A Hybrid Machine Learning Approach for Continuous Risk Management in Business Process Reengineering Projects

RAFFAK Hicham[1], LAKHOUILI Abdallah[2] , MANSOURI Moahmed[3]
Faculty of Sciences and Techniques University, Hassan 1 st Settat, Morocco[1]
Faculty of Sciences and Techniques University, Hassan 1 st Settat, Morocco[2]
National School of Applied Sciences University, Hassan 1 Berrechid, Morocco[3]

***Abstract*—This study proposes a hybrid machine learning approach for continuous risk management in Business Process Reengineering (BPR) projects. This approach combines supervised and unsupervised learning techniques, integrating feature selection and preprocessing through Principal Component Analysis (PCA), clustering with K-means, and visualization with t-SNE. The labeled data are then used as input for predictive modeling with XGBoost, optimized using Particle Swarm Optimization (PSO), Grey Wolf Optimizer (GWO), and Grid Search algorithms.PCA reduces data dimensionality, simplifying analysis and improving model performance. K-means and t-SNE are employed for data clustering and visualization, enabling the identification of risk segments and uncovering hidden patterns. XGBoost, a powerful boosting algorithm, is utilized for predictive modeling due to its efficiency, accuracy, and ability to handle missing values. Optimization techniques further enhance XGBoost's performance by fine-tuning its hyperparameters. The approach was applied to a risk database from the automotive sector, demonstrating its practical applicability. Results show that PSO achieves the lowest mean squared error (MSE) and root mean squared error (RMSE), followed by GWO and Grid Search. Mahalanobis distance yields more accurate clustering results compared to Euclidean, Manhattan, and Cosine distances. This hybrid machine learning approach significantly enhances risk detection, evaluation, and mitigation in BPR projects, offering a robust framework for proactive decision-making.**

***Keywords*—BPR; Risk management; PCA; K-means; XGBoost; PSO; GWO**

I. Introduction

In today's fast-paced and competitive business environment, organizations, companies, and enterprises consist of a series of organized and interconnected business processes and activities arranged sequentially, requiring effective and efficient management to achieve strategic objectives [1]. Business Process Management (BPM) provides a systematic approachto managing work and achieving goals [2]. Furthermore, due to the dynamic nature of business, organizations often evolve through growth, transformation, or expansion into new markets [3]. This evolution impacts business processes, which must be adjustedto align with the company's needs. Since the First Industrial Revolution, when Henry Ford introduced the assembly line, business processes have played a crucial role in managing and improving productivity [4]. Consequently, the science of processes has emerged, introducing numerous tools and techniques, such as Business Process Reengineering (BPR), as powerful methods to improve process efficiency and productivity [5]. Additionally, as dynamic components, business processes are influenced by external events and other internal processes within the same organization [6]. Thus, Business Process Management has evolved from the initial concept of Business Process Reengineering to awell-established management approach [7]. These strategies have improved the monitoring and control of efficiency, productivity, profitability, service levels, and other business objectives [8]. As companies grow, transform, or expand, the efficiency of business processes can be affected, sometimes requiring a redesign of processesto adapt to business changes [9].

With process automation and digital transformation, many manual tasks have been converted into digital platforms, suchas workflow management systems, thereby increasing productivity, efficiency, and effectiveness [10]. Automationhas provided organizations with an abundance of data and detailed records [11]. Rapid advancements in information technology, automation, and digital transformation have elevated expectations regarding the purpose of processes, even before considering their improvement or reengineering [12].

The integration of hybrid methods combining supervised and unsupervised learning offers promising prospects for continuous risk management in BPR projects. Supervised learning, which relies on labeled data, enables the creation of accurate predictive models for identifying and quantifying risks [13]. On the other hand, unsupervised learning, which does not require labeled data, excels at uncovering hidden structures and unknown patterns within the data, offering a deep understanding of potential risks. The combination of these two approaches leverages the complementary advantages of each method.

To strengthen this approach, a hybridization of unsupervised algorithms such as K-means and t-SNE with supervised algorithms like XGBoost, optimized by methods such as PSO, GWO, and Grid Search, can be used for risk management in BPR projects. K-means and t-SNE are particularly effective for clustering and visualizing data, enabling the identification of emerging risk segments and anomalies in operational data in real-time [14]. XGBoost is known for its performance in terms of precision and speed in classification and regressiontasks [15].

Optimizing XGBoost with techniques such as Particle Swarm Optimization (PSO), Grey Wolf Optimizer (GWO), and Grid Search further enhances the accuracy and robustness of predictive models [16]. PSO and GWO are particularly effective in searching for optimal solutions within complex parameter spaces, while Grid Search provides an exhaustive method for exploring possible parameter combinations.

In Section II, a literature review of the various concepts addressed in this study will be presented. Subsequently, in Section III, we will introduce the proposed approach along with the computational conditions of our model. Section IV will focus on the results and their discussion before concluding the study in Section V.

II. Literature Review

A. Business Process Re-engineering

Since its emergence in the early 1990s, Business Process Reengineering (BPR) has attracted considerable interest. Both scholars and industry professionals have extensively discussed its significance, methodologies, impacts, and success factors [9]. BPR emerged as a groundbreaking strategy aimed at fundamentally rethinking and restructuring business operations to achieve substantial improvements in metrics such as cost, quality, service, and speed [10]. Reengineering involves a radical redesign of business processes, characterized by an extensive overhaul of the organization's processes, technologies, management systems, organizational structures, and core values. The goal is to realize significant performance enhancements throughout the organization. For BPR to succeed, it must be integrated with other organizational components, leverage advanced technology, and employ various methodologies. BPR cannot thrive in isolation. Information Technology (IT) plays a critical role in BPR by providing the tools needed for exceptional organizational achievements, though its role is often misunderstood [12].

For any implementation team, the ultimate objective is to achieve a high success rate in their projects. However, the outcomes of business process reengineering initiatives have been mixed, often due to the adoption of best practices or industry benchmarks from various sectors without fully understanding the specific needs of the target industry. Notably, approximately 70% of such projects fail, largely due to the absence of an appropriate framework or methodology [8]. Nonetheless, numerous factors influence a project's outcome. These factors serve as critical indicators in predicting the project's trajectory or assessing its potential for success. BPR inherently involves risks, and its successful implementation relies on several critical success factors [17].

The successful implementation of BPR relies on several key factors, offering valuable practical insights [18]. Change management and organizational culture play a central role, emphasizing effective communication, robust motivation and reward systems, employee empowerment, continuous training and development, and a collaborative work environment. Similarly, managerial competence and support are essential, requiring strong leadership, expertise in risk management, active engagement and support from senior management, as well as an appropriate organizational structure. The BPR process itself must align seamlessly with organizational goals through strategic planning, effective project management, proper methodological application, productive consultation, and a clear BPR vision. Finally, IT capabilities are indispensable, encompassing a robust IT infrastructure, enhanced IT functionality, and the alignment of IT systems with BPR strategies to ensure successful implementation.

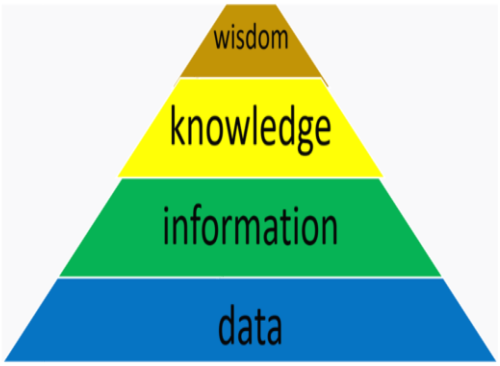

Fig. 1. Data transformation pyramid.

Agile development principles are progressively replacing traditional tools, leading to a significant transformation in engineering and management methods [19]. This evolution creates new requirements for the design and management of knowledge bases.

Data, as illustrated in Fig. 1, even in their raw state, form the foundation of layered processing systems. Their quality directly impacts their ability to generate added value, particularly in optimizing processes and improving product development [20]. An effective database should, therefore, act as a key resource to support management tools and performance indicators essential for structured and informed development [21].

However, the lack of sufficiently complete data and usable knowledge poses a major challenge. This gap hinders strategic decision-making, team coordination, and the overall optimization of product lifecycle management [22].

Data exploitation in risk management, while advantageous, faces significant challenges. The application of risk management tools such as PFMEA generates volumes of data that are often unmanageable for engineers, with heterogeneous sources and varied formats. This lack of standardization, combined with poorly structured data entry, results in duplicates and irrelevant data. Historically, risk evaluation relied on human expertise, but this approach is biased by personal experiences, leading to inconsistent and ambiguous risk identification and management across projects. This complicates the reuse of historical data and highlights the importance of data quality and standardization for effective processing [23].

B. Principal Component Analysis

Reducing the number of dimensions in large, high-dimensional datasets is crucial for effective analysis. This process can either serve as the primary objective for visualizing complex data or act as a preliminary step before further analysis, such as clustering. Principal Component Analysis (PCA) is one

of the earliest and most renowned techniques for dimensionality reduction. Initially introduced by Pearson in 1901 and later independently refined by Hotelling in 1933, where the concept of "principal components" was formally established, PCA is also known by several other names, including the Karhunen-Loeve method, eigenvector analysis, and empirical orthogonal functions. PCA remains one of the most widely used methods for creating low-dimensional representations of multivariate data (Fig. 2) [24].

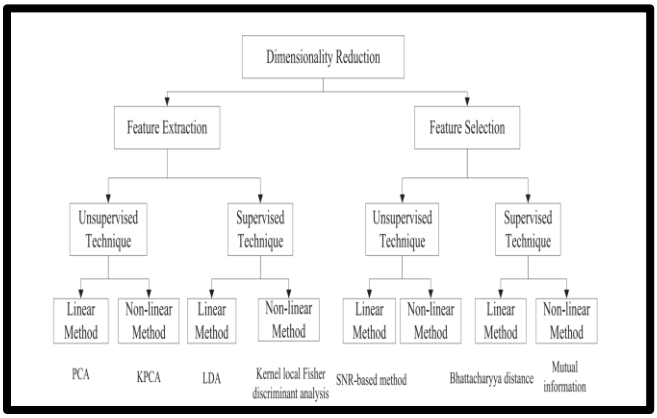

Fig. 2. Typical nomenclature of dimensionality reduction techniques.

PCA is a linear technique because it constructs components as linear combinations of the original variables (features). Despite its linearity, PCA can preserve the non-linearity of the data, making it effective for visualization purposes. The process involves iteratively calculating the direction of maximum variance and then projecting the data onto a perpendicular hyperplane. This method quickly identifies a few orthogonal directions that capture most of the data's variability, resulting in a low-dimensional representation. When all principal components are considered, the process can be visualized as a rotation in the space of the original variables. For a thorough exploration and historical context of principal component analysis, refer to [25].

C. Clustering

Clustering divides a group of individuals into several categories based on their similarities, where the differences among individuals within the same category should be as small as possible [26]. The most representative clustering methods are based on geometric distance measurement. Clustering techniques can analyze complex input data patterns and suggest solutions that might not be evident otherwise. They reveal customer typologies, enabling highly effective marketing strategies [27].

The goal of classification is to build a function or model based on the characteristics of the entire dataset and then categorize each object into a known object class. Classification has a wide range of applications, such as medical diagnosis, credit scoring, image pattern recognition, target market positioning, defect detection, efficiency analysis, graphic processing, insurance fraud analysis, [26].

Prediction involves using knowledge generated from historical and current data to deduce future data trends. While classification is used to predict classes, analysts often want to predict certain values of missing or unknown data. In other words, the desired prediction outcome corresponds to numerical data [26].

a) Positioning Euclidean distance, Manhattan distance, Mahalanobis distance and Cosine similarity: The Euclidean distance and Manhattan distance are both specific cases of the Minkowski distance. Let X and Y be two data samples, each consisting of T, elements, defined as follows:

$$X = [X_1, X_2, \cdots, X_T] \quad ; Y = [Y_1, Y_2, \cdots, Y_T] \tag{1}$$

The Minkowski distance of order p (where p is an integer) between two samples X and Y is defined by the following equation:

$$d(X,Y) = \left(\sum_{i=1}^{n} |X_i - Y_i|^p\right)^{\frac{1}{p}} \tag{2}$$

This distance metric evaluates the difference between two data samples as vectors in a multi-dimensional space, with the order p determining the emphasis on individual component differences. As p increases, larger differences in components have a more pronounced effect on the overall distance. Specifically, when p=1 and p=2, the Minkowski distance simplifies to the Manhattan distance and the Euclidean distance, respectively. As p approaches infinity, it converges to the Chebyshev distance (Fig. 3) [28].

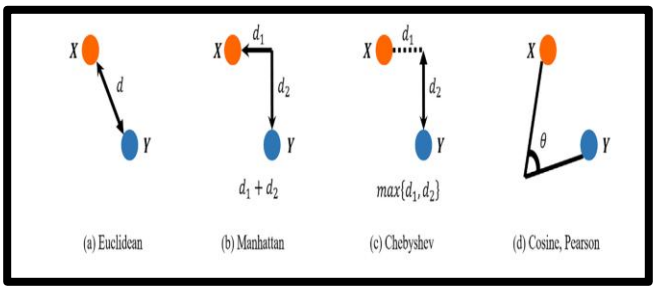

Fig. 3. Diagram of several distance measures.

The Euclidean distance is a measure of the straight-line distance between two points in Euclidean space, forming the basis of classical geometry. It is widely used as a similarity measure, particularly suitable for cases where there is no inherent correlation between different features [27]. y default, unless specified otherwise, the Euclidean distance is often employed. However, it is sensitive to the scale of the features, which can skew the results if the units are not consistent. Hence, it is generally necessary to normalize or standardize the data before applying this distance measure [28].

The Manhattan distance, also known as the city block distance or taxicab distance, calculates the distance between two points as the sum of the absolute differences of their Cartesian coordinates. The term "Manhattan distance" arises from the grid-like street layout of Manhattan, where the shortest path between two points involves a series of right-angle turns [29]. For example, in a study on personalized visual comfort control in buildings, individual user preferences and energy consumption profiles were analyzed. Collaborative user preferences were computed based on the Manhattan distance between the target occupant and others, leading to recommended adjustments in light intensity.

The Mahalanobis distance, as defined in Eq. (2), measures the distance between two vectors X and Y from the same distribution, using the covariance matrix S [30].

$$d(X,Y) = \sqrt{(X-Y)^{\top} S^{-1} (X-Y)} \quad (3)$$

where S is the covariance matrix. This distance metric extends the Euclidean distance by incorporating correlations between data points through the covariance matrix S. It is particularly effective for datasets with reduced features, although the covariance matrix can introduce unwanted redundancies. The Mahalanobis distance remains stable against projections or scaling of the data, making it useful for identifying outliers. For instance, Westermann et al. utilizedthe Mahalanobis distance to filter outliers from building energy data, classifying points farthest from the center of a multivariate Gaussian distribution as outliers [31].

Cosine similarity measures the similarity between vectors by focusing on their direction and angle rather than magnitude. The cosine similarity between two vectors X and Y is defined as:

$$cos(\theta) = \frac{\sum_{i=1}^{T} X_i Y_i}{\sqrt{\sum_{i=1}^{T} (X_i)^2} \times \sqrt{\sum_{i=1}^{T} (Y_i)^2}} \quad (4)$$

where θ is the angle between X and Y. A smaller angle indicates a higher similarity between the two vectors. The value of this equation ranges between -1 and 1. Based on this equation, the cosine distance can be defined, ranging from 0 to 2.

A study proposed a modified cosine similarity measure for initializing input weights in a building energy consumption prediction model, defined as follows [32]:

$$cos'(\theta) = \frac{\sum_{i=1}^{T} (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{T} (x_i - \bar{X})^2} \times \sqrt{\sum_{i=1}^{T} (y_i - \bar{Y})^2}} \quad (5)$$

where $\bar{X}$ and $\bar{Y}$ are the mean values of X and Y, respectively.

Instead of using Euclidean distance, which is highly sensitive to magnitudes, the modified cosine similarity coefficient is used to initialize the weights connecting the input neurons and the hidden neurons in the extreme learning machine, thereby improving its generalization ability [33].

b) K-means algorithm: K-means is one of the clustering algorithms [26]. It takes the number of clusters as a parameter and partitions the data into the specified number of clusters so that the similarity within each cluster is high [34]. K-means is an iterative approach that calculates the centroid values before each iteration. It requires precise numbers of clusters k, as the initial cluster center can change, which may lead to unstable data grouping [35]. Data points are moved between different clusters based on the centroids calculated in each iteration [36]. The process is repeated until the sum of distances cannot be decreased further. The advantages of K-means include its speed and scalability: it is one of the fastest clustering models and can efficiently handle large datasets with many records and numerous input clustering fields [26]. The K-means algorithm is presented in Algorithm 1.

Algorithm 1: K-means Algorithm.

1. Initially, based on the value of k, k random points are chosen as initial centroids.
2. The distances from each data point to the previously chosen centroids are calculated.
3. The distance values are compared, and each data point is assigned to the centroid with the shortest Euclidean distance.
4. The previous steps are repeated. The process stops if the clusters obtained are the same as those in the previous iteration.

c) Evaluation criteria: Unsupervised evaluation criteria are based on internal clustering information, such as the distance between objects within a cluster and the centroid of that cluster [37]. These criteria often rely on the simplest clustering definition, which states that objects within the same cluster should be as close as possible, and objects fromdifferent clusters should be as far apart as possible. To determine if a clustering respects this intuitive definition, distance measures are calculated between cluster representatives and residual objects. These unsupervised measures evaluate both the compactness and separability of clusters. Since the definitionof cluster quality is not formally defined, numerous criteria evaluate results differently. Some criteria are used directly as an objective function and optimized by a clustering algorithm. Others are too costly to evaluate during algorithm execution and are calculated after its application.

- Silhouette Coefficient (CS):

The silhouette coefficient evaluates the compactness of clusters and their separability [35]. It can be calculated for each object, each cluster, and the entire clustering. For an object x,it is defined as:

$$CS(x) = \frac{b_x - a_x}{max(a_x, b_x)} \quad (6)$$

where a_x is the average distance between object x and all other objects in the same cluster, and b_x is the average distance between xx and all objects not in the same cluster. The coefficient $C(x)$ ranges between -1 and 1. A positive value (a_x<b_x) indicates that objects in the same cluster as x are closer to xx than objects in other groups. For a cluster, the silhouette coefficient is the average of the coefficients of objects in that cluster:

$$CS(C_i) = \frac{1}{|C_i|} \sum_{x \in C_i} CS(x) \quad (7)$$

- Clustering Evaluation

The silhouette coefficient for clustering is equal to the average of the silhouette coefficients of its clusters:

$$CS(C) = \frac{1}{K} \sum_{i=1}^{K} CS(C_i) \quad (8)$$

The silhouette coefficient ranges between -1 and 1, with a positive value indicating that the clusters are very compact and well-separated. It should be noted that calculating this index is relatively time-consuming because many distance calculations are required for its evaluation.

D. The T-Distributed Stochastic Neighbor Embedding

Nonlinear techniques offer major advantages in processing nonlinear and complex datasets. The t-distributed stochastic neighbor embedding (t-SNE), among these techniques, has become a benchmark method for dimensionality reduction and data visualization across various fields [38]. Its applications encompass a wide range of domains, including microbiome data, single-cell RNA sequencing, bird song analysis, computational fluid dynamics, genomic data, and remote sensing images, among others. The t-SNE algorithm projects complex datasets onto a 2D or 3D space while preserving the local structure of the original high-dimensional space. However, while t-SNE excels in data visualization, it lacks an intrinsic mechanism to map new data points onto the low-dimensional representation, limiting its use in classification and regression tasks [39, 40].

E. The Extreme Gradient Boosting

It is built on gradient boosting trees, Extreme Gradient Boosting (XGBoost) is an algorithm delivering significant performance improvements over traditional gradient boosting methods. Based on the Classification and Regression Tree (CART) theory, XGBoost stands out as an effective tool for addressing regression and classification problems [41]. Furthermore, XGBoost is a flexible computing library that incorporates innovative algorithms with Gradient Boosting Decision Trees (GBDT) methods [42].

The objective function of XGBoost, post-optimization, comprises two components: one for model deviation and another regularization term to mitigate overfitting. Consider D = {(x_i,y_i)} as a dataset containing nn samples and m features, where the predictive model is an additive ensemble of k base models. The prediction for a sample is expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \varphi, \qquad \varphi = \{ f(x) = w_s(x) \mid (s: \mathbb{R}^m \to T, w_s \in \mathbb{R}^T) \} \tag{10}$$

where $\hat{y}_i$ is the predicted label for the i-th sample, x_i is the i-th sample $f_k(x_i)$ is the predicted score, and φ represents the set of regression trees. which is a tree structure parameter of s, f(x) and w representing the weight of leaves and the number of leaves. The objective function in XGBoost is a combination of the traditional loss function and a term for model complexity, described by:

$$\text{Obj} = \sum_{i=1}^{m} l(\hat{y}_i, y_i^{(t-1)}) + f_i(x_i)) + \Omega(f_k) \qquad \Omega(f_k) = \gamma T + \frac{1}{2}\lambda w^2. \tag{11}$$

In this formula, the first term, $l(\hat{y}_i, y_i)$ is the traditional loss function, and the second term, $\Omega(f_k)$, accounts for the model's complexity. Here, γ and λ are parameters used to tune the tree's complexity, helping smooth the final learning weights and preventing overfitting.

F. Optimization Techniques

a) GWO algorithm: The Grey Wolf Optimizer (GWO) is an optimization method inspired by the social hierarchy and hunting strategies of grey wolves. This algorithm mimics the leadership and cooperative behavior observed in wolf packs. The wolf pack is classified into four types of wolves [43, 44]:

Algorithm 2: GWO Algorithm

(a) Alphas: The leaders of the pack, responsible for decision-making and guiding the group.
(b) Betas: These wolves act as deputies to the alphas, assisting in decision-making and other critical tasks.
(c) Deltas: Subordinate to both alphas and betas, these wolves still hold authority over omegas and include roles such as scouts, sentinels, elders, hunters, and caretakers.
(d) Omegas: The lowest ranking members of the pack, often serving as scapegoats, and must submit to all other wolves in the hierarchy.

b) PSO algorithm: Particle Swarm Optimization (PSO), developed by Kennedy and Eberhart, is inspired by the study of bird flocking behavior during their search for food [45]. The algorithm updates each particle's velocity and position by considering the best position found by any particle in the swarm and each particle's personal best position within the search space. The PSO procedure encompasses five key steps:

Algorithm 3: PSO Algorithm

(1) initialization,
(2) evaluation,
(3) updating the particle's best position (Pbest),
(4) updating the global best position (Gbest),
(5) updating the particles' velocity and position. Particles adjust their trajectories based on Pbest and Gbest, progressively converging towards the optimal solution.

Furthermore, XGBoost is a flexible computing library that incorporates innovative algorithms with Gradient Boosting Decision Trees (GBDT) methods [42].

III. Proposed Approach

Fig. 4 illustrates the structure of the proposed framework for a hybrid machine learning (ML) approach aimed at risk assessment and classification in the context of an operational process reengineering project. This framework integrates feature selection and preprocessing, unsupervised learning (UL), and supervised learning (SL) paradigms as its core components.

In the preprocessing phase, the primary objectives are attribute transformation, composite attribute splitting, dimensionality reduction, and attribute rank analysis. Principal component analysis (PCA) is the pivotal tool employed for feature selection. The UL tools employed include the k-means algorithm and t-SNE for clustering and assigning target labels to the dataset. K-means identifies the number of clusters that correspond to different risk impact levels within the dataset. t-SNE provides detailed visualizations, facilitating the understanding of clusters, patterns, and relationships between risk input parameters, a key objective in data mining. Additionally, t-SNE maps each data point to a specific cluster (target class), thereby generating a target vector for the dataset. The main purpose of the UL stage is to transform the previously unlabeled risk dataset into a labeled dataset suitable for SL.

In the SL phase, XGBoost is employed to perform regression, classification, and risk forecasting using the labeled dataset obtained from the UL stage. To further enhance the model's accuracy, grid search, particle swarm optimization, and the grey wolf optimizer algorithms are used to optimize the parameters of the XGBoost model, with the results from each

method compared. The implementation of the design is carried out in three stages: data collection, description and preprocessing, rank analysis, clustering visualization, and evaluation of the overall approach.

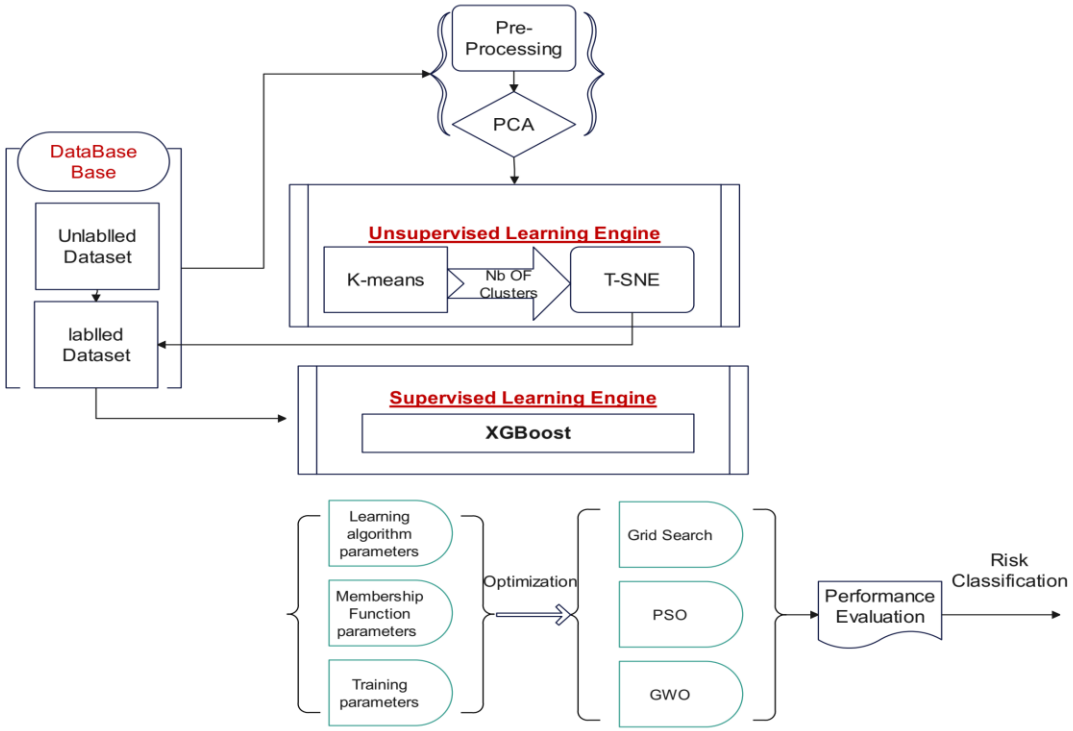

Fig. 4. Proposed approach.

IV. Results and Discussion

We applied the program to a risk database based on PFMEA register of a company operating in the automotive sector. The database consists of 11 attributes, 4 of which are numerical: the severity, occurrence, and detectability factors, and the fourth is the risk impact index, which is the RPN (Risk Priority Number), used to categorize and predict the risk.

The other categorical attributes are the risk title, owner, description, concerned part, detection tools, action to be applied, and description.

PCA is used in this code to reduce the dimensionality of the input data. The principal components capture the majority of the variance in the data while reducing the number of dimensions, which simplifies the analysis and can improve the performance of machine learning models.

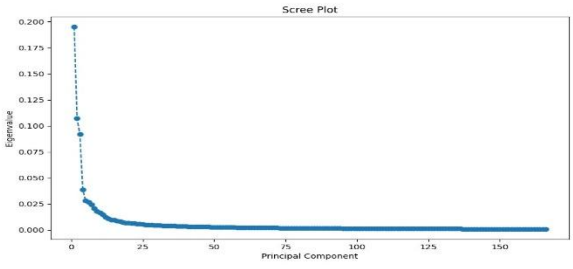

Fig. 5. Variance explained by principal components in PCA.

The Fig. 5 illustrates the eigenvalues associated with each principal component, showing the amount of variance captured by each component from the data. The graph reveals a sharp decline in eigenvalues from the first principal component to the second and third, indicating that the initial components capture most of the data's variance. After the initial drop, the eigenvalues level off, suggesting that the subsequent components contribute minimally to the total variance. This pattern generally indicates that only the first principal components are significant for explaining the dataset's variability, with the others being less important.

TABLE I. K-Means vs. Agglomerative Clustering Silhouette Scores

Algorithm	Silhouette Score
K-means	0.17712200253115506
Agglomerative Clustering	0.15189647725476427

The silhouette score is considered as a criterion for evaluating the distances between the data points and the clustering. Based on this score and by comparing the four distances: Euclidean, Manhattan, Mahalanobis, and Cosine, the Manhattan and Mahalanobis distances were shown to be more effective. However, given the advantages of the Mahalanobis distance, the authors opted for Mahalanobis distance due to its ability to consider variable correlations, which can be crucial depending on the context and nature of the data.

As shown in the Table I, when comparing the two algorithms, K-means and Agglomerative Clustering, K-means proves to be more effective with a silhouette score of 0.17712200253115506.

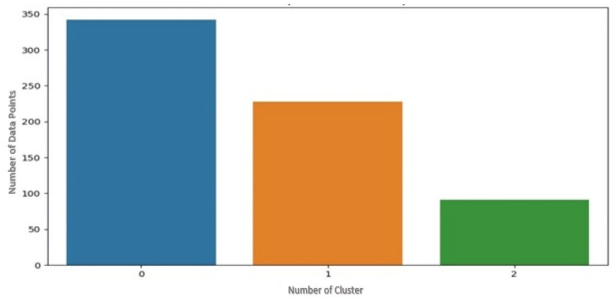

Fig. 6. Distribution of data points by cluster.

The clustering performed by the k-means algorithm identified three clusters (Fig. 6), corresponding to three levels of risk. However, based on human evaluation and as outlined in the PFMEA grid, four levels of risk are distinguished. Cluster 0 has the highest number of data points, with slightly over 300 points, indicating it is the most populated cluster. Cluster 1 follows with around 225 data points, and Cluster 2 has the fewest data points, with slightly less than 150 points. This distribution suggests that the data points are not evenly distributed among the clusters, with Cluster 0 containing the majority of the data points, Cluster 1 having a moderate amount, and Cluster 2 containing the least.

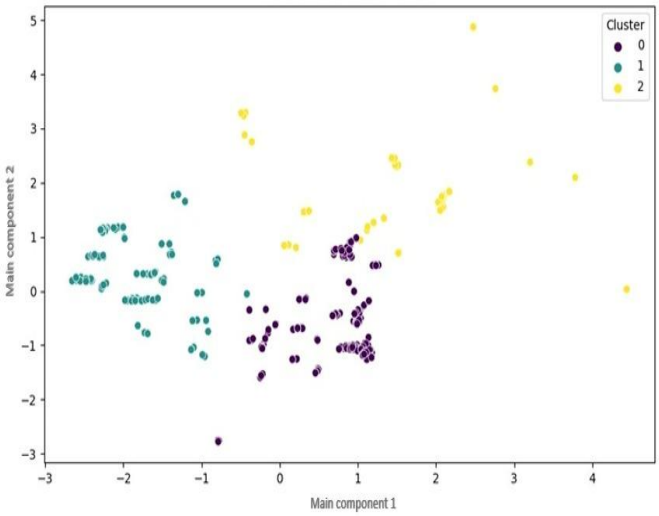

Fig. 7. Cluster visualization using ACP.

To better understand the relationship between the data, two dimensionality reduction and visualization methods were employed: PCA, a linear technique, and t-SNE, a non-linear method. For the first method (Fig. 7), although the clusters are relatively well-separated, some proximity between clusters 0 and 1 is observed.

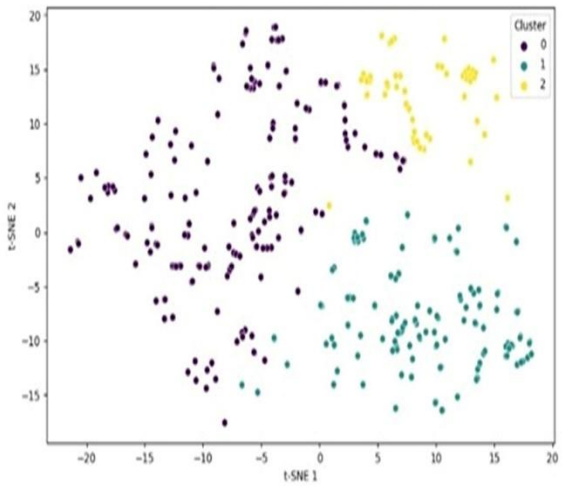

Fig. 8. Cluster visualization using t-SNE.

On the other hand, t-SNE, which leverages non-linear relationships to optimize the local representation of the data, allows for a clearer separation, particularly between clusters 0 and 2, but at the cost of less intuitive axis interpretation (Fig.8). In conclusion, PCA is suitable for data with a linear structure, while t-SNE is more effective in identifying complex relationships in non-linear data. For our dataset we adopt the t-SNE algorithm.

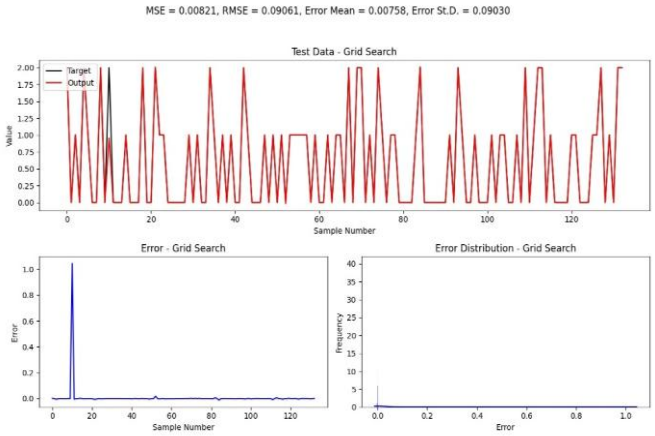

Fig. 9. Grid search optimization results: test data, error, and error distribution.

XGBoost is particularly appreciated for its performance and efficiency in machine learning competitions, as well as its ability to handle data with missing values and reduce overfitting through regularization. XGBoost is therefore a relevant choice for this type of analysis due to its robustness and ability to provide accurate results on complex datasets. In our study, we optimized XGBoost using the three techniques: GWO, PSO, and Grid Search, and evaluated the performance of our model based on two error metrics: MSE and RMSE.

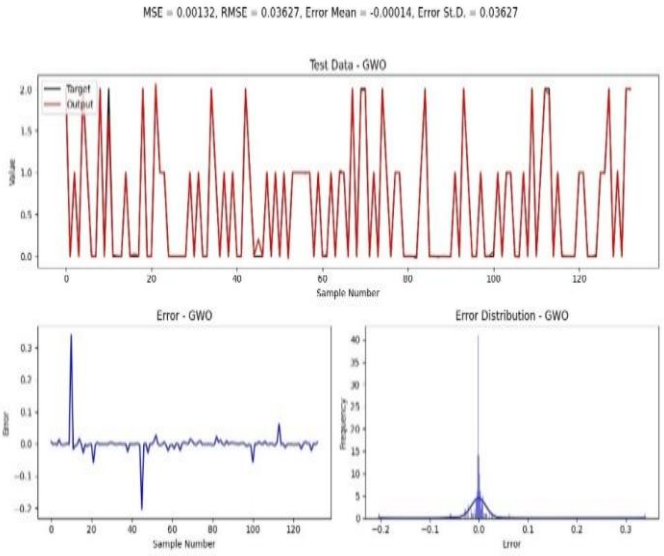

Fig. 10. GWO optimization results: test data, error, and error distribution.

The analysis of the three optimization methods—PSO (Fig. 11), GWO (Fig. 10), and Grid Search (Fig. 9)—reveals significant differences in their performance. The PSO model shows the best results with an MSE of 0.00119 and RMSE of 0.03446, indicating high accuracy as the predicted values closely follow the target values, and the error distribution is tightly centered around zero.

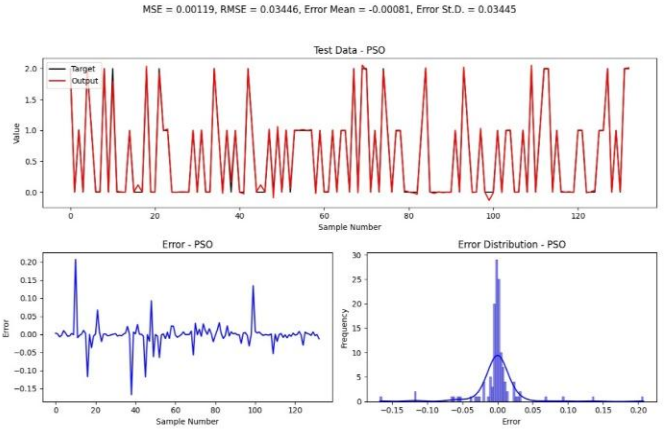

Fig. 11. PSO optimization results: test data, error, and error distribution.

The GWO model follows, with slightly higher MSE of 0.00132 and RMSE of 0.03627, showing more variation and a wider error spread compared to PSO, but still demonstrating good performance. The Grid Search model exhibits the highest error metrics MSE of 0.00821 and RMSE of 0.09061, with significant deviations and a broad error distribution, indicating poor alignment between predicted and target values and the least consistent performance among the three. Therefore, PSO is the most effective optimization method, followed by GWO, while Grid Search is the least effective.

V. Conclusion

The approach presented in this study, designed for continuous risk management in BPR projects, offers a comprehensive method that combines supervised and unsupervised machine learning techniques for effective risk management. The study present the dynamic nature of business processes and the importance of adopting strategies tailored to organizational changes. This hybrid approach leverages the strengths of unsupervised learning to uncover hidden relationships within the data and supervised learning for predictive modeling.

Algorithms such as K-means and t-SNE for clustering and visualization, and XGBoost for classification and regression, in this approach, aims to provide a robust framework for risk assessment. The optimization of XGBoost using advanced techniques like PSO, GWO and Grid Search further enhances the model's accuracy and reliability.

The article underscores the importance of using Mahalanobis distance due to its ability to consider variable correlations, which is crucial for accurate risk assessment. The authors applied their methodology to a real-world risk database in the automotive sector, demonstrating the practical applicability and effectiveness of their approach.

The results indicate that PSO outperforms other optimization methods in terms of accuracy, followed by GWO and Grid Search. The study concludes that adopting a hybrid machine learning approach can significantly improve the detection, evaluation, and mitigation of risks in BPR projects, ultimately contributing to the success of such initiatives.

In summary, the article emphasizes the value of combining supervised and unsupervised learning techniques, along with advanced optimization methods, to manage risks in BPR projects effectively. This integrated approach not only enhances predictive accuracy but also provides valuable insights into the underlying risk patterns, facilitating proactive and informed decision-making.

References

[1] G Harmon, P. Business Process Change: A Business Process Management Guide for Managers and Process Professionals; Morgan Kaufmann, Publishers: Burlington, MA, USA, 2019.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] P., Ehrlich, H.-C., Steinke, T.: ZIB structure prediction pipeline: composing a com- plex biological workflow through web services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148_1158. Springer, Heidelberg (2006). doi:10.1007/11823285_121.

[3] Nisar, Q.A., Ahmad, S. and Ahmad, U. (2014) 'Exploring factors that contribute to success of business process reengineering and impact of business process reengineering on organizational performance: a qualitative descriptive study on banking sector at Pakistan', Asian Journal of Multidisciplinary Studies, Vol. 2, No. 6, pp.219–224 http://ajms.co.in/sites/ajms/index.php/ajms/article/viewFile/405/365.

[4] Tsakalidis, G.; Vergidis, K. Towards a Comprehensive Business Process Optimization Framework. In Proceedings of the 2017 IEEE 19th Conference on Business Informatics (CBI), Thessaloniki, Greece, 24–27 July 2017; IEEE: Piscataway, NJ, USA, 2017; Volume 1, pp. 129–134. doi: 10.1109/CBI.2017.39Y.

[5] Ghanadbashi, S. and Ramsin, R. (2016) 'Towards a method engineering approach for business process reengineering', IET Software, Vol. 10, No. 2, pp.27–44, doi: 10.1049/ietsen.2014.0223.

[6] Erim, A. and Vayvay, O. (2010) 'Is the business process reengineering (BPR) proved itself to be a trustable change management approach for multinational corporations' case studies from the literature', Journal of Aeronautics and Space Technologies, Vol. 4, No. 4, pp.23–30.

[7] Bhaskar, H.L. (2016) 'A critical analysis of information technology and business process reengineering', Int. J. Productivity and Quality Management, Vol. 19, No. 1, pp.98–115. doi:10.1504/IJPQM.2016.078018.

[8] Alghamdi, H.A., Alfarhan, M.A. and Abdullah, A.L. (2014) 'BPR: evaluation of existing methodologies and limitations', International Journal of Computer Trends & Technology, Vol. 7, No. 4, pp.224–227 [online]http://www.ijcttjournal.org/Volume7/number-4/IJCTTV7P154.pdf. doi: 10.14445/22312803/IJCTT-V7P154.

[9] Bhaskar, H.L. (2018) 'Business process reengineering framework and methodology: a critical study', Int. J. Services and Operations Management, Vol. 29, No. 4, pp.527–556. doi:10.1504/IJSOM.2018.090456.

[10] Yin, G. (2010) 'BPR application', Modern Applied Science, Vol. 4, No. 4, pp.96–101. doi: http://dx.doi.org/10.5539/mas.v4n4p96.

[11] Hammer, M. and Champy, J. (1993) 'Reengineering the corporation: a manifesto for business revolution', Business Horizons, Vol. 36, No. 5, pp.90–91, ISBN: 9781857880977.

[12] Eke, G.J. and Achilike, A.N. (2014) 'Business process reengineering in organizational performance in Nigerian banking sector', Academic Journal of Interdisciplinary Studies, Vol. 3, No. 5, pp.113–124, doi: http://dx.doi.org/10.5901/ajis.2014.v3n5p113.

[13] Mlay, S.V., Zlotnikova, I. and Watundu, S. (2013) 'A quantitative analysis of business process reengineering and organizational resistance: the case of Uganda', The African Journal of Information Systems, Vol. 5, No. 1, pp.1–26.

[14] Maaten, L. v. d., & Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579-2605.

[15] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.

[16] Holland, J. H. (1992). Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. MIT Press.

[17] Jamali, G., Abbaszadeh, M.A., Ebrahimi, M. and Maleki, T. (2011) 'Business process reengineering implementation: developing a causal model of critical success factors', International Journal of e-Education, e-Business, e-Management and e Learning, Vol. 1, No. 5, pp.354–358. doi:10.7763/IJEEEE.2011.V1.58.

[18] Hicham, R., Abdallah, L., Mohamed, M. (2024). Agile Framework that Integrates Continuous Risk Management for the Implementation of BPR. In: Ezziyyani, M., Kacprzyk, J., Balas, V.E. (eds) International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD'2023). AI2SD 2023. Lecture Notes in Networks and Systems, vol 930. Springer, Cham. https://doi.org/10.1007/978-3-031-54318-0_24A. Cheriyadat, and L. M. Bruce, "Why principal Component analysis is not an Appropriate feature extractionmethod for hyperspectral data," in Proceedings of the IEEE Geosci. and remote Sens. Symp., 2003, pp. 3420–3422.

[19] Baumeister, J., Seipel, D., & Puppe, F. (2009). Agile development of rule systems. Dans Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches (1nd Edi., Vol. 1, pp. 253-272). United States of America: IGI Global.

[20] Jifa, G., & Lingling, Z. (2014). Data, DIKW, Big Data and Data Science. Procedia Computer Science, 31, 814-821. doi:10.1016/j.procs.2014.05.332.

[21] Choi, T.-M., Chan, H. K., & Yue, X. (2017). Recent Development in Big Data Analytics for Business Operations and Risk Management. IEEE Transactions on Cybernetics, 47(1), 81-92. doi: 10.1109/tcyb.2015.2507599.

[22] Zhang, Y., Ren, S., Liu, Y., & Si, S. (2017). A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products. Journal of Cleaner Production, 142, 626-644. doi: 10.1016/j.jclepro.2016.07.123.

[23] Hicham, R.; Abdallah, L.;Mohamed, M. Risk Management and Assessment Hybrid Framework for Business Process Reengineering Projects: Application in Automotive Sector. Eng 2024, 5, 1360–1381.https://doi.org/10.3390/eng5030071.

[24] X. Jia, and J. A. Richards, "Segmented principal components transformation for efficient hyperspectral remotesensing image display and classification," IEEE Trans. Geosci. Remote Sens., Vol. 37, no. 1, pp. 538–542, 1999.

[25] Cheng, Ying, Ken Chen, Hemeng Sun, Yongping Zhang, and Fei Tao. "Data and knowledge mining with big data towards smart production." Journal of Industrial Information Integration 9 (2018) : 1-13.

[26] Tsiptsis, Konstantinos K., and Antonios Chorianopoulos. "Data mining techniques in CRM : inside customer segmentation." (2011).

[27] F. Iglesias, W. Kastner, Analysis of similarity measures in times series clustering for the discovery of building energy patterns, Energies 6 (2) (2013) 579–597.

[28] Fan, C., Yan, D., Xiao, F., Li, A., An, J., & Kang, X. (2020, October). Advanced data analytics for enhancing building performances: From data-driven to big datadriven approaches. In Building Simulation (pp. 1-22). Tsinghua University Press.

[29] P. Kar, A. Shareef, A. Kumar, K.T. Harn, B. Kalluri, S.K. Panda, ReViCEE: A recommendation based approach for personalized control, visual comfort & energy efficiency in buildings, Build. Environ. 152 (2019) 135–144.

[30] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The mahalanobis distance, Chemometrics and intelligent laboratory systems 50 (1) (2000) 1–18.

[31] R. Ruiz de la Hermosa González-Carrato, Wind farm monitoring using Mahalanobis distance and fuzzy clustering, Renewable Energy 123 (2018) 526–540.

[32] P. Westermann, C. Deb, A. Schlueter, R. Evins, Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data, Appl. Energy 264 (2020) 114715, https://doi.org/10.1016/j.apenergy.2020.114715.

[33] Y. Xu, M. Zhang, L. Ye, Q. Zhu, Z. Geng, Y.L. He, Y. Han, A novel prediction intervals method integrating an error & self-feedback extreme learning machine with particle swarm optimization for energy consumption robust prediction, Energy 164 (2018) 137–146.

[34] Ao Li, Cheng Fan, Fu Xiao, Zhijie Chen, Distance measures in building informatics: An in-depth assessment through typical tasks in building energy management, Energy and Buildings, Volume 258, 2022, 111817, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2021.111817.

[35] Christy, A. Joy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa. "RFM ranking–An effective approach to customer segmentation." Journal of King Saud University-Computer and Information Sciences 33, no. 10 (2021) : 1251-1257.

[36] Syakur, M. A., B. K. Khotimah, E. M. S. Rochman, and Budi Dwi Satoto. ""Integrationk-means clustering method and elbow method for identification of the best customer profile cluster."" In IOP conference series : materials science and engineering, vol. 336, no. 1, p. 012017. IOP Publishing, 2018.

[37] M. K. Pakhira, S. Bandyopadhyay, et U. Maulik. Validity index for crisp and fuzzy clusters. Pattern Recognition, 37(3) :487 — 501, 2004.

[38] L. Kaufman et P. Rousseeuw. Finding Groups in Data An Introduction to Cluster Analysis. Wiley Interscience, New York, 1990.

[39] Maaten, L. v. d., & Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579-2605.

[40] Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S., & Kluger, Y. (2019). Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. Nature Methods, 16(3), 243-245.

[41] Wu, J., Xia, H., Cheng, Q., & Li, L. (2017). Application of Machine Learning Algorithms to Predict Central Neuropathic Pain in People with Spinal Cord Injury. Journal of Pain Research, 10, 1627-1634.

[42] Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. the 22nd ACM SIGKDD International Conference. 2016: 785-794.

[43] S. Shoghian, M. Kouzehgar, A comparison among wolf pack search and four other optimization algorithms, Int. J. Comput. Inf. Eng. 6 (2012) 1619–1624.

[44] S. Mirjalili, S.M. Mirjalili, A. Lewis, Grey wolf optimizer, Adv. Eng. Softw.69(2014)46–61, http://dx.doi.org/10.1016/J.ADVENGSOFT.2013.12.007.

[45] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Proceedings of the International Conference on Neural Networks, 1995, pp. 1942–1948.