DOI:10.34918/87329

Agnieszka BANAŚ[1,*], Marco SAVARESE[2], Rossella TUPLER[3], Joanna POLAŃSKA[4]

Chapter 3. TRANSCRIPTOMICS-BASED MUSCULAR DYSTROPHY PATIENT STRATIFICATION WITH THE USE OF MACHINE LEARNING

3.1. Introduction

Muscular dystrophies are a group of genetic disorders characterized by progressive muscle wasting and degeneration. There are several different types of muscular dystrophies with distinct patterns of muscle development and genetic causes. By identifying relevant features, personalized treatment strategies can be provided [1].

Over the past few years, RNA sequencing (RNA-seq) has emerged as a powerful genomic tool for the identification of genetic variants that are associated with various diseases, including rare genetic disorders like muscular dystrophies. RNA-seq is a powerful tool for unravelling the complexity of gene regulation and understanding the dynamics of RNA molecules in tissues [2].

Machine learning is a subset of artificial intelligence that focuses on developing algorithms and models capable of learning and making decisions. It has the ability to analyze complex datasets and discover hidden patterns and associations. When combined with RNA-seq, machine learning can be leveraged to develop clustering models for muscular dystrophy patient stratification, leading to improved diagnosis and treatment strategies. The aim of this study is to create a preprocessing pipeline that includes a series of steps to prepare the data for effective analysis and identifying molecular signatures.

[1] Faculty of Automatic Control, Silesian University of Technology, Gliwice, Poland.
[2] Folkhälsan Research Center, University of Helsinki.
[3] Dipartimento di Scienze Biomediche,Università degli Studi di Modena e Reggio Emilia.
[4] Department of Data Science and Engineering, Faculty of Automatic Control, Silesian University of Technology, Gliwice, Poland.
[*] Corresponding author: agniban056@student.polsl.pl.

3.2. Dataset

The dataset consists of 61,587 gene expression profiles from 197 patients with muscular dystrophies, obtained using the RNA-seq method with HiSeq4000 and NovaSeq platforms. RNA-seq is high-throughput sequencing technique used to analyze the transcriptome of an organism. It provides insight into the RNA molecules present in a biological sample, enabling the study of gene expression. It quantifies the amount of RNA molecules produced from that gene and provides valuable information about its transcriptional activity [2]. The range of expression levels in RNA-seq dataset spans several orders of magnitude, reflecting the varying levels of gene expression between different genes and patients. At the lower end of the range are genes with very low or negligible levels of expression. On the other hand, genes that are highly expressed can have expression levels that are several orders of magnitude higher.

A description of the clinical features is provided in Tables 1, 2 and 3.

Table 1

Patients phenotype

Phenotype	Number of patients
Titinopathy	63
Myopathy	32
IBM	28
Amputee control	12
CK control	7
Myopathy (HNRNPA1)	6
DMD	5
SMPX myopathy	4
FSHD	3
Actinin-2 myopathy	2
LGMD D2	1
Control cells	1
Becker	1
OPDM	1

Table 2

Patients sex

Sex	Number of patients
Male	91
Female	48
Not known	28

Table 3

Age at biopsy

Age at biopsy	Number of patients
<30	23
30 – 50	37
>50	78
Not known	29

3.3. Methods

The preprocessing pipeline in the context of machine learning refers to a series of steps and techniques applied to raw data to transform it into a format suitable for training a machine learning model. Each step in the pipeline serves a specific purpose and contributes to the overall data preparation process. The pipeline created is shown in Fig. 1.

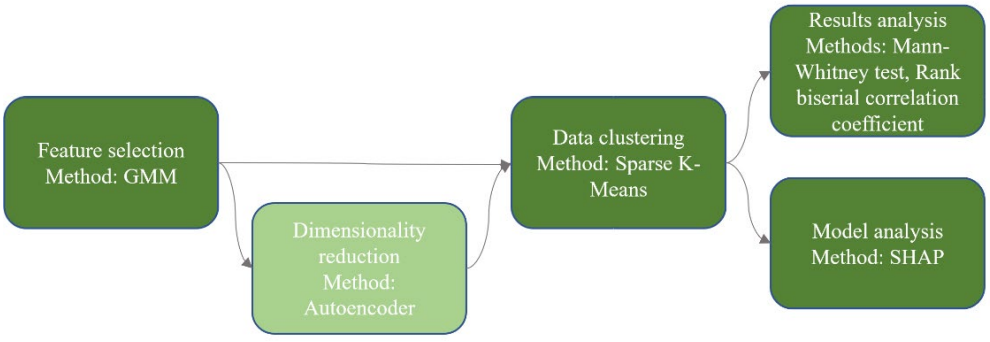

Fig. 1. Preprocesing pipeline
Rys. 1. Potok przetwarzania

3.3.1. Feature selection

The first step in a preprocessing pipeline, especially in high-dimensional datasets, is feature selection. It refers to the process of identifying the subset of features that are most relevant to the analysis. In the context of feature selection for muscular dystrophy, we decided to reject genes that are not actively expressed or have low levels of expression. Hence, the expression levels of genes were converted into binary values representing the presence or absence of gene expression. The binary heatmap depicted in Fig. 2 represents gene expression activity, where active genes are denoted by 1, and inactive genes are denoted by 0.

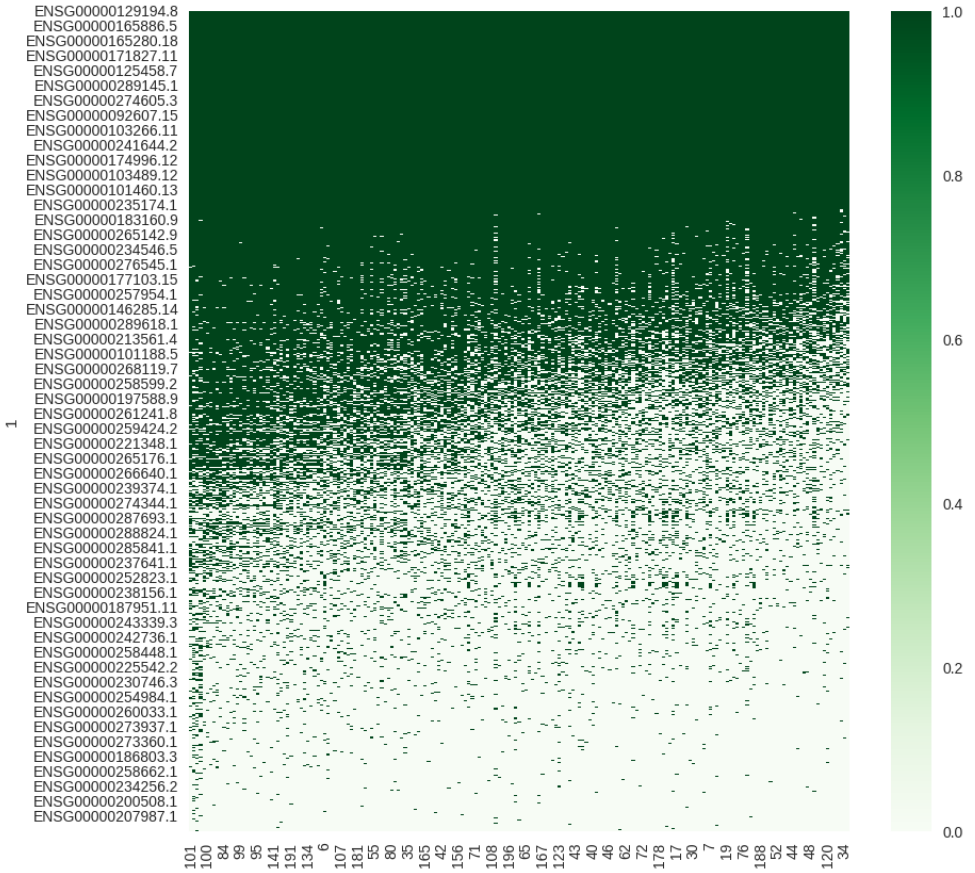

Fig. 2. Binary form of gene expression
Rys. 2. Binarna postać ekspresji genów

The Gaussian Mixture Model (GMM) was then applied to the data to reject certain components if they did not provide meaningful information. It assumes that the data points are generated from a mixture of Gaussian distributions. Determining the optimal number of components is an important step in the analysis of gene expression data. For this purpose, Bayesian Information Criterion (BIC) was used. The BIC is based on the principle of penalized likelihood, which takes into account both the likelihood of the

data given the model and a penalty term for model complexity. The GMM with the lowest BIC value is usually considered to be the optimal model. The visual representation of the GMM results is shown on Fig. 3.

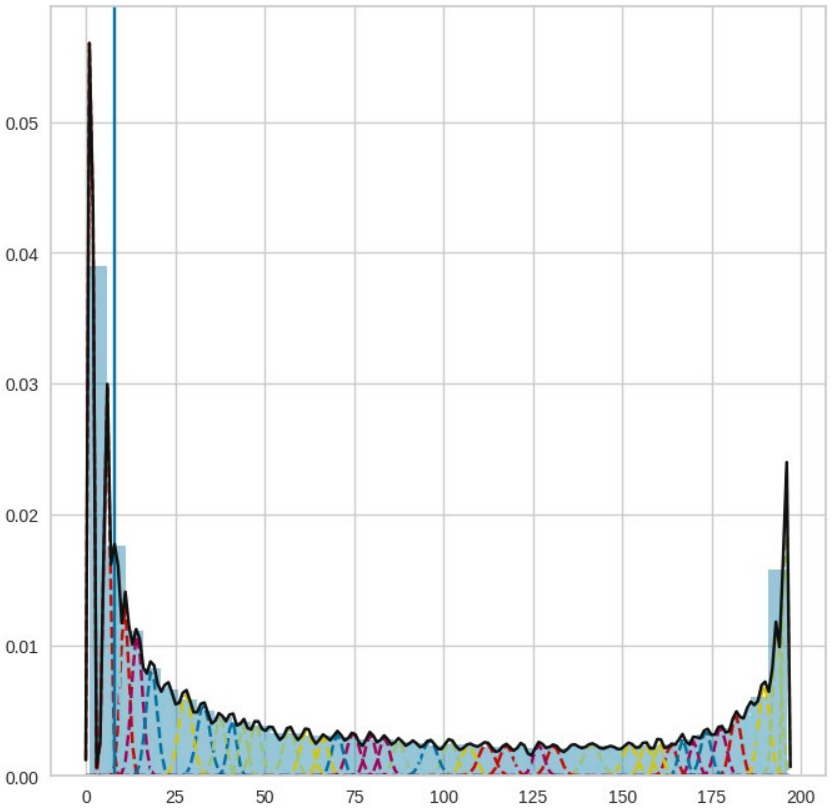

Fig. 3. GMM result
Rys. 3. Wynik działania GMM

After applying the gene rejection criteria based on low expression levels, the dataset was reduced from its original size of 61,587 genes to 32,606 genes.

3.3.2. Clustering

Determining the optimal number of clusters is a critical step in clustering analysis. The Calinski-Harabasz index and the Davies-Bouldin index were used to evaluate the clustering performance and to determine the optimal number of clusters. As the number of clusters increased, the Davies-Bouldin index increased, indicating worsened cluster separation and compactness. The highest Calinski-Harabasz index reflects a clustering solution with the most distinct clusters. Based on the indices values, it was observed that the solution with two clusters yielded the best between cluster separation and compactness within each cluster.

Clustering is a fundamental technique in unsupervised machine learning that involves grouping similar data points together based on their intrinsic characteristics. It is a process of dividing a dataset into subsets or clusters, where data points within each cluster have certain similarities and are dissimilar to data points in other clusters.

Sparse k-means involves selecting a sparse set of relevant features that contribute most to each cluster. By selecting a sparse set of features, the algorithm reduces the dimensionality of the data, which can improve accuracy and scalability compared to classical k-means [3]. The application of sparse k-means clustering to the gene expression dataset produced distinct and interpretable clustering results. UMAP (Uniform Manifold Approximation and Projection) was used for visualization purposes in the analysis of the dataset. The main UMAP principle is to find a low-dimensional representation which will preserve these neighborhoods as much as possible [4]. The using of UMAP to the patients gene expression dataset resulted in a plot on Fig. 4 that effectively captured the underlying structure and relationships of the data. The UMAP plot revealed distinct clusters representing different groups within the patient population. By applying UMAP to a dataset affected by batch effects, it is possible to visualize and potentially correct for these unwanted variations. However, the second plot provides evidence that there is no batch effect, so there is no need to apply a batch correction technique.

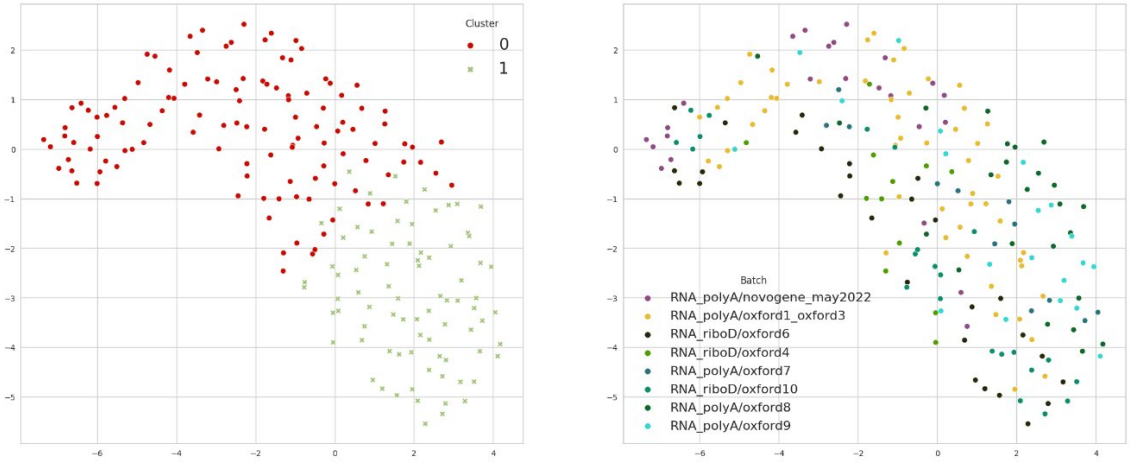

Fig. 4. UMAP visualisation
Rys. 4. Wizualizacja UMAP

The improved separation between clusters and reduced dimensionality through feature sparsity contributed to a more efficient and interpretable clustering solution.

3.3.3. Statistical analysis

To identify the most diversifying genes within specific clusters, the Mann-Whitney test was utilized. By performing the Mann-Whitney test on a feature of interest within different clusters, was determined whether the feature exhibits significant variation that distinguishes the clusters.

The Rank Biserial Correlation Coefficient can be used to assess the effect size in the context of clustering analysis. A higher absolute value of the coefficient indicates a stronger cluster separation based on the feature.

Figure 5 shows box plots illustrating the gene expression levels of of the most diversifying genes across different clusters, including MYH7, TNNI1 and ACTN2. These muscular genes play important roles in muscle contraction. Both cardiac and skeletal muscle disorders can result from a defect of MYH7 [5]. The TNNI1 gene is switched on during skeletal muscle myogenesis and is co-expressed during early stages of development [6]. ACTN2 is highly abundant in cardiac and skeletal muscle, where it plays several functional roles in the sarcomeres [7]. Each box corresponds to a different cluster and the vertical axis represents the gene expression values. The box represents the interquartile range of values within each cluster, with the median indicated by a line inside the box. By examining the box plots, we can observe variations in gene expression levels between clusters.

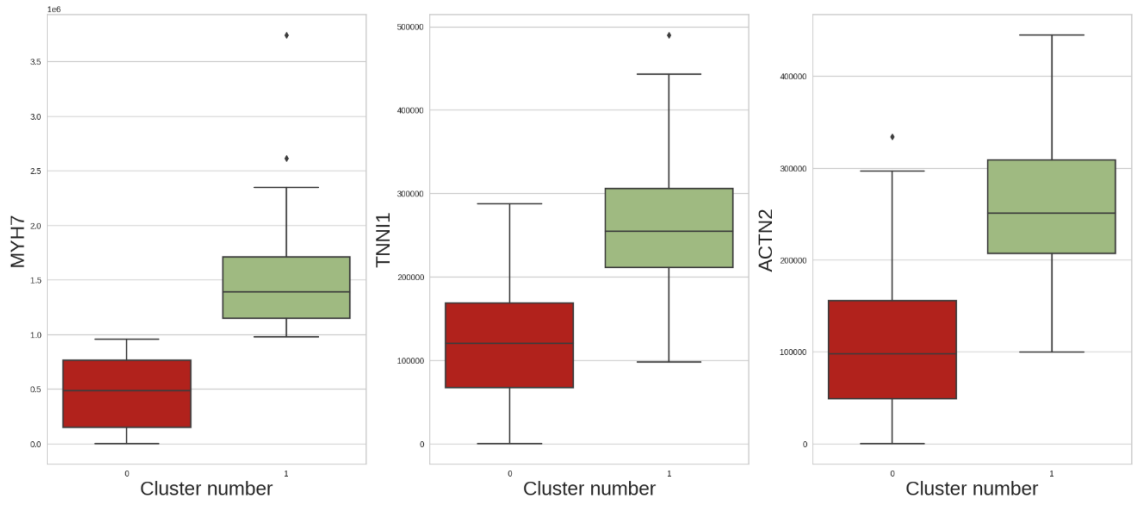

Fig. 5. Gene expression values across clusters
Rys. 5. Wartości ekspresji genów pomiędzy klastrami

3.3.4. Dimensionality reduction

Dimensionality reduction is an additional step that was included in the preprocessing pipeline to reduce the number of features in the dataset while preserving the most important information. Autoencoders were used as part of the analysis. The autoencoder neural network was trained on the dataset to learn a compressed representation of the input data. By minimizing the reconstruction error between the original input and the reconstructed output, the autoencoder effectively captured the most important features but yield a different number of clusters than what was previously determined. The analysis using the autoencoder resulted in the identification of 5 distinct clusters within the dataset which are demonstrated in Fig. 6 after UMAP transformation.

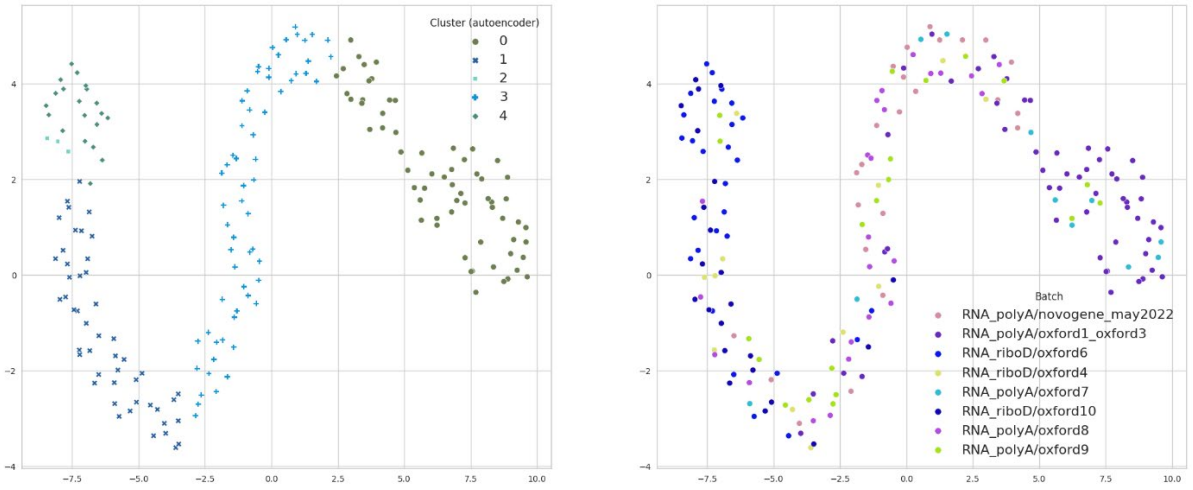

Fig. 6. UMAP space after applying the autoencoder
Rys. 6. Przestrzeń UMAP po użyciu autenkodera

The Sankey diagram shows on Fig. 7 the distribution of genes within clusters, comparing the results with and without dimensionality reduction techniques. It visually represents the flow of genes between clusters in both scenarios, highlighting any similarities in gene assignments and transitions.

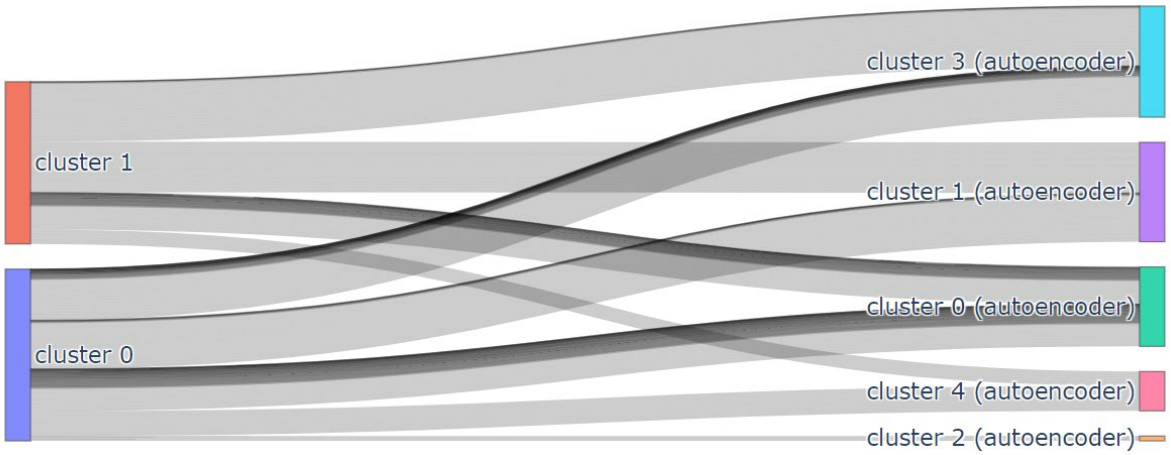

Fig. 7. Sankey diagram
Rys. 7. Diagram Sankey

3.3.5. Model analysis

Using the SHAP algorithm, we can calculate the impact of each feature on a given prediction by evaluating its contribution to the difference between the predicted outcome and the expected average prediction. SHAP (Shapley Additive exPlanations) is a model-agnostic interpretability technique used to explain the predictions of machine learning models [8]. Positive SHAP values indicate features that push the prediction towards the positive class, while negative values indicate features that push it towards the negative class. Fig. 8. Illustrates the results of the SHAP algorithm using Random Forest as the prediction model. By analyzing the SHAP values of the genes, it is possible to identify the genes that have the most significant effect on the separation of the clusters.

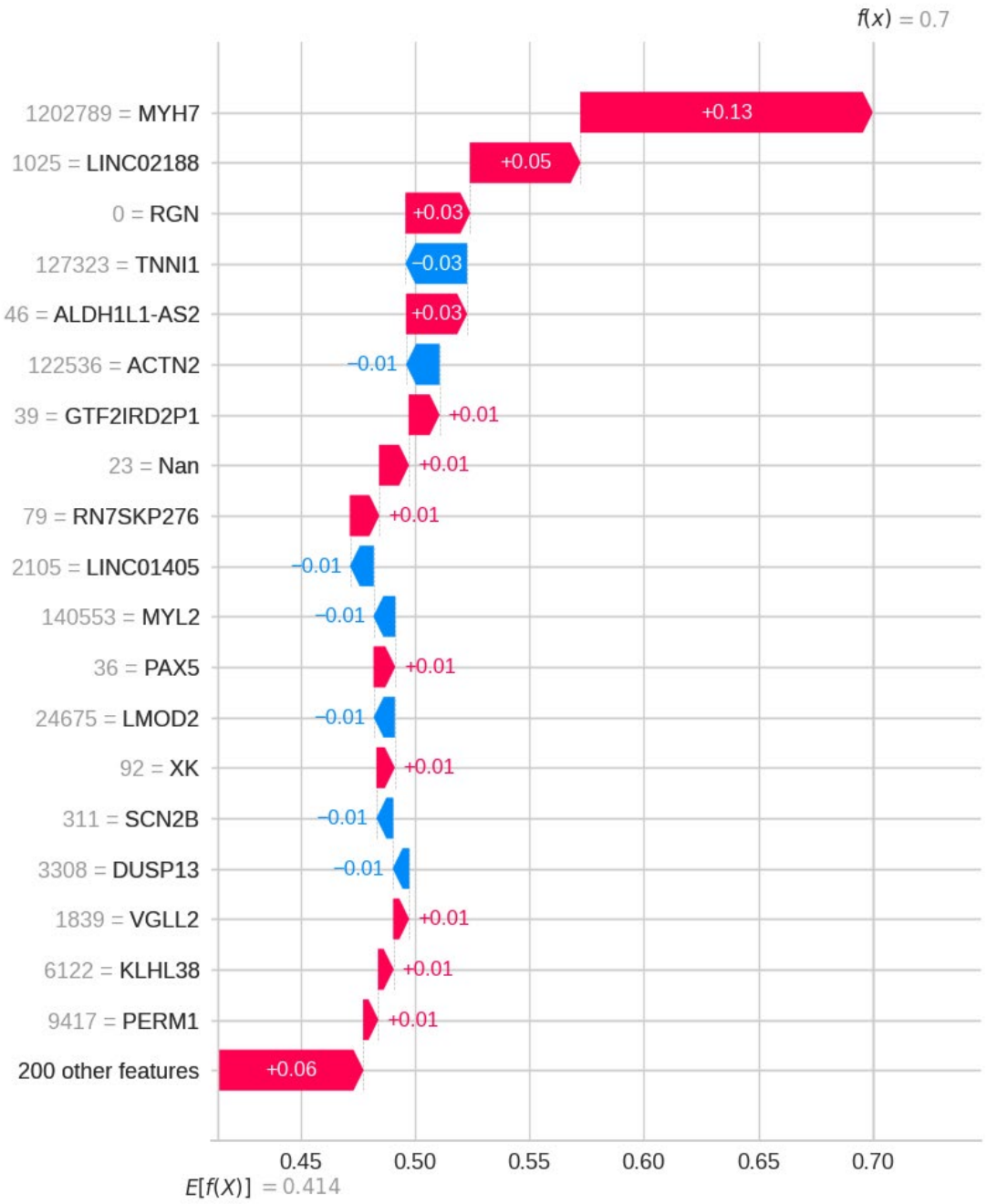

Fig. 8. SHAP
Rys. 8. SHAP

3.4. Conclusions

In conclusion, clustering analysis of the patients dataset revealed two distinct subgroups based on genetic profiles. The analysis of muscular genes revealed diverse expression patterns across different clusters. Specifically, the genes MYH7, TNNI1, or ACTN2 exhibited varying expression levels among the clusters. Further investigation

into the specific genetic mutations and biomarkers associated with each cluster could potentially lead to personalized treatment strategies and improved prognostic indicators for patients with muscular dystrophies.

Bibliography

1. Prof Alan EH Emery FRCP, The Lancet, *The muscular dystrophies* (2002) **359**:687–695.
2. Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, Ali Mortazavi, Genome Biology, *A survey of best practices for RNA-seq data analysis* (2016) **13**.
3. Daniela M. Witten, Robert Tibshirani, J Am Stat Assoc, *A framework for feature selection in clustering* (2010) **490**: 713–726.
4. Alex Diaz-Papkovich, Luke Anderson-Trocmé, Simon Gravel, Journal of Human Genetics, *A review of UMAP in population genetics* (2021) **66**: 85–91.
5. C. Fiorillo, G. Astrea, M. Savarese, D. Cassandrini, G. Brisca, F. Trucco, M. Pedemonte, R. Trovato, L. Ruggiero, L. Vercelli, A. D'Amico, G. Tasca, M. Pane, M. Fanin, L. Bello, P. Broda, O. Musumeci, C. Rodolico, S. Messina, G. L. Vita, M. Sframeli, S. Gibertini, L. Morandi, M. Mora, L. Maggi, A. Petrucci, R. Massa, M. Grandis, A. Toscano, E. Pegoraro, E. Mercuri, E. Bertini, T. Mongini, L. Santoro, V. Nigro, C. Minetti, F. M. Santorelli, C. Bruno on behalf of the Italian Network on Congenital Myopathies, Orphanet Journal of Rare Diseases, *MYH7-related myopathies: clinical, histopathological and imaging findings in a cohort of Italian patients* (2016) **11:** 11–91.
6. Antony J. Mullen, Paul J.R. Barton, Gene, *Structural characterization of the human fast skeletal muscle troponin I gene (TNNI2)* (2000) **242:** 313–320.
7. Marco Savarese, Johanna Palmio, Juan José Poza, Jan Weinberg, Montse Olive, Ana Maria Cobo, Anna Vihola, Per Harald Jonson, Jaakko Sarparanta, Federico García-Bragado, Jon Andoni Urtizberea, Peter Hackman, Bjarne Udd, Annals of Neurology, *Actininopathy: A new muscular dystrophy caused by ACTN2 dominant mutations* (2019) **6:** 899–906.
8. Guy Van den Broeck, Anton Maximilian Schleich, Dan Suciu, Journal of Artificial Intelligence Research, *On the Tractability of SHAP Explanations* (2022) **74:** 51–886.

TRANSCRIPTOMICS-BASED MUSCULAR DYSTROPHY PATIENT STRATIFICATION WITH THE USE OF MACHINE LEARNING

Abstract

Muscular dystrophies are a set of genetic disorders characterized by progressive muscle wasting and degeneration. Each type of muscular dystrophy is associated with specific gene mutations that have an effect on the function of the muscles. We aim to identify patient subpopulations and find their molecular signatures. The dataset consists of 61,587 RNA-Seq-based gene expression profiles from 197 patients with different muscular dystrophies. The preprocessing pipeline was created to filter noisy data that can lead to misinterpretation of the study results.

First, genetic data were transformed to present/absent form, and then the Gaussian Mixture Model of their occurrence distribution was used to identify genes with high absences. Then, the sparse K-means algorithm was applied to partition samples into clusters with Calinski-Harabasz and Davies Bouldin indices used to find the cluster number. Differences in gene expression across clusters were detected by the Mann--Whitney test and rank biserial correlation coefficient serving as effect size measure. The large effect size was observed in 219 features, mainly associated with muscular genes such as MYH7, TNNI1, or ACTN2. UMAP transformation was used to visualize results, and the 2D graph spanned over the limited feature domain was built to confirm obtained patient splitting.

Keywords: muscular dystrophy, unsupervised learning, sparse K-means, UMAP