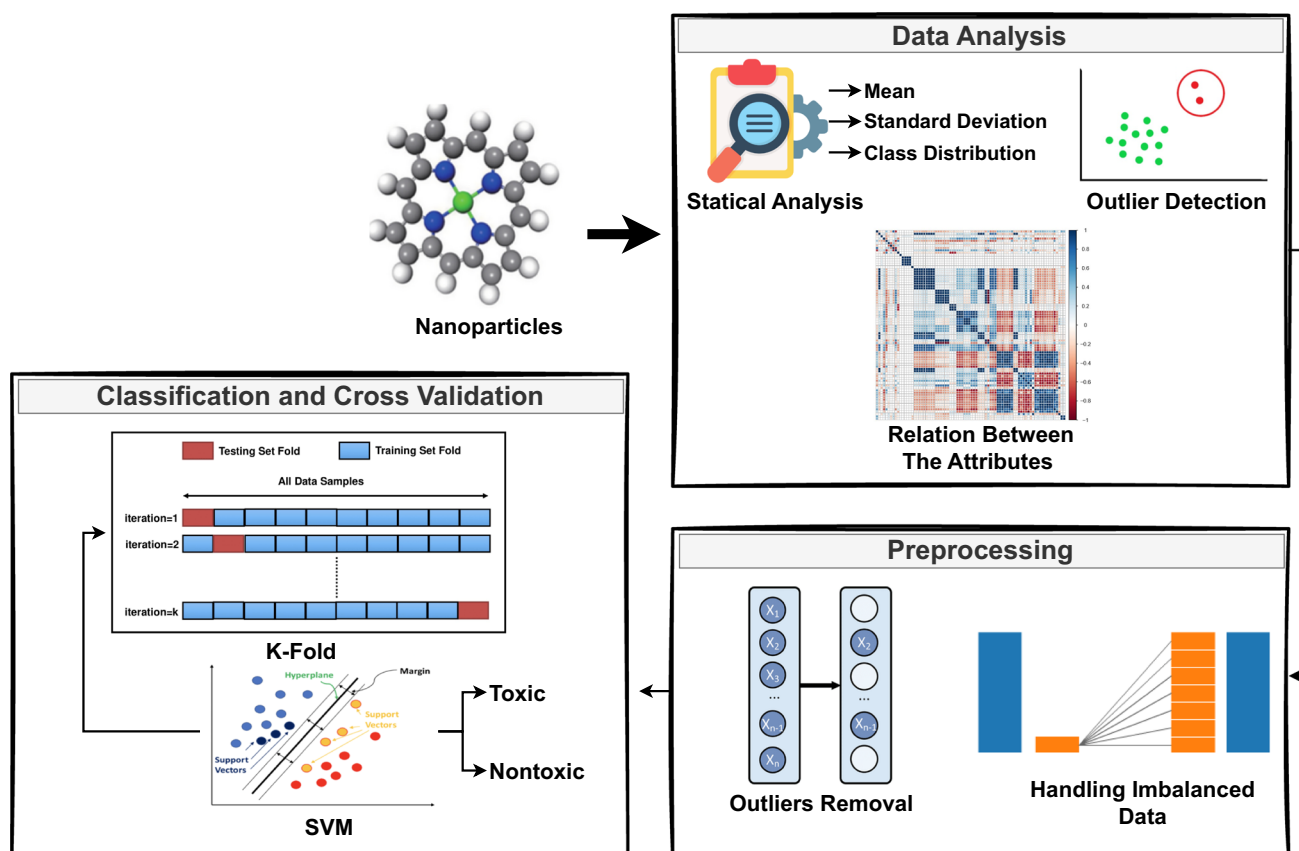


Fig. 1 The proposed ML toxicity detection of metal oxide nanoparticles model block diagram

points' neighbors. Core points are those that have at least the required number of neighbors inside an epsilon-sized neighborhood. The points that do not fit into any clusters are considered outliers. DBSCAN can identify outliers that have exceptionally high and low values when compared to static approaches like mean and standard deviation. The statistical approaches failed to detect the outlier below and above a specified threshold (Celik et al. 2011). However, outliers can occasionally be any data that does not appear frequently, not just those at extremes. These outliers as well as those with exceptionally high and low values can be efficiently found using DBSCAN.

DBSCAN is crucial to this paper because it is utilized as a tool for identifying outliers in the dataset. It works by separating core points, border points, and outliers based on how densely the data points are distributed throughout the feature space. By de-emphasizing less important features, feature selection algorithms can indirectly reduce their impact on the model. Because less significant features are frequently more affected by outliers, the impact of outliers can be somewhat reduced by removing the irrelevant features. DBSCAN is primarily used in this paper to visualize the impact of feature selection on the performance of the subsequent classification. It assists in identifying cases that might differ significantly

from the typical data distribution and offers insightful information about the robustness and sensitivity of the selected features

3.1.3 Relation between the attributes

Identifying the relationship between the attributes is considered an important task in data analysis. Correlation analysis is one of the used techniques to detect the relationship in the data. These relationships help us to get a better understanding of the data. The strength of a relationship between two attributes is mainly based on three criteria. These criteria are direction, dispersion strength, and form. However, the causality between two attributes (variables) cannot be determined if these two attributes are independent. Numerically, the correlation coefficient is used to determine the relationship between attributes. The positive sign of the coefficient indicates the two attributes are positively correlated, while the sign indicates a negative correlation. If the value is zero that means no correlation. The strongest linear correlation is indicated by a correlation coefficient of -1 or $+1$. The Pearson correlation coefficient, abbreviated as ' r ' is the most often used metric of correlation. A two continuous variables' linear relationship is measured by the Pearson correlation

coefficient. The formula for calculating the Pearson correlation coefficient (r) between two variables Z and Y is as follows:

$$r = \frac{\sum_{j=1}^n (Z_j - \bar{Z})(Y_j - \bar{Y})}{\sqrt{\sum_{j=1}^n (Z_j - \bar{Z})^2 \sum_{j=1}^n (Y_j - \bar{Y})^2}} \quad (4)$$

where Z_j and Y_j are the j th data point of the two variables Z and Y , respectively. \bar{Z} and \bar{Y} are the averages of Z and Y , respectively. n is the number of data points in the sample.

3.2 Metal oxide nanoparticles preprocessing phase

In this phase, possible solutions for the problems discovered from the previous phase are proposed. These problems are imbalanced datasets and outliers due to irrelevant features. The first problem is handled by the Synthetic Minority Over-sampling Technique (SMOTE). By de-emphasizing less important features, feature selection algorithms can indirectly reduce their impact on the model. Because less significant features are frequently more affected by outliers, the impact of outliers can be somewhat reduced by removing the irrelevant features. Thus, for the second problem, a novel enhanced binary sine tree-seed-based feature selection algorithm is proposed to select the best relevant features.

3.2.1 Handling imbalanced data

As shown in Fig. 7, there is an unequal distribution of classes. There is 51 for toxic nanoparticles and 443 for nontoxic nanoparticle. An imbalanced dataset is what this issue is known as when there are more occurrences of the first class than the other class. The class with lower instances is called the minority class. This class has higher penalty errors. Additionally, it may result in improper classification results. In order to achieve overall accuracy, supervised machine learning algorithms like SVM, neural networks, and decision trees are frequently employed. This obtained accuracy may provide a privilege to the majority class (Tharwat et al. 2017). Thus, there is a need to address the class imbalance problem. In this paper, one of the re-sampling methods to reduce the sparseness of the minority class and to achieve more balanced samples in each class is employed. SMOTE is one of the re-sampling techniques. It is used to overcome the problem of a few instances of the minority class in the metal oxide nanoparticles dataset. The dataset is then divided into a training set and a testing set, with the training set being utilized only for SMOTE analysis. To successfully balance the class distribution within the training set, this method requires creating synthetic samples for the minority class. This step is critical for mitigating class imbalance difficul-

ties and improving the model's ability to reliably classify instances from the minority class during training.

3.2.2 Outliers removal

The data need to be processed further in order to decrease the amount of outliers. To reduce their dimensionality and improve the accuracy of their clustering and classification, the data in this section are submitted to a feature selection algorithm. In order to increase classification accuracy, feature selection is a crucial step. The main objectives of using a feature selection algorithm are to improve classification accuracy, reduce the feature set, decrease the number of outliers, save storage space, speed up computation, and improve data comprehension. The two main categories of feature selection algorithms are filter-based and wrapper-based algorithms. In filter-based feature selection algorithms, a subset of features is selected based on the "intrinsic" importance of the features as determined by univariate statistics. In contrast, the effectiveness of wrapper-based feature selection algorithms is evaluated using the classifier performance. Wrapper-based feature selection algorithms thus successfully address the real problem.

One of the most recent swarm intelligence algorithms is employed in this paper to determine the ideal feature subset for building the most accurate classifier model. A V-shaped transfer function is offered as an enhanced binary version of the sine tree-seed algorithm. The likelihood of changing each dimension of a person's location from one to zero, and vice versa, is known as the transfer function. In this paper, the position of each person is changed from a real number to either zero or one by the V-shaped transfer function. Equation (5) defines the mathematical illustration of the applied V-shaped transfer function.

$$V = T(Y_{i,j}) = \left\lfloor \frac{Y_{i,j}}{\sqrt{1 + (Y_{i,j})^2}} \right\rfloor \quad (5)$$

where $Y_{i,j}$ is the position of i th tree at j th dimension.

To utilize the traditional STSA for the feature selection problem, an enhanced binary version of STSA is proposed, namely EBSTSA. In the proposed EBSTSA, two main modifications are added. These modifications are updating the positions of seeds and the fitness function. The position of seeds is changed from continuous to binary discrete values, where each individual is either zero or one. The zero value means that this feature will not be selected, while the one value means that this feature will be considered. The size of each seed position equals the size of the feature subset. The V-shaped transfer function is employed to convert each seed position to discrete binary space. Equation (6) shows the updating position of the seed to the V-shaped transfer function.

$$ST_{i,j} = \begin{cases} 1 & \text{if } (T(Y_{i,j})) \geq R \\ 0 & \text{Otherwise} \end{cases} \quad (6)$$

where R is random number in $[0, 1]$.

During the optimization process, each seed position is assessed using a predetermined fitness function. Two evaluation factors are taken into account in this paper. Maximum classification accuracy and a low number of the selected features serve as the evaluation criterion. The goodness of the seed position, which maximizes the classification accuracy while minimizing the feature subset, is assessed using the fitness function. Equation (7) illustrates the employed fitness function. The initial value of α in this paper is set to 0.01, and knn , a supervised classifier, is used to calculate CA with three neighbors and a “Euclidean” distance metric. Algorithm (1) shows the pseudocode of the proposed EBSTSA-based feature selection algorithm. In this algorithm, M is the population size, T_{max} is the maximum number of iterations, L and U are the lower and the upper boundary of the search space, $r_{i,j}$ is a random number generated in $[0, 1]$ for each j th dimension and i th iteration, CT is the control parameter, F_b is the best obtained fitness value so far, $ST_{i,j}$ is the position of seed of i th tree in j th dimension, $Y_{rand,j}$ is the position of random selected tree j th dimension, and $ST_{i,j}^*$ is the best position of the seed.

$$Fs = (1 - \alpha) \times CA + \alpha \times \left(\frac{L}{D}\right) \quad (7)$$

$$L = \sum_{j=0}^D ST_j \quad (8)$$

where L is the length of the selected features, CA is the classification accuracy, D is the dimension length, and α is the weight parameter weight.

Input: $M, D, T_{max}, L_j, U_j, CT$

Output: F_b and $ST_{i,j}^*$

Initialize randomly trees positions $Y_{i,j}$ using

$Y_{i,j} = L_j + r_{i,j} \times (U_j - L_j)$;

Initialize $\varepsilon = 2 \times (1 - \frac{M}{D \times 1000})$;

Set $i=0$;

for $i < T_{max}$ **do**

Evaluate each tree position using Eq. (7);

Order the fitness values from best to worst and set the initial best fitness F_s with its corresponding position;

Choose the number of seeds produced for each tree (N_s)

$N_s = L + (U - L) \times \cos(0.5 * \frac{M}{D \times 1000} * \pi) + 1$;

if $Rand < 0.5$ **then**

Update the seed position using

$r \times Y_{rand,j} + (1 - r) \times F_b, rand < 0.5 * CT$;

else

if $Rand \geq CT$ **then**

Update the seed position using

$ST_{i,j} = r_{i,j} \times Y_{i,j} + \varepsilon(Y_{rand,j} - r_{i,j} \times Y_{i,j}) \times (\sin(\pi \times \cos(r_{i,j}))), rand \geq CT$;

else

Update the seed position using

$ST_{i,j} = Y_{i,j} + \varepsilon(F_b - r_{i,j} \times Y_{i,j}) \times (\sin(\pi \times \cos(r_{i,j}))), rand < CT$;

end

end

end

end

Update the seed position using Eq. (6) ;

if $F_s(ST_{i,j}) < F_b$ **then**

Update F_b and $ST_{i,j}^*$;

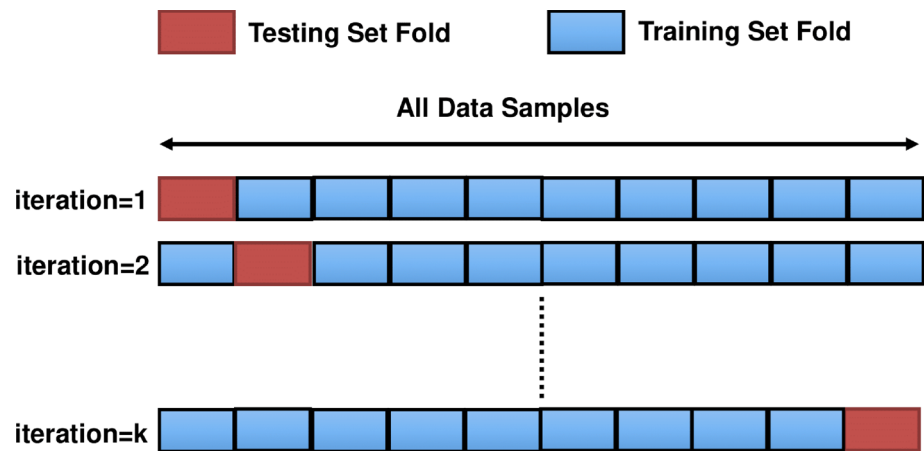
end

end

Print F_b and $ST_{i,j}^*$;

Algorithm 1: The Pseudocode of the Proposed EBSTSA

Fig. 2 Explanation of k -fold cross-validation method



3.3 Classification and cross-validation

The support vector machine (SVM) with k -fold cross-validation (k -FCV) method is employed in this experiment, where the experiment is repeated k times with the citation (Peco Chacon and García Márquez 2023). The entire data samples are divided into roughly k equal-sized subgroups in the k -FCV. One subset is used for training each k times, while another $k - 1$ subset is utilized for testing each k times. The average error rate is determined to produce a single estimation after the iteration reaches k . The value of k in this paper is set to 3. Figure 2 illustrates the k -fold cross-validation method that was employed.

4 Results and analysis

Three key experiments are carried out in this section. Analyzing the key characteristics of the metal oxide nanoparticles dataset is the goal of the first experiment in Sect. 4.1. Additionally, it aims to discover the main problems associated with this dataset. This experiment is divided into three parts. The first sub-experiment attempts to assess the statistical properties of the adopted dataset. The second sub-experiment is to determine whether there are any anomalies (outliers) in the adopted metal oxide nanoparticles dataset. Finally, in the third sub-experiment, the relation between attributes is detected and analyzed for better data understanding. The second experiment in Sect. 4.2 aims to tackle the problems associated with the metal oxide nanoparticle dataset. This experiment is divided into two sub-experiments. The first sub-experiment aims to overcome the imbalanced dataset problem. The second sub-experiment aims to remove the outliers. In this sub-experiment, the performance of the proposed enhanced binary version of the based feature selection algorithm is evaluated. Additionally, the performance of the proposed EBSTSA is compared with other well-known and recent binary feature selection algorithms. Finally, the over-

all performance of the proposed toxicity detection of the metal oxide nanoparticles model is tested in the third main experiment using the support vector machine (SVM) and K -fold cross-validation approach. It should be noted that all experiments are performed on the same PC with the same configuration parameters in order to obtain an unbiased comparison of all conducted experiments. The detailed configuration settings are as follows: Intel Core i3 processor; Windows 7 operating system; 2 GB RAM; and Matlab R2013a.

4.1 Metal oxide nanoparticles data analysis results

In this experiment, several measurements are adopted to evaluate the main characteristics of the metal oxide nanoparticle dataset. These measurements are mean, standard deviation, class distribution, and correlation values. Additionally, a visualization of the DBSCAN clustering results is conducted.

4.1.1 Statistical analysis results

Figure 3 shows the mean and the standard deviation of meta-oxide nanoparticle data. As can be seen, the standard deviation is very high compared with the mean. Thus, it is a great indicator of outliers. Additionally, as the standard deviation exceeds the mean, the distribution of data can be assumed to be non-Gaussian. With a significance level of 0.05, the Kolmogorov-Smirnov (K-S) test was executed under the assumption that the dataset had a normal distribution. The K-S test was conducted using Python with the Pandas, Matplotlib, and Scipy libraries. The test yielded a K-S statistic of roughly 0.9612 and a corresponding p-value of roughly 0. The null hypothesis is rejected since the p-value is smaller than the selected significance level. This is strong evidence that the data does not have a normal distribution. The K-S statistic demonstrates that the actual cumulative distribution function of the data and the theoretical cumulative distribution function of a normal distribution differ signif-

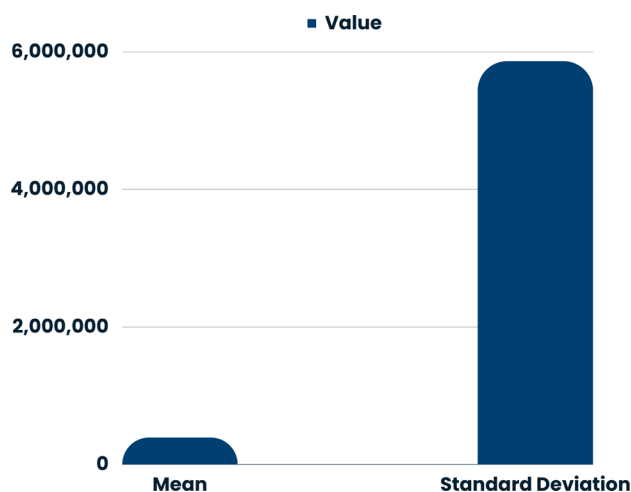


Fig. 3 Mean and standard deviation of the dataset

icantly. According to the K-S test results and the normal distribution assumption with a significance level of 0.05, the data are not likely to be normally distributed. In metal oxide nanoparticle data, the number of toxic nanoparticle records is 51 and the number of non-toxic nanoparticle records is 443. As can be seen, the number of toxic nanoparticle records is smaller than the number of non-toxic nanoparticle records, which indicates that the two classes are not well separable. This problem of unequal distribution of classes is called an imbalanced dataset. The imbalanced dataset can negatively affect classification performance. Next, the commonly used methods to tackle imbalanced classes are employed.

4.1.2 Outliers detection results

Figure 4 shows the clustering result of DBSCAN using squared Euclidean distance metric with the minimum number of points equal to 3 and the neighborhood distance epsilon equal to 1.5. As can be seen, there are many outliers exist in the dataset. The number of outliers is 270. Thus, this is an indication of the existence of outliers in the metal oxide nanoparticles dataset, which are highly required to be removed. An additional experiment in Fig. 5 was carried out to evaluate the performance of the proposed EBSTSA feature selection method in outlier detection. The purpose of this experiment was to compare the number of outliers found before and after the feature selection method. The Z-scoring method was employed with the outlier threshold greater than 3 or less than -3 (Number of standard deviations away from the mean). Each data point in the entire 494×79 matrix was evaluated for outliers using the z-scoring method prior to feature selection. Following that, the proposed EBSTSA-based feature selection algorithm and outlier detection method were applied to the dataset. As can be seen from the findings, the number of outliers decreased after the proposed EBSTSA

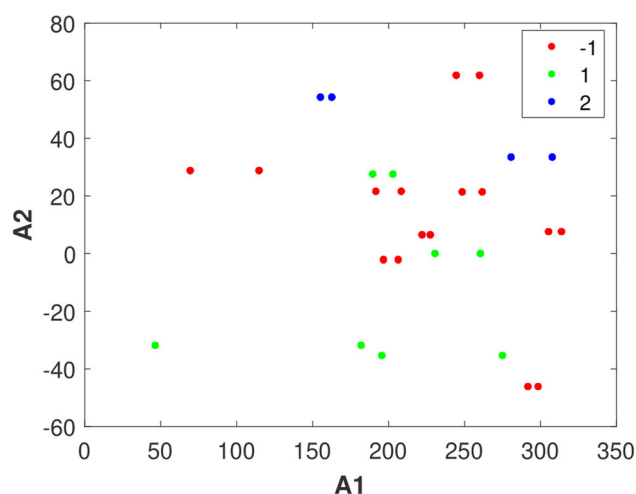


Fig. 4 DBSCAN result using squared Euclidean Distance Metric

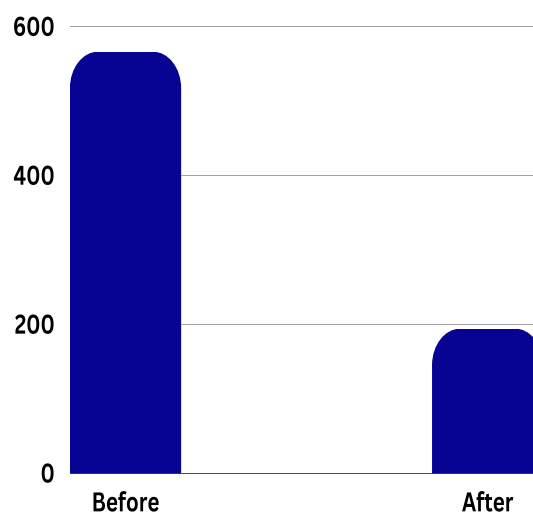


Fig. 5 The number of outliers based on using the Z-score method before and after applying the proposed EBSTSA

was applied. This finding sheds light on the algorithm's impact on outlier identification performance, implying that the feature selection algorithm reduced the amount of data points labeled as outliers.

4.1.3 Relation between attributes results

Figure 6 shows samples of the correlation between the first four attributes. It compares the correlation between material type, core size, hydro size, and surface charge concerning the rest of the 79 attributes. As can be seen, there are many correlation coefficients greater than 0.5. In other words, many redundant attributes can be removed.

According to the first phase of the proposed model, there is a good balance between the physicochemical, molecular/atomic descriptors, and assay-related. More specifically, the hydrodynamic diameter and the size of the original

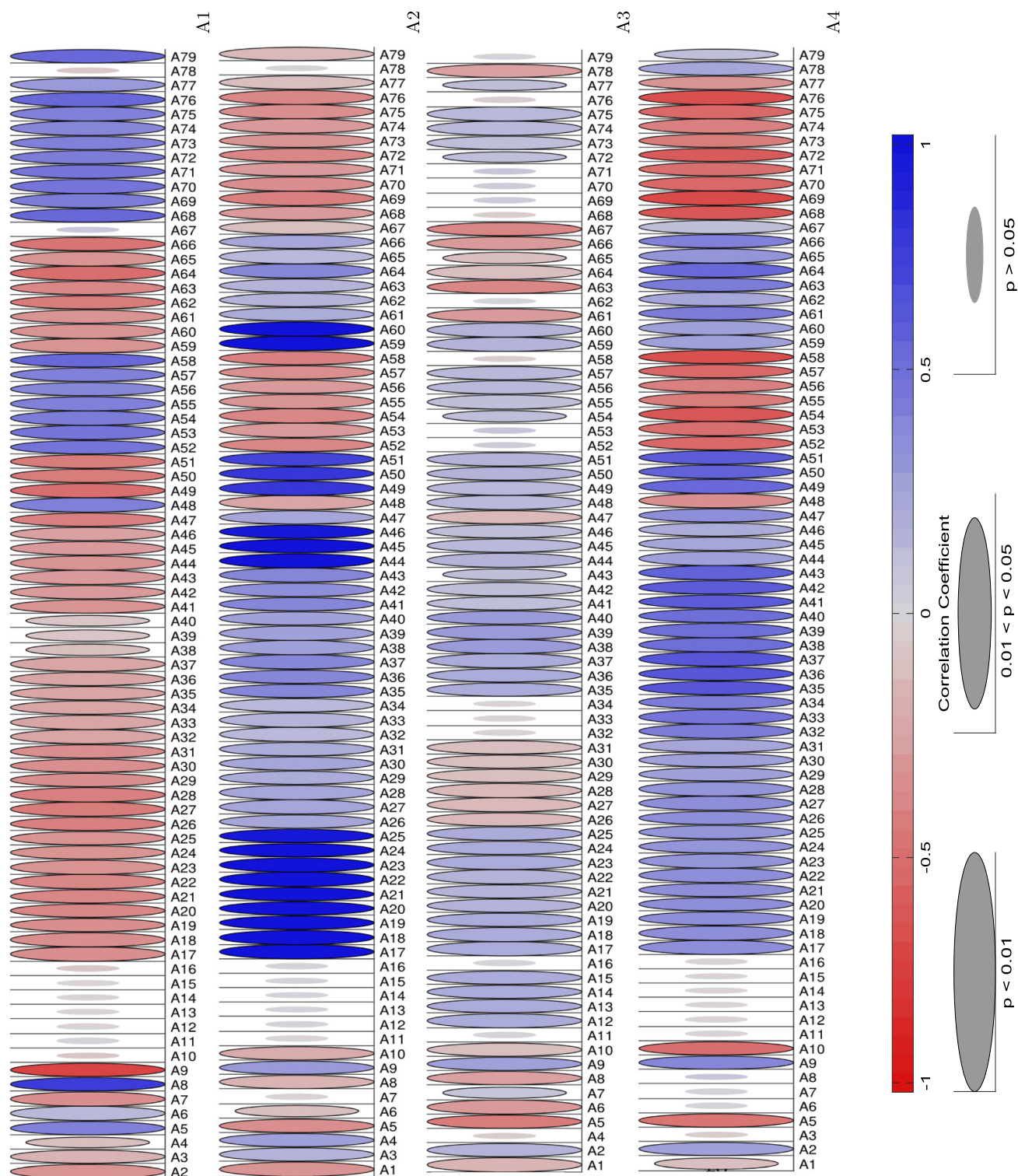


Fig. 6 The correlation of A1, A2, A3, and A4 verse the whole rest of attributes

nanoparticles' cores as determined by electron microscopy are combined in the physicochemical descriptors to describe size-related nanoparticles. In earlier times, nanoparticle core size was associated with cytotoxicity (Lag et al. 2018; Ali et al. 2019), and smaller particles being more dangerous. Due to the fact that, on a given mass basis, there would be a much higher number of small particles compared to bigger ones, this was paired with the exposure dose test parameter (Huk et al. 2014). Because dose–response relationships are at the heart of toxicity evaluation, the exposure dose also plays an essential role in the cytotoxicity of metal oxide nanoparticles; however, in nanoparticles, the correlations are not necessarily linear. Particle agglomeration is responsible for high nanoparticle concentrations. It was discovered that the link between cytotoxicity and nanoparticle hydrodynamic size is greater than their core size (Nel et al. 2009; Abakumov et al. 2018). Furthermore, because of its linkage to the energy bandgap (EG), Ec has been linked to nanoparticle toxicity. Because nanoparticles can cause oxidative stress and severe pulmonary inflammation in animals, EG coincides with the cellular redox potential (Zhang et al. 2012). Ec is the lowest unoccupied molecular orbital (LUMO) involved in electron transfers to and from the surface of MexOy when the valence band (EV) is captured. The electron moves straight from the aqueous electron donor to the conduction band if the cellular redox potential exceeds the conduction band edge of the MexOy nanoparticles.

4.2 Metal oxide nanoparticles data preprocessing results

The outcomes of the employed methods to handle the imbalanced dataset problem and the outliers problem are conducted in this experiment. Several measures are also employed to assess and compare the performance of the proposed EBSTSA-based feature selection algorithm. These are the rank-sum test's mean, standard deviation, and *p*-values.

4.2.1 Handling imbalanced data results

Figure 7 illustrates the data class distribution before and after the SMOTE re-sampling method. As is evident, there has been a significant improvement in the class distribution as new, synthetic samples from the minority class (toxic nanoparticles) have been produced. As a result, the classification performance can be greatly enhanced.

Because dose–response relationships are at the heart of hazardous quality assessment, the reality of exposure dose plays a vital role in metal oxide nanoparticles cytotoxicity, even though the correlations are not linear in the case of nanoparticles. Furthermore, high nanoparticle concentrations cause particle agglomeration. DNA ladder bands on agarose gel electrophoresis appeared as a result of Au NPs'

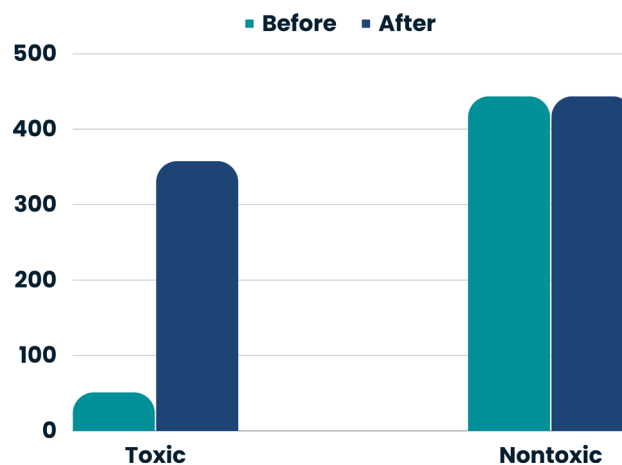


Fig. 7 Classes distribution of the data before and after using SMOTE

inactivation of the DNA repair mechanism (Loan et al. 2018). From Nishanth et al. (2011), it also found Ag, Al, carbon black, Ag coated with carbon, and Au NPs in addition to a variety of NPs. According to the stability and dissociation potential of the metal atoms as well as their distance from one another on the surface of nanoparticles, the coordination number of metal atoms is proportional to these factors. The ionic or covalent chemical bonding inside the NPs and the possible route of toxicity linked to NP breakdown and ion release are indicated by the coordination number. Small values of this descriptor demonstrated the propensity of metal atoms to dissociate from the NP surface, releasing ions into the environment and potentially producing toxicological consequences (Burk et al. 2018).

4.2.2 Outlier removal results

Twenty-three benchmark datasets from the UCI machine learning repository are utilized to assess the robustness and reliability of the proposed EBSTSA-based feature selection algorithm. The datasets utilized are labeled D1 through D23. Table 2 has a thorough explanation of the adopted datasets.

Table 3 compares the mean and standard deviation obtained from the proposed EBSTSA-based feature selection with binary particle swarm optimization-based feature selection (BPSO) (Mirjalili and Lewis 2013), quantum binary artificial bee colony-based feature selection (QBABC) (Barani and Nezamabadi-pour 2018), binary salp swarm algorithm based feature selection (BSSA) (Sayed et al. 2018), binary chicken swarm optimization based feature selection (BCSO) (Hafez et al. 2015), binary butterfly optimization algorithm-based feature selection (BBOA) (Arora and Anand 2018), binary ant lion optimization-based feature selection (BALO) (Emary et al. 2016), and binary atom search optimization-based feature selection (BASO) (Too and Rahim 2020). The parameter settings for the adopted algorithms are shown in

Table 2 Benchmark datasets description

ID	Dataset	No. features	No. Instances	No. classes	Missing values
D1	Iris	4	150	3	No
D2	Sonar	60	208	2	No
D3	Diabetes	8	786	2	No
D4	Wisconsin Diagnosis Breast Cancer	32	596	2	No
D5	TicTacToe	8	768	2	No
D6	Glass Identification	10	214	6	No
D7	Ovarian Cancer	216	4000	2	No
D8	Arrhythmia	279	452	13	No
D9	Chess	36	3196	2	No
D10	Poker Hand	10	25010	10	Yes
D11	Abalone	8	4177	28	Yes
D12	Waveform	21	5000	3	No
D13	Zoo	18	101	7	No
D14	Credit Approval	15	690	2	Yes
D15	Mice Protein Expression	82	1080	2	Yes
D16	Parkinsons	22	195	2	No
D17	Lung Cancer	32	57	3	No
D18	SPECTF Heart	45	267	2	Yes
D19	Thoracic Surgery	17	470	2	Yes
D20	Statlog Heart	17	470	2	No
D21	Indian liver patient	13	270	2	Yes
D22	Germen Credit	24	1000	2	No
D23	Letter	16	5000	26	No

the Table 4. The rest of the parameters, such as dimension length, population size, maximum number of iterations, and fitness function, are set to the same value for all algorithms. This is being done in order to compare these algorithms fairly. As with the comparable benchmark dataset, the dimension length is set to have the same number of features. The maximum number of iterations is set to 30, and the population size is set to 50. The selection of hyperparameter values in this paper follows their default settings, mimicking the original configuration of the algorithm. It must be underlined how crucial hyperparameters are in deciding how well an algorithm performs. These parameters are crucial in determining how well any optimization algorithm performs and behaves. Fine-tuning these parameters can have a considerable impact on the algorithm's ability to find optimal solutions to various problem domains. When these parameters are altered, different problem domains will have different results in terms of convergence rate, solution quality, and robustness. Tuning these parameters is deemed to be outside the scope of this paper in this particular research setting. This decision was made to ensure that the proposed EBSTSA algorithm's improvements are properly credited to its inherent design and to maintain a steady baseline for evaluation. Thus, this paper only focuses on boosting the performance of EBSTSA. The

selection of the control parameter of EBSTSA is based on the try and error method. The proposed EBSTSA-based feature selection algorithm, as can be shown, produced the best outcomes for 13 out of 23 benchmark datasets. They are D2, D4, D7, D9, D10, D12–D15, D17, D18, D20, and D22. Due to its superior performance across six benchmark datasets, BBOA is ranked second. BSSA is likewise in the third position. The performance of STSA can be impacted by selecting a suitable transfer function, as can be seen from all the results attained. This is may due to that the proposed EBSTSA has few parameters to set, despite BCSO and QBABC having many parameters to be tuned. Setting the optimal values for these parameters can effectively improve the algorithm's performance. In this paper, EBSTSA has only a parameter called the control parameter *CT*. *CT* is set to 0.6 same as the value in Jiang et al. (2020), as we found that this value based on trial and error is the optimal value to boost the performance of the original STSA.

For further investigation of the performance of the proposed EBSTSA, one well-known statistical test called *p*-values of the Wilcoxon rank-sum is used. Table 5 compared the performance of EBSTSA-based feature selection with BPSO, QBABC, BCSO, BALO, BASO, BSSA, and BBOA-based feature selection algorithms for metal oxide

Table 3 EBSTSA vs. other metaheuristic swarm algorithms for 23 benchmark datasets in terms of mean and standard

	EBSTSA		BPSO		QBABC		BCSO		BALO		BASO		BSSA		BBOA	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
D1	0.0331	0.0000	0.0328	0.0023	0.0371	0.0019	0.0381	0.0069	0.0339	0.0022	0.0365	0.0029	0.0371	0.0032	0.0318	0.0037
D2	0.1027	0.0166	0.1114	0.0068	0.1084	0.0072	0.1170	0.0021	0.1079	0.0087	0.1143	0.0069	0.1094	0.0043	0.1081	0.0082
D3	0.2417	0.0035	0.2133	0.0055	0.2172	0.0043	0.2208	0.0015	0.2167	0.0031	0.2163	0.0055	0.2181	0.0024	0.2117	0.0062
D4	0.0271	0.0016	0.0322	0.0005	0.0279	0.0010	0.0284	0.0015	0.0273	0.0009	0.0330	0.0015	0.0279	0.0013	0.0300	0.0013
D5	0.1886	0.0080	0.1676	0.0063	0.1214	0.0029	0.1242	0.0022	0.1219	0.0019	0.1747	0.0056	0.1244	0.0018	0.1218	0.0022
D6	0.2039	0.0144	0.1755	0.0055	0.1958	0.0121	0.2140	0.0119	0.1887	0.0174	0.1920	0.0133	0.1881	0.0144	0.1789	0.0103
D7	0.0261	0.0051	0.0371	0.0032	0.0475	0.0049	0.0571	0.0033	0.0482	0.0026	0.0428	0.0078	0.0468	0.0021	0.0437	0.0061
D8	0.2846	0.0045	0.3498	0.0130	0.2457	0.0217	0.3714	0.0173	0.2410	0.0104	0.3143	0.0072	0.3480	0.0207	0.3411	0.0134
D9	0.0013	0.0001	0.1373	0.0016	0.0492	0.0008	0.0482	0.0013	0.0487	0.0009	0.1229	0.0301	0.0483	0.0012	0.0504	0.0005
D10	0.0176	0.0004	0.4075	0.0012	0.3989	0.0014	0.4006	0.0015	0.2004	0.0012	0.3742	0.0119	0.4003	0.0006	0.3996	0.0018
D11	0.0308	0.0001	0.6214	0.0025	0.6240	0.0013	0.6290	0.0016	0.0249	0.0032	0.6231	0.0035	0.0242	0.0021	0.6167	0.0018
D12	0.0094	0.0007	0.1807	0.0029	0.1565	0.0033	0.1617	0.0018	0.1573	0.0022	0.1879	0.0104	0.1595	0.0028	0.1635	0.0036
D13	0.0027	0.0006	0.0687	0.0061	0.0252	0.0070	0.0352	0.0043	0.0368	0.0043	0.0634	0.0166	0.0352	0.0043	0.0245	0.0066
D14	0.0058	0.0003	0.1224	0.0021	0.1149	0.0059	0.1155	0.0051	0.1135	0.0046	0.1267	0.0040	0.1158	0.0026	0.1177	0.0019
D15	0.0000	0.0000	0.0066	0.0007	0.0086	0.0006	0.0124	0.0012	0.0099	0.0011	0.0007	0.0019	0.0001	0.0008	0.0067	0.0008
D16	0.0021	0.0003	0.0627	0.0103	0.0466	0.0068	0.0498	0.0073	0.0471	0.0020	0.0628	0.0099	0.0482	0.0053	0.0020	0.0040
D17	0.0003	0.0006	0.1412	0.0136	0.2166	0.0221	0.2309	0.0175	0.1996	0.0343	0.1049	0.0906	0.1878	0.0338	0.1841	0.0335
D18	0.0000	0.0000	0.0013	0.0000	0.0042	0.0004	0.0165	0.0025	0.0040	0.0000	0.0008	0.0007	0.0045	0.0013	0.0043	0.0009
D19	0.0000	0.0000	0.0087	0.0034	0.0002	0.0037	0.0307	0.0022	0.0018	0.0019	0.0019	0.0007	0.0002	0.0033	0.0000	0.0014
D20	0.0070	0.0002	0.1625	0.0025	0.1254	0.0025	0.1377	0.0029	0.1307	0.0044	0.1679	0.0113	0.1272	0.0111	0.1296	0.0047
D21	0.0112	0.0002	0.2259	0.0036	0.2262	0.0034	0.2641	0.0052	0.2406	0.0068	0.2295	0.0044	0.2431	0.0089	0.0102	0.0032
D22	0.0108	0.0002	0.2379	0.0097	0.2280	0.0023	0.2250	0.0013	0.2269	0.0027	0.2299	0.0064	0.2261	0.0019	0.2258	0.0055
D23	0.0003	0.0001	0.0587	0.0033	0.0010	0.0049	0.0762	0.0048	0.0044	0.0062	0.0441	0.0342	0.0002	0.0061	0.0585	0.0046

Table 4 Parameter settings for EBSTSA, BALO, BBOA, QBABC, BASO, BBSA, BCSO, and BPSO-based feature selection algorithms

Algorithm	Parameter(s)	Description	Value
EBSTSA	The control parameter (CT)	This parameter is used to determine how the new seeds in an optimization algorithm are generated	0.6
BALO	The exploration and exploitation balancing factor (I)	This parameter establishes the trade-off between exploration and exploitation. Higher values encourage exploration, whereas lower values place more emphasis on exploitation	1
BBOA	The switching probability	The probability of a butterfly changing its search mode is determined by this parameter. More exploration is encouraged by a larger switch probability, while the use of the current search mode is encouraged by a lower number	0.8
	The power exponent	The movement of butterflies toward potential solutions is influenced by the power exponent. During the search process, it influences how much exploration and exploitation is done	0.1
	The sensory modality	This describes the particular sensory input pathway that a butterfly employs to interpret its surroundings	0.01
QBABC	The number of food sources	This shows how many food sources the colony has access to. Every food source stands for a possible solution to the optimization problem	5
	The colony size	It refers to the colony's total number of artificial bees. The colony size can improve the algorithm's capacity to explore	10
	The number of limit trials	This parameter specifies how many times a bee can try to improve a food source before switching to a different one	5
BASO	Depth weight (Alpha)	Alpha affects the algorithm's exploration-exploitation balance. It establishes the scope of the search	50
	Multiplier weight (Beta)	The fitness function's impact on the search process is under the control of beta. It helps direct the search in the direction of promising areas	0.2

Table 4 continued

Algorithm	Parameter(s)	Description	Value
BSSA	The exploration and exploitation balancing factor ($c1$)	This parameter establishes the trade-off between exploration and exploitation. Higher values encourage exploration, whereas lower values place more emphasis on exploitation	2
BCSO	The number of chicken updated	The number of hens selected for each iteration to update their positions is determined by this parameter	10
	The percent of roosters population size	It indicates the proportion of roosters in the population. Roosters are essential in determining how the entire swarm moves	0.15
	The percent of hens population size	This parameter shows what proportion of the population is made up of hens. Hens provides a behavioral cue for the swarm by adhering to roosters	0.7
	The percent of mother hens population size	It represents the proportion of the population who serve as mother hens. Leading and directing the swarm is the responsibility of mother hens	0.05
BPSO	The inertial weight	It regulates how the particle's present velocity affects how it moves across search space. Higher values place more of an emphasis on exploration, while lower values encourage exploitation	1
	Personal learning coefficient	This parameter regulates how a particle's personal best position affects its movement	1.5
	The inertia weight damping ratio	It controls the inertial weight reduction across iterations, assisting the algorithm's convergence to a solution	0.9
	Global learning coefficient	The effect of the global best location on a particle's movement within the swarm is controlled by this parameter	2.0

nanoparticles dataset. It is clear that the proposed EBSTSA achieved the maximum accuracy while maintaining the highest stability. The results are statistically significant, as can also be shown from the comparison in terms of p -values.

From the results, the selected attributes from the EBSTSA-based feature selection algorithm are A25, A28, A44, A46, A50, A57, A59, A61, A68, A70, A72, A77, and A79. As

can be seen, the number of selected attributes is 13 out of 79 with a percentage of 16.64%. Thus, it can be observed that the EBSTSA-based feature selection algorithm's outstanding ability to locate the required minimal set of attributes maximizes classification accuracy while reducing the dimension size. This EBSTSA-based feature selection algorithm enables the selection of nano descriptors that quantify the

Table 5 EBSTSA vs. the other metaheuristic swarm algorithms for metal oxide nanoparticles dataset

	Mean	SD	<i>p</i> -values
EBSTSA			
BPSO	0.01314	0.00214	<0.05
QBABC	0.02625	0.00295	<0.05
BCSO	0.02816	0.00401	<0.05
BALO	0.02614	0.00436	<0.05
BASO	0.01484	0.00483	<0.05
BSSA	0.02414	0.00405	<0.05
BBOA	0.01935	0.00339	<0.05

distinctive properties of the surface atoms based on different parameters, including as potential energies and coordination numbers. In evaluating the risk of manufactured nanomaterials, the development of nanotoxicity prediction models is becoming more and more crucial. Additionally, metal oxide nanoparticles may subject organisms, including people, to long-term stress. As a result, understanding the connection between the properties of metal oxide nanoparticles and their toxicity is crucial for evaluating the hazards of ENMs. The results also showed that the proposed EBSTSA-based feature selection algorithm is capable of identifying relationships between the physicochemical characteristics, laboratory conditions, and bioactivity of the nanomaterials and their toxicity endpoints. According to the National Institutes of Health Material DB, classifications such as “Material type,” “Manufacturer,” and “Manufacturing method” exist for the nanoparticles used in publications in Information DB. The physicochemical characteristics of nanoparticles, including their core size, shape, surface charge, hydrodynamic size, specific surface area, and dissolving influence on the toxicity of nanomaterials, are stored in the PCHEM DB (Parashar et al. 2020). Nanoparticle toxicity is influenced by the exposure duration, conditions, and length. The toxicity of nanoparticles was evaluated on several cell lines. Several studies have seen the impact of SW-CNTs on human cell lines including human HEK cells, human HEL 293 cells, human macrophage cells, human A549 cells, and human epithelial-like Hela cells. The application and toxicology of nanoparticles are impacted by their small size and higher surface area-to-volume ratio. Consensus about dosimetry for nanoparticle uptake, distribution, and toxicity is based on the shape, size or surface area of the particles (Li et al. 2014; Asmatulu et al. 2022).

To ensure that the obtained Wilcoxon rank-sum test's *p*-values statistically significant findings are reliable and not just chance occurrences, another experiment is conducted. Table 6 shows the multiple hypothesis testing corrections methods. These correction methods are Hommel, Bon-

ferroni, Benjamini-Hochberg (abbreviated as ‘FDR-BH’), Holm, and Benjamini-Yekutieli (abbreviated as ‘FDR-BY’). The significance level in this experiment is set to 0.05, which is the threshold for addressing a result statistically significant. In hypothesis testing, *p*-values show the probability of obtaining outcomes that are as extreme as, or more severe than, the observed results under the null hypothesis. When performing multiple tests, the probability of incorrectly rejecting a null hypothesis (Type I error) increases. Correction methods are employed to alleviate this (Demšar 2006). In this experiment, the results are carried out for 25 independent runs to get meaningful statistical measures. The results revealed an interesting trend: most of the conducted results across the adopted 23 benchmark datasets, have *p*-values less than the traditional threshold of 0.05. This observation is the result of using the Wilcoxon rank-sum test in conjunction with multiple hypothesis testing correction methods. These adjusted *p*-values imply that the proposed EBSTSA-based feature selection algorithm is statistically significance when compared to BALO, BBOA, QBABC, BASO, BSSA, and BCSO-based feature selection algorithms. Furthermore, the results revealed a stark difference between the positive and negative ratings observed throughout 25 independent runs. This difference, with *R*+ about 25 times greater than *R*- on the majority of the adopted datasets, highlights the algorithm's outstanding performance under a wide range of settings. This means that the EBSTSA algorithm is not only robust but also consistently effective at separating the most relevant features. It also indicates that the EBSTSA-based feature selection algorithm outperforms its competitors across a wide range of datasets, further boosting its reputation as a dependable and potent methodology for feature selection in data analysis and modeling endeavors. Moreover, the results revealed that the Holm, Bonferroni, and FDR-BY adjustment *p*-value methods have the most powerful behavior. Another finding, the results show no significant difference between proposed EBSTSA algorithm compared to BCSO and BALO for D4, compared to QABC and BBOA for D7, compared to BPSO, QABC, BCSO, BALO, BASO, BCSO, BALO, BSSA, and BBOA for D8, compared to BALO using FDR-BY and Bonferroni for D12, compared to BALO using Bonferroni and FDR-BY for D14, compared to BALO for D15, compared to BPSO using Bonferroni and FDR-BY for D18, compared to BPSO and BASO for D19, and compared to BPSO and BASO for D20.

The next experiment proves that using the proposed EBSTSA feature selection can significantly reduce the number of outliers. Figure 8 shows the DBSCAN result using squared Euclidean distance metric with the same configuration as in the previous experiment in Fig. 4 (the minimum number of points equals 3 and the neighborhood distance epsilon equals 1.5). As can be observed the number of outliers reduced from 270 to 0. Additionally, the results show

Table 6 Adjusted p -values for the Wilcoxon Rank-Sum Test with a Level of Significance = 0.05 for EBSTSA vs. BPSO, EBSTSA vs. BSSA, EBSTSA vs. QBABC, EBSTSA vs. BCSO, EBSTSA vs. BALO, EBSTSA vs. BASO, EBSTSA vs. BBOA

		BPSO	QBABC	BCSO	BALO	BASO	BSSA	BBOA
D1	R+	25	25	25	25	25	25	25
	R−	0	0	0	0	0	0	0
	Unadjusted	1.78E−09	2.68E−09	1.24E−09	1.63E−09	3.63E−09	1.23E−09	3.17E−08
	Bonferroni	1.25E−08	1.88E−08	8.65E−09	1.14E−08	2.54E−08	8.58E−09	2.22E−07
	Holm	8.58E−09	8.58E−09	8.58E−09	8.58E−09	8.58E−09	8.58E−09	3.17E−08
	Hommel	5.34E−09	5.45E−09	4.84E−09	4.89E−09	7.26E−09	4.84E−09	3.17E−08
	FDR-BH	3.11E−09	3.75E−09	3.11E−09	3.11E−09	4.24E−09	3.11E−09	3.17E−08
	FDR-BY	8.07E−09	9.73E−09	8.07E−09	8.07E−09	1.10E−08	8.07E−09	8.23E−08
D2	R+	19	25	18	23	22	22	18
	R−	6	0	7	2	3	3	7
	Unadjusted	5.41E−08	5.42E−11	4.99E−06	1.03E−09	3.44E−09	3.69E−09	2.14E−07
	Bonferroni	3.79E−07	3.79E−10	3.49E−05	7.21E−09	2.41E−08	2.58E−08	1.50E−06
	Holm	1.62E−07	3.79E−10	4.99E−06	6.18E−09	1.72E−08	1.72E−08	4.28E−07
	Hommel	1.62E−07	3.79E−10	4.99E−06	6.18E−09	1.38E−08	1.48E−08	4.28E−07
	FDR-BH	7.57E−08	3.79E−10	4.99E−06	3.61E−09	6.46E−09	6.46E−09	2.50E−07
	FDR-BY	1.96E−07	9.84E−10	1.29E−05	9.35E−09	1.67E−08	1.67E−08	6.47E−07
D3	R+	25	25	25	25	25	25	25
	R−	0	0	0	0	0	0	0
	Unadjusted	1.40E−09	1.40E−09	1.38E−09	1.40E−09	1.40E−09	1.40E−09	2.01E−09
	Bonferroni	9.80E−09	9.80E−09	9.66E−09	9.80E−09	9.80E−09	9.80E−09	1.41E−08
	Holm	9.66E−09	9.66E−09	9.66E−09	9.66E−09	9.66E−09	9.66E−09	9.66E−09
	Hommel	2.01E−09	2.01E−09	2.01E−09	2.01E−09	2.01E−09	2.01E−09	2.01E−09
	FDR-BH	1.63E−09	1.63E−09	1.63E−09	1.63E−09	1.63E−09	1.63E−09	2.01E−09
	FDR-BY	4.24E−09	4.24E−09	4.24E−09	4.24E−09	4.24E−09	4.24E−09	5.21E−09
D4	R+	25	0	12	14	23	0	11
	R−	0	25	13	11	2	25	14
	Unadjusted	1.40E−09	1.32E−09	0.281	0.51099	3.04E−06	1.34E−09	0.19348
	Bonferroni	9.80E−09	9.24E−09	1	1	2.13E−05	9.38E−09	1
	Holm	9.24E−09	9.24E−09	0.58044	0.58044	1.22E−05	9.24E−09	0.58044
	Hommel	7.00E−09	6.60E−09	0.51099	0.51099	1.22E−05	6.70E−09	0.4215
	FDR-BH	3.27E−09	3.27E−09	0.327833	0.51099	5.32E−06	3.27E−09	0.270872
	FDR-BY	8.47E−09	8.47E−09	0.850025	1	1.38E−05	8.47E−09	0.702332
D5	R+	25	25	25	25	25	25	25
	R−	0	0	0	0	0	0	0
	Unadjusted	2.77E−09	1.29E−09	1.27E−09	1.32E−09	1.73E−09	1.28E−09	1.34E−09
	Bonferroni	1.94E−08	9.03E−09	8.89E−09	9.24E−09	1.21E−08	8.96E−09	9.38E−09
	Holm	8.89E−09	8.89E−09	8.89E−09	8.89E−09	8.89E−09	8.89E−09	8.89E−09
	Hommel	2.77E−09	2.60E−09	2.60E−09	2.64E−09	2.77E−09	2.60E−09	2.68E−09
	FDR-BH	2.77E−09	1.88E−09	1.88E−09	1.88E−09	2.02E−09	1.88E−09	1.88E−09
	FDR-BY	7.18E−09	4.86E−09	4.86E−09	4.86E−09	5.23E−09	4.86E−09	4.86E−09

Table 6 continued

		BPSO	QBABC	BCSO	BALO	BASO	BSSA	BBOA
D6	R+	21	25	25	25	25	25	18
	R−	4	0	0	0	0	0	7
	Unadjusted	4.65E−05	1.27E−09	1.09E−09	1.35E−09	1.36E−09	1.36E−09	0.005243
	Bonferroni	3.26E−04	8.89E−09	7.63E−09	9.45E−09	9.52E−09	9.52E−09	0.036703
	Holm	9.30E−05	7.63E−09	7.63E−09	7.63E−09	7.63E−09	7.63E−09	0.005243
	Hommel	9.30E−05	3.81E−09	3.27E−09	4.05E−09	4.08E−09	4.08E−09	0.005243
	FDR-BH	5.43E−05	1.90E−09	1.90E−09	1.90E−09	1.90E−09	1.90E−09	0.005243
	FDR-BY	1.41E−04	4.94E−09	4.94E−09	4.94E−09	4.94E−09	4.94E−09	0.013595
D7	R+	23	17	23	25	25	25	13
	R−	2	8	2	0	0	0	12
	Unadjusted	0.000545	0.039496	2.71E−07	3.56E−09	1.36E−09	1.33E−09	0.87605
	Bonferroni	0.003816	0.276472	1.90E−06	2.49E−08	9.52E−09	9.31E−09	1
	Holm	0.001635	0.078992	1.08E−06	1.78E−08	9.31E−09	9.31E−09	0.87605
	Hommel	0.001635	0.078992	1.08E−06	1.78E−08	8.16E−09	7.98E−09	0.87605
	FDR-BH	0.000763	0.046078	4.74E−07	8.31E−09	4.76E−09	4.76E−09	0.87605
	FDR-BY	0.001979	0.119475	1.23E−06	2.15E−08	1.23E−08	1.23E−08	1
D8	R+	20	18	15	11	23	21	13
	R−	5	7	10	14	2	4	12
	Unadjusted	0.012982	0.061349	0.043732	0.38777	5.44E−05	0.005489	0.83612
	Bonferroni	0.090874	0.429443	0.306124	1	3.81E−04	0.038427	1
	Holm	0.06491	0.184047	0.174928	0.77554	3.81E−04	0.032937	0.83612
	Hommel	0.06491	0.184047	0.131196	0.77554	3.81E−04	0.032937	0.83612
	FDR-BH	0.030291	0.085888	0.076531	0.452398	3.81E−04	0.0192136	0.83612
	FDR-BY	0.078541	0.222696	0.198433	1	9.87E−04	0.04981812	1
D9	R+	25	25	25	24	25	23	8
	R−	0	0	0	1	0	2	17
	Unadjusted	1.40E−09	1.41E−09	5.85E−09	3.70E−07	4.61E−09	1.45E−07	0.011027
	Bonferroni	9.80E−09	9.87E−09	4.10E−08	2.59E−06	3.23E−08	1.02E−06	0.077189
	Holm	9.80E−09	9.80E−09	2.34E−08	7.40E−07	2.31E−08	4.35E−07	0.011027
	Hommel	8.40E−09	8.46E−09	2.34E−08	7.40E−07	1.84E−08	4.35E−07	0.011027
	FDR-BH	4.94E−09	4.94E−09	1.02E−08	4.32E−07	1.02E−08	2.03E−07	0.011027
	FDR-BY	1.28E−08	1.28E−08	2.65E−08	1.12E−06	2.65E−08	5.26E−07	0.028591
D10	R+	24	22	25	23	25	25	25
	R−	1	3	0	2	0	0	0
	Unadjusted	1.81E−08	1.11E−06	9.07E−07	2.08E−06	2.01E−09	3.98E−08	3.22E−09
	Bonferroni	1.27E−07	7.77E−06	6.35E−06	1.46E−05	1.41E−08	2.79E−07	2.25E−08
	Holm	9.05E−08	2.72E−06	2.72E−06	2.72E−06	1.41E−08	1.59E−07	1.93E−08
	Hommel	9.05E−08	2.08E−06	1.81E−06	2.08E−06	1.21E−08	1.59E−07	1.93E−08
	FDR-BH	4.22E−08	1.30E−06	1.27E−06	2.08E−06	1.13E−08	6.97E−08	1.13E−08
	FDR-BY	1.10E−07	3.36E−06	3.29E−06	5.39E−06	2.92E−08	1.81E−07	2.92E−08

Table 6 continued

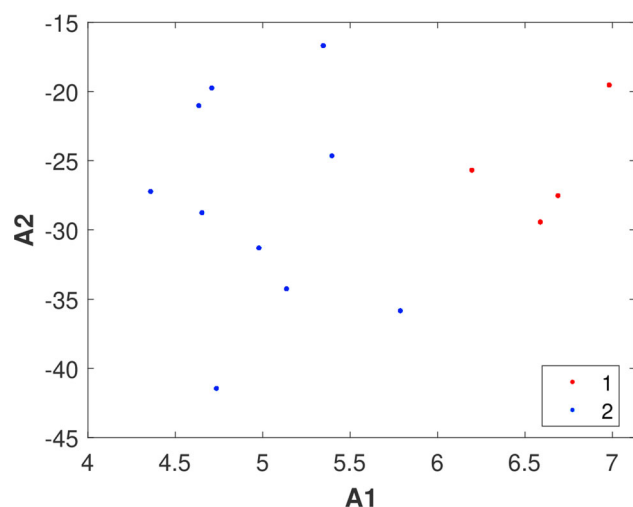
		BPSO	QBABC	BCSO	BALO	BASO	BSSA	BBOA
D11	R+	25	24	25	25	25	21	21
	R−	0	1	0	0	0	4	4
	Unadjusted	1.42E−09	1.15E−08	1.17E−09	1.39E−09	1.41E−09	1.06E−05	3.02E−05
	Bonferroni	9.94E−09	8.05E−08	8.19E−09	9.73E−09	9.87E−09	7.42E−05	2.11E−04
	Holm	8.34E−09	3.45E−08	8.19E−09	8.34E−09	8.34E−09	2.12E−05	3.02E−05
	Hommel	5.68E−09	3.45E−08	4.68E−09	5.56E−09	5.64E−09	2.12E−05	3.02E−05
	FDR-BH	2.49E−09	1.61E−08	2.49E−09	2.49E−09	2.49E−09	1.24E−05	3.02E−05
	FDR-BY	6.44E−09	4.17E−08	6.44E−09	6.44E−09	6.44E−09	3.21E−05	7.83E−05
D12	R+	24	25	25	16	25	23	19
	R−	1	0	0	9	0	2	6
	Unadjusted	3.20E−08	1.78E−09	1.42E−09	0.036126	1.42E−09	3.58E−08	0.000384
	Bonferroni	2.24E−07	1.25E−08	9.94E−09	0.252882	9.94E−09	2.51E−07	0.002689
	Holm	1.28E−07	9.94E−09	9.94E−09	0.036126	9.94E−09	1.28E−07	0.000768
	Hommel	9.60E−08	8.90E−09	7.10E−09	0.036126	7.10E−09	1.07E−07	0.000768
	FDR-BH	5.01E−08	4.15E−09	4.15E−09	0.036126	4.15E−09	5.01E−08	0.000448
	FDR-BY	1.30E−07	1.08E−08	1.08E−08	0.093669	1.08E−08	1.30E−07	0.001162
D13	R+	25	25	25	24	25	23	8
	R−	0	0	0	1	0	2	17
	Unadjusted	1.40E−09	1.41E−09	5.85E−09	3.70E−07	4.61E−09	1.45E−07	0.011027
	Bonferroni	9.80E−09	9.87E−09	4.10E−08	2.59E−06	3.23E−08	1.02E−06	0.077189
	Holm	9.80E−09	9.80E−09	2.34E−08	7.40E−07	2.31E−08	4.35E−07	0.011027
	Hommel	8.40E−09	8.46E−09	2.34E−08	7.40E−07	1.84E−08	4.35E−07	0.011027
	FDR-BH	4.94E−09	4.94E−09	1.02E−08	4.32E−07	1.02E−08	2.03E−07	0.011027
	FDR-BY	1.28E−08	1.28E−08	2.65E−08	1.12E−06	2.65E−08	5.26E−07	0.028591
D14	R+	25	25	24	16	19	25	23
	R−	0	0	1	9	6	0	2
	Unadjusted	1.42E−09	1.40E−09	3.20E−08	0.036126	0.00038418	1.42E−09	3.58E−08
	Bonferroni	9.94E−09	9.80E−09	2.24E−07	0.252882	2.69E−03	9.94E−09	2.51E−07
	Holm	9.80E−09	9.80E−09	1.28E−07	0.036126	7.68E−04	9.80E−09	1.28E−07
	Hommel	7.10E−09	7.00E−09	9.60E−08	0.036126	7.68E−04	7.10E−09	1.07E−07
	FDR-BH	3.31E−09	3.31E−09	5.01E−08	0.036126	4.48E−04	3.31E−09	5.01E−08
	FDR-BY	8.59E−09	8.59E−09	1.30E−07	0.093669	1.16E−03	8.59E−09	1.30E−07
D15	R+	25	24	25	16	21	25	25
	R−	0	1	0	9	4	0	0
	Unadjusted	9.67E−11	1.23E−08	1.16E−08	0.64837	1.49E−06	9.69E−11	9.72E−11
	Bonferroni	6.77E−10	8.61E−08	8.12E−08	1	1.04E−05	6.78E−10	6.80E−10
	Holm	6.77E−10	4.64E−08	4.64E−08	0.64837	2.98E−06	6.77E−10	6.77E−10
	Hommel	4.84E−10	3.69E−08	3.48E−08	0.64837	2.98E−06	4.85E−10	4.86E−10
	FDR-BH	2.27E−10	1.72E−08	1.72E−08	0.64837	1.74E−06	2.27E−10	2.27E−10
	FDR-BY	5.88E−10	4.46E−08	4.46E−08	1	4.51E−06	5.88E−10	5.88E−10

Table 6 continued

		BPSO	QBABC	BCSO	BALO	BASO	BSSA	BBOA
D16	R+	25	25	25	25	25	25	24
	R−	0	0	0	0	0	0	1
	Unadjusted	3.95E−10	2.00E−10	1.10E−10	3.95E−10	4.71E−10	3.11E−10	2.33E−06
	Bonferroni	2.77E−09	1.40E−09	7.70E−10	2.77E−09	3.30E−09	2.18E−09	1.63E−05
	Holm	1.58E−09	1.20E−09	7.70E−10	1.58E−09	1.58E−09	1.56E−09	2.33E−06
	Hommel	7.90E−10	6.28E−10	5.65E−10	7.90E−10	9.42E−10	7.07E−10	2.33E−06
	FDR-BH	5.50E−10	5.50E−10	5.50E−10	5.50E−10	5.50E−10	5.50E−10	2.33E−06
	FDR-BY	1.42E−09	1.42E−09	1.42E−09	1.42E−09	1.42E−09	1.42E−09	6.04E−06
D17	R+	22	24	22	24	25	21	23
	R−	3	1	3	1	0	4	2
	Unadjusted	6.09E−07	6.38E−09	1.61E−06	3.09E−06	1.41E−09	1.12E−05	0.000141
	Bonferroni	4.26E−06	4.47E−08	1.13E−05	2.16E−05	9.87E−09	7.84E−05	9.94E−04
	Holm	3.05E−06	3.83E−08	6.44E−06	9.27E−06	9.87E−09	2.24E−05	1.42E−04
	Hommel	3.05E−06	3.83E−08	6.18E−06	9.27E−06	9.87E−09	2.24E−05	1.42E−04
	FDR-BH	1.42E−06	2.23E−08	2.82E−06	4.33E−06	9.87E−09	1.31E−05	1.42E−04
	FDR-BY	3.68E−06	5.79E−08	7.31E−06	1.12E−05	2.56E−08	3.39E−05	3.68E−04
D18	R+	4	25	25	25	10	25	25
	R−	21	0	0	0	15	0	0
	Unadjusted	0.041218	6.32E−11	5.64E−11	7.79E−11	0.000537	6.59E−11	9.47E−11
	Bonferroni	0.288526	4.42E−10	3.95E−10	5.45E−10	0.003764	4.61E−10	6.63E−10
	Holm	0.041218	3.95E−10	3.95E−10	3.95E−10	0.001075	3.95E−10	3.95E−10
	Hommel	0.041218	1.90E−10	1.89E−10	2.34E−10	0.001075	1.98E−10	2.84E−10
	FDR-BH	0.041218	1.33E−10	1.33E−10	1.33E−10	0.000627	1.33E−10	1.33E−10
	FDR-BY	0.106872386	3.44E−10	3.44E−10	3.44E−10	0.001626	3.44E−10	3.44E−10
D19	R+	3	25	25	25	4	25	25
	R−	22	0	0	0	21	0	0
	Unadjusted	0.080996	7.45E−11	7.16E−11	4.43E−12	0.087999	9.67E−12	1.76E−11
	Bonferroni	0.566972	5.22E−10	5.01E−10	3.10E−11	0.615993	6.77E−11	1.23E−10
	Holm	0.161992	2.86E−10	2.86E−10	3.10E−11	0.161992	5.80E−11	8.80E−11
	Hommel	0.087999	2.24E−10	2.15E−10	3.10E−11	0.087999	5.28E−11	8.80E−11
	FDR-BH	0.087999	1.04E−10	1.04E−10	3.10E−11	0.087999	3.38E−11	4.11E−11
	FDR-BY	0.228168	2.70E−10	2.70E−10	8.04E−11	0.228168	8.78E−11	1.06E−10
D20	R+	4	25	25	25	1	25	25
	R−	21	0	0	0	24	0	0
	Unadjusted	0.94455	6.72E−12	5.99E−11	1.79E−11	0.33706	1.80E−11	2.77E−12
	Bonferroni	1	4.70E−11	4.19E−10	1.25E−10	1	1.26E−10	1.94E−11
	Holm	0.94455	4.03E−11	1.80E−10	8.95E−11	0.67412	8.95E−11	1.94E−11
	Hommel	0.94455	3.60E−11	1.80E−10	7.16E−11	0.67412	7.20E−11	1.94E−11
	FDR-BH	0.94455	2.35E−11	8.39E−11	3.15E−11	0.393236667	3.15E−11	1.94E−11
	FDR-BY	1	6.10E−11	2.17E−10	8.17E−11	1	8.17E−11	5.03E−11

Table 6 continued

		BPSO	QBABC	BCSO	BALO	BASO	BSSA	BBOA
D21	R+	25	25	25	22	25	22	22
	R−	0	0	0	3	0	3	3
	Unadjusted	1.38E−09	1.52E−09	7.94E−09	7.26E−07	1.38E−09	1.08E−06	4.97E−06
	Bonferroni	9.66E−09	1.06E−08	5.56E−08	5.08E−06	9.66E−09	7.56E−06	3.48E−05
	Holm	9.66E−09	9.66E−09	3.18E−08	2.18E−06	9.66E−09	2.18E−06	4.97E−06
	Hommel	6.90E−09	7.60E−09	3.18E−08	1.62E−06	6.90E−09	2.16E−06	4.97E−06
	FDR-BH	3.55E−09	3.55E−09	1.39E−08	1.02E−06	3.55E−09	1.26E−06	4.97E−06
	FDR-BY	9.20E−09	9.20E−09	3.60E−08	2.64E−06	9.20E−09	3.27E−06	1.29E−05
D22	R+	25	25	25	25	25	25	25
	R−	0	0	0	0	0	0	0
	Unadjusted	1.40E−09	1.36E−09	1.37E−09	1.37E−09	1.41E−09	1.38E−09	1.41E−09
	Bonferroni	9.80E−09	9.52E−09	9.59E−09	9.59E−09	9.87E−09	9.66E−09	9.87E−09
	Holm	9.52E−09	9.52E−09	9.52E−09	9.52E−09	9.52E−09	9.52E−09	9.52E−09
	Hommel	1.41E−09	1.41E−09	1.41E−09	1.41E−09	1.41E−09	1.41E−09	1.41E−09
	FDR-BH	1.41E−09	1.41E−09	1.41E−09	1.41E−09	1.41E−09	1.41E−09	1.41E−09
	FDR-BY	3.66E−09	3.66E−09	3.66E−09	3.66E−09	3.66E−09	3.66E−09	3.66E−09
D23	R+	25	25	25	25	25	25	25
	R−	0	0	0	0	0	0	0
	Unadjusted	1.39E−09	1.39E−09	1.38E−09	1.40E−09	1.41E−09	1.38E−09	1.39E−09
	Bonferroni	9.73E−09	9.73E−08	9.66E−09	9.80E−09	9.87E−09	9.66E−09	9.73E−09
	Holm	9.66E−09	1.39E−08	9.66E−09	9.66E−09	9.66E−09	9.66E−09	9.66E−09
	Hommel	2.78E−09	1.39E−08	2.76E−09	2.80E−09	2.82E−09	2.76E−09	2.78E−09
	FDR-BH	1.65E−09	1.39E−08	1.65E−09	1.65E−09	1.65E−09	1.65E−09	1.65E−09
	FDR-BY	4.27E−09	3.60E−08	4.27E−09	4.27E−09	4.27E−09	4.27E−09	4.27E−09

**Fig. 8** DBSCAN result using squared Euclidean distance metric after handling outliers

that there are only two clusters which are the actual clusters (toxic or non-toxic).

4.3 Prediction of toxicity effect of metal oxide nanoparticles results

In this experiment, the performance of the overall toxicity detection of the metal oxide nanoparticles dataset is assessed using the metrics of misclassification error rate, specificity, sensitivity, f1-score, and computational CPU time.

Figure 9 compared the error rate produced on average of 3 folds using SVM with Radial Basis Function (RBF) kernel function before applying each phase separately and after applying the feature selection algorithm using EBSTSA and handling imbalanced dataset using SMOTE. It should be mentioned that RBF is used because this kernel function has fewer parameters than the sigmoid and the polynomial kernel functions (Tharwat et al. 2017). Thus, RBF is used to produce the optimal estimation. These parameters are the regularization parameter C which is used to control the number of vectors and thus the classification accuracy and σ which affects the partitioning outcome in the feature space. Choosing these two parameters can significantly influence classification performance. The very smaller value of σ results in over-fitting which means more variance and less bias. The larger value of σ results in increasing the number of

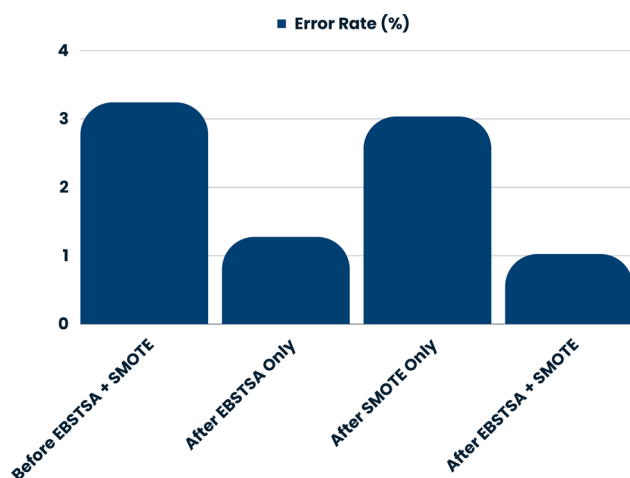


Fig. 9 The average error rate before and after applying SMOTE and EBSTSTA

support vectors thus increasing the number of misclassified samples and maximizing the kernel value. Meanwhile, the small value of C results in a maximized margin, increasing the number of support vectors and increasing the number of misclassified samples. On the other hand, the larger value of C will result in minimizing the width of the margin and decreasing the number of support vectors, thus decreasing the number of misclassified samples (Tharwat et al. 2017). In this paper, C is set to 10, and σ is set to 0.05 because it was found that these values are the optimal values minimizing the number of support vectors consequently decreasing the number of misclassified samples. Figure 9 shows that a balanced distribution of the classes was achieved by combining the proposed EBSTSA feature selection algorithm with the SMOTE sampling method, which resulted in a 1.02 average error rate and 98.98% classification accuracy. Additionally, it can be seen that using EBSTSA rather than only SMOTE results in a notable improvement in the model's performance. This indicates that employing the feature selection algorithm with the adopted dataset was successful.

Figure 10 compares the sensitivity, specificity, and f1-score before applying each phase separately and after applying both feature selection and handling an imbalanced dataset on average 3 folds. It should be noted that the sensitivity is used to measure "When it's detected as toxic, how often does it predict toxically?". As can be seen, after applying feature selection and handling an imbalanced dataset, the sensitivity result reaches 98.87%. These results can act as a director for potential impediments to the demonstration and give future users the sign of the reliability of their predictions. Specificity is used to measure "When it's detected as non-toxic, how often does it predict non-toxic?". As can be observed, after handling the two problems with the metal oxide dataset and reducing the size of the features set, the specificity result becomes 100%. The F1-score is the har-

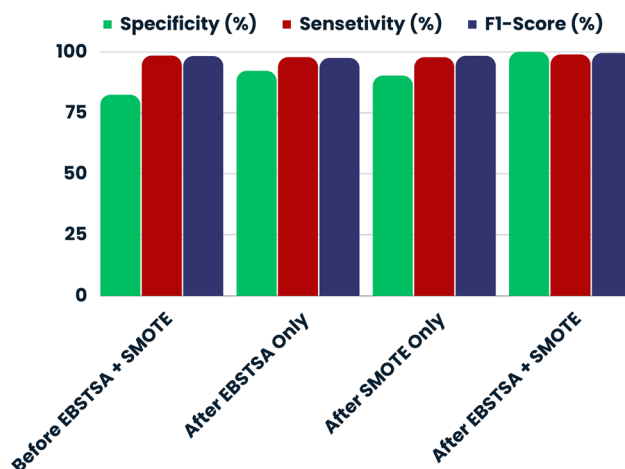


Fig. 10 The average sensitivity, specificity, and F1-score before and after applying SMOTE and EBSTSTA

monic mean of recall and precision. Better performance is indicated by higher numbers on the scale, which ranges from 0 to 1. The F1-score is particularly sensitive to imbalances between precision and recall because the harmonic mean gives lower values more weight than other averaging methods, including the arithmetic mean. As can be observed from Fig. 10, after utilizing the proposed EBSTSA and SMOTE, the overall proposed toxicity detection of the metal oxide nanoparticles model obtained a 99.47% of f1-score. Another finding, from this figure, it can be observed that each part of the model is crucial, as the results improved after applying SMOTE and remarkably improved after applying the proposed EBSTSA-based feature selection algorithm. Thus, it can be indicated that some features of the adopted dataset are irrelevant and can significantly affect the performance of the proposed toxicity detection model. These results are consistent with the obtained results from Fig. 9. These findings permit read access based on the chemical similarity of particular MexOy and the use of ATP or LDH assays to predict cytotoxicity.

Finally, Fig. 11 compared the computational time in seconds before and after each phase. As can be seen, after handling the existence of irrelevant features and imbalanced data problems, the CPU computational time increased. This happens due to after applying SMOTE, the number of samples in the train set is increased. However, this time is acceptable compared with the model detection accuracy. By compiling the relevant descriptors (physicochemical, atomistic, and molecular) for each nanoparticle and developing a first grouping hypothesis that connects an endpoint to certain behavior and reactivity features, the proposed model may thus be applied to NPs. Additionally, the samples were divided into groups and employed in accordance with the applicability of the model by using computational techniques and data gap-filling.

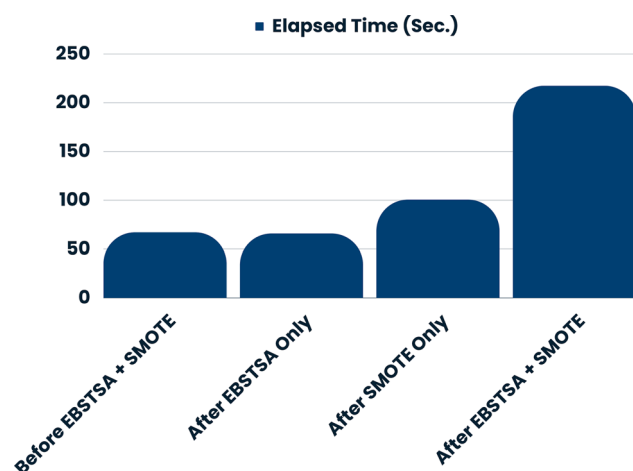


Fig. 11 The CPU computational time summary

The results also demonstrated the classifications' predictability and robustness. The test parameters, as specified in the experimental section, were computed using the experimentally determined and predicted values of the viability of the BEAS-2B and RAW264.7 cells, respectively. Y-randomization showed resilience and validity with good agreement. Based on the calculated APD value of 2.645 (normalized test set range: 0–0.591), all predictions in our scenario were deemed to be trustworthy. These findings can serve as a guide for possible algorithmic restrictions and show whether predictions made by future users are likely to be accurate.

An additional experiment is conducted in Fig. 12 to show how effective the proposed model is. In this experiment, the ROC (Receiver Operating Characteristic) curve is demonstrated. The ROC curve serves as an example of a binary classification model's efficacy. For various classification criteria, it shows the true positive rate (sensitivity) vs the false positive rate (1 - specificity). In other words, it shows how effectively the model can distinguish between the two classes. This figure's upper left corner, where the ROC curve is located, shows how well the model can distinguish between the two groups. Additionally, it shows that the model has better classification model features, including higher sensitivity (true positive rate) and lower false positive rate.

Table 7 shows the distribution of classes across the training and testing sets before and after applying SMOTE for the training set. Additionally, the metrics for each fold, including the f1-score, sensitivity, specificity, error rate, and computational time are given. In this experiment, the results were compared before and after applying the SMOTE method to the training set. Specifically, the proposed model that solely utilized the EBSTSA-based feature selection algorithm was compared with the proposed model that employed both the EBSTSA-based feature selection and SMOTE. This comparison aimed to assess the effectiveness of integrating

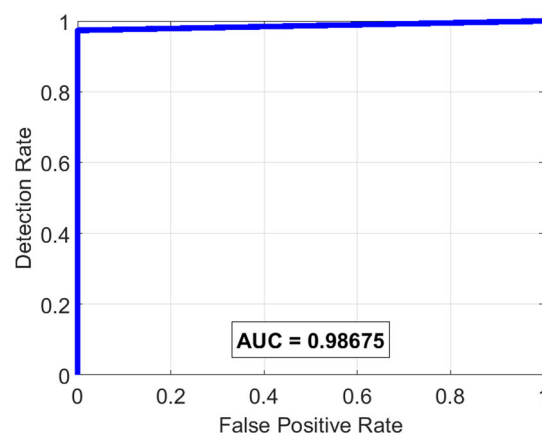


Fig. 12 The ROC curve of the proposed model

SMOTE into the model. The results clearly show that the proposed model is effective in determining the toxicity of metal oxide nanoparticles, proving its suitability for this prediction task. It is obtained on average 1.02% of error rate, 100% of specificity, 98.87% of sensitivity, and 99.47% of f1-score with 217.12 average elapsed time in seconds. A notable improvement in performance measures can be observed after applying the SMOTE to the training dataset. Several important metrics, such as the error rate, specificity, sensitivity, and F1-score, all show improvement. This indicates that the use of SMOTE improves the model's ability to classify instances correctly, especially in situations when class imbalances exist. Moreover, the results of this paper have essential effects on the study of nanotoxicology. The proposed model's proven ability to successfully identify the toxicity of metal oxide nanoparticles represents a significant improvement in prediction abilities. This development has the potential to simplify and improve toxicity evaluations of nanoparticles, giving crucial information about their potential dangers and safety profiles. In particular, the model's ability to identify toxic effects can result in more accurate hazard assessments, allowing researchers to recognize and rank potentially hazardous nanoparticles for additional research. Additionally, it can help to develop specialized safety standards and laws for their manufacture and use, as well as the design of safer nanoparticles.

5 Conclusion and future work

In this paper, a validated and easily applicable machine-learning toxicity detection model for the prediction of the cytotoxicity of metal oxide nanoparticles is proposed. The proposed model was tested using a dataset of 24 distinct metal oxide nanoparticles, including their physicochemical and structural characteristics (17 commercial and 7 domestically generated). Out of the 79 descriptors provided as input,

Table 7 The class distribution of each class for train and test sets with the error rate, specificity, sensitivity, and F1-score per fold

K-Fold	Train		Test		Error Rate (%)	Specificity (%)	Sensitivity (%)	F1-Score (%)	Elapsed time (s)
	Toxic	Nontoxic	Toxic	Nontoxic					
1	Before After	296 296	17 17	147 147	1.40 2.44	94.12 100.00	97.00 97.28	96.41 98.62	76.71 214.35
2	Before After	295 295	17 17	148 148	0.61 0.00	100.00 100.00	97.01 100.00	97.12 100.00	61.08 218.44
3	Before After	295 295	17 17	148 148	1.80 0.61	82.35 100.00	99.31 99.32	98.79 99.66	58.59 218.58
				Avg.Before Avg.After	1.27 1.02	92.16 100.00	97.77 98.87	97.44 99.47	65.46 217.12

13 were determined to be statistically significant. Two experimental parameters—the core and hydrodynamic size of the nanoparticles—two assay-related parameters—exposure dose and assay type (ATP or LDH)—as well as three computational descriptors—conduction band energy, average coordination number of the metal atoms on the surface, and v-Metal atoms surface or force vector surface normal component of the metal atoms—are on the list. Physicochemical descriptor libraries and average C.N. can be used to identify EC. Metal surface atoms and V. Molecular dynamics software, such as LAMMPS, can be used to compute the surface of metal atoms. Based on the chemical similarity of particular metal oxide nanoparticles and the application of the ATP or LDH assays to predict cytotoxicity, the proposed toxicity detection model permits read access. In comparison to other well-known and recent swarm-based feature selection algorithms, the experimental results demonstrated the reliability and effectiveness of the proposed toxicity detection model based on a novel EBSTSA-based feature selection algorithm over 23 benchmark datasets. The overall proposed model obtained an average 1.02% of error rate, 100% of specificity, 98.87% of sensitivity, and 99.47% of f1-score. Additionally, the results revealed that machine learning algorithms can produce encouraging results in the field of predicting the toxicity of metal oxide nanoparticles. Due to the promising result, the proposed model may be utilized to improve risk assessments and informed decision-making in industries and biological applications that use metal oxide nanoparticles.

Additional research on the proposed model's scalability and generalizability across a wider range of metal oxide nanoparticles, taking into account different compositions, sizes, and surface coatings. In addition, integrating multi-modal data sources, such as physicochemical properties, gene expression patterns, and cellular responses, will be taken into account to evaluate the performance of the proposed model. Moreover, further research into employing different swarm intelligence algorithms can also be taken into consideration.

Author Contributions Design, implementation, data analysis, methodology, and writing the original draft were performed by Gehad Ismail Sayed. Material preparation, data collection, and analysis are performed by Heba Alshater. Supervision and editing are done by Aboul Ella Hassanien. All authors read and approved the final manuscript.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Data availability The data that support the findings of this paper come from a meta-analysis of a dataset by Zhang et al. (2012) retrieved from the S2NANO (www.s2nano.org) database on the cytotoxicity of 24 MexOy NPs to murine myeloid (RAW264.7) and human bronchial epithelial (BEAS-2B) cell lines using a single parameter lactate dehydrogenase (LDH) and adenosine triphosphate (ATP). To facilitate accessibility within the scientific community and to interested stake-

holders, the fully detailed model and guiding material on its use have been made publically available as a web service (<https://cellviability.cloud.nanosolveit.eu/>).

Declarations

Conflict of interest The authors stated that they do not have any conflicts of interest.

Ethical approval None.

Informed consent None.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abakumov M, Semkina A, Skorikov A, Vishnevskiy D, Ivanova A, Mironova E, Davydova G, Majouga A, Chekhonin V (2018) Toxicity of iron oxide nanoparticles: size and coating effects. *J Biochem Mol Toxicol* 32:1–25
- Ali AF, Mostafa A, Sayed GI, Elfattah MA, Hassanien AE (2016) Nature inspired optimization algorithms for ct liver segmentation. *Medical imaging in clinical applications: algorithmic and computer-based approaches*. Springer, Cham, pp 431–460
- Ali S, Rizk M, Hamed M, Aboul-ela E, El-Rigal N, Aly H, Abdel-Hamid A (2019) Assessment of titanium dioxide nanoparticles toxicity via oral exposure in mice: effect of dose and particle size. *Biomarkers* 24:492–498
- Arora S, Anand P (2018) Binary butterfly optimization approaches for feature selection. *Expert Syst Appl* 116:1–20
- Asmatulu E, Andalib MN, Subeshan B, Abedin F (2022) Impact of nanomaterials on human health: a review. *Environ Chem Lett* 20(4):2509–2529
- Barani F, Nezamabadi-pour H (2018) Bqiabc: a new quantum-inspired artificial bee colony algorithm for binary optimization problems. *J AI Data Mining* 6(1):133–143
- Burk J, Sikk L, Burk P, Manshian B, Soenen S, Scott-Fordsmand J, Tamm T, Tamm K (2018) Fe-doped zno nanoparticle toxicity: assessment by a new generation of nanodescriptors. *Nanoscale* 10:21985–21993
- Cabral F, Santana B, Lange C, Batista B, Seabra A, Ribeiro M (2023) Pluronic f-127 hydrogels containing copper oxide nanoparticles and a nitric oxide donor to treat skin cancer. *Pharmaceutics* 15(7):1971
- Cai X, Lee A, Ji Z, Huang C, Chang C, Wang X, Liao Y, Xia T, Li R (2019) Reduction of pulmonary toxicity of metal oxide nanoparticles by phosphonate-based surface passivation. *Particle Fibre Toxicol* 14:1–11
- Celik M, Dadaser-Celik F, Dokuz A (2011) Anomaly detection in temperature data using dbscan algorithm. In 2011 International symposium on innovations in intelligent systems and applications, pp. 91–95, Istanbul, Turkey
- Chavez-Calderon A, Paraguay-Delgado F, Orrantia-Borunda E, Luna-Velasco A (2016) Size effect of sno2 nanoparticles on bacteriotoxicity and their membrane damage. *Chemosphere* 165(19):33–40
- Davidson D, Derk R, He X, Stueckle T, Cohen J, Pirela S (2016) Direct stimulation of human fibroblasts by nCeO₂ in vitro is attenuated with anamorphous silica coating. *Part Fibre Toxicol* 25:13–23
- Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 7:1–30
- Divya C, Muthuvinothini A (2015) Nanoscience and nanotechnology. *Edorium J Nanotechnol* 2:1–3
- Elreedy D, Atiya AF, Kamalov F (2023) A theoretical distribution analysis of synthetic minority oversampling technique (smote) for imbalanced learning. *Mach Learn*. <https://doi.org/10.1007/s10994-022-06296-4>
- Emary E, Zawbaa H, Hassanien A (2016) Binary ant lion approaches for feature selection. *Neurocomputing* 213:54–65
- Gamal A, Sayed GI, Darwish A, Hassanien AE (2017) A new proposed model for plant diseases monitoring based on data mining techniques. *Plant bioinformatics: decoding the phyta*. Springer, Cham, pp 179–195
- Gambhir RP, Rohiwal SS, Tiwari AP (2022) Multifunctional surface functionalized magnetic iron oxide nanoparticles for biomedical applications: a review. *Appl Surf Sci Adv* 11:100303
- Gombolay GY, Gopalan N, Bernasconi A, Nabbout R, Megerian JT, Siegel B, Hallman-Cooper J, Bhalla S, Gombolay MC (2023) Review of machine learning and artificial intelligence (ml/ai) for the pediatric neurologist. *Pediatric Neurol* 141:42–51
- Hafez A, Zawbaa H, Emary E, Mahmoud H, Hassanien A (2015) An innovative approach for feature selection based on chicken swarm optimization. In: 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR), pp. 19–24
- Hameed S, Shah S, Iqbal J, Numan M, Muhammad W, Junaid M, Shah S, Khursheed R, Umer F (2019) Cannabis sativa mediated synthesis of gold nanoparticles and its biomedical properties. *Bioinspired Biomimetic Nanobiomater* 9(2):1–8
- Huk A, Izak-Nau E, Reidy B, Boyles M, Duschl A, Lynch I, Dusinska M (2014) Is the toxic potential of nanosilver dependent on its size? *Particle Fibre Toxicol* 65:1–11
- Ijaz M, Alfian G, Syafrudin M, Rhee J (2018) Hybrid prediction model for type 2 diabetes and hypertension using dbscan-based outlier detection, synthetic minority over sampling technique (smote), and random forest. *Appl Sci* 8(8):13–25
- Irimi F (2022) Health and environmental safety of nanomaterials: O data, where art thou? *NanoImpact* 25:100378
- Jiang J, Xu M, Meng X, Li K (2020) Stsa: a sine tree-seed algorithm for complex continuous optimization problems. *Physica A* 537:1–19
- Khan A, Fan X, Salam A, Azhar W, Ulhassan Z, Qi J, Liaquat F, Yang S, Gan Y (2023) Melatonin-mediated resistance to copper oxide nanoparticles-induced toxicity by regulating the photosynthetic apparatus, cellular damages and antioxidant defense system in maize seedlings. *Environ Pollut* 316:120639
- Labouta H, Asgarian N, Rinker K, Cramb D (2019) Meta-analysis of nanoparticle cytotoxicity via data-mining the literature. *Am Chem Soc Nano* 13:1583–1594
- Lag M, Skuland T, Godymchuk A, Nguyen T, Pham H, Refsnes M (2018) Nanoparticle-induced cytokine responses in beas-2b and hbec3-kt cells: significance of particle size and signalling pathways in different lung cell cultures. *Basic Clin Pharmacol Toxicol* 122:620–632
- Li R, Ji Z, Chang C, Dunphy D, Cai X, Meng H (2014) Surface interactions with compartmentalized cellular phosphates explain rare earth oxide nanoparticle hazard and provide opportunities for safer design. *Am Chem Soc Nano* 8(2):1771–1783

- Loan T, Do L, Yoo H (2018) Platinum nanoparticles induce apoptosis on raw 264.7 macrophage cells. *J Nanosci Nanotechnol* 18(2):861–864
- McGee C, Storey S, Clipson N, Doyle E (2017) Soil microbial community responses to contamination with silver, aluminium oxide and silicon dioxide nanoparticles. *Ecotoxicology* 26(3):449–458
- Mirjalili S, Lewis A (2013) S-shaped versus v-shaped transfer functions for binary particle swarm optimization. *Swarm Evolut Comput* 9:1–14
- Nel A, Madler L, Velegol D, Xia T, Hoek E, Somasundaran P, Klaessig F, Castranova V, Thompson M (2009) Understanding biophysicochemical interactions at the nano-bio interface. *Nat Mater* 8(7):543–557
- Nishanth R, Jyotsna R, Schlager J, Hussain S, Reddanna P (2011) Inflammatory responses of raw 264.7 macrophages upon exposure to nanoparticles: role of ros-nf_ b signaling pathway. *Nanotoxicology* 5(4):502–516
- Osmond-McLeod M, Osmond R, Oytam Y, McCall M, Feltis B, Mackay-Sim A (2013) Surface coatings of zno nanoparticles mitigate differentially a host of transcriptional, protein and signalling responses in primary human olfactory cells. *Part Fibre Toxicol* 23:10–54
- Parashar A, Shukla A, Singh R (2020) Metal oxides nanoparticles via sol-gel method: a review on synthesis, characterization and applications. *J Mater Sci Mater Electron* 31:3729–3749
- Peco Chacon AM, García Márquez FP (2023) Support vector machine and k-fold cross-validation to detect false alarms in wind turbines. Sustainability: cases and studies in using operations research and management science methods. Springer, Cham, pp 81–97
- Sayed GI, Hassanien AE (2023) Prediction of CO₂ emission in cars using machine learning algorithms. The power of data: driving climate change with data science and artificial intelligence innovations. Springer, Cham, pp 85–97
- Sayed G, Khoriba G, Haggag M (2018) A novel chaotic salp swarm algorithm for global optimization and feature selection. *Appl Intell* 48:3462–3481
- Tharwat A, Moemen Y, Hassanien A (2017) Classification of toxicity effects of biotransformed hepatic drugs using whale optimized support vector machines. *J Biomed Inform* 68:132–149
- Too J, Rahim A (2020) Binary atom search optimisation approaches for feature selection. *Connect Sci* 32(4):406–430
- Wan Xiang, Wang Wenqian, Liu Jiming, Tong Tiejun (2014) Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med Res Methodol* 14:1–13
- Yazdipour AB, Masoorian H, Ahmadi M, Mohammadzadeh N, Ayyoubzadeh SM (2023) Predicting the toxicity of nanoparticles using artificial intelligence tools: a systematic review. *Nanotoxicology* 17(1):62–77
- Zhang H, Ji Z, Xia T, Meng H, Low-Kam C, Liu R, Pokhrel S, Lin S, Wang X, Liao Y-P et al (2012) Use of metal oxide nanoparticle band gap to develop a predictive paradigm for oxidative stress and acute pulmonary inflammation. *ACS Nano* 6(5):4349–4368

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.