Novel Methodology for Detecting Effective Features in Patients with Multiple Myeloma

Małgorzata Szatkowska[1], Waldemar Wójcik[1], Wojciech Legieć[2], Iryna Lurie[3], Serge Olszewski[4], Volodymyr Lytvynenko[5], Mariia Voronenko[5]

[1] Lublin University of Technology, ul. Nadbystrzycka 38D, Lublin Voivodeship, 20-618, Lublin, Poland
[2] Centrum Onkologii Ziemi Lubelskiej imienia sw Jana z Dukli, Doktora Kazimierza Jaczewskiego 7, 20-090 Lublin, Poland
[3] Ben-Gurion University of Negev, David Ben Gurion Blvd 1, Beer Sheva, 8410501, Izrael
[4] Taras Shevchenko National University of Kyiv, 64/13, Volodymyrska Street, City Kyiv, 01601, Ukraine
[5] Kherson National Technical University, Beryslavske Shose, 24, Kherson, 73008, Ukraine

Abstract

Multiple myeloma (MM) is a malignant condition characterized by the uncontrolled growth of abnormal plasma cells and the extensive destruction of bone tissue, leading to symptoms such as pain and bone fractures. This disease is caused by chromosomal abnormalities and abnormalities in the surrounding tissue microenvironment. In this study, we present a novel comprehensive technology for selecting effective lineament in a collected dataset of patients with MM and removing irrelevant lineament from this data. This research presents classical and inductive technologies based on the K-means, C-means, and Bayesian hierarchical Technology clustering (BHC) technologies. The main technology used in this study was the BHC technology, and the impact of four internal measure (silhouette, Dunn index, Calinski-Harabasz index, entropy) on clustering effectiveness was investigated. The overall use of the proposed noise elimination technique in conjunction with the inductive approach significantly improves the quality of clustering complex objects. The proposed clustering technology can be beneficial for extracting relevant lineament from the results of laboratory tests for patients with multiple myeloma in several aspects.

Keywords [1]

Multiple myeloma, lineament selection, K-means, C-means, Inductive clustering, Bayesian hierarchical Technology clustering, denoise, Data imputation, silhouette, Dunn index, Calinski-Harabasz index, entropy.

1. Introduction

With the elaboration of computer technology, their use in medical diagnostics continues to grow. Although the physical examination by a physician is still a valuable diagnostic technique, it is now standard practice to use a variety of modern diagnostic instruments and devices, especially when analyzing the results of laboratory tests such as haematological, cytological, biochemical, and immunological tests. This is especially true for extensive imaging studies such as CT and MRI, the

MoMLeT+DS 2023: 5th International Workshop on Modern Machine Learning Technologies and Data Science, June 3, 2023, Lviv, Ukraine

EMAIL: m.szatkowska@pollub.pl (M. Szatkowska); waldemar.wojcik@pollub.pl (W. Wójcik); Legiec.wojciech@gmail.com (W.Legieć); lurieira@gmail.com (I. Lurie); olszewski.serge@gmail.com (S. Olszewski); immun56@gmail.com (V. Lytvynenko); mary.voronenko@gmail.com (M. Voronenko)

ORCID: 0009-0004-5549-8464 (M. Szatkowska); 0000-0002-0843-8053 (W. Wójcik); 0000-0003-30935248 (W. Legieć); 0000-0001-8100-1846 (I. Lurie); 0000-0003-4499-8485 (S. Olszewski); 0000-0002-1536-5542 (V. Lytvynenko); 0000−0002−5392−5125 (M.Voronenko)

evaluation of which can vary depending on the radiologist's experience and working conditions (amount of tests performed, stress level, fatigue).

In many persons, the diagnosis may be too radical (over diagnosis), which may lead them to believe they have a disease of their own, or conversely, miss some lesions (hypodiagnosis). In this context, it becomes natural to look for solutions that can provide a more objective interpretation of research results. Thanks to the use of specialized technologies and advanced digital data processing techniques, the diagnosis can be established more quickly and objectively, and the role of the oncologist will be to control and verify this process.

Bone marrow cancer is a malignant disease that develops as a result of cell mutation. The pathological process is also called myeloma disease or sarcoma. During this pathology, the tissue undergoes a mutation, which stops functioning. As a result, other organs also stop working properly, which in general has a negative impact on the entire human body.

In this study, the author decided to focus on multiple myeloma [1]. Multiple myeloma (MM) is a tumor disease with uncontrolled proliferation of clonal plasma cells and extensive skeletal bone damage, accompanied by pain and bone fractures, which is caused by chromosomal abnormalities and stromal microenvironment pathology. The disease is also characterized by the attendance of monoclonal protein in the blood and/or urine.

MM accounts for 1% of all cancers and 10-13% of hematological tumors. MM accounts for 2% of deaths in all malignant tumors. The disease occurs in all countries of the world in people of all races. In Western countries and the United States, the incidence of MM is 5-10 cases per 100,000 people per year. For example, about 4,000 new cases of MM are diagnosed each year in Italy and 20,000 in the United States. The mortality rate is 4.1 cases per 100,000 person per year. The incidence among the Japanese and Chinese is much lower at 1 per 100,000 person. About 2,000 people fall ill each year and an equal amount die [1, 2]. MM is a disease of the elderly. The median age is approximately 70 years; only 37% of person are less than 65 years old by the time the disease is diagnosed. At the age of 65-74 years, 26% of people fall ill; at the age of 75 years and older, 37% fall ill. Persons at the age of 80 have the disease 10 times more often than 50-year-olds. The rate of people under 40 years old doesn't exceed 2-3%, and under 30 years old - 0.3%. Men fall ill more often (about 60%) than women. The annual incidence of MM in persons aged 65-74 years is about 31 cases per 100,000 people, and at the age of 75 years and older - up to 46 cases [1, 2]. In the future, the amount of elderly MM persons is likely to increase, which is associated with improved survival rates due to the use of new drugs and hematopoietic stem cell transplantation, as well as the increase in life expectancy of the global population as a whole.

Geographically, it varies widely in different regions of the world and is highest in industrialized areas of Australia, New Zealand, Europe and North America [3]. The incidence of MM in the United States averages 4-5 new cases and can reach 9-10 cases per 100,000 population per year among the African American population [4, 5]. In contrast, in East Asian countries, particularly in Japan, this rate is lower and does not exceed 1.2 cases per 100,000 person per year [6]. Countries with a low incidence of MM include South Korea (1.4) [7], China (1.3) [8] and Taiwan (1.8 cases per 100,000 population per year) [9]. So far, MM remains an incurable pathology, and therefore the main goal of treatment is to prolong overall survival (OS). This figure varies from country to country and depends on the quality of care provided.

The incidence of myeloma showed a strong correlation with mortality rates in countries with very low incidence rates (less than 1/100,000; $\rho = 0.95$, $p < 0.0001$), indicating a significantly shorter survival time in these countries. However, as the incidence rates increased, the correlation between incidence and mortality gradually decreased. In countries with incidence rates ranging from 1 to 3 per 100,000, the correlation coefficient dropped to 0.58, and in countries with incidence rates greater than 3 per 100,000, it further decreased to 0.36.

In the field of medicine, especially in oncology, modern data collection technologies have enabled the generation of massive datasets containing thousands or more lineament. However, the high dimensionality of these datasets poses challenges for selecting discriminatory lineament due to the curse of dimensionality. While several population-based techniques to feature selection have been proposed, few studies address the fact that there can be multiple optimal subsets of lineament for the task of selecting relevant lineament.

We propose a feature selection technique that utilizes cluster assay of lineament. This technique leverages knowledge about correlation to cluster lineament, incorporating this knowledge into the coding technique and search process. The objective is to identify different subsets of lineament that exhibit very similar or identical classification performance. In addition, we propose the use of both traditional iterative clustering techniques such as K-means, C-means, and Bayesian hierarchical Technology clustering, as well as their inductive counterparts, to further enhance the feature selection process.

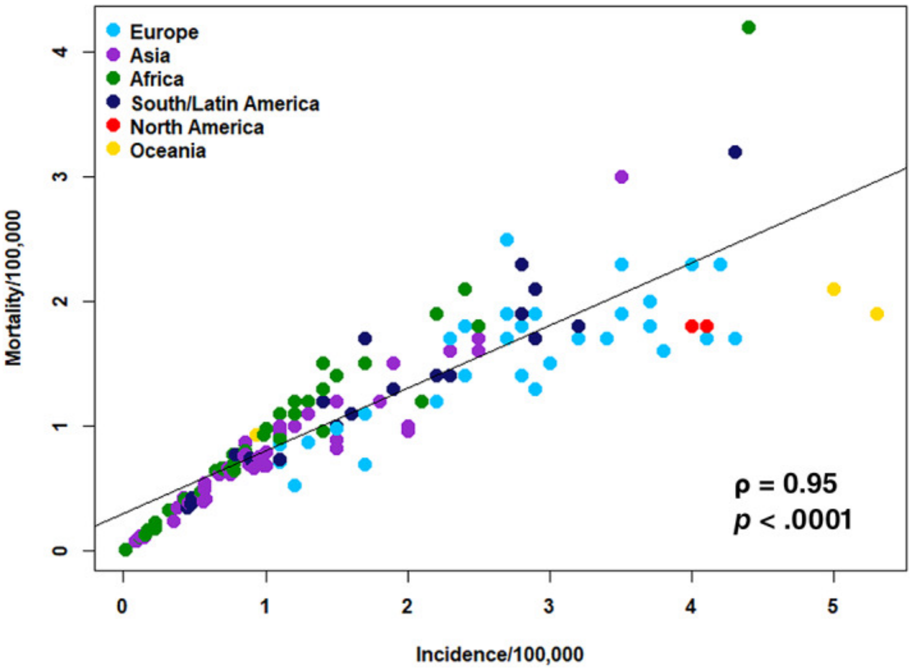


Figure 1: Relationship between mortality and morbidity in countries with a population of more than 1 million (n = 150) [9].

The main obstacles to early diagnosis of multiple myeloma are early detection and diagnosis. Therefore, it is essential that people with suspected multiple myeloma be estimated as soon as possible so that a treatment regimen can be established as quickly as possible. Otherwise, the disease will progress rapidly, leading to death.

Feature selection is a decisive preprocessing technique for reducing data dimensionality. In the field of medical diagnostics, it is essential to identify the most significant risk factors associated with a disease. By identifying the most influential lineament, unnecessary and redundant lineament can be eliminated from the disease dataset, leading to faster and more accurate results.

Prior to applying any technology to the data, it is advisable to remove noisy and inconsistent data to improve the accuracy of the results and save time. While reducing the dimensionality of the dataset is significant in real-world applications, the selection of the most significant lineament significantly decreases complexity exponentially [10].

In recent years, intelligent feature selection techniques have been widely applied to healthcare datasets to extract valuable information. Clinical databases use feature selection techniques for the assay and prediction of various chronic diseases, including diabetes, cancer, heart disease, strokes, hypertension, thalassemia, and more [11]. Given the abundance of redundant and irrelevant lineament in medical databases, an efficient feature selection technique is necessary to identify relevant lineament associated with the disease.

However, it is worth noting that non-hierarchical clustering technologies also have limitations when applied to medical data lineament.

Sensitivity to initial conditions. The results of non-hierarchical technologies can heavily depend on the chosen initial conditions or random initialization. Different runs of the technology may lead to different clusters and interpretations of results. This can complicate result in repeatability and reproducibility, especially when working with massive datasets or complex structures.

Dependency on hyperparameter selection. Non-hierarchical technologies require the selection of various hyperparameters, such as the amount of clusters or the distance metrics used. Incorrect choice of these hyperparameters can lead to misinterpretation or distortion of results. Finding optimal values for hyperparameters can be a challenging task, especially when working with medical data where explicit knowledge of the true cluster structure may be lacking.

Scalability issue. Non-hierarchical technologies can face scalability issues when processing massive medical datasets. The computational complexity of the technologies can be high, especially with a massive amount of lineament or records. This can lead to performance limitations and longer technology execution times.

Lack of hierarchical information. Unlike hierarchical technologies, non-hierarchical technologies do not preserve the hierarchical structure of clusters. This means that information about the relationships and hierarchy between clusters may be lost. This can be a imperfection if hierarchical information is significant for data interpretation or further assay.

These drawbacks need to be considered when choosing a feature clustering technique for medical data and attention should be paid to adequacy. Feature selection techniques can be such types:

- Filter techniques: These techniques analyze the intrinsic properties of data, disregarding the classifier.
- Wrapper techniques: These techniques employ classifiers to assess the performance of a given feature subset.
- Embedded techniques: These techniques integrate the feature selection process directly into the training of the classifier.

Most of these techniques can perform two primary operations: ranking and subset selection. In the ranking operation, the significance of each individual feature is estimated, typically without considering potential interactions between elements in the overall feature set. In the subset selection operation, a final subset of relevant lineament is generated. In some cases, these two operations are executed sequentially, while in others, only the selection operation is performed. Generally, subset selection is always controlled, while the level of control in ranking techniques may vary.

In this case, the investigated data lacks class labels. Conducting feature clustering in medical data is reasonable, even in the absence of class labels and well-defined target lineament. The reasons and necessity for such an approach are as follows:

Exploring data structure. Clustering allows for the exploration of data structure and the identification of internal patterns and relationships between lineament. Even without specific target labels, clustering can help identify groups of similar observations and uncover common characteristics in the data. This can be useful for generating hypotheses, understanding relationships between changeables, and informing further research.

Detecting new patient subgroups. Clustering can help discover new subgroups of patients with multiple myeloma who share common medical characteristics, even without explicit target lineament. This may lead to the discovery of new disease subtypes or conditions that could have clinical or prognostic significance. This approach can be particularly valuable in phenotype classification studies or personalized medicine research. Data preprocessing for subsequent assay. Feature clustering can serve as a stage of data preprocessing, especially in the absence of class labels. It can help decrease the dimensionality, identify the significant lineament, and prepare the data for further assay or modeling. Such an approach can contribute to improving assay efficiency, enhancing result interpretation, and reducing the influence of noise or uninformative lineament.

Risk group identification and risk factors. Clustering can help identify patient groups at an increased risk of developing specific diseases or conditions.

This can provide a foundation for further investigations into risk factors and the elaboration of personalized techniques to prevention and treatment. Overall, conducting feature clustering in medical data has several significant excellences:

- Discovery of new medical subgroups: Cluster assay can help identify patient subgroups with common medical characteristics that may indicate new medical subtypes of diseases or conditions. This can be useful for more accurate diagnosis, personalized treatment, and providing more effective and personalized healthcare.

- Revealing relationships and significant factors: Cluster assay enables the identification of hidden relationships and significant factors that may be associated with the elaboration or prediction of specific diseases. This can contribute to a deeper understanding of diseases, the identification of new risk factors or predictors, and help in the elaboration of novel techniques to prevention and treatment.
- Support for decision-making in medicine: Feature clustering can be valuable for decision-making in clinical practice. By identifying characteristics associated with specific outcomes or predictions, technologies or tools can be developed to assist doctors in making informed decisions about diagnosis, treatment, and patient management.
- Simplification and interpretation of data: Cluster assay can help decrease data dimensionality by identifying the most informative and distinguishing lineament. This can significantly simplify data assay, enhance understanding and interpretation of results, and facilitate data visualization.
- Advancement of personalized medicine: Feature clustering can serve as a foundation for the elaboration of personalized medicine, where treatment and care can be tailored to the unique characteristics and needs of each patient. This involves identifying subgroups and designating optimal treatment strategies.

Both hierarchical and non-hierarchical clustering can be used for feature clustering. However, non-hierarchical clustering technologies have several excellences over hierarchical technologies when addressing feature selection tasks in massive medical datasets:

High performance: Non-hierarchical technologies typically operate faster than hierarchical technologies, especially when processing massive datasets. This is particularly significant when working with medical data, which often contain a massive amount of lineament and records.

Flexibility and control: Non-hierarchical technologies offer more flexibility in controlling the clustering process and selecting settings according to specific needs and tasks. You can choose the amount of clusters, distance metrics used, clustering merging and splitting technologies, and more. This allows for more precise customization of the feature selection process and adaptation to specific requirements of medical research.

Better scalability: Non-hierarchical clustering technologies usually scale better when working with massive datasets. Hierarchical technologies may encounter high computational complexity and require massive memory volumes when analyzing massive medical datasets, leading to limitations and performance issues.

Better robustness to outliers: Non-hierarchical clustering technologies are typically more robust to outliers and noise in the data, as they do not construct a complete hierarchical structure. This is particularly significant in the assay of medical data, where anomalies or incorrect values may be present.

However, it should be noted that the choice between non-hierarchical and hierarchical clustering technologies depends on the specific context and task.

Hierarchical technologies are linked to the creation of dendrograms. In agglomerative technologies, each object is initially treated as a separate cluster, and these clusters are progressively merged as the technology progresses.

After all, the technology of hierarchical cluster assay is used for a limited amount of lineament and is not applicable for massive data sets due to the complexity of the agglomerative technology and excessively massive dendrograms. In iterative technologies, the data is divided into several clusters at once, the amount of which is estimated over certain measure. Then, elements are moved between clusters to optimize a specific measure, such as minimizing variability within clusters [11].

Nevertheless, iterative clustering technologies, such as k-means, exhibit a amount of limitations:

1. The guarantee of achieving the global minimum of the overall sum of squared deviations is not provided, only one of the local minima.
2. The outcome is influenced by the initial selection of cluster centers, which makes designating the optimal choice uncertain.
3. Prior knowledge of the amount of clusters is required.

One of the main drawbacks of existing iterative technologies is their high subjectivity. To enhance the objectivity of clustering, inductive techniques based on the group technique of data processing [11] can be employed. These techniques involve processing data through two equally influential subsets and making the final decision regarding object partitioning into clusters based on the combined use of outside measure for relevance and internal measure for assessing clustering quality. Therefore, developing technologies and clustering techniques based on inductive modeling techniques to solve the problem of cancer subtype identification is an significant task.

Non-hierarchical clustering technologies also have some drawbacks when conducting feature clustering in medical data:

- Sensitivity to initial conditions: The results of non-hierarchical technologies can heavily depend on the selected initial conditions or random initialization. Different technology runs can lead to different clusters and interpretation of results. This can complicate result repeatability and reproducibility, especially when working with massive datasets or complex structures.
- Dependency on the choice of hyperparameters: Non-hierarchical technologies require the selection of various hyperparameters, such as the amount of clusters or distance metrics used. Incorrect selection of these hyperparameters can lead to misinterpretation or distortion of results. Finding optimal values for hyperparameters can be a challenging task, especially when working with medical data, where explicit knowledge of the true cluster structure may be lacking.
- Scaling problem: Non-hierarchical technologies may encounter scalability issues when processing massive medical datasets. The computational complexity of technologies can be high, especially with a massive amount of lineament or records. This can result in performance limitations and increased technology execution time.
- Lack of hierarchical information: Unlike hierarchical technologies, non-hierarchical technologies do not preserve the hierarchical structure of clusters. This means that information about the relationships and hierarchy between clusters may be lost. This can be a imperfection if hierarchical information is significant for data interpretation or further assay.

These drawbacks should be taken into account when choosing a technique for clustering medical data lineament and attention should be paid to adequacy.

The main objective of this study is to present a novel comprehensive technology for selecting effective lineament in a collected dataset of multiple myeloma patient test results and removing irrelevant lineament from this data. This research presents classical and inductive technologies based on K-means, C-means, and Bayesian Hierarchical clustering technologies. Comparative studies of the presented technologies have been conducted. This approach can help identify influential lineament in the dataset. Changeables that change the cluster structure during technology realization are identified and selected as significant changeables in the dataset. Additionally, by identifying effective changeables in clustering, it is possible to select cluster labels based on the identified changeables.

2. Related Works

Currently, there are numerous techniques to feature selection, some of which are outlined below. In [12], the authors presented a new feature selection technology for symbolic attributes based on measuring the distance between feature values. In [13], the author estimated the effectiveness of feature selection techniques based on inter-class and probability distances in the preprocessing stage for constructing decision trees. This research showed that, overall, the proposed technique outperforms the use of probability measures. Researchers in [14] proposed a new feature selection technology to enhance the accuracy of classification techniques, utilizing fuzzy entropy measure for selecting relevant lineament. In [15], the authors applied fuzzy approximation operators for feature selection. The authors of [16] used a hybrid approach, combining genetic technologies and generalized regression neural networks, for selecting a subset of lineament. In [17], a novel approach to feature selection was introduced, utilizing 2,1-norm minimization and noise elimination. [18] presented a hybrid feature selection technology combining mutual information and rough sets. An ant colony optimization (CO) technology was applied for feature selection and elimination in electromyography signal classification in [19].

The application of genetic technologies and particle swarm optimization in a hybrid feature selection technique was demonstrated in [20]. [21] proposed a feature selection technology based on Case-based Reasoning (CBR) that incorporated feature selection reduction techniques and cluster assay. [22] proposed an unsupervised feature selection technology based on the Regularized Self-Representation (RSR) technique, selecting the most significant lineament for clustering and classification tasks.

A hybrid feature selection technology combining a binary quantum-inspired gravitational search technology with the k-nearest neighbor classifier was presented in [23], showing favorable results compared to other techniques. [24] proposed a hybrid feature selection technology using particle swarm optimization and correlation information. [25] introduced a combined technology based on Ant Colony Optimization (ACO) and Bee Colony Optimization (BCO) for selecting significant lineament in a dataset, demonstrating high effectiveness. Feature selection technologies based on random projections, singular value decomposition, and K-Means clustering were proposed in [26,27], where lineament were clustered using the K-Means technology.

As evident from the compilation of research in this field, various techniques have been developed using evolutionary technologies for feature selection, such as [12, 13, 14, 20, 24], but the K-means technique was not utilized. Although evolutionary techniques have computational complexity and long execution times, the technology proposed in this work can be effective due to its short operational time. In [28, 29], an integrated technology for selecting effective lineament in a dataset and removing irrelevant lineament was proposed. It was based on the K-means clustering technology. In this technique, changeables that alter the cluster structure during the technology realization are identified and selected as significant changeables in the dataset. Additionally, for designating effective changeables, cluster labels were selected based on the identified changeables. The results showed that the proposed technology achieved higher classification efficiency and could eliminate irrelevant and redundant lineament more effectively than other techniques.

A drawback of evolutionary technologies is that they require exploring numerous solution domains and spaces to achieve an optimal answer, which can be time-consuming. In contrast, the designation of solution domains in this approach is selective, and there is no need to explore the entire situation.

This work consists of five sections. Section 3 describes the problem statement. The proposed technique is presented in Section 4. Section 5 presents the experimental results, and Section 6 provides discussions. Section 6 concludes the work.

3. Problem Statement

A flowchart of the identification of experimental data obtained in the examination of persons with multiple myeloma is presented in Figure 1.

The study incorporates the rules of inductive modeling into the process of inductive clustering, which encompasses the following steps [30]:

Missing data recovery.

Normalization of the lineament of the objects under study, i.e., bringing them to the same diapason with a common feature median.

Noise removal from the data.

Division of the original dataset into two equally sized subsets.

Designation of an outside measure or set of significance measure for selecting the optimum clustering for the two equally sized subsets.

Selection or elaboration of a base clustering technology used as a component of the inductive technology for objective clustering.

The process of extracting relevant lineament consists of the following stages: data preprocessing and feature selection, techniques of identifying relevant lineament, verification of results, visualization, and description of clusters.

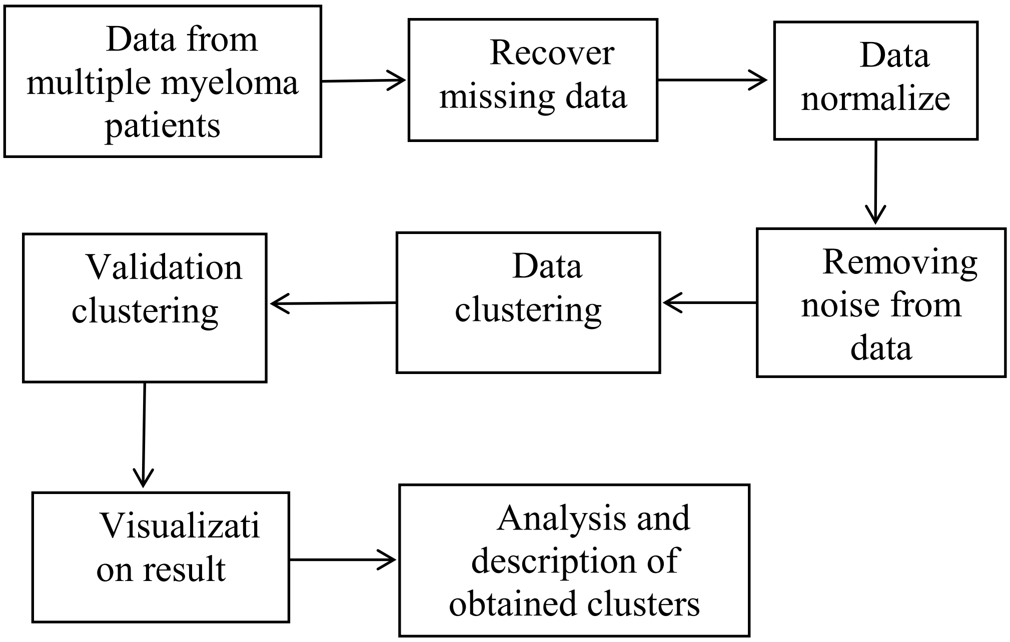


Figure 2: Procedure for Identification of Signs Obtained in the Examination of Persons with Multiple Myeloma.

In most cases, laboratory datasets are multidimensional and contain noise and missing values. This work employs six clustering techniques:

a) Classical clustering techniques: k-means and its fuzzy version, c-means, and Bayesian hierarchical clustering.

b) Inductive clustering techniques: inductive k-means, its fuzzy version, c-means, and inductive Bayesian hierarchical clustering.

The evaluation of results is done using the Dunn index, the Calinski-Harabasz index, entropy, and graphical visualization using the Silhouette index.

The goal of the work is to develop inductive technologies for feature clustering based on k-means, c-means, and hierarchical Bayesian clustering technologies, as well as to estimate the quality of the obtained results. Hierarchical technologies are associated with constructing dendrograms. In agglomerative technologies, all objects are initially considered separate clusters and are merged during the technology's execution.

The hierarchical clustering is suitable for a a bit of amount of objects and is not suitable for massive datasets due to the complexity of the agglomerative technology and the resulting massive dendrograms. In iterative clustering technologies, the data is immediately divided into multiple clusters, and the amount of clusters is estimated based on certain conditions. Then, elements are moved between clusters to optimize a specific measure, such as minimizing within-cluster variability [30].

Nevertheless, iterative clustering technologies, specifically the k-means technology, possess certain limitations:

- The global minimum of the overall sum of squares is not guaranteed to be achieved, only one of the local minima.
- The results depend on the initial selection of cluster centers, and the optimal selection is unknown.
- The amount of clusters must be known in advance.

High subjectivity is one of the key drawbacks of existing iterative technologies. Increasing the objectivity of clustering is possible by using inductive modelling techniques for complex systems based on the inductive data processing approach [31]. In this approach, data processing is performed on two equally sized subsets, and the final decision regarding the nature of object separation into clusters is made based on the combined use of outside relevance measure and internal clustering quality evaluation measure.

Each technique has its own excellences and limitations and is oriented towards specific types of data. High subjectivity is a significant obstacle in existing procedures, meaning that qualitative processing on one sample collection does not yield an equivalent level of results on another comparable dataset. Enhancing the objectivity of clustering is possible by applying inductive modelling techniques for complex systems based on the Group Technique of Data Handling (GMDH) [32, 33]. In this approach, two subsets of equal size ensure data processing, and the final judgment is based on the nature of object partitioning into clusters according to outside relevance rules and internal clustering quality assessment recommendations.

The technology of inductive clustering [31, 33, 34] obeys the rules:

1. The heuristic self-organization strategy, which involves sequentially amounting different increasingly complex candidate technologies in order to select the best technologies based on a specific outside measure or a group of recommendations for evaluating the data grouping measure.

2. The postulate of outside extension, which aims to request "fresh knowledge" for objective technology verification repeatedly.

3. The non-finality policy of decisions, involves generating not just a single result but a set of intermediate results with subsequent selection of the best ones.

Implementing these policies in a modified form serves as a prerequisite for creating an inductive framework for the objective clustering of complex data.

Therefore, the elaboration of hybrid object clustering systems based on inductive modelling techniques for complex systems is a relevant task both theoretically and practically. There are numerous clustering technologies available. Some divide a dataset into a known amount of groups, while others automatically designate the amount of clusters. One of the objectives of this work is to conduct a comparative study on the effectiveness of applying inductive clustering technology based on the K-means, C-means, and Bayesian hierarchical clustering (BHC) technologies.

Thus, the elaboration of technologies and clustering techniques for feature selection based on inductive modelling techniques to solve the problem of extracting effective lineament and removing irrelevant lineament from laboratory test results in multiple myeloma is a relevant task.

4. Materials and Techniques

4.1. Data

The data consists of a collection from 21 persons with varying stages of multiple myeloma severity. Data was gathered from 213 cytological, haematological, immunological, and biochemical tests, with a total of 525 examinations conducted during the screening process.

4.2. Missing data imputation

To address the issue of missing data, there are several techniques that can be used for imputation. Here are some of them [36, 37]:

Mean imputation: This technique involves replacing missing values with the mean value of the column. It is suitable when the data is normally distributed and there are no significant outliers. Mean imputation can be performed for the entire column or only for specific rows with missing values.

Median imputation: Similar to mean imputation, the median can be used to fill in missing values instead of the mean. The median is more robust to outliers and may be preferred if the data contains outliers or is not normally distributed.

Interpolation: Interpolation is used to fill in missing values based on neighboring values. There are various interpolation techniques available, such as linear interpolation, cubic interpolation, or nearest neighbor interpolation. The choice of interpolation technique depends on the data characteristics and context.

Regression technologies: If you have other lineament that can be used to predict the missing values, you can build a regression technology where the missing values are the dependent changeable and the other lineament are independent changeables. Then, use this technology to predict and fill in the missing values.

Multiple imputations: Multiple imputations is a statistical technique that generates multiple possible values for each missing value. These values can then be used for data assay or modelling. Multiple imputations is based on modelling and random generation of values, taking into account the relationships between changeables.

Machine learning techniques: Machine learning techniques such as random forests or gradient boosting can be used for imputing missing values. You can use other lineament in the data to train a technology and predict the missing values.

Each approach has its excellences and limitations, and the choice of technique depends on the specific dataset and the nature of the missing data.

The choice of a specific technique for filling in missing numerical data depends on several factors, such as:

Data characteristics: Examine the distribution and properties of the data. For example, if the data is heavily skewed or contains outliers, techniques that are robust to outliers may be preferable. If the data exhibits temporal dependence, techniques that account for this dependence may be more suitable.

Context of assay: Consider the purpose of the assay and the specific nature of the data. Certain techniques may be more appropriate for particular types of assay or modelling.

The proportion of missing values: If the missing values constitute a significant portion of the data, removing or imputing them with constants may lead to result distortion. In such cases, the use of machine learning techniques or multiple imputation techniques may be preferable.

Available resources: Some techniques may require greater computational resources or expert knowledge. Ensure that the chosen technique is feasible given the available resources.

Results verification: It is significant to assess how the chosen imputation technique affects the final data assay or technology. Different techniques can yield different results, and it is crucial to ensure that the imputation does not distort the final conclusions.

In general, it is recommended to conduct multiple experiments with different techniques and compare their results. This will help in selecting the most suitable technique for filling in missing numerical data based on the specific dataset and assay task.

In this study, the k-nearest neighbor technology (k-NN) was used. K-NN is an automatic object classification technique. The main rule is that an object is assigned to the class that is most common among its neighbors [38].

The neighbors are selected based on a set of objects whose classes are already known, and, using the key value k for this technique, it is designated which class is the most common among them. The k-nearest neighbor technology is based on the assumption that if objects are close in terms of n-1 properties, they are also close in terms of the n-th property.

Filling in missing values in a data table using the k-nearest neighbor technique works as follows: first, among all rows in the table, k rows that are most "similar" to the row containing the missing value are identified. The measure of "similarity" between rows (objects) is given by the Euclidean distance between rows in the column (property) space. The smaller the Euclidean distance between objects in the property space, the more "similar" they are to each other.

The column containing the predicted value is referred to as the target column. To obtain a prediction for the unknown element's value of the target property, the values of the target property from the k nearest neighbors are averaged, weighted by the inverse of the Euclidean distance to the row containing the missing value.

$$a_{ij} = \frac{\sum_{l=1}^{k} a_{lj} C_l}{\sum_{l=1}^{k} C_l}, \tag{1}$$

where C_l – weight (competence) of the l-th nearest neighbor, inversely proportional to the Cartesian distance r_{li} between l-th and j-th lines

$$r_{li} = \sum_{p \neq l} \left(a_{lp} - a_{ip} \right)^2 \tag{2}$$

Specifically, the k-nearest neighbor (k-NN) technology substitutes the missing value with the target property value of the object that is most similar to the predicted object.

The main feature that distinguishes the k-NN technique from others is the absence of a training stage in this technology. One of the main excellences of this approach is the ability to update the training dataset without retraining the classifier. This property can be useful, for example, in cases where the training dataset is frequently augmented with new data, and retraining takes too much time. The main drawback of the k-nearest neighbor technique is the time-consuming nature of the classification stage.

4.3. Removing Noise from Data

For many arrays of experimental data aimed at uncovering the relationship between diverse characteristics of the studied phenomena, the relationship between quantitative measures of similarity between two characteristics and the category amount to which a selected characteristic belongs has a fundamentally stochastic nature [40, 41]. If the category amount is represented as an ordered set of amounts, then the relationship between the quantitative similarity measure and the category amount S[n] can be viewed as a stochastic process with an uncertain probability distribution. The attendance of numerous undesignated parameters when describing such arrays using probability theory techniques gives rise to various challenges in constructing continuous predictive technologies for these processes [42]. If we treat these ordered arrays as generalized signals with noise:

$$S\left[n\right]=x\left[n\right]+\xi\left[n\right] \qquad (3)$$

where $x\left[n\right]$ is the discrete values of a smooth defining function that describes the shape of the signal, and $\xi\left[n\right]$ is the discrete values of a symmetric random process; applying standard discrete signal processing techniques to them allows for the extraction of the defining component and the construction of a predictive technology on the array $x\left[n\right]$. One of the commonly used techniques for processing noisy signals is its transformation using a moving average technology.

Moving average is a discrete sequence of data constructed by averaging several consecutive values of another discrete sequence. In our case, it is the investigated signal $S\left[n\right]$. It can be seen as a type of mathematical convolution. If we represent the original sequence as $y_1,\ldots,y_n$, then its two-sided moving average is given by the following expression:

$$z_t=\frac{1}{2k+1}\sum_{-k}^{k}y_{t+j}, \quad t=k+1,k+2,\ldots,n-k. \qquad (4)$$

Thus, $z_{k+1},\ldots,z_{n-k}$ forms a new sequence that is based on the average values of the original time series, $\left\{y_t\right\}$. Similarly, the one-sided moving average $\left\{y_t\right\}$ is given by the following expression:

$$z_t=\frac{1}{k+1}\sum_{j=0}^{k}y_{t-j}, \quad t=k+1,k+2,\ldots,n. \qquad (5)$$

Moving averages are used in two main ways:

Two-sided (weighted) moving averages are applied to filter a discrete sequence, suppressing additive noise, in order to estimate or extract the underlying trend [43, 44].

One-sided (weighted) moving averages are used as simple forecasting techniques for time series. Typically, the noisy discrete sequence of data consists of a smooth underlying trend and additive symmetric noise: $y_t=f\left(t\right)+\varepsilon_t$, where $f\left(t\right)$ is a smooth and continuous function of t, and $\left\{\varepsilon_t\right\}$ is the additive noise with zero means. In this case, the power of the additive noise significantly exceeds the power of the smooth trend. Suppressing the additive noise and estimating $f\left(t\right)$ is referred to as filtering, and the two-sided moving average is one way to accomplish this.

$$\hat{f}\left(t\right)=\frac{1}{2k+1}\sum_{j=-k}^{k}y_{t+j}, \quad t=k+1,k+2,\ldots,n-k. \qquad (6)$$

The idea behind using moving averages for filtering is that experimental data or observations, presented as an ordered sequence, are likely to be close in value. Thus, averaging points that are located near an observation provides a reasonable estimate of the trend at that observation. The moving average eliminates the stochastic component of the data, leaving the smooth trend component.

Moving averages do not allow for the estimation of $f(t)$ near the ends of a time series (in the first and last k periods). This can pose difficulties when the trend estimation is used for forecasting or analyzing the most recent data. Each average consists of $2k+1$ observations, and sometimes it is referred to as a $(2k+1)$ moving average filter or smoother. The massive the value of k, the flatter and smoother the estimate of $f(t)$ will be. A smooth estimate is usually desirable, but a flat estimate is biased, especially near peaks and troughs in $f(t)$. When $\{\varepsilon_t\}$ is a white noise sequence (i.e., independent and identically distributed with zero mean and variance σ^2, the bias is given by $E\left[\hat{f}(x)\right]-f(x)\approx\frac{1}{6}f''(x)k(k+1)$ and the variance is given by $V\left[\hat{f}\right]\approx\sigma^2/(2k+1)$. Thus, there is a trade-off between increasing bias (with massive k) and increasing variance (with smaller k).

4.4. Inductive clustering technologies

4.4.1. Normalization

The data was normalized based on its characteristics using the following formula:

$$x'_{ij}=\frac{x_{ij}-med_j}{\max\left(\left|x_{ij}-med_j\right|\right)}, \tag{7}$$

where x_{ij} is the value of the attribute i in column j, x'_{ij} is the normalized value of this attribute, med_j j is the median of column j. The choice of this normalization technique was designated by the fact that as a result, the set of data attributes in all columns had the same median with a maximum diapason of variation of attributes from -1 to 1, while the data volume for each column falling into the interquartile distance (50%) is the massive compared to other normalization techniques.

The formula used for data normalization considered the value of attribute i in column j (x_{ij}), the normalized value of this attribute (x'_{ij}), and the median of column j (med_j). This normalization technique was chosen to ensure that all data attributes across columns had the same median, with the attribute values ranging from -1 to 1. Additionally, this technique maximized the data volume within the interquartile diapason (50%) for each column, making it more favorable compared to other normalization techniques.

4.4.2. Division into Equally Spaced Sets

The technology for dividing the original set of objects Ω into 2 equally powerful disjoint subsets ΩA and ΩB consists of the following steps [45]:

- calculation of $n\cdot(n-1)2$ pairwise distances between objects in the original data sample;
- selection of a pair of objects $X_s,X_p$the distance between which is minimal:

$$d\left(X_S,X_p\right)=\min_{i,j}d\left(X_i,X_j\right) \tag{8}$$

- distribution of the object X_s into a subset Ω^A, and the object X_p into a subset Ω^B;
- repeating steps 2-3 for the remaining objects. If the amount of objects is odd, the last object is distributed into both subsets.

4.4.3. Inductive k-means technology

The k-means technology is a machine learning technology designed to solve the clustering problem. It is a non-hierarchical and iterative clustering technique that has gained popularity for its simplicity, ease of realization, and high-quality results. The technology was first developed independently by mathematicians Hugo Steinhaus [45] and Stuart Lloyd [46] in the 1950s, and it gained further attention with the publication of McQueen's work [47] in 1967.

The k-means technology is based on the expectation-maximization (EM) technology, which is also used for Gaussian mixture technologies. The main idea behind the k-means technology is to randomly assign data points to clusters and then iteratively update the cluster centroids based on the mean of the data points assigned to each cluster. In each iteration, the data points are reassigned to the cluster with the closest centroid based on a chosen distance metric.

The objective of the k-means technology is to divide a set of n observations into k clusters, where each observation is assigned to the cluster with the closest centroid based on a chosen distance metric. The aim is to create clusters that minimize the distance between each observation and its assigned cluster centroid.

Step 1. Start

Step 2. Formation of the initial set Ω of studied objects. Presentation of the data in the form of a matrix $\Omega = \{x_{ij}\};\ i = \overline{1,n};\ j = \overline{1,m}$, where n is the amount of rows or the amount of objects under investigation, m is the amount of columns or the amount of lineament characterizing the objects.

Step 3. Data preprocessing - data normalization:

- median normalization (Feature Median) is obtained by calculating the median of all data attributes:

$z_{ij} = \left(x_{ij} - med_j\right) / mad_j$

where $x_{ij}\left(z_{ij}\right)$ is the i-th observation in the j-th changeable (the i-th normalized observation in the j-th changeable), $med_j = \underset{i}{med}\left(x_{ij}\right)$ is the median for the j-th changeable, and $mad_j = \underset{i}{mad}\left(x_{ij}\right)$ is the mean absolute deviation for the j-th changeable.

- normalization using a standardized score (z-score) is a measure of the relative spread of the observed or measured value, which shows how many standard deviations is its spread of the relative average value. This is a dimensionless statistic used to compare values of different dimensions or a measurement scale.

$$z_{ij} = \frac{x_{ij} - \overline{X}}{S_{x_{ij}}} \tag{9}$$

where $\overline{X}$ is the average value, $S_{x_{ij}}$ is the standard deviation of the i-th observation in the j-th changeable. The best normalization technique depends on the data that will be normalized. Typically, the Z-score is very common to normalize the data [48].

Step 4. Dividing Ω into two equally powerful subsets in accordance with the above technology. The resulting subsets Ω^A and Ω^B can be formally represented as follows:

$$\begin{gathered} \Omega^A = \{x_{ij}^A\};\ \Omega^B = \{x_{ij}^B\}; \\ i = \overline{1,n_A = n_B};\ n_A + n_B = n;\ j = \overline{1,m} \end{gathered} \tag{10}$$

Step 5. Choosing the initial amount of clusters $k = k_{\min}$.

Step 6. Configuring the k-means clustering technology.

For each equidistant subset:

Step 7. Sequential clustering and cluster fixing.

Step 8. Calculation of the internal measure for the quality of clustering.

Silhouette: $SWC = \frac{1}{K}\sum_{j=1}^{K} S_{x_j}$

Dunn Index: $DI\left(k\right)=\min_{i\in k} s$

Calinski – Harabasz Index: $QC_{CH}=\dfrac{QCB\cdot\left(N-K\right)}{QCW\cdot\left(K-1\right)}\rightarrow \max$

Entropy: $\left[PE=\dfrac{\sum_{q=1}^{Q}\sum_{k=1}^{K}\ln\left(u_{qk}\right)}{Q}, PE\in\left[0,\ln Kk\right] \right]$

Step 9. Calculation of the outside balance measure: $ECB=\sqrt{\dfrac{\left(IC_A-IC_B\right)^2}{\left(IC_A+IC_B\right)^2}}\rightarrow opt$

Step 10. If the value of the balance measure reaches the optimum, then:

Step 11 Fixes the resulting clustering.

otherwise the amount of clusters increases by 1 and steps 5–9 are repeated

Step 12. Designating the optimal amount of clusters k_{opt} .

Step 13. Clustering data (the set Ω of objects under study), fixing the clusters.

Step 14. Validation of the results of clustering.

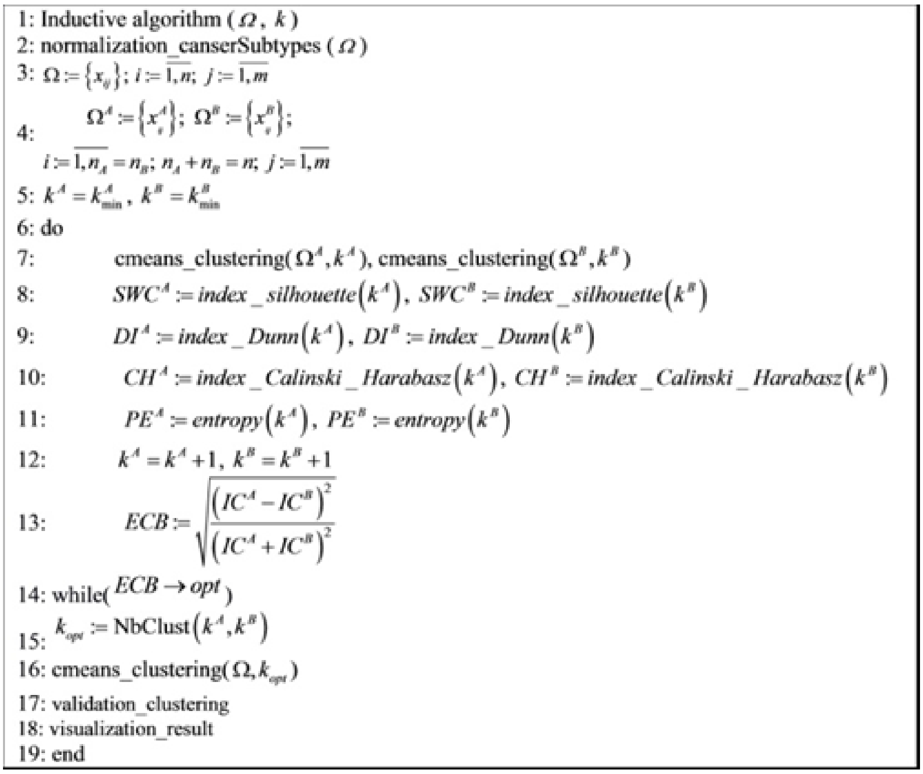
Step 15. Visualize the results of clustering.

Step 13. The End

4.4.4. Inductive Fuzzy c-Means Technology

The fuzzy c-means clustering technique (also known as fuzzy clustering, soft k-means, or c-means) is used to partition a given set of elements into a specified amount of fuzzy sets. It can be considered as an enhanced version of the k-means technology, where the degree of membership (or responsibility) of each element to each cluster is calculated.

The original c-means technology was developed in 1973 [49] and further improved in 1981 [50]. The pseudocode of the fuzzy c-means clustering technology is presented in Figure 2.



```
Inductive algorithm ( $\Omega$ , $k$ )
normalization_canserSubtypes ( $\Omega$ )
$\Omega:=\left\{x_{ij}\right\};\ i:=\overline{1,n};\ j:=\overline{1,m}$
$\Omega^A:=\left\{x_{ij}^A\right\};\ \Omega^B:=\left\{x_{ij}^B\right\};$
$i:=\overline{1,n_A}=n_B;\ n_A+n_B=n;\ j:=\overline{1,m}$
$k^A=k_{\min}^A,\ k^B=k_{\min}^B$
do
        cmeans_clustering( $\Omega^A,k^A$ ), cmeans_clustering( $\Omega^B,k^B$ )
        $SWC^A:=index\_silhouette\left(k^A\right),\ SWC^B:=index\_silhouette\left(k^B\right)$
        $DI^A:=index\_Dunn\left(k^A\right),\ DI^B:=index\_Dunn\left(k^B\right)$
        $CH^A:=index\_Calinski\_Harabasz\left(k^A\right),\ CH^B:=index\_Calinski\_Harabasz\left(k^B\right)$
        $PE^A:=entropy\left(k^A\right),\ PE^B:=entropy\left(k^B\right)$
        $k^A=k^A+1,\ k^B=k^B+1$
        $ECB:=\sqrt{\dfrac{\left(IC^A-IC^B\right)^2}{\left(IC^A+IC^B\right)^2}}$
while( $ECB\rightarrow opt$ )
$k_{opt}:=\text{NbClust}\left(k^A,k^B\right)$
cmeans_clustering( $\Omega,k_{opt}$ )
validation_clustering
visualization_result
end
```

Figure 3: Pseudocode of the C-means technology for solving feature clustering problems.

4.4.5. Inductive Bayesian Hierarchical Technology

The Bayesian hierarchical clustering (BHC) technology, proposed in [51], differs from other hierarchical clustering technologies that use fixed distance measures like Euclidean or Manhattan distances. Instead, BHC utilizes a probabilistic distance measure, where the distance represents the probability of data elements belonging to a particular cluster. This probabilistic approach is crucial during the iterative merging of clusters to form new clusters within the hierarchical structure [50].

BHC is an technology for hierarchical agglomerative clustering that employs a Bayesian probabilistic distance measure. It follows a bottom-up approach, starting with all data elements in separate clusters and iteratively merging them until fusion occurs. The merging process is guided by pre-computed probabilities using Bayes' theorem. The output is a dendrogram illustrating the hierarchical structure derived from the input dataset.

The current focus lies in addressing the challenges associated with clustering complex high-dimensional data in the attendance of high levels of noise. In this study, high-dimensional data refers to data where the dimensionality of the feature space is equal to or significantly greater than the amount of objects being analyzed. Along with high dimensionality, the data exhibit specific characteristics such as the level and specificity of the noise component, arising from biological processes or imperfections in the data generation system.

The increasing demand for accurate detection and identification systems across various conditions has led to a growing interest in extracting information from complex high-dimensional data. While numerous clustering technologies exist, each with its own excellences and imperfections, their subjectivity poses a significant drawback. Achieving high-quality clustering on one dataset does not guarantee similar results on another dataset. To enhance the objectivity of clustering, inductive techniques based on the group technique of data processing [51] can be employed. These techniques involve processing data through two equally influential subsets and making the final decision on object partitioning into clusters based on the combined use of outside relevance measure and internal clustering quality assessment. Therefore, the elaboration of hybrid technologies and techniques for clustering objects based on inductive modeling of complex systems remains a pressing issue in both theory and practice.

The realization of the technology includes the following steps:

In more detail, the technology is as follows:

Step 1. Start

Step 2. Data preprocessing to decrease the dimension of the feature space using the Shannon entropy:

$$H = -K \cdot \sum_{i=1}^{n} p_i \ln(p_i)$$

Step 3. Formation of the initial set Ω of studied objects.

Step 4. Dividing Ω into two equally powerful subsets in accordance with the above technology. The resulting subsets Ω^A and Ω^B can be formally represented as follows:

$$\Omega^A = \left\{x_{ij}^A\right\} \; ; \quad \Omega^B = \left\{x_{ij}^B\right\} \; ;$$

$$i = \overline{1, n_A} = n_B; \quad n_A + n_B = n; \quad j = \overline{1, m}$$

Step 5. Configuring the BHC clustering technology.

Step 6. For the amount of clusters $k \in [k_{\min}, k_{\max}]$:

Step 6.1 Sequential clustering and cluster fixing for $\Omega^A = \left\{x_{ij}^A\right\}$; $\Omega^B = \left\{x_{ij}^B\right\}$

Step 6.2 Calculation of internal clustering quality measure for $\Omega^A = \left\{x_{ij}^A\right\}$; $\Omega^B = \left\{x_{ij}^B\right\}$:

Index Silhouette	$SWC = \frac{1}{K} \sum_{i=1}^{K} S_{x_j} \to \max$
Index Dunn	$DI(k) = \min_{i \in k}$

Index Calinski – Harabasz	$QC_{CH} = \frac{QCB \cdot (N-K)}{QCW \cdot (K-1)} \rightarrow \max$
Index Entropy	$\left[PE = \frac{\sum_{q=1}^{Q} \sum_{k=1}^{K} \ln\left(u_{qk}\right)}{Q}, \ PE \in [0, \ln Kk] \right]$

Step 6.3 Calculation of outside balance measure $ECB = \sqrt{\frac{\sum_{i=1}^{p}\left(IC_A^i - IC_B^i\right)^2}{\sum_{i=1}^{p}\left(IC_A^i + IC_B^i\right)^2}} \rightarrow \min, \ p$.

Step 6.4 If the value of the balance measure is not minimum, then Step 6.1 is repeated. - 6.3. otherwise

Step 7. Fixing the minimum value of the outside balance measure.

Step 8. Designating the optimal amount of clusters k_{opt}

Step 9. Clustering data (the set Ω of objects under study), fixing the clusters.

Step 10. The End

4.5. Clustering Quality Assessment

As measure for the quality of clustering were used:
Index Silhouette [52]:

$$SWC = \frac{1}{K}\sum_{i=1}^{K} S_{x_j} \rightarrow \max , \tag{11}$$

where K represents the amount of clusters, S_{x_j} denotes the optimal membership of the element x_j in cluster p. Silhouette refers to a technique for interpreting and checking consistency within data clusters. The silhouette value is a measure of how similar the object is in its own cluster (cohesion) compared to other clusters (separation). The silhouette diapasons from -1 to +1, where a high value indicates that the object is in good agreement with its own cluster and is poorly aligned with neighboring clusters. If most objects are of high importance, then a clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may be too many or too few clusters. The best partition is characterized by the maximum SWC, which is achieved when the distance inside the cluster is a bit of and the distance between the elements of neighboring clusters is massive.

Index Dunn [53]. It is a metric for evaluating clustering technologies. Compares intercluster dissolution with cluster diameter. The higher the index value, the better the clustering. The purpose of this index is to identify clusters that are compact, with a a bit of difference between cluster members and well-separated, where the objects of different clusters are quite far apart from each other than the dispersion within the cluster. For this purpose of clusters, a higher Dunn index indicates better clustering. One of the imperfections of using this index is the high computational cost, as the amount of clusters and the dimension of the data increase.

$$DI(k) = \min_{i \in k} \tag{12}$$

Calinski-Harabasz Index [54]:

$$QC_{CH} = \frac{QCB \cdot (N-K)}{QCW \cdot (K-1)} \rightarrow \max \tag{13}$$

where *N* represents the total amount of objects in the dataset, and K denotes the amount of clusters being considered.

The Calinski-Harabasz Index, also referred to as the Variance Ratio Measure, calculates the ratio between the sum of the between-cluster dispersion and the within-cluster dispersion for all clusters. The highest index value indicates the most optimal cluster structure.

Entropy [55]:

$$\left[PE = \frac{\sum_{q=1}^{Q}\sum_{k=1}^{K}\ln\left(u_{qk}\right)}{Q},\ PE \in \left[0, \ln Kk\right] \right] \qquad (14)$$

Entropy is a quantitative measure of the organization or disorder within a system. The entropy of a partition reaches its minimum value when the system is highly organized (in the case of a perfect partition, the entropy is zero). In other words, the higher the degree of membership of an element to a specific cluster (and the lower its membership to other clusters), the lower the entropy value, indicating a more accurate clustering.

5. Experiment and Results

5.1. Characterization pre-processing results

As a result of applying the missing data imputation technologies, normalization, and moving average filtering described in Section 4.3, three datasets were obtained, as shown in Figure 3.

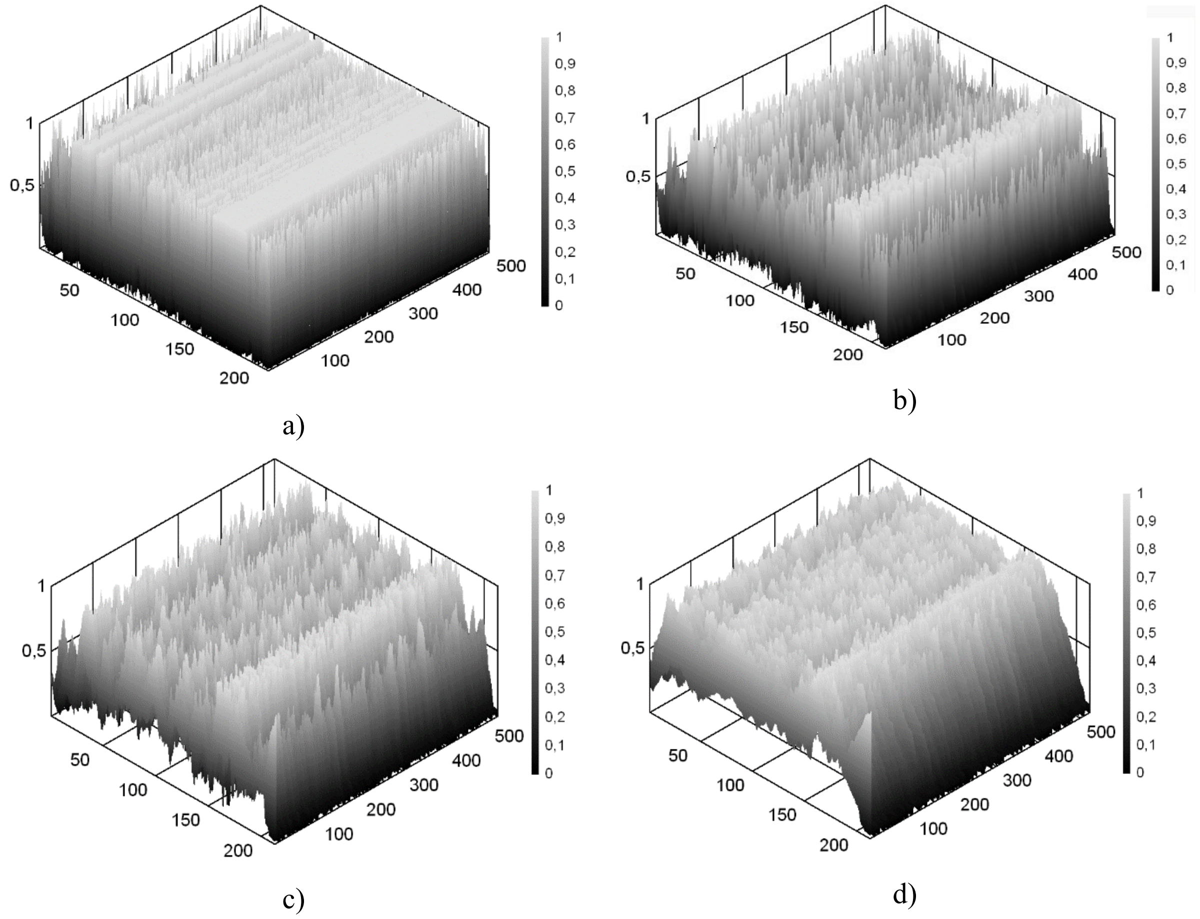


Figure 3: Data surface before and after moving average preprocessing. a) represents the original set of descriptor vectors, b) with moving average noise suppression using window 3 averaging point, c) with moving average noise suppression using window 9 averaging point, and d) with moving average noise suppression using window 27 averaging point.

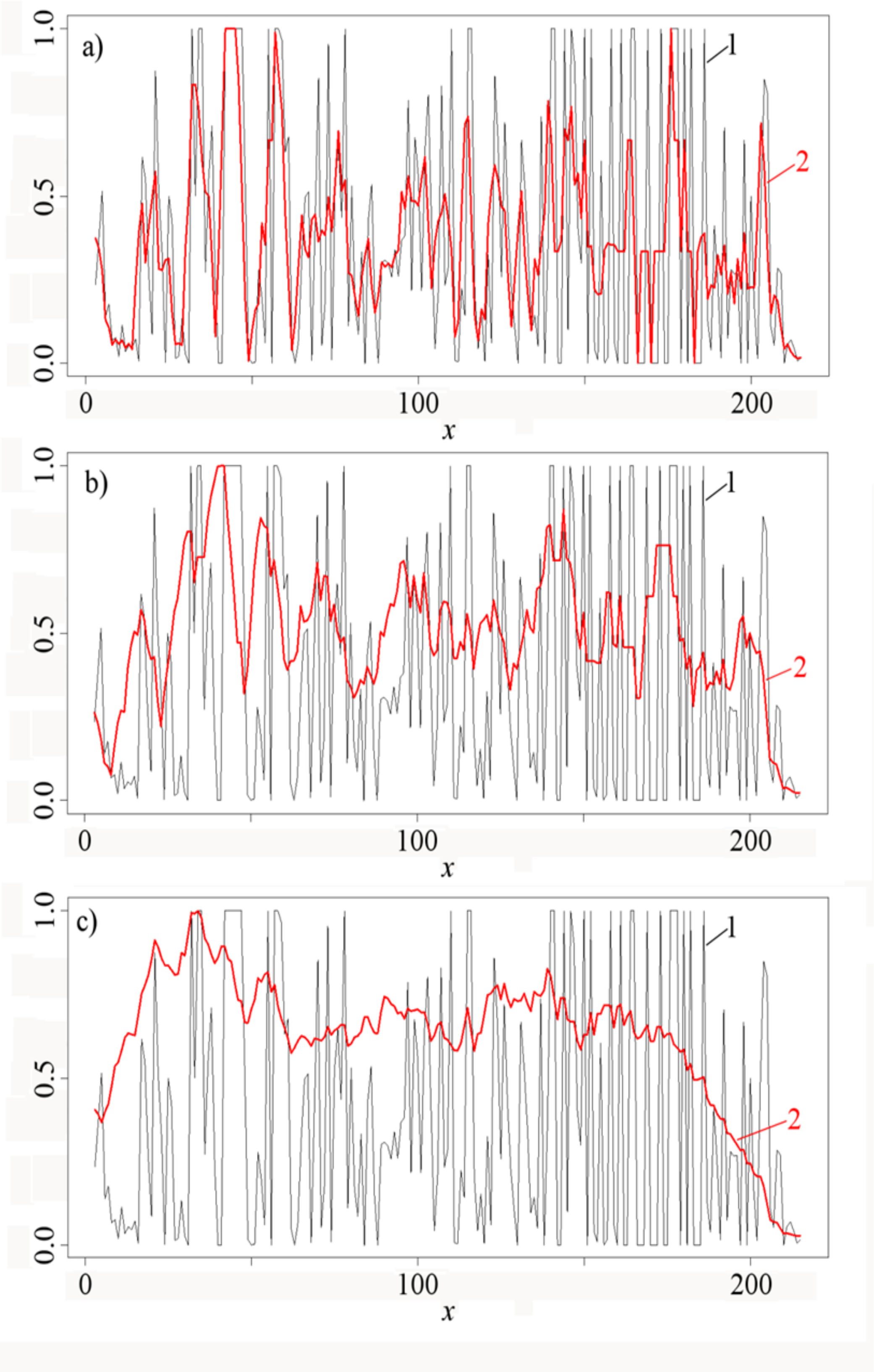


Figure 4: The result of suppressing symmetric additive noise using the moving average technique at different window values. Graph a) corresponds to window value 3, graph b) to 9, and graph c) to 27. Curve 1 in all graphs corresponds to the original data set, curve 2 corresponds to the data set with suppressed noise at the corresponding value of the moving average window.

5.2. Clustering Results

Fig. 5. Results of symmetric additive noise suppression by moving average technique at different values of the window. Graph a) corresponds to window value 3, graph b) to 9, and graph c) to 27. Curve 1 in all graphs corresponds to the original data set, curve 2 corresponds to the data set with suppressed noise at the corresponding moving average window value.

After data normalization and data reduction, combined moving average techniques in non-inductive k-means, c-means, and hierarchical inductive clustering technologies for three 213 × 525 datasets The results are presented in Table 1.

Table 1
Results of designating the amount of clusters by k-means, c-means and hierarchical Bayesian clustering technologies and assessment of clustering quality

Technology	K-means	C-means	Bayesian Hierarchical Clustering
Size	213x525	213x525	213x525
Technique for designating the amount of Clusters	Elbow Technique	Silhouette Technique	Elbow Technique
Amount of Clusters	3	8	3
Silhouette	0.139	0.082	0.115
Dunn Index	0.124	0.126	0.127
Calinski-Harabasz Index	62.263	25.150	50.590
Entropy	1.042077	1.82242	1.011493

After data normalization and data reduction, combined moving average techniques in inductive k-means, c-means, and hierarchical inductive clustering technologies for three data sets divided into sets A and B as matrices of size 107x525. The results are presented in Table 2.

Table 2
Results of designating the amount of clusters by k-means, c-means, and inductive hierarchical Bayesian clustering technologies and clustering quality assessment

Technology	K-means Inductive		C-means Inductive		Bayesian Hierarchical Clustering	
Data	Set A	Set B	Set A	Set B	Set A	Set B
Size	107x525	107x525	107x525	107x525	107x525	107x525
Technique for designating the amount of Clusters	Inductive Technique	Inductive Technique	Inductive Technique	Inductive Technique	Inductive Technique	Inductive Technique
Amount of Clusters	3	3	2	2	4	4
Silhouette	0.150	0.082	0.240	0.234	0.155	0.079
Dunn Index	0.240	0.232	0.213	0.170	0.240	0.232
Calinski-Harabasz Index	30.956	30.889	45.897	43.931	22.723	22.7902
Entropy	1.085	1.064	0.688	0.688	1.152	1.348
ECB	0.001		0.022		0.003	

Table 3
Placement of lineament in clusters, using non-inductive clustering techniques

Technology	Elow Technique	Feature

K-means	Cluster 1	x1 x2 x3 x4 x14 x15 x16 x17 x18 x20 x21 x28 x30 x31 x33 x38 x39 x46 x55 x60 x61 x62 x63 x70 x71 x72 x73 x74 x75 x76 x79 x92 x93 x94 x98 x99 x100 x107 x109 x111 x122 x124 x125 x145 x146 x149 x151 x152 x155 x160 x161 x169 x170 x171 x172 x173 x174 x179 x180 x181 x182 x183 x184 x185 x190 x194 x195 x196 x198 x241 x245 x249 x250 x252 x253 x255 x258 x259 x260 x261 x262 x266 x267 x268 x271
	Cluster 2	x19 x34 x35 x36 x37 x40 x41 x42 x44 x47 x48 x54 x64 x65 x66 x68 x80 x81 x82 x83 x84 x86 x87 x88 x90 x91 x95 x96 x97 x101 x103 x104 x105 x113 x115 x116 x117 x118 x119 x120 x121 x126 x127 x144 x150 x156 x157 x158 x159 x162 x163 x164 x165 x166 x167 x168 x186 x187 x188 x200 x201 x203 x204 x205 x207 x208 x209 x210 x211 x212 x213 x215 x216 x217 x218 x220 x221 x222 x223 x224 x225 x227 x228 x229 x231 x251 x263 x264 x265
	Cluster 3	x5 x6 x7 x8 x9 x10 x11 x12 x13 x22 x23 x24 x25 x26 x27 x57 x58 x59 x77 x78 x108 x147 x148 x153 x154 x191 x192 x193 x197 x246 x247 x248 x254 x273 x275 x277 x281 x282 x283
	Shijhouette Technique	Feature
	Cluster 1	x1 x4 x14 x21 x28 x60 x61 x62 x71 x72 x73 x74 x75 x76 x78 x79 x107 x108 x109 x146 x147 x148 x152 x170 x171 x182 x183 x184 x185 x193 x194 x195 x196 x197 x198 x245 x249 x253 x254 x255 x258 x260 x261 x262 x267 x268 x271 x273
	Cluster 2	x2 x3 x15 x16 x17 x18 x19 x20 x30 x31 x33 x34 x38 x39 x55 x63 x68 x70 x84 x86 x91 x92 x93 x94 x95 x98 x99 x100 x101 x105 x111 x113 x115 x116 x122 x124 x125 x126 x127 x144 x145 x149 x150 x151 x155 x160 x161 x162 x163 x164 x168 x169 x172 x173 x186 x190 x205 x241 x250 x251 x252 x259 x263 x265 x266
	Cluster 3	x5 x6 x7 x8 x9 x10 x11 x12 x13 x22 x23 x24 x25 x26 x27 x57 x58 x59 x77 x153 x154 x191 x192 x246 x247 x248 x275 x277 x281 x282 x283
	Cluster 4	x35 x36 x37 x40 x41 x42 x48 x54 x64 x65 x66 x80 x81 x82 x83 x87 x88 x90 x96 x97 x103 x104 x117 x118 x119 x120 x121 x156 x157 x158 x159 x165 x166 x167 x187 x188 x216 x217 x218 x220 x221 x222 x223 x224 x225 x227 x228 x229 x231 x264
	Cluster 5	x174 x179 x180 x181
	Cluster 6	x44 x46 x47
	Cluster 7	x200 x201 x203 x204
	Cluster 8	x207 x208 x209 x210 x211 x212 x213 x215
C-means	Elow Technique	
	Cluster 1	x2 x15 x16 x17 x20 x30 x31 x38 x39 x63 x79 x94 x98 x111 x113 x124 x144 x149 x155 x160 x161 x172 x173 x179 x185 x200 x201 x203 x204 x205 x209 x220
	Cluster 2	x3 x18 x19 x33 x34 x35 x36 x37 x40 x41 x42 x44 x46 x47 x48 x54 x55 x64 x65 x66 x68 x80 x81 x82 x83 x84 x86 x87 x88 x90 x91 x92 x95 x96 x97 x101 x103 x104 x105 x115 x116 x117 x118 x119 x120 x121 x122 x125 x126 x127 x150 x151 x156 x157 x158 x159 x162 x163 x164 x165 x166 x167 x168 x174 x186 x187 x188 x207 x208 x210 x211 x212 x213 x215 x216 x217 x218 x221 x222 x223 x224 x225 x227 x228 x229 x231 x241 x250 x251 x252 x263 x264 x265 x266
	Cluster 3	x1 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x21 x22 x23 x24 x25 x26 x27 x28 x57 x58 x59 x60 x61 x62 x70 x71 x72 x73 x74 x75 x76 x77 x78 x93 x99 x100 x107 x108 x109 x145 x146 x147 x148 x152 x153 x154 x169 x170

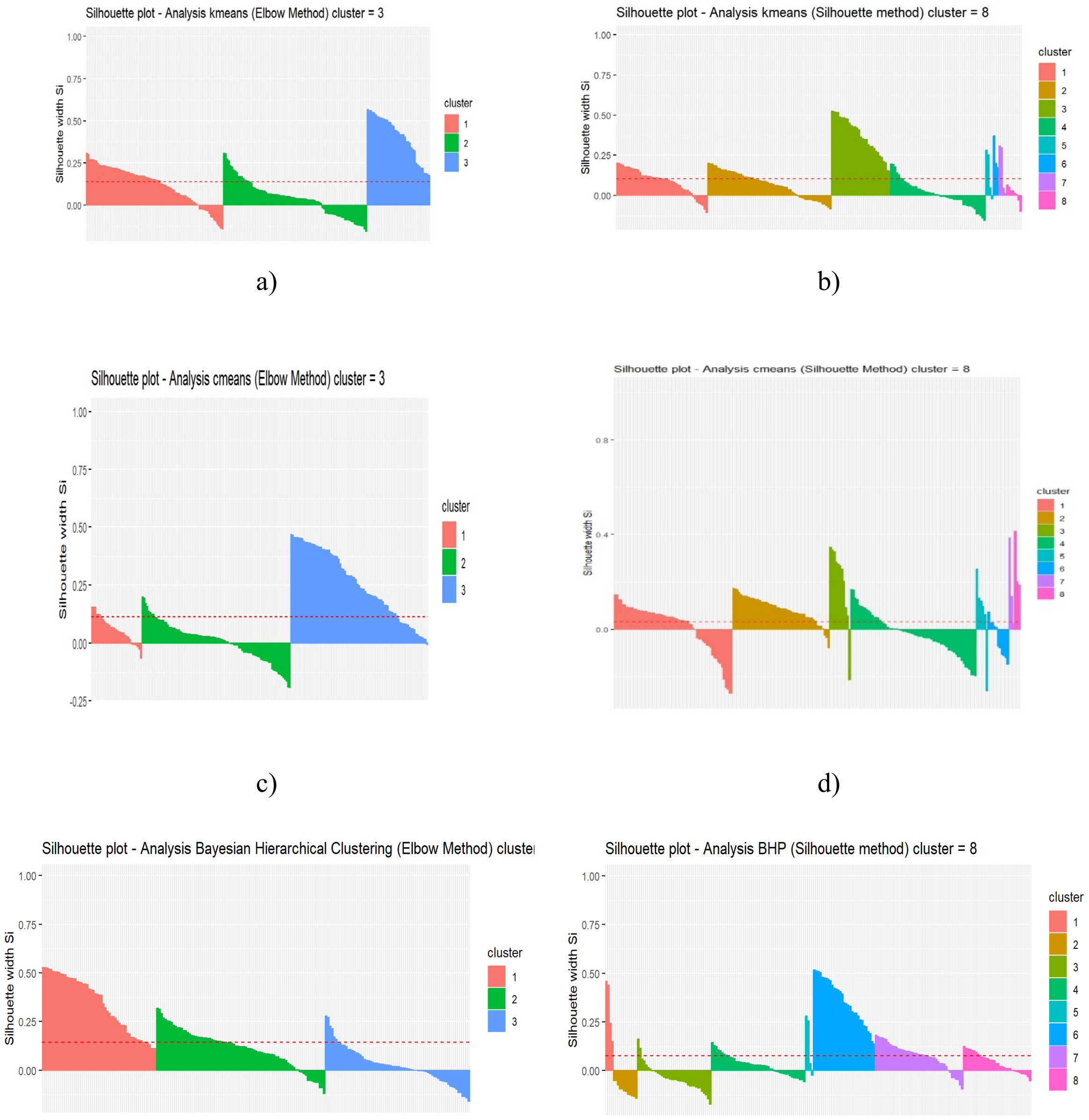
		x171 x180 x181 x182 x183 x184 x190 x191 x192 x193 x194 x195 x196 x197 x198 x245 x246 x247 x248 x249 x253 x254 x255 x258 x259 x260 x261 x262 x267 x268 x271 x273 x275 x277 x281 x282 x283
	Shijhouette Technique	
	Cluster 1	x1 x4 x5 x8 x9 x13 x14 x21 x22 x23 x26 x27 x28 x30 x57 x58 x59 x60 x61 x62 x71 x72 x73 x74 x75 x76 x77 x78 x79 x108 x109 x111 x152 x154 x171 x180 x181 x182 x183 x184 x193 x194 x195 x196 x197 x198 x245 x246 x247 x248 x249 x253 x254 x255 x260 x261 x267 x268 x271 x273 x275 x283
	Cluster 2	x2 x3 x15 x16 x17 x18 x19 x20 x31 x33 x34 x38 x39 x63 x68 x70 x91 x92 x93 x94 x95 x98 x99 x100 x101 x107 x122 x124 x125 x126 x144 x145 x149 x150 x151 x155 x161 x162 x163 x185 x204 x205 x241 x250 x251 x252 x258 x259 x262 x263 x266
	Cluster 3	x6 x7 x10 x11 x12 x190 x191 x192 x277 x281 x282
	Cluster 4	x35 x36 x37 x40 x41 x42 x47 x48 x54 x64 x65 x66 x80 x81 x82 x83 x87 x88 x90 x96 x97 x103 x104 x105 x113 x115 x116 x117 x118 x119 x120 x121 x127 x156 x157 x158 x159 x164 x165 x166 x167 x186 x187 x188 x200 x201 x203 x207 x208 x212 x213 x215 x216 x217 x218 x220 x221 x222 x223 x224 x225 x227 x228 x229 x231 x264
	Cluster 5	x24 x25 x146 x147 x148 x153
	Cluster 6	x44 x46 x55 x84 x86 x160 x172 x173 x174 x179 x265
	Cluster 7	x168 x169 x170
	Cluster 8	x209 x210 x211
Bayesian Hierarchical Clustering	Elow Technique	
	Cluster 1	x34 x35 x36 x37 x40 x41 x42 x47 x48 x54 x64 x65 x66 x80 x81 x82 x83 x86 x87 x88 x90 x96 x97 x103 x104 x105 x117 x118 x119 x120 x121 x127 x156 x157 x158 x159 x163 x164 x165 x166 x167 x168 x186 x187 x188 x201 x203 x207 x208 x209 x210 x211 x212 x213 x215 x216 x217 x218 x220 x221 x222 x223 x224 x225 x227 x228 x229 x231 x251 x263 x264 x265
	Cluster 2	x1 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x21 x22 x23 x24 x25 x26 x27 x28 x57 x58 x59 x60 x76 x77 x78 x108 x109 x146 x147 x148 x152 x153 x154 x170 x183 x184 x191 x192 x193 x196 x197 x246 x247 x248 x249 x254 x255 x261 x268 x271 x273 x275 x277 x281 x282 x283
	Cluster 3	x2 x3 x4 x15 x16 x17 x18 x19 x20 x30 x31 x33 x38 x39 x44 x46 x55 x61 x62 x63 x68 x70 x71 x72 x73 x74 x75 x79 x84 x91 x92 x93 x94 x95 x98 x99 x100 x101 x107 x111 x113 x115 x116 x122 x124 x125 x126 x144 x145 x149 x150 x151 x155 x160 x161 x162 x169 x171 x172 x173 x174 x179 x180 x181 x182 x185 x190 x194 x195 x198 x200 x204 x205 x241 x245 x250 x252 x253 x258 x259 x260 x262 x266 x267
	Shijhouette Technique	
	Cluster 1	x5 x6 x7 x8 x9 x10 x11 x12 x13 x22 x23 x24 x25 x26 x27 x57 x58 x59 x77 x147 x153 x154 x191 x192 x246 x247 x248 x275 x277 x281 x282 x283
	Cluster 2	x167 x168 x169
	Cluster 3	x86 x87 x88 x90 x163 x164 x165 x218 x220 x221 x222
	Cluster 4	x93 x94 x98 x99 x100 x122 x124 x125 x145 x146
	Cluster 5	x2 x3 x18 x19 x20 x31 x33 x34 x37 x38 x39 x44 x46 x63 x68 x80 x83 x84 x91 x92 x95 x101 x105 x111 x113 x115 x116 x117 x121 x126 x127 x144

	x150 x151 x160 x161 x162 x173 x186 x190 x200 x203 x204 x205 x207 x241 x250 x251 x252 x263 x265 x266
Cluster 6	x35 x36 x40 x41 x42 x47 x48 x54 x64 x65 x66 x81 x82 x96 x97 x103 x104 x118 x119 x120 x156 x157 x158 x159 x166 x187 x188 x201 x208 x209 x210 x211 x212 x213 x215 x216 x217 x223 x224 x225 x227 x228 x229 x231 x264
Cluster 7	x1 x4 x14 x15 x16 x17 x21 x28 x30 x55 x60 x61 x62 x70 x71 x72 x73x74 x75 x76 x78 x79 x107 x108 x109 x148 x149 x152 x155 x170 x171 x172 x182 x183 x184 x185 x193 x194 x195 x196 x197 x198 x245 x249 x253 x254 x255 x258 x259 x260 x261 x262 x267 x268 x271 x273
Cluster 8	x174 x179 x180 x181

Table 4
Placement of lineament in clusters, using inductive clustering techniques

K-means	Feature
Cluster 1	x2 x3 x4 x15 x16 x17 x18 x19 x20 x30 x31 x33 x38 x39 x44 x46 x55 x61 x62 x63 x68 x70 x71 x72 x73 x74 x75 x79 x84 x91 x92 x93 x94 x95 x98 x99 x100 x101 x107 x111 x113 x115 x116 x122 x124 x125 x126 x144 x145 x149 x150 x151 x155 x160 x161 x162 x169 x171 x172 x173 x174 x179 x180 x181 x182 x185 x190 x194 x195 x198 x200 x204 x205 x241 x245 x250 x252 x253 x258 x259 x260 x262 x266 x267
Cluster 2	x34 x35 x36 x37 x40 x41 x42 x47 x48 x54 x64 x65 x66 x80 x81 x82 x83 x86 x87 x88 x90 x96 x97 x103 x104 x105 x117 x118 x119 x120 x121 x127 x156 x157 x158 x159 x163 x164 x165 x166 x167 x168 x186 x187 x188 x201 x203 x207 x208 x209 x210 x211 x212 x213 x215 x216 x217 x218 x220 x221 x222 x223 x224 x225 x227 x228 x229 x231 x251 x263 x264 x265
Cluster 3	x1 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x21 x22 x23 x24 x25 x26 x27 x28 x57 x58 x59 x60 x76 x77 x78 x108 x109 x146 x147 x148 x152 x153 x154 x170 x183 x184 x191 x192 x193 x196 x197 x246 x247 x248 x249 x254 x255 x261 x268 x271 x273 x275 x277 x281 x282 x283
C-means	
Cluster 1	x1 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x17 x21 x22 x23 x24 x25 x26 x27 x28 x30 x57 x58 x59 x60 x61 x62 x70 x71 x72 x73 x74 x75 x76 x77 x78 x79 x92 x93 x94 x99 x100 x107 x108 x109 x145 x146 x147 x148 x149 x152 x153 x154 x169 x170 x171 x172 x179 x180 x181 x182 x183 x184 x190 x191 x192 x193 x194 x195 x196 x197 x198 x245 x246 x247 x248 x249 x253 x254 x255 x258 x259 x260 x261 x262 x266 x267 x268 x271 x273 x275 x277 x281 x282 x283
Cluster 2	x2 x3 x15 x16 x18 x19 x20 x31 x33 x34 x35 x36 x37 x38 x39 x40 x41 x42 x44 x46 x47 x48 x54 x55 x63 x64 x65 x66 x68 x80 x81 x82 x83 x84 x86 x87 x88 x90 x91 x95 x96 x97 x98 x101 x103 x104 x105 x111 x113 x115 x116 x117 x118 x119 x120 x121 x122 x124 x125 x126 x127 x144 x150 x151 x155 x156 x157 x158 x159 x160 x161 x162 x163 x164 x165 x166 x167 x168 x173 x174 x185 x186 x187 x188 x200 x201 x203 x204 x205 x207 x208 x209 x210 x211 x212 x213 x215 x216 x217 x218 x220 x221 x222 x223 x224 x225 x227 x228 x229 x231 x241 x250 x251 x252 x263 x264 x265
Cluster 3	x5 x6 x7 x8 x9 x10 x11 x12 x13 x22 x23 x24 x25 x26 x27 x57 x58 x59 x77 x153 x154 x191 x192 x246 x247 x248 x275 x277 x281 x282 x283
Cluster 4	x35 x36 x37 x40 x41 x42 x47 x48 x54 x64 x65 x66 x80 x81 x82 x83 x87 x88 x90 x96 x97 x103 x104 x117 x118 x119 x120 x121 x156 x157 x158 x159 x165 x166 x167 x187 x188 x201 x203 x207 x208 x209 x210 x211 x212 x213 x215 x216 x217 x218 x220 x221 x222 x223 x224 x225 x227 x228 x229 x231 x264

5.3. Visualization of results

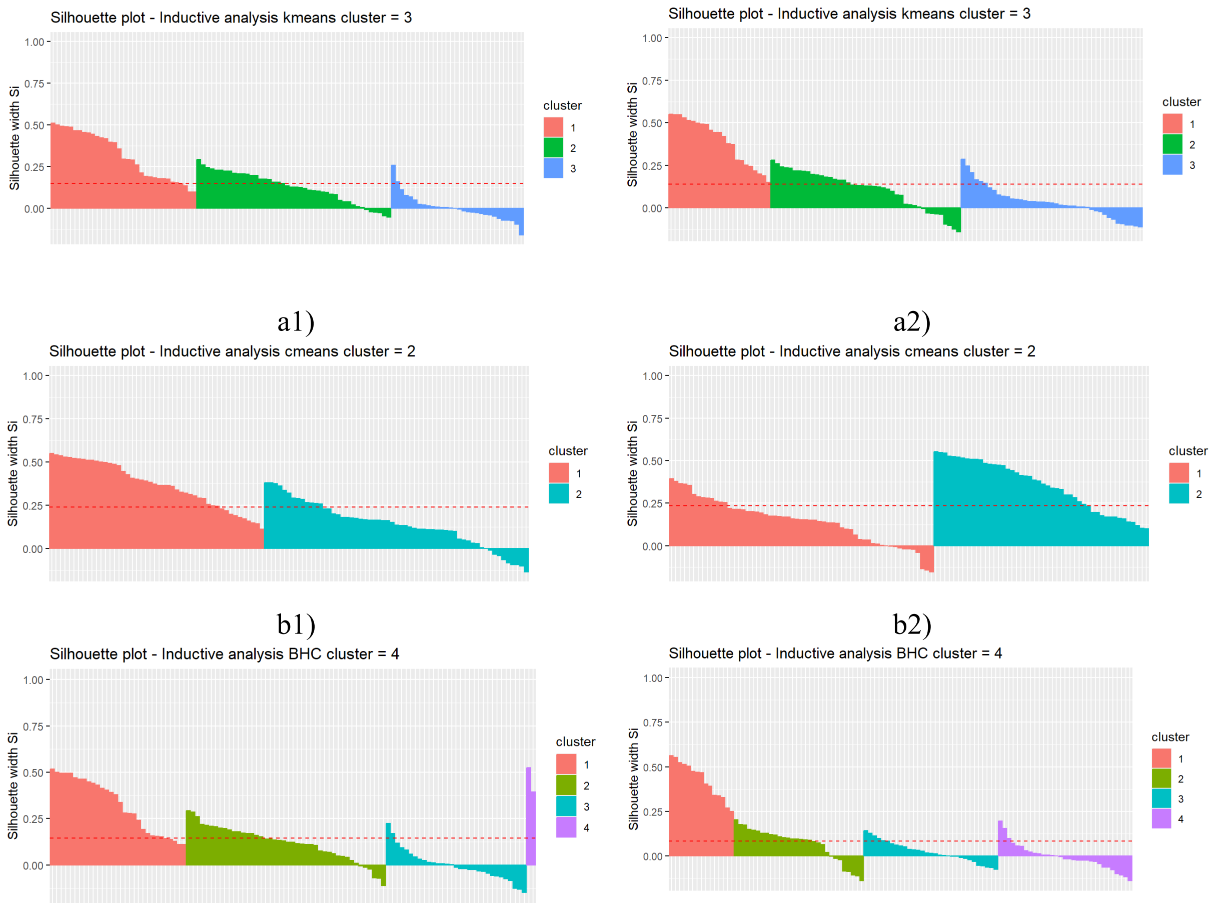


a) b)

c) d)

e) f)

Figure 6: Silhouette diagrams obtained using classical k-means, c-means and BHC clustering techniques: (a) silhouette k-means, k=3, designation of the amount of clusters was performed using Elbow Technique; (b) silhouette k-means, k=8 designation of the amount of clusters was performed using Silhouette Technique; (c) silhouette c-means, k=3, designation of the amount of clusters was performed using Elbow Technique; d) c-means silhouette, k=8 designation of the amount of clusters was performed using Silhouette Technique; e) Bayesian Hierarchical Clustering silhouette, k=3 designation of the amount of clusters was performed using Elbow Technique; f) Bayesian Hierarchical Clustering silhouette, k=8 designation of the amount of clusters was performed using Silhouette Technique



a1) a2)

b1) b2)

c1) c2)

Figure 7: Silhouette diagrams obtained using inductive clustering technologies k-means, c-means and BHC, (in the inductive technologies the amount of clusters is designated automatically): (a1 and a2) silhouette k-means, k=3; (b1 and b2) silhouette k-means, k=2; (c1 and c2) silhouette c-means, k=4.

6. Discussion

Comparative studies of non-inductive clustering technologies and their evaluation using the measure for assessing the quality of clustering (Tabl.1) showed that the highest quality partitioning at k=3. This applies to all three clustering technologies. As for inductive analogues of the technologies under study (Tabl.2), here the evaluation was performed for each technology simultaneously on two equal and relevant sets (A and B), which does not contradict the further association of their selected relevant clusters (Tab.4.).

It is significant to mention that, due to the nature of inductive technologies, the designation of the amount of clusters was performed automatically. Each technology utilized stochastic indicators or values of the membership function, as well as the inductive probabilistic hierarchical clustering technology.

The inductive probabilistic hierarchical clustering technology employed boundary probabilities to identify which clusters should be merged in order to prevent overflow. Essentially, it estimated the probability that all the data in a potential union originated from the same mixture component and compared this probability to the significantly massive amount of hypotheses at the lower levels of the clustering hierarchy.

As shown in Table 2, when partitioned into 3 clusters, the Silhouette index yielded values of 0.150 and 0.082 for the respective sets A and B. Using the c-means technology, 2 clusters were obtained, as evidenced by the Dunn Index, Calinski-Harabasz Index, and Entropy Index values, indicating that the boundaries between clusters are highly fuzzy. The inductive BHC technology identified 4 clusters, which is confirmed by the values of Silhouette index (0.155 and 0.079) respectively on the relevant sets A and B).

Regarding the Silhouette index of Figs. 5 and 6, it should be noted that "questionable" clusters are characterized by negative values, which requires additional research.

7. Conclusions

The proposed clustering technology can be useful for identifying relevant lineament in the results of laboratory tests for persons with multiple myeloma. It demonstrates high performance of the developed inductive technologies, namely k-means, c-means, and Bayesian hierarchical clustering based on the inductive modeling of complex systems. The main technology used in this study was the Bayesian hierarchical clustering technology, and the impact of four internal measure (silhouette, Dunn index, Calinski-Harabasz index, entropy) on clustering effectiveness was investigated. Additionally, the application of the moving average technology for noise elimination in the data was proposed for the first time. The overall use of the proposed noise elimination technique in conjunction with the inductive approach significantly improves the quality of clustering complex objects. The excellence of the proposed technologies lies in their stability, achieved by using an outside balance measure for two identical samples.

The proposed clustering technology can be beneficial for extracting relevant lineament from the results of laboratory tests for persons with multiple myeloma in several aspects:

Identification of influential lineament: The proposed clustering techniques allow for the identification of groups of similar objects based on their characteristics. Lineament that significantly alter the cluster structure or separate objects into different groups can be considered significant and relevant. This helps identify lineament that may play a crucial role in the diagnosis, prognosis, or classification of multiple myeloma.

Removal of irrelevant lineament: The proposed clustering techniques can help identify lineament that do not contribute significantly to the cluster structure or fail to separate objects into distinct groups. Such lineament can be deemed irrelevant and excluded from further assay. This decreases the dimensionality of the data and simplifies result interpretation.

Selection of cluster labels: Clustering can aid in identifying clusters that exhibit distinct characteristics or behaviors. Extracting relevant lineament can assist in choosing appropriate labels for these clusters, facilitating more accurate result interpretation with potential clinical implications.

Overall, the proposed clustering technology allows for the systematic assay of laboratory test results for persons with multiple myeloma, the identification of significant lineament, and the simplification of data interpretation. This can enhance the understanding of the disease, the elaboration of diagnostic and prognostic technologies, and support decision-making in clinical practice.

8. References

[1] J.-L. Harousseau, M. Dreyling, on behalf of the ESMO Guidelines Working Group. Multiple myeloma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann. Oncology 2010, 21(Suppl. 5): v155–7

[2] S.S. Bessmeltsev, K.M. Abdukadyrov, Mnozhestvennaya miyeloma. Sovremennyy vzglyad na problemu [Multiple myeloma. Current view of the problem]. Almaty, 2007.

[3] A. Jemal, et al., Cancer statistics. CA Cancer J. Clin 2007, 57(1), pp.43–66. doi: 10.3322/canjclin.57.1.43. PMID: 17237035.

[4] J. Cid Ruzafa, et al., Patient population with multiple myeloma and transitions across different lines of therapy in the USA: an epidemiologic model. Pharmacoepidemiol Drug Saf 2016, 25(8), pp. 871–9. doi: 10.1002/pds.3927. PMID: 27476979.

[5] P.S. Rosenberg, et al., Future distribution of multiple myeloma in the United States by sex, age, and race/ ethnicity. Blood 2015, 125(2), pp. 410–2. doi: 10.1182/blood-2014-10-609461. PMID: 25573972.

[6] K. Yamabe, et al., Epidemiology and burden of multiple myeloma in Japan: a systematic review. Value Health 2015, 18(7): A449, doi: 10.1016/j. jval.2015.09.1129. PMID: 26532529.

[7] J. Hong, et al., Recent advances in multiple myeloma: a Korean perspective. Korean J Intern Med 2016, 31(5):820–34, doi: 10.3904/kjim.2015.408. PMID: 27604794.

[8] X.C. Chen, et al., Epidemiological differences in haematological malignancies between Europe and China. Lancet Oncol 2014;15(11):471–2, doi: 10.1016/ S1470-2045(14)70441-3. PMID: 25281463.

[9] J.H. Chen, et al., Prevalence and mortality-related factors of multiple myeloma in Taiwan. PLoS One 2016;11(12):e0167227, doi: 10.1371/ journal.pone.0167227. PMID: 27907052.

[10] H. Ludwig, et al., Multiple Myeloma Incidence and Mortality Around the Globe; Interrelations Between Health Access and Quality, Economic Resources, and Patient Empowerment. Oncologist. 2020 Sep;25(9):e1406-e1413. doi: 10.1634/theoncologist.2020-0141. Epub 2020 May 7. PMID: 32335971; PMCID: PMC7485361.

[11] Kh. Tadist, et al., Feature selection techniques and genomic big data: a systematic review. Journal of Big Data 27.08.2019.

[12] M. Alweshah, et al., Coronavirus herd immunity optimizer with greedy crossover for feature selection in medical diagnosis. Knowledge-Based Systems, 235 (2022) 107629.

[13] C. Stan, D. Waltz, Towards memory based reasoning. Commun ACM. 1986, 29(12), pp.1213–1228.

[14] S. Piramuthu, Evaluating feature selection techniques for learning in data mining applications. Eur J Oper Res. 2004, 156(2), pp. 483–494, doi:10.1016/S0377-2217 (02)00911-6.

[15] J.D. Shie, S.M. Chen, Feature subset selection based on fuzzy entropy measures for handling classification problems. Appl Intell. 2008, 28(1), pp.69–82.

[16] S. Zhao, E.C. Tsang, On fuzzy approximation operators in attribute reduction with fuzzy rough sets. J Inf Sci. 2008, 178(16), pp. 3163–3176.

[17] I.A. Gheyas, L.S. Smith, Feature subset selection in massive dimensionality domains. PatternRecognit. 2010, 43(1), pp.5–13.

[18] F. Nie, et al. Efficient and robust feature selection via joint 2, 1-norms minimization. Adv Neural Information Process Syst. 2010, pp.1813–1821.

[19] S. Foithong, O. Pingern, B. Atachoo, Feature subset selection wrapper based on mutual information and rough sets. Expert Syst Appl. 2011, 39(1), pp. 574–584.

[20] H. Huang, et al. Ant colony optimization–based feature selection for surface electromyography signals classification. Comput Biol Med. 2011, 42(1), pp. 30–38.

[21] P. Ghamisi, et al., Feature selection based on hybridization of genetic technology and particle swarm optimization. IEEE Geosci Remote Sens Lett. 2015,12(2), pp. 309–313.

[22] G.N. Zhu, et al., An integrated feature selection and cluster assay techniques for case-based reasoning. Eng Appl Artif Intel. 2015, 39, pp.14–22. doi:10.1016/j.engappai.2014.11.006.

[23] P. Zhu, et al., Unsupervised feature selection by regularized self-representation. Pattern Recognit, 2015,48(2), pp. 438–446.

[24] F. Barani, et al., Application of binary quantum-inspired gravitational search technology in feature subset selection. Appl Intell. 2017, 47(2), pp. 304–318.

[25] P. Moradi, et al., A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. Appl Soft Comput. 2016, 43, pp.117–130.

[26] P. Shunmugapriya, et al., A hybrid technology using ant and bee colony optimization for feature selection and classification (AC-ABC hybrid). Swarm Evol Comput. 2017, 36, pp. 27–36.

[27] C. Boutsidis, et al., Unsupervised feature selection for the k-means clustering problem. NIPS. 2009, pp.153–161.

[28] C. Boutsidis, et al. Randomized dimensionality Reduction for k-means clustering. IEEE Trans. Inf Theory. 2015, 61(2). pp.1045–1062.

[29] F. Moslehi, A. Haeri , A novel feature selection approach based on clustering technology, Journal of Statistical Computation and Simulation, 2021,91:3, pp. 581-604, doi: 10.1080/00949655.2020.1822358

[30] S. Babichev, M. A. Taif, V. Lytvynenko, Estimation of the inductive model of objects clustering stability based on the k-means technology for different levels of data noise. Radio electronics, computer science, management. Zaporozhye: NAS of Ukraine, 2016, no. 4, pp. 54-60.

[31] M. E Celebi, et al., A comparative study of efficient initialization techniques for the k-means clustering technology. Expert Systems with Applications. 40 (1), pp. 200-210. arXiv: 1209.1960.

[32] H. R. Madala, Inductive Learning Technologies for Complex Systems Modeling / H. R. Madala, A. G. Ivakhnenko. - CRC Press, 1994. –365 p.

[33] A.G. Ivakhnenko, Objective clusterization on the basis of the theory of self-organization of models. Soviet J. Automat. Inform. Sci. 20(5), pp. 1–9 (1987).

[34] V. Stepashko, Inductive modeling from historical perspective. In: Proceedings of the 12th international scientific and technical conference on computer sciences and information technologies, CSIT 2017, vol. 1, pp. 537–542 (2017).

[35] V.S. Stepashko, Theoretical aspects of GMDH as a technique of inductive modelling. Managing Systems and Machines 2, pp. 31-38 (2003) [In Russian].

[36] Zh. Hu, et al., An Evolving Cascade System Based on a Set of Neo - Fuzzy Nodes. International Journal of Intelligent Systems and Applications (IJISA) 8(9), 2016, pp. 1-7.

[37] F.D. Mwale, et al., Infilling of Missing Rainfall and Streamflow Data in the Shire River Basin, MalawiA Self Organizing Map Approach. Physics and Chemistry of the Earth, Parts A/B/C 50–52 (2012), pp. 34–43, doi:10.1016/j.pce.2012.09.006.

[38] F. B. Hamzah, et al., Imputation Techniques for Recovering Streamflow Observation: A Techniqueological Review. Edited by Fei Li. Cogent Environmental Science 6, no. 1 (January 1, 2020): 1745133. doi:10.1080/23311843.2020.1745133

[39] H. Lee, K. Kwangmin, Interpolation of Missing Precipitation Data Using Kernel Estimations for Hydrologic Modeling. Advances in Meteorology 2015, pp. 1–12. doi:10.1155/2015/935868.

[40] J. Chen, et al., Jackknife Variance Estimation for Nearest-Neighbor Imputation. Journal of the American Statistical Association 96, no. 453 (March 2001), pp.260–269. doi:10.1198/016214501750332839.

[41] M. G. Kendall, A. Stuart, J. K. Ord, Kendall's advanced theory of statistics, vol. 3, Hodder Arnold, London, 1983.

[42] D. Ladiray, et al., Seasonal adjustment with the X-11 technique, vol. 158 of Lecture notes in statistics, Springer-Verlag, 2001.

[43] S. Makridakis, et al., Forecasting: techniques and applications, 3rd edn, John Wiley & Sons, New York, 1998.

[44] J. Spencer, On the graduation of the rates of sickness and mortality, Journal of the Institute of Actuaries 38, 1904, pp. 334–343.

[45] R. J. Hyndman International Encyclopedia of Statistical Science, ed. Miodrag Lovric, Springer. pp.866-869.

[46] L.V. Sarycheva, Objective cluster assay of the data on the basis of the Group Technique of Data Handling//Problem of Management and Informatics, 2008, no. 2, pp. 86-104. [In Russian].

[47] H.Steinhaus, Sur la division des corp materiels en parties.Bull. Acad Polon Sci 1.804 (1956):801.

[48] S. P. Lloyd, Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, SP: Least squares quantization in PCM. IEEE Trans. Inform. Theor. (1957/1982).

[49] J. MacQueen, Some techniques for classification and assay of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, no. 14, 1967.

[50] M. Melnik, Fundamentals of applied statistics. Moscow: Energoatomizdat, 1983, 416 p.

[51] J.C. Dunn. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters // Journal of Cybernetics, 1973, 17 09 (t. 3, No. 3), pp. 32–57, ISSN 0022-0280, doi: 10.1080 / 01969727308546046.

[52] K.A. Heller, Z. Ghahramani, (2005). Bayesian Hierarchical Clustering. Proceedings of the 22nd international conference on machine learning, pp.297-304. Retrieved from https://doi.org/10.1145/1102351.1102389.

[53] N. Lowing, R. Bomalaski, D. Mitra, (2017). Bayesian Hierarchical Clustering. Nicholas Lowing & Ryan Bomalaski Group 3 CSE 5290 Dr.

[54] A. Ivakhnenko, Group technique of data handling as competitor to the technique of stochastic approximation / A. Ivakhnenko // Soviet Automatic Control,1968, vol. 3, pp. 64–68.

[55] L. Kaufman, P. Rousseeuw, Finding Groups in Data. An Introduction to Cluster Assay. Wiley, 2005, https://doi.org/ 10.1002/9780470316801.

[56] J.C. Bezdek, et al., Optimal fuzzy partitions: A heuristic for estimating the parameters in a mixture of normal dustrubutions//IEEE Transactions on Computers, 1975, pp. 835–838, https://doi.org/10.1109/T-C.1975.224317.

[57] T. Calinski, et al., A dendrite technique for cluster assay. Communication in statistics, 1974, no.3, pp. 1–27, https://doi.org/ 10.1080/03610927408827101.

[58] S. Ch. Sripada, M. S. Rao, Comparison of purity and entropy of k-means clustering and fuzzy c means clustering, Indian journal of computer science and engineering; Vol 2 no.3 June 2011, ISSN:0976-5166.