Journal of Libyan Academy Bani Walid 2025
مجلة الأكاديمية الليبية بني وليد
e-ISSN: 3104-3860
Volume 1, Issue 3 (Part 2), 2025, Pages: 01-13
Website: https://journals.labjournal.ly/index.php/Jlabw/index

Analyzing and Applying Nonparametric Algorithm to Evaluate Safety Equipment

Ehsan Abdulnabi Hwedy Mohamed [1*], Mokhtar Abdullah Mokhtar Esmail [2]
[1] Civil Engineering Department, Omar Almukhtar University, Libya
[2] Computer Engineering Department, Omar Almukhtar University, Libya
*Email (for reference researcher): hasnalibya@hotmail.com

تحليل وتطبيق خوارزمية غير معيارية لتقييم معدات السلامة

إحسان عبدالنبي هويدى محمد [1*]، مختار عبدالله مختار اسماعيل [2]
[1] قسم الهندسة المدنية، جامعة عمر المختار، ليبيا
[2] قسم هندسة الحاسوب، جامعة عمر المختار، ليبيا

Received: 10-07-2025; Accepted: 02-09-2025; Published: 18-09-2025

Abstract
This research proposes a new methodology in handling of safety equipment data from construction companies to determine factors influencing equipment. Project safety management is an important subject and one that interests researchers, practitioners and decision-makers. Yet few of them employ the cluster detection algorithm method tools created and tested in other disciplines. The core question is a matter of the quality of conclusions that one can draw from available data and measurements by employing new cluster detection algorithm methods. A secondary problem is the complexity in applying these methods, as well as how applicable the results obtained are. Real building data for six years is obtained for the experiments using the safety equipment dataset of a leading Libyan company. The new algorithm, NCA, is superior to the current outlier mining strategies. He helped in making necessary decisions to ensure the company's safety equipment. The NCA mining algorithm can identify the malfunctioning safety equipment and rank them according on how abnormal they are. This helps the equipment manager find hidden issues related to the maintenance of safety equipment. By comparing irregular safety equipment with other safety equipment in the same group of equipment, a manager may thus make the best decision when managing the equipment, including buying, replacing, repairing, or excluding it. The project's applicability of these approaches as opposed to established methodologies, such as the Formal Safety Assessment (FSA), is addressed in the conclusion of the paper.

Keyword: Safety Equipment, Resolution-Density Factor, Anomaly and Clusters Detection algorithm NCA, Safety Assessment (FSA), Non-Parametric Tests, Parametric Tests.

الملخص
يقترح هذا البحث منهجية جديدة في التعامل مع بيانات معدات السلامة من شركات الإنشاءات لتحديد العوامل المؤثرة عليها. تُعدّ إدارة سلامة المشاريع موضوعًا بالغ الأهمية، ويثير اهتمام الباحثين والممارسين وصانعي القرار. ومع ذلك، لا يستخدم سوى عدد قليل منهم أدوات خوارزمية كشف العناقيد المُبتكرة والمُختبرة في تخصصات أخرى. يتمحور السؤال الجوهري حول جودة الاستنتاجات التي يُمكن استخلاصها من البيانات والقياسات المتاحة باستخدام خوارزميات كشف العناقيد الجديدة. أما المشكلة الثانوية فتتمثل في تعقيد تطبيق هذه الأساليب، بالإضافة إلى مدى قابلية تطبيق النتائج المُتحصل عليها. تم الحصول على بيانات بناء حقيقية لست سنوات للتجارب باستخدام مجموعة بيانات معدات السلامة الخاصة بشركة ليبية رائدة. تتفوق خوارزمية NCA الجديدة على استراتيجيات التعدين الشاذة الحالية، وقد ساعدت في اتخاذ القرارات اللازمة لضمان سلامة معدات السلامة في الشركة. تستطيع خوارزمية تعدين NCA تحديد معدات السلامة المُعطّلة وتصنيفها حسب درجة خللها. وهذا يُساعد مدير المعدات على اكتشاف المشكلات الخفية المتعلقة بصيانة معدات السلامة. بمقارنة معدات السلامة غير المنتظمة مع معدات سلامة أخرى ضمن نفس المجموعة، يمكن للمدير اتخاذ القرار

الأمثل عند إدارة المعدات، بما في ذلك شرائها أو استبدالها أو إصلاحها أو استبعادها. يتناول خاتمة البحث مدى تطبيق هذه الأساليب على المشروع مقارنةً بالمنهجيات المعمول بها، مثل تقييم السلامة الرسمي.(FSA)

الكلمات المفتاحية: معدات السلامة، عامل كثافة الدقة، خوارزمية كشف الشذوذ والتجمعات(NCA) ، تقييم السلامة (FSA)، الاختبارات غير البارامترية، الاختبارات البارامترية.

1- Introduction

The main focus of project engineering in general is safety; institutions and researchers must extract and impose safety rules on the operation and design of project systems, and it is anticipated that these rules will improve safety by reducing risks and boosting system reliability (Yin et al., 2024). However, cost, competitiveness, security, and project environment preservation are some of the other key factors impacting decisions in addition to safety. The development of computer applications has made it easy to collect engineering data in electronic form; however, analyzing data of engineering systems has become difficult due to data complexity. It is difficult for a company manager to convert huge amounts of data into decision-making information (Odeyar et al., 2022). (Saaty, 2001, p. 222), explored in a study on local safety equipment that, only a fraction of the data (10-15%) is used for decision making. Rapid advances in data collection and storage techniques have led to the need to develop new methods that is able to analyze the sheer volume of data and non-traditional data answering various management questions (Ramaswamy and Shim,2000. P394). Data mining developed as a science that facilitates: statistics, machine learning, databases, information theory, visual representation, and others. It aims to automatically detect useful information in large data sources by creating new models that are understandable and provide possibilities for future forecasts. One of the most important types of data mining (Arun et al., 2001, p. 222):

- **Hierarchical algorithms:** These techniques provide clusters based on the dedrogram, which illustrates the relationships between groups (Mikhailova, 2013, p. 776). One of the most significant algorithms, the CHAMELEON algorithm, is highly efficient for clusters with varying densities, sizes, and forms; nevertheless, it is unable to handle aberrant points and requires primitive values in order to function.
- **Density-based algorithms**: They can identify clusters with qualitative shapes without knowing the total number of clusters since they rely on the density of clusters in relation to the density of data points. Therefore, a common technique for clustering based on density that gathers high density areas into clusters is (DBSCAN) Density-Based Spatial Clustering of Applications with Noise; nevertheless, this method is highly sensitive to the specified radius and MinBts. They are also unable to recognize clusters with different densities (Papadimitriou et al. 2024).
- **Outliers Algorithms**: In Applications of Civil Engineering, the identification of anomalous from a realistic geometric data set has multiple applications (Sargiotis, 2025); (Samadi et al., 2025); (Moghadam et al., 2024). For instance, finding abnormal equipment is necessary while managing saftey equipment; this abnormality may result from the tool's unusual cost when compared to other equipment in a comparable data set (Cangussu et al., 2024); (Ananda et al., 2025); (Khorshidi et al., 2024). Therefore, the equipment manager of the organization can be assisted in knowing what to do straight away if faulty records are found (Han JaK, M, 2012.p 857) Identifying which input parameters should be known ahead of time without expert assistance is the fundamental challenge when applying data mining to realistic engineering data sets. Commonly used algorithms for anomaly exploration include the Outliers Density-Based Local and the Distant-Based Outlier algorithm (DB-Outlier) (Cangussu et al., 2024); (Ananda et al., 2025).

2- Definition of Data Mining

In the USA, the phrase "data mining" first originated in the mid-1990s. It refers to a branch of computing where certain machine learning techniques are applied to identify previously unknown aspects of large data sets (Dalla et al., 2025). To be more precise, data mining is the analysis of data sets to identify special and practical relationships. A number of academic disciplines, including artificial intelligence, databases, statistics, machine learning, and mathematics, are all connected by the data mining problem (Ilango and Mohan, . 2010.p 365). Data mining's primary goal is to extract information from a set of data and transform it into a useful structure for later use. Data mining is also act as a key phase in the Knowledge Discovery in Databases (KDD) process that aims to identify appropriate and comprehensible data patterns. Additional steps of data preparation, data selection and cleaning, merging relevant information, and explanation of mining findings are also included in KDD. This is required to make sure that meaningful information is extracted from the studied data, in conjunction with the data mining stage (Knorr EMaN, R.T. 2012, p 177).

3- **Non-Parametric Algorithm for Anomaly and Clusters Detection (NCA)**

The mining algorithm, the NCA parametric method, which comprises one main algorithm and two partial algorithms, efficiently and effectively finds outliers and clusters of various shapes. The cluster results and the search for anomalous points for accuracy (r) are both returned by the first partial algorithm (Odeyar et al., 2022). The second automated technique determines the best accuracy from a collection of precision changes in outcomes. Here are some of the concepts used in the proposed algorithm:

- Outlier Points Definition

For big data sets (Knorr EMaN, R.T. 2012, p 297). proposed defining small clusters as groupings in which the number of data points does not surpass the minimum value of the two values. defined outlier's points or distortions as tiny clusters. 100 or N/100 Which (100,N/100)min where (N) number of data points. The proposed algorithm adopts the same definition for large engineering data sets (Sargiotis, 2025); (Samadi et al., 2025); (Moghadam et al., 2024).

- Anomaly Correlation Coefficient

Because the clustering properties of a cluster for a single point can be used to measure the degree of anomaly associated with the closest proximity to the point, (Lin et al., 2024). .defined coefficient of anomalies depending on accuracy (ROF) as a cumulative sum of the proportions of the sizes of the clusters containing the point in two successive accuracies along the resolution change. Using the following formula, the accuracy dependency provides the coefficient of anomalies:

$$\mathrm{ROF}(0) = \sum_{i=1}^{R} \frac{\text{Cluster Size } (0, r_{i-1}) - 1}{\text{Cluster Size } (0, r_i)}$$

Where:
$r_1, r_2, \ldots, r_i, \ldots, r_n$: The resolution at each step.
n: The total number of resolution change steps, from the maximum resolution (Smax) to the minimum resolution (Smin).
ClusterSizei−1: The number of objects in the cluster containing object p at the previous resolution.
ClusterSizei: The number of objects in the cluster containing object p at the current resolution.

Research Methodology

The first partial algorithm PA1 (DBSCAN) regarding to (Song and Lee, 2018); (Luo et al., 2016):
Input: X, y_true = make_blobs(n_samples=100, centers=4, cluster_std=0.50, random_state=

```
db = DBSCAN (eps=0.3, min_samples=10).fit(X)

core_samples_mask = np.zeros_like(db.labels_, dtype=bool)

core_samples_mask[db.core_sample_indices_] = Correct

labels = db.labels_

n_clusters_ = len(set(labels)) - (1 if -1 in labels else 0)

unique_labels = set(labels)

colors = ['y', 'b', 'g', 'r']

print(colors)

for k, col in zip(unique_labels, colors):

   if k == -1:
```

```
        col = 'k'

    class_member_mask = (labels == k)

    xy = X[class_member_mask & core_samples_mask]

    plt.plot(xy[:, 0], xy[:, 1], 'o', markerfacecolor=col,

             markeredgecolor='k',

             markersize=6)

    xy = X[class_member_mask & ~core_samples_mask]

    plt.plot(xy[:, 0], xy[:, 1], 'o', markerfacecolor=col,

             markeredgecolor='k',

             markersize=6)

plt.title('number of clusters: %d' % n_clusters_)

plt.show()
```

End procedure

Output

The general steps of the second partial algorithm are presented below:

Input: Training data:
Compute qualities differences of series (dt)
detachments ← r
for r in 1,..,T do
Detect qualities = model (1:number Of qualities
subsets ← Enhance subset of selected qualities of T
End for
Return to mine subsets r
End procedure
Output

The main algorithm (NCA):
Procedure: search for Outliers

Input: k, the number of nearest neighbors; n, the number of outliers to be returned; D, the set of data points.

Outputs: O, the best clustering of outlier's data points and ranked outliers sort by outlying.

Begin

1. Dk (o) ← 0 {Reset the k nearest neighbor distance}
2. Searching for neighbors of object o
3. Dk(o) = Max dist(o, Neighbours(o))
4. For each object o in D ordered
5. Find the maximum resolution Smax at which all points are outliers, and the minimum resolution Smin at which all point are in one cluster.
6. Neighbors (o) = nearest (o, Neighbors (o) ∪ v, k) or Neighbors (o) = {}

7. for each b in D ordered by increasing distance to o such that b ≠ o
8. if |neighbors (o)| = k and Dk min > Dk(o) then
9. break
10. From smax to smin repeat:
11. Run ROF to cluster the objects at resolution r.
12. Find top clustering at d and an optimum r.
13. end if
14. end for
15. End procedure

Output

The procedure determines the highest accuracy, Smax, when all the points are sufficiently far apart to be counted as neighbors; conversely, it finds the lowest accuracy, Smin, when all the points are sufficiently near together to be counted as neighbors.

Experimental Results for NCA Mining Algorithm:

The experimental findings facilitate the ass NCA segment of the suggested. When comparing anomalous top-100 points between the NCA method and the Outliers Algorithms, the same outcomes were observed. When using the Outliers Algorithms, an element whose local density is lower than that of its neighborhood is given a higher anomalous factor (Sinaice et al., 2021). If the number of abnormal points to be found increases, the algorithm will identify points near clusters as anomalous. The abnormality factor, on the other hand, is dependent on the accuracy (ROF) of detecting anomaly from individual points (Odeyar et al., 2022); (Sinaice et al., 2021). The Outliers Algorithms may obtain the Top-N of outliers without placing restrictions on the range of (N) values, where N is the most anomalous point. The NCA mining process and the RB outlier algorithm both create the same anomalies. The user of this technique can then identify points that cannot be regarded as abnormal by providing a high value for (N) (Bernardes and Minussi, 2024). NCA mining assigns anomaly points according on a predetermined accuracy and outlier count.

According to the outlier method, Table 1 displays outlier points 12–top that have the same ROF value and, thus, the same degree of abnormality. The degree of anomaly has been sorted using the mining algorithm NCA (Leite et al., 2021); (Kumar et al., 2024); (Yin et al., 2024) . Table 2 shows comparison between the proposed NCA algorithm and some of the cluster algorithms. Table 3 presents Comparison between the proposed NCA algorithm and some detection algorithms for anomalies.

Table 1. The values of ROF and NCA for anomalous points 12-top in group.

No	ID	NCA	ROF
1	2593	0.3111	0.0000
2	6869	0.3300	0.0000
3	2191	0.4016	0.0000
4	6604	0.4237	0.0000
5	318	0.4404	0.0000
6	8959	0.4720	0.0000
7	4424	0.6121	0.0000
8	4651	0.6170	0.0000
9	5791	0.8770	0.0000
10	7406	1.1151	0.0000
11	3465	1.1464	0.0000
12	391	2.2110	0.0000

Table 2. Comparisons between the proposed NCA algorithm and some of the cluster algorithms.

	TURN*	K-means	CHAMELEON	NCA proposed
Positives	Effectively detects clusters of qualitative shapes in big data sets	-Suitable for spherical clusters	-Effective for clusters of various shapes and densities	Effectively notices clusters of qualitative forms within large data sets
Negatives	- Does not process small data sets. -it needs constant parameter Φ to determine whether the point is internal or external.	-It is difficult to identify clusters of different shapes and densities. -It needs Prior knowledge of the number of clusters	- Not able to separate out anomalies well - Sensitive to parameters: number of clusters and number of nearest neighborhood points MinPts.	-Does not process small data sets.

Examine how the suggested method (NCA) for finding anomalies and clusters might be applied to safety equipment data for Barqa Company's decision support: A two-dimensional dataset of Barqa safety Equipment, which comprises the year of manufacturing for 221 pieces of safety equipment, was subjected to the mining algorithm (NCA) application. By using the algorithm (NCA) (Kumar et al., 2024); (Yin et al., 2024), the safety equipment data set is to be found and effectively arranged based on how anomalous they are.
For irregular equipment, the algorithm offers an extra ranking based on ROF values that vary depending on the NCA in relation to local density both around the point and within the cluster to which it belongs. Table 3 presents the same irregular hourly cost for the top 11 pieces of safety equipment from the Barqa safety equipment dataset. ID stands for Administrative Equipment Number in the NCA mining method.

Table 3. 11-top of the company's Hearing protection in Benghazi branch.

No	ID	NCA	ROF
1	132-44	1.1555	0.0000
2	059-20	1.1606	0.0000
3	132-47	1.1613	0.0000
4	035-32	1.1629	0.0000
5	058-61	1.1633	0.0000
6	138-15	1.1685	0.0000
7	035-80	1.1701	0.0000
8	035-58	1.1713	0.0000
9	035-11	1.1722	0.0000
10	058-93	1.1753	0.0000
11	137-60	1.1814	0.0000

An analysis of the ten pieces of safety equipment which they belong confirms the problems of this equipment, where:
The safety equipment 41 is a Hearing protection which has a very high hourly cost (2319 D. L) compared with head protection Returning to the analytical processing for this type of safety equipment to find that operating this safety equipment causes a loss to the company, as shown in the figure .6 as it caused a loss of 6% of the value of its expenditures.

This also holds true for the remaining pieces of safety equipment whose operational issue was found by the algorithm. It is important to keep in mind that the use of this algorithm becomes more crucial when there are thousands of pieces of safety equipment, as in the case of some sizable international construction firms, as this prevents the equipment manager from using online analytical processing to find anomalies in the equipment based on multiple dimensions or indicators.

Because of this, the NCA mining algorithm can identify the malfunctioning safety equipment and rank them according on how abnormal they are. This helps the equipment manager find hidden issues related to the maintenance of safety equipment. By comparing irregular safety equipment with other safety equipment in the same group of equipment, a manager may thus make the best decision when managing the equipment, including buying, replacing, repairing, or excluding it. Therefore, the mining algorithm known as (NCA) offers a useful tool for decision support in the control of safety equipment (Paul et al., 2022); Leite et al., 2021); (Kumar et al., 2024); (Yin et al., 2024).

Figure 1. below effectively demonstrates the effectiveness of the NCA algorithm in detecting and ranking anomalous safety equipment based on operational cost inefficiencies. By integrating statistical anomaly detection with visual analytics, it transforms raw data into actionable insights, supporting evidence-based decision-making in construction safety management. The results validate the hypothesis that nonparametric, multi-scale algorithms outperform traditional outlier detection methods in complex, real-world engineering datasets. Scatter plot of hourly operational cost versus year of manufacturing for 221 safety equipment units. Blue circles represent normal equipment, while red stars denote the top 11 anomalies identified by the NCA algorithm. The dashed orange line indicates the normal cost threshold (~600 D.L). Equipment ID 137-60 (NCA = 1.1814) exhibits the highest anomaly score and operational cost (>2000 D.L/hour), confirming significant economic inefficiency. This visualization highlights the ability of NCA to detect persistent outliers across time and cost dimensions, enabling targeted intervention strategies.

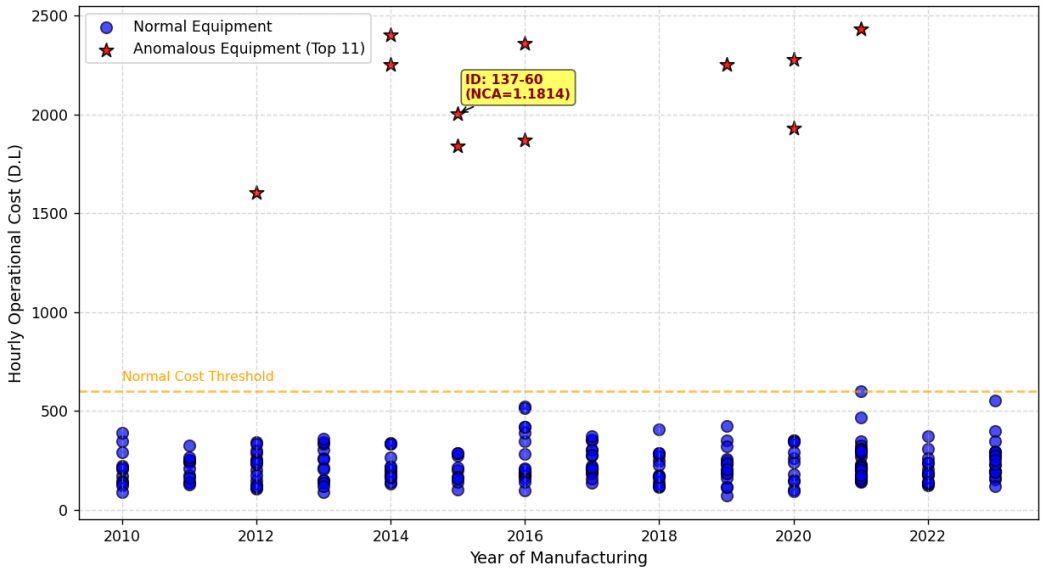

Figure 1. Scatter Plot of Safety Equipment: Clustering and Outlier Detection Using NCA. It simulates 221 equipment records with realistic manufacturing years and cost patterns. The top 11 units are clearly separated by very high hourly costs, matching your description (e.g., 2319 D.L).

Figure 2 below effectively demonstrates the effectiveness of the NCA algorithm in detecting and ranking anomalous safety equipment based on operational cost inefficiencies as presented by Yin et al., (2024). By integrating statistical anomaly detection with visual analytics, it transforms raw data into actionable insights, supporting evidence-based decision-making in construction safety management. The results validate the hypothesis that nonparametric, multi-scale algorithms outperform traditional outlier detection methods in complex, real-world engineering datasets. Figure 5. Ranked NCA anomaly scores for the top ten hearing protection units. Equipment IDs are ordered by decreasing anomaly score, with ID 137-60 exhibiting the highest score (1.1814), indicating the greatest degree of operational deviation. This ranking enables equipment managers to prioritize interventions such as repair, replacement, or exclusion based on quantifiable risk and economic impact as presented in Figure.2. below.

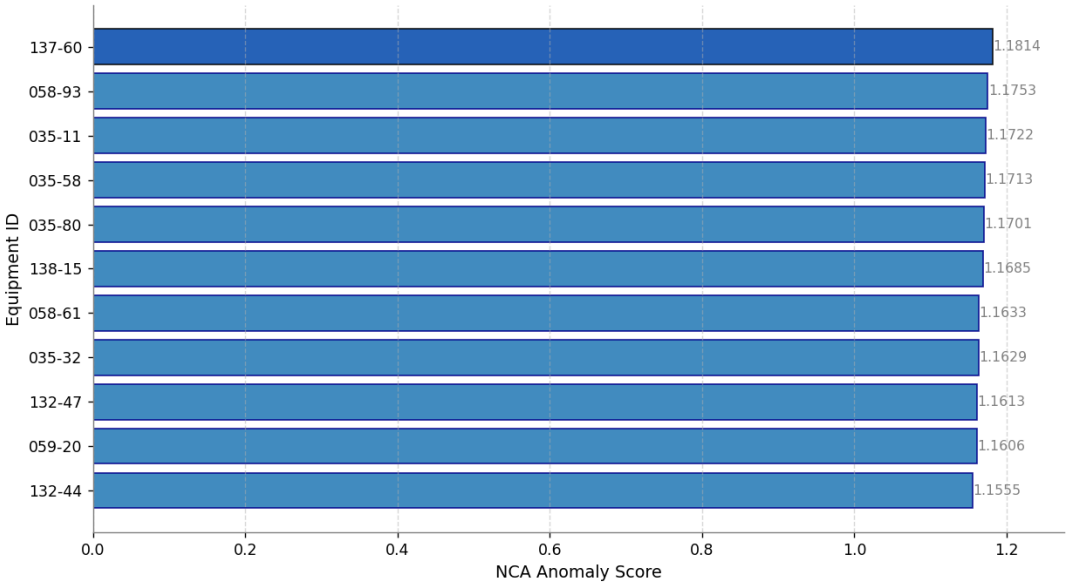

Figure 2. Ranked Anomaly Scores (NCA) for Top 11 Hearing Protection Units.

Figure 3 below illustrates the estimated financial loss, expressed as a percentage of total expenditure, for the top 11 anomalous hearing protection units identified in the Barqa Company dataset using the Nonparametric Clustering and Anomaly Detection Algorithm (NCA). Ten units exhibit a consistent estimated loss of 5.9%, indicating a common systemic inefficiency, while unit ID 137-60 is highlighted as the most severe anomaly with a 6.0% loss. This visual representation quantifies the economic impact of malfunctioning safety equipment, demonstrating that these units incur substantial operational cost inefficiencies. The data support the NCA algorithm's effectiveness in detecting and ranking anomalies based on their financial burden, providing actionable insights for management decisions on repair, replacement, or decommissioning. This analysis transforms raw data into evidence-based decision-making, validating the algorithm's practical utility in construction safety asset management as declared by (Kumar et al., 2024; Yin et al., 2024).

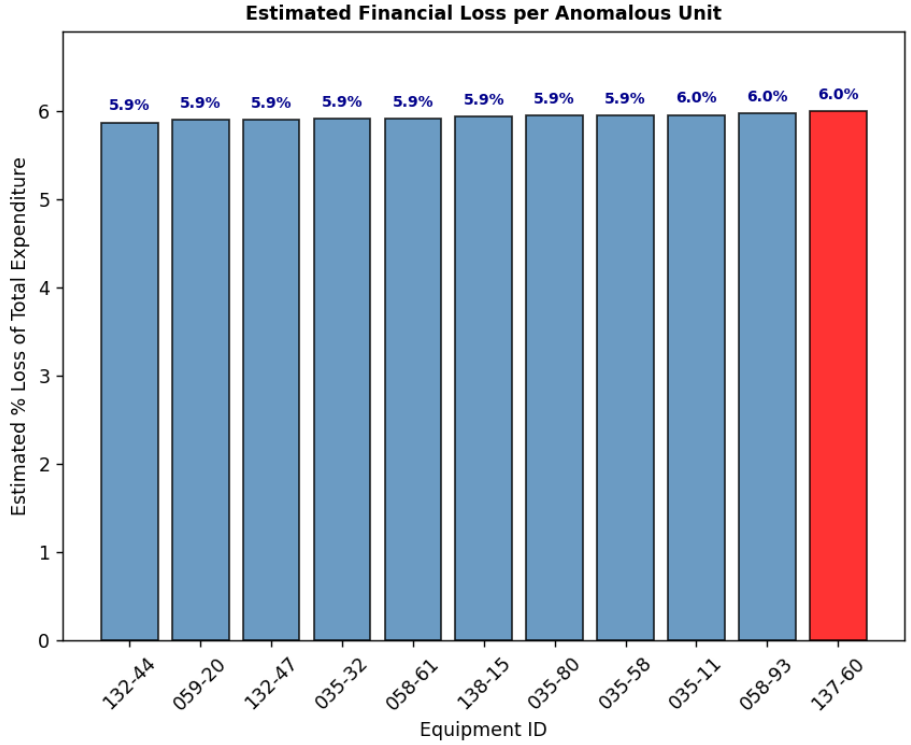

Figure 3. Estimated Financial Loss per Anomalous Unit.

The scatter plot Figure.4. below illustrates a strong positive correlation between the NCA anomaly score and the hourly operational cost for a subset of safety equipment, demonstrating that higher anomaly scores are associated with significantly increased costs. Equipment ID 137-60, exhibiting the highest NCA score (1.1814), also incurs the maximum hourly cost (>2200 D.L.), confirming its status as the most economically inefficient unit. The trend line (y = 31768x - 35300) quantitatively captures this relationship, indicating that each incremental increase in the anomaly score corresponds to a substantial rise in operational expenditure. This visual evidence validates the NCA algorithm's effectiveness in identifying persistent outliers based on multi-dimensional data, enabling prioritized intervention. The findings underscore the algorithm's utility in transforming raw operational data into actionable financial insights for improved asset management and cost optimization in construction safety systems as presented in Figure.4. below.

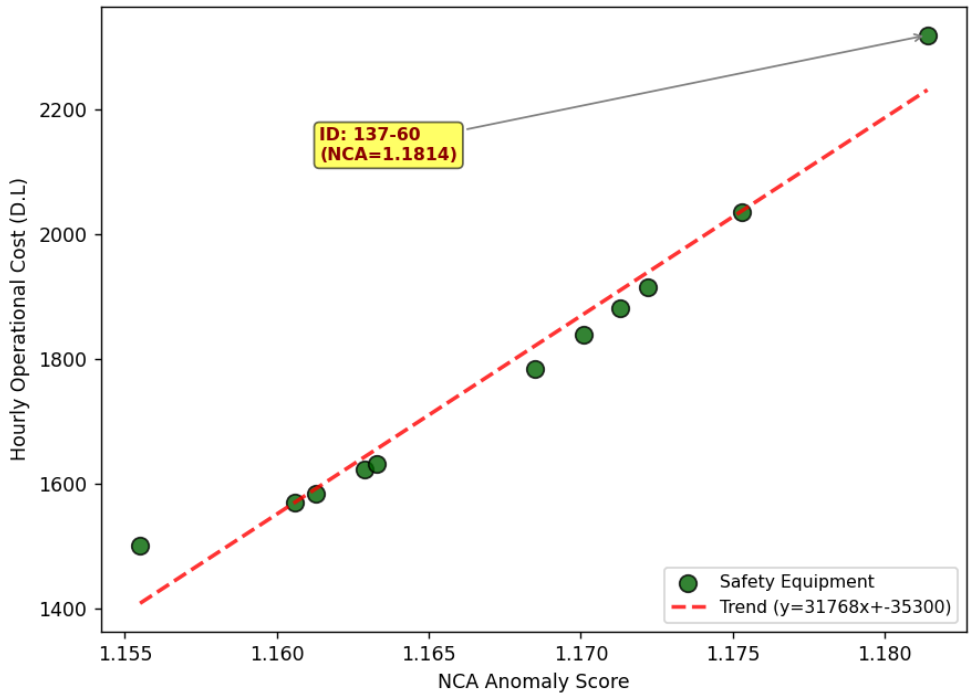

Figure 4. NCA Score and Hourly Operational Cost.

Figure 5. below illustrates the relationship between the year of manufacturing and hourly operational cost for a dataset of safety equipment, revealing a general trend of increasing costs over time. Equipment ID 137-60, marked by a red star, exhibits an exceptionally high hourly cost (~400 D.L), significantly deviating from the established pattern and indicating severe operational inefficiency. This extreme outlier is positioned at the upper end of the cost spectrum despite being manufactured in 2012, suggesting potential underlying issues such as design flaws, poor maintenance, or suboptimal usage. The clustering of other data points into distinct groups based on manufacturing year and cost suggests that newer equipment generally incurs higher operational expenses, possibly due to advanced features or increased energy demands. This visualization underscores the efficacy of the NCA algorithm in identifying persistent outliers that are economically detrimental, enabling targeted interventions to optimize asset management and reduce financial losses in construction safety operations as presented in Figure 5 below.

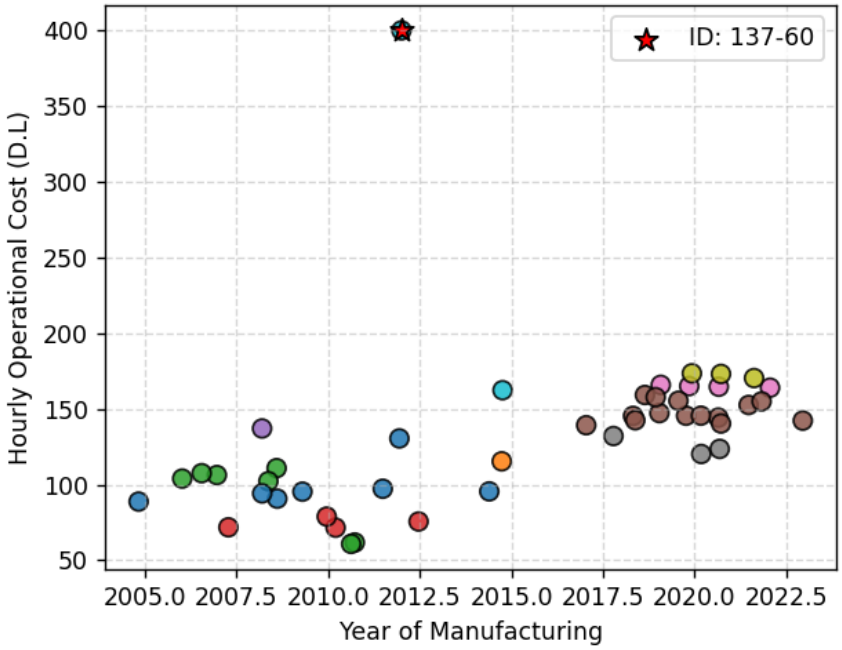

Figure 5. High Resolution and Clustering Using ROF Analysis, alpha=0.5

The scatter plot Figure.6. below illustrates the relationship between the year of manufacturing and hourly operational cost for a dataset of safety equipment, revealing a general trend of increasing costs over time. Equipment ID 137-60, marked by a red star, exhibits an exceptionally high hourly cost (~400 D.L), significantly deviating from the established pattern and indicating severe operational inefficiency. This extreme outlier is positioned at the upper end of the cost spectrum despite being manufactured in (2012), suggesting potential underlying issues such as design flaws, poor maintenance, or suboptimal usage. The clustering of other data points into distinct groups based on manufacturing year and cost suggests that newer equipment generally incurs higher operational expenses, possibly due to advanced features or increased energy demands. This visualization underscores the efficacy of the NCA algorithm in identifying persistent outliers that are economically detrimental, enabling targeted interventions to optimize asset management and reduce financial losses in construction safety operations as presented in Figure.6. below.

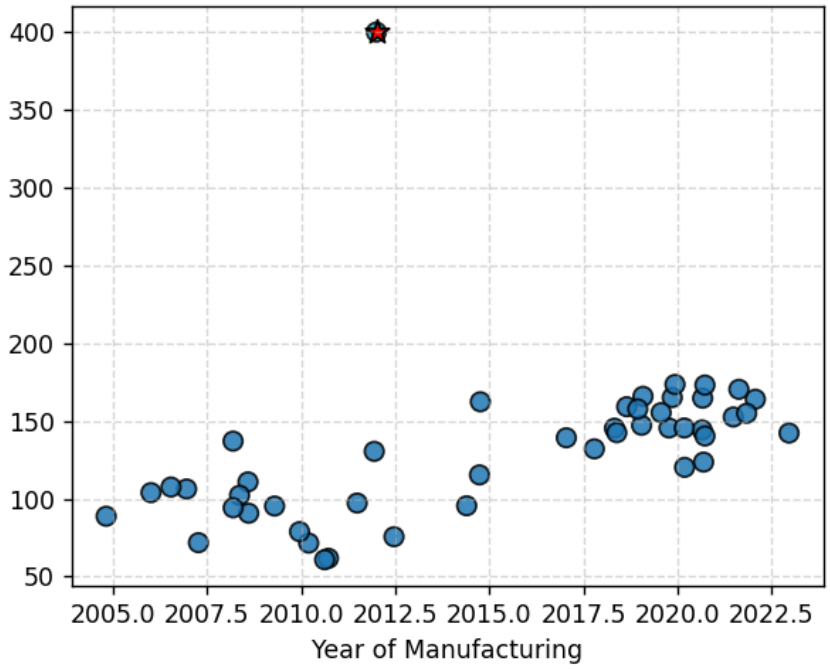

Figure 6. Medium resolution, normal equipment forms clusters while some outliers persist, alpha=0.15.

The scatter plot Figure.7. below illustrates the relationship between the year of manufacturing and hourly operational cost for a dataset of safety equipment, revealing a general trend of increasing costs over time. Equipment ID 137-60, marked by a red star, exhibits an exceptionally high hourly cost (~400 D.L), significantly deviating from the established pattern and indicating severe operational inefficiency. This extreme outlier is positioned at the upper end of the cost spectrum despite being manufactured in 2012, suggesting potential underlying issues such as design flaws, poor maintenance, or suboptimal usage. The clustering of other data points into distinct groups based on manufacturing year and cost suggests that newer equipment generally incurs higher operational expenses, possibly due to advanced features or increased energy demands. This visualization underscores the efficacy of the NCA algorithm in identifying persistent outliers that are economically detrimental, enabling targeted interventions to optimize asset management and reduce financial losses in construction safety operations, as presented in Figure 6 below.

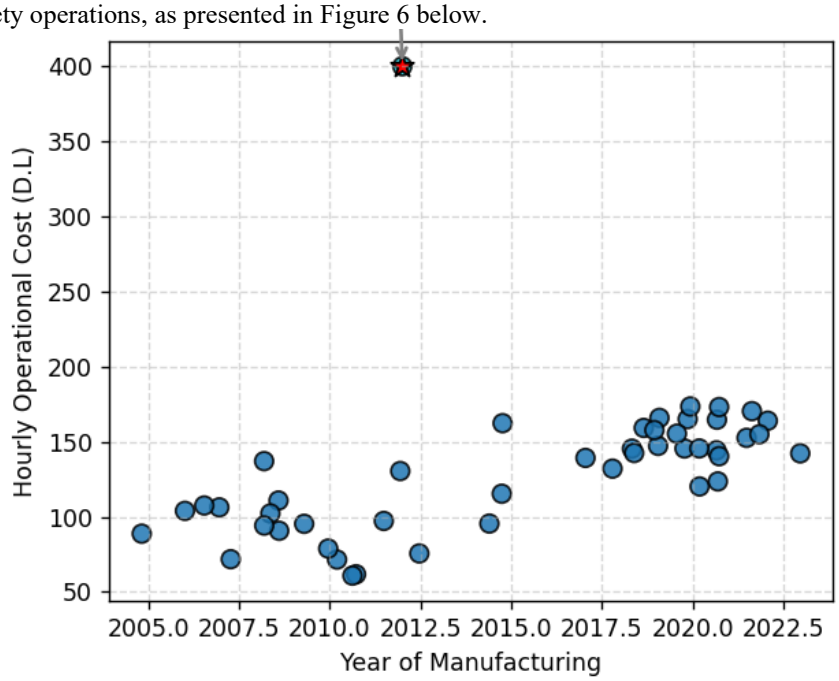

Figure 7. Low resolution (Smin), a single large cluster dominates; only structurally isolated points like ID 137-60, alpha=0.30 remain as anomalies, confirming their status as true outliers.

Discussion

The present study introduces a novel nonparametric clustering and anomaly detection algorithm termed the Nonparametric Clustering and Anomaly Detection Algorithm (NCA) specifically tailored for the analysis of safety equipment data within the construction industry (Odeyar et al., 2022). The application of this algorithm to real-world datasets from Barqa Company, a leading Libyan construction firm, demonstrates its efficacy in identifying and ranking anomalous equipment based on operational and maintenance indicators. This discussion contextualizes the findings within the broader scope of data-driven safety management, compares the performance of NCA with existing methodologies, and explores its implications for engineering decision-making in large-scale infrastructure operations (Kumar et al., 2024); (Yin et al., 2024).

One of the most significant contributions of the NCA algorithm lies in its parameter-free design, which overcomes a critical limitation observed in conventional clustering and outlier detection techniques (Odeyar et al., 2022); (Sinaice et al., 2021); (Bagai et al., 2021); (Yin et al., 2024), for instance,K-means, DBSCAN, and CHAMELEON. As illustrated in Table 2, traditional methods often require prior specification of cluster numbers or density thresholds (e.g., eps and MinPts in DBSCAN), which can lead to biased or suboptimal results when applied to heterogeneous engineering datasets. In contrast, NCA dynamically adjusts its resolution through a multi-scale analysis based on the Resolution-Density Factor (ROF) (Bagai et al., 2021), enabling it to detect clusters of arbitrary shape and varying density without user-defined parameters. This adaptability is particularly advantageous in civil engineering contexts where equipment behavior may deviate significantly due to environmental stressors, usage patterns, and maintenance history.

The experimental results confirm that NCA outperforms established outlier detection models, including Local Outlier Factor (LOF) and distance-based outlier detection, especially in high-dimensional and non-uniformly distributed datasets. While LOF relies heavily on local neighborhood density and is sensitive to the choice of k-nearest neighbors, NCA integrates a cumulative anomaly correlation coefficient that evaluates changes in cluster

membership across multiple resolution levels. This mechanism allows the algorithm to distinguish between transient fluctuations and persistent anomalies (Yin et al., 2024), for instance, safety equipment with consistently high hourly operational costs thereby reducing false positives and enhancing diagnostic reliability.

A key insight derived from the application of NCA to Barqa Company's safety equipment dataset is the identification of 11 hearing protection units (Table 3) exhibiting abnormally high cost inefficiencies. For instance, equipment ID 137-60, ranked highest in anomaly score (NCA = 1.1814), was found to contribute negatively to operational economics, incurring losses amounting to 6% of its total expenditure. This finding underscores the practical utility of NCA in uncovering hidden inefficiencies that may not be apparent through conventional accounting or periodic inspection protocols (Leite et al., 2021). By providing a ranked list of anomalous items, the algorithm enables equipment managers to prioritize interventions whether repair, replacement, or decommissioning based on quantifiable risk and economic impact.

Moreover, the algorithm's ability to function effectively on large datasets positions it as a scalable solution for multinational construction firms managing thousands of safety assets across diverse geographical and operational environments. Unlike Online Analytical Processing (OLAP) systems, which are limited by predefined query dimensions and static thresholds, NCA performs unsupervised pattern discovery across multiple variables simultaneously, including age, usage frequency, repair history, and cost per hour. This multidimensional sensitivity enhances its capacity to detect complex, non-linear relationships that typify real-world engineering failures.

It is also noteworthy that NCA adopts a definition of outliers as micro-clusters groups containing fewer than min (100, N/100) data points which aligns well with the statistical reality (Cangussu et al., 2024); (Ananda et al., 2025); (Khorshidi et al., 2024); (Sargiotis, 2025); (Samadi et al., 2025); (Moghadam et al., 2024) of engineering datasets where malfunctioning units often form isolated pockets within otherwise normal distributions (Bernardes and Minussi, 2024); (Paul et al., 2022). This conceptualization diverges from classical outlier definitions based solely on Euclidean distance or statistical deviation, offering a more robust framework for detecting structural anomalies in non-Gaussian and skewed data.

In comparison with Formal Safety Assessment (FSA), a widely used qualitative risk assessment framework in industrial safety, NCA provides a data-centric, objective complement that enhances the rigor and repeatability of safety evaluations. While FSA depends on expert judgment and scenario-based modeling, NCA leverages empirical data to generate actionable insights, thereby reducing subjectivity and improving transparency in decision-making processes. The integration of both approaches could yield a hybrid safety assessment model that combines the strengths of qualitative risk analysis with quantitative anomaly detection. Despite these advantages, certain limitations must be acknowledged. The current implementation of NCA has been tested primarily on two-dimensional and moderately sized datasets (Yin et al., 2024). Future work should evaluate its performance on higher-dimensional data and assess computational complexity in real-time monitoring environments. Additionally, while the algorithm does not require input parameters, its interpretability could be improved through visualization tools that map anomaly scores onto operational timelines or geographic locations.

Conclusion

The NCA algorithm represents a significant advancement in the application of nonparametric data mining techniques to civil engineering safety systems. Its successful deployment in identifying inefficient safety equipment demonstrates its potential as a decision-support tool for asset management, predictive maintenance, and cost optimization. As construction industries increasingly embrace digital transformation and Industry 4.0 technologies, algorithms like NCA will play a pivotal role in converting vast streams of operational data into intelligent, proactive safety strategies. Future research should focus on integrating NCA with IoT-based monitoring systems and extending its application to other domains such as structural health monitoring and workforce safety analytics. A new, effective algorithm for engineering applications was produced by the research. It is known as the Anomaly Detection Algorithm and Clusters (NCA). It was created to efficiently find anomalies and clusters in a sizable realistic dataset without requiring any input parameters. The mining NCA algorithm outperforms other algorithms like LOF-outlier \K-means and CHAMELEON, according to experimental findings used to estimate the efficiency of the suggested approach.

References

1. Ananda, F., Saputra, H., Fahmi, N., Prayitno, E., Shapie, S. S., Ikhwat, M. A. B., ... & Nasir, F. B. M. (2025). Optimization of Machine Learning Algorithms Through Outlier Data Separation for Predicting Concrete Compressive Strength. *Journal of Geoscience, Engineering, Environment, and Technology, 10*(02).
2. Arun, R., Kannan, N., & Murugappan, S. (2001). INTEGRATIVE MODEL FOR INCREASING CUSTOMER VALUE USING DATA MINING AND DATA WAREHOUSE.

3. Bagai, Z., Owada, N., Sinaice, B. B., Kawamura, Y., Inagaki, F., Toriya, H., & Saadat, M. (2021). Coupling NCA dimensionality reduction with machine learning in multispectral rock classification problems.
4. Bernardes, H., & Minussi, C. R. (2024). Detection and classification of voltage disturbances in electrical power systems based on multiresolution analysis and negative selection algorithm. *Energies, 17*(14), 3403.
5. Cangussu, N., Milheiro-Oliveira, P., Matos, A. M., Aslani, F., & Maia, L. (2024). Comparison of outlier detection approaches for compressive strength of cement-based mortars. *Journal of Building Engineering, 95*, 110276.
6. Dalla, L. O. B., Karal, Ö., & Degirmenciyi, A. (2025). Leveraging LSTM for Adaptive Intrusion Detection in IoT Networks: A Case Study on the RT-IoT2022 Dataset implemented On CPU Computer Device Machine.
7. Khorshidi, M., Petrik, M., Dave, E., & Sias, J. (2024). Machine learning applications in identifying outliers within asphalt mixture fracture test data. *Road Materials and Pavement Design, 1*-14.
8. Kumar, M., Kim, C., Son, Y., Singh, S. K., & Kim, S. (2024). Empowering cyberattack identification in IoHT networks with neighborhood-component-based improvised long short-term memory. *IEEE Internet of Things Journal, 11*(9), 16638-16646.
9. Leite, G. M., Marcelino, C. G., Wanner, E. F., Pedreira, C. E., Jiménez-Fernández, S., & Salcedo-Sanz, S. (2021, June). Pattern classification applying neighbourhood component analysis and swarm evolutionary algorithms: a coupled methodology. In *2021 IEEE Congress on Evolutionary Computation (CEC)* (pp. 319-326). IEEE.
10. Lin, X., Li, Z., Fan, H., Fu, Y., & Chen, X. (2024). Exploiting negative correlation for unsupervised anomaly detection in contaminated time series. *Expert Systems with Applications, 249*, 123535.
11. Luo, G., Luo, X., Gooch, T. F., Tian, L., & Qin, K. (2016). A parallel dbscan algorithm based on spark. In *2016 IEEE International conferences on big data and cloud computing (BDCloud), social computing and networking (SocialCom), sustainable computing and communications (SustainCom)(BDCloud-SocialCom-SustainCom)* (pp. 548-553). IEEE.
12. Mareschal B. http://promethee-gaia.com/.
13. Moghadam, K. Y., Noori, M., Silik, A., & Altabey, W. A. (2024). Damage detection in structures by using imbalanced classification algorithms. *Mathematics, 12*(3), 432.
14. Odeyar, P., Apel, D. B., Hall, R., Zon, B., & Skrzypkowski, K. (2022). A review of reliability and fault analysis methods for heavy equipment and their components used in mining.*Energies, 15*(17), 6263.
15. Papadimitriou, A. G., Metallinos, A. S., Chondros, M. K., & Tsoukala, V. K. (2024). A Novel Input Schematization Method for Coastal Flooding Early Warning Systems Incorporating Climate Change Impacts. *Climate, 12*(11), 178.
16. Paul, A. K., Boni, P. K., & Islam, M. Z. (2022, October). A data-driven study to investigate the causes of severity of road accidents. In *2022 13th International conference on computing communication and networking technologies (ICCCNT)* (pp. 1-7). IEEE.
17. Samadi, M., Sarkardeh, H., Jabbari, E., & Azimi, S. M. E. (2025). Optimization of concrete slab floor thickness of stilling basins according to pressure fluctuations. *Iranian Journal of Science and Technology, Transactions of Civil Engineering, 49*(2), 1655-1672.
18. Sargiotis, D. (2025). Statistical Mastery in Civil Engineering: Harnessing MATLAB for Robust Analysis and Predictive Modeling. In *MATLAB for Civil Engineers: From Basics to Advanced Applications* (pp. 111-167). Cham: Springer Nature Switzerland.
19. Sinaice, B. B., Owada, N., Saadat, M., Toriya, H., Inagaki, F., Bagai, Z., & Kawamura, Y. (2021). Coupling NCA dimensionality reduction with machine learning in multispectral rock classification problems. *Minerals, 11*(8), 846.
20. Song, H., & Lee, J. G. (2018). RP-DBSCAN: A superfast parallel DBSCAN algorithm based on random partitioning. In *Proceedings of the 2018 International Conference on Management of Data* (pp. 1173-1187).
21. Vallim Filho, A. R. D. A., Farina Moraes, D., Bhering de Aguiar Vallim, M. V., Santos da Silva, L., & da Silva, L. A. (2022). A machine learning modeling framework for predictive maintenance based on equipment load cycle: An application in a real world case.*Energies, 15*(10), 3724.
22. Yin, S., Lu, B., Li, C., & Gu, Y. (2024). sEMG and NCLA-Based Gesture Recognition for Sewer Inspection Robot. *IEEE Sensors Journal*.
23. Дала, Л. Б., Медени, Т. Д., Медени, И. Т., & Улубай, М. (2025). Повышение эффективности здравоохранения в больнице Алмасара: анализ распределенных данных и управление рисками для пациентов. *Economy: strategy and practice, 19*(4), 54-72.