Application of artificial intelligence for easier monitoring of alleles frequencies in population of the pregnant women (patients), associated with hereditary thrombophilia

Marko Živanović, M.Sc
Academy of Technical and Art Applied Studies Belgrade,
The School of Electrical and Computer Engineering, Belgrade, Serbia
markoz@gs.viser.edu.rs

Nebojša Bogdanović, Ph.D.
Institute of Molecular Biophysics, The Southeastern Center for Microscopy of MacroMolecular Machines (SECM4), Florida State University, Tallahassee, FL, USA
nbogdanovic@fsu.edu

Stefan Erčić, M.Sc.
Grammar school (Savremena Gimnazija, Zemun, Belgrade) Serbia
stefercic@gmail.com,

Vanja Luković, Ph.D
University of Kragujevac, Faculty of Technical Sciences Čačak
vanja.lukovic@ftn.kg.ac.rs

Katarina Živojinović, M.Sc.
Family Medica, polyclinics & surgeries
kzivojinovic@familymedica.rs

Abstract **— Thrombophilia is a hereditary condition characterized by an increased tendency for blood clot formation, affecting approximately 8–11% of the European population. It is typically inherited in an autosomal dominant pattern and includes around 10 subtypes, categorized based on genetic. The condition often leads to complications in pregnant women, including spontaneous abortion or fetal deformities, and increases the risk of heart attack, stroke, pulmonary embolism, and deep vein thrombosis. The Generalized Predictive Tool (GPT) aims to monitor allele frequencies associated with hereditary thrombophilias and predict pregnancy complications, enhancing early detection and management. A clustering analysis of the dataset identified three distinct clusters: Cluster 1, the largest, with 1,946 instances; Cluster 0, with 778 instances; and Cluster 2, the smallest, with 35 instances. The imbalance in cluster sizes reflects the dataset's inherent structure or the complexity of the clusters. The study employed unsupervised machine learning, with K-means outperforming hierarchical clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Gaussian Mixture Model (GMM). This analysis offers valuable insights into sub-group characteristics and presents opportunities for predictive modeling, particularly for thrombophilia, to enhance risk management during pregnancy.**

Keywords-keyword Thrombophilia; Pregnancy K-means, DBSCAN, GMM,; Silhouette Score, Davies-Bouldin Index and Calinski-Harabasz Index Artificial intelligence;

I. INTRODUCTION

Thrombophilia is a group of disorders in which blood has an increased tendency to clot. It may be caused by inherited or acquired conditions. Secondary disorders include heparin-induced thrombocytopenia, antiphospholipid antibody syndrome, neoplasia, oral contraceptive use, obesity, smoking and surgery. Primary disorders or genetic causes of thrombophilia include Factor V Leiden mutation, deficiency of antithrombin III, protein C or S, histidine-rich glycoprotein deficiency and prothrombin-related thrombophilia.

Thrombophilia is associated with risk of deep venous thrombosis and/or venous thromboembolism. Thrombosis may occur in uncommon sites such as the splanchnic, cerebral, and retinal veins. However, the clinical expression of hereditary thrombophilia varies widely. Some individuals never develop thrombosis, others may remain asymptomatic until adulthood and others have recurrent thromboembolism before 30 years of age.

Patients heterozygous for the Factor V Leiden or FII (Prothrombin G20210A) mutation are at a mild risk of thrombosis and 3.8 and 4.9 times, respectively more prone to a first blood clot. However, if the patient is the carrier of both heterozygous mutations, then the risk becomes higher and increases by up to 20 times. Homozygous patients with FII and FV mutations are extremely rare in the general population [1].

Factor V Leiden thrombophilia is the most common inherited form of thrombophilia. The prevalence in the US and European general populations is 3-8% for one copy of the factor V Leiden mutation; about 1:5000 persons have two copies of the mutation [2]. Moderate protein S deficiency is

estimated to affect 1:500 individuals. Severe deficiency is rare and its prevalence is unknown [3]. Moderate protein C deficiency affects about 1:500 individuals. Severe deficiency occurs in about 1:4000000 newborns [4]. Prothrombin-related thrombophilia is the second most common genetic form of thrombophilia, occurring in about 1.7-3% of the European and US general populations [5]. Hereditary antithrombin III deficiency has a prevalence of 1:500-5000 in the general population [6]. Thrombophilia has autosomal dominant, autosomal recessive, or X-linked inheritance [7].

The clustering results in this study indicate three distinct clusters within the dataset. Cluster 1, which contains the dominant number of instances (1946), is the largest cluster in the analyzed dataset. Cluster 0 is smaller, with 778 instances, while Cluster 2 represents the smallest group, containing only 35 instances. This distribution reflects a significant imbalance, with Cluster 1 containing many more instances than the other clusters. The clustering model used for this analysis is based on unsupervised machine learning, which is particularly useful since the dataset does not contain a target column.

The unsupervised model identified three clusters: 0, 1, and 2. Among the various clustering methods tested, K-means showed the best performance. Other methods, such as hierarchical clustering, DBSCAN, and GMM, provided useful insights but did not outperform K-means in this case. This clustering analysis helps reveal the inherent structure within the dataset and is crucial for understanding the characteristics of different sub-groups, especially in relation to the hereditary thrombophilia condition.

This study applies unsupervised machine learning to analyze allele frequency patterns in hereditary thrombophilia. It introduces a Generalized Predictive Tool (GPT) for monitoring allele frequencies and predicting pregnancy complications. Clustering analysis identified three patient groups, with K-means outperforming other methods. Using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, the study validates K-means for thrombophilia classification. These findings enhance genetic risk assessment and patient management, particularly during pregnancy.

II. FREQUENCY IN POPULATION

Almost half of all hereditary forms are represented in the disorder Factor V Leiden (FVL). The disorder is caused by resistance to active protein C, impairing its ability to degrade factors Va and VIIIa. Inheritance is autosomal dominant. This disorder is thought to affect 8% of Europeans. In contrast, only 3% was found in the population living in Africa, China, Japan, and other parts of Asia, as well as among African Americans. The influence on the course and outcome of pregnancy largely depends on the fact whether it is a homozygous or heterozygous disorder, but also on other acquired disorders and risk factors that the individual possesses [8]. The percentage of homozygosity is up to 1% of cases, but it manifests a much more pronounced phenotypic severity of the condition. From a genetic perspective, the disorder is thought to be based on a mutation on chromosome 1q 23 where arginine is changed to glutamine [9].The most common complications can be categorized by their occurrence during specific periods of pregnancy, distinguishing between maternal and fetal complications [10].

In complications of pregnancy, caused as potential consequences of genetic thrombophilia include:
1. early pregnancy losses
2. repeated first trimester pregnancy loss
3. second and third trimester pregnancy loss
4. preeclampsia
5. placental abruption
6. premature birth
7. intrauterine fetal death (fetus mortus in utero - FMU)
8. fetal growth restriction (fetal growth restriction - FGR)
9. venous thromboembolism (VTE) [11].

One study reported that the risk of some complication in pregnancy is 3.8 times higher in women who are Factor V Leiden positive, compared to women who do not have this mutation. The live birth rate in FVL positive women was only 11% compared to 49% in women with a normal genotype. In this case too, the results between different studies are contradictory. A population study that included more than 2,000 pregnant women showed that the percentage of live births was 89% after recurrent miscarriages, 98% after a single late loss, while the pre-index pregnancy percentages were 28%, 49% and 30% [12]. In a large family study that included patients with documented venous thromboembolism or premature atherosclerosis and possession of FV Leiden or prothrombin G20210A mutations, and their first-line relatives, live birth rates in second pregnancy after first loss were 77% in carriers and 76% in non-carriers. -carriers after the first early miscarriage. After a late miscarriage in the first pregnancy, the live birth rate in the second pregnancy was 68% in carriers and 80% in non-carriers [13].

The pooled results of 10 studies that included a large sample of over 20,000 women and were able to be analyzed showed a slightly increased risk of pregnancy loss for women carrying the Factor V Leiden mutation, but not for those carrying the prothrombin G20210A mutation [14].

III. MATERIALS AND METHODS

This study utilized a dataset of 2760 samples, with 90% of the patients being pregnant women aged 18 to 40, spanning data from 2018 to 2024. The analysis was conducted using the Python programming language and the Power BI tool. All clustering methods are thoroughly presented in the Clustering section.

IV. CLUSTERING

Clustering is an unsupervised learning technique used to group data based on their similarity. The idea of clustering is to identify latent groups within a dataset, where objects within the same group (cluster) are more similar to each other than to objects in other groups. Clustering has a wide range of applications in various domains, including biomedicine, market analysis, natural language processing, pattern

recognition, and social networks [15]. A cluster C is formally defined as a subset of data X where the intra-cluster similarity is maximized, as expressed in (1):

$$\frac{1}{|C|}\sum_{x_i,x_j\in C} sim(x_i,x_j), \tag{1}$$

indicating that points within the same cluster are highly similar based on the similarity measure $sim(x_i,x_j)$, Additionally, the inter-cluster similarity is minimized, as shown in (2):

$$\min_{x_i\in C_a, x_j\in C_b} sim(x_i,x_j), \quad a\neq b \tag{2}$$

ensuring that points from different clusters are as dissimilar as possible. This dual criterion forms the foundation of clustering, emphasizing cohesion within clusters and separation between them. Where C_a and C_b are distinct clusters [16]. This dual criterion ensures that clusters are internally cohesive while remaining well-separated from one another.

A. *K-means method*

The K-means algorithm is one of the most well-known algorithms for partitioning clustering, with the goal of dividing data into k clusters in such a way that it minimizes the sum of squared distances between points and cluster centers [17]. Formally, the algorithm optimizes the following function:

$$\min_{\{C_1,C_2,\ldots,C_k\}}\sum_{i=1}^{k}\sum_{x\in C_i}\left\|x-\mu_i\right\|^2 \tag{3}$$

Where C_i are the clusters, x is a point belonging to cluster C_i and μ_i is the centroid of the cluster, defined as the mean of the points within the cluster, defined as the mean of the points within the cluster, as shown in formula (4):

$$\mu_i=\frac{1}{|C_i|}\sum_{x\in C_i} x \tag{4}$$

The K-means algorithm is used for analyzing genetic data represented by numerical values for different genes (Fig 1). Yellow color represented. For each genotype, a mapping set of values ("mut/mut", "wt/mut", "wt/wt") was used, which allows the data to be processed numerically. Cluster 2, marked in yellow, represents the number of samples without mutations. Cluster 1, marked in green, is predominantly homozygous, while Cluster 0, marked in purple, consists of a heterozygous combination of alleles.

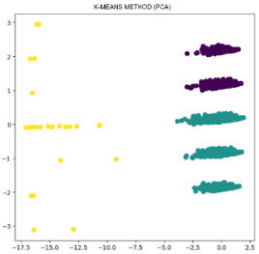

Figure 1. K-means method

B. *Elbow method*

The Elbow method is a commonly used technique for determining the optimal number of clusters in k-means clustering. The goal of this method is to find the number of clusters k that best represents the data, minimizing the variance within the clusters (Fig 2). The process is based on analyzing the change in within-cluster variance as the number of clusters increases, and the point at which the improvement begins to decrease indicates the optimal number of clusters [18].

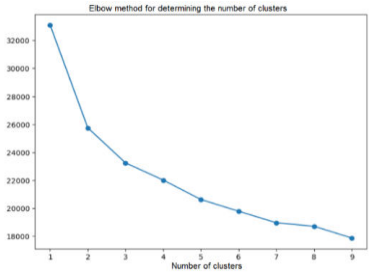

Figure 2. Elbow method for determing the number of clusters

First, the values of $WCSS(k)$ are calculated for different numbers of clusters k. Then, a graph is created where the x-axis represents the number of clusters and the y-axis represents the within-cluster variance.

$$WCSS(k)=\sum_{i=1}^{k}\sum_{x_j\in C_i}\left\|x_j-\mu_i\right\|^2 \tag{5}$$

In the graph, the point where there is a sharp decrease in WCSS values is sought, after which the decrease becomes less steep[19]. This point is known as the elbow, and it indicates the optimal number of clusters. The number is provided in point 3 of our work.

C. *GMM*

The GMM is a probabilistic model that uses a mixture of several normal (Gaussian) distributions to model a data set. GMM is applied in data clustering when we assume that the data comes from multiple different normal distributions, but we do not know exactly which distribution each data point originates from (Fig 3). This model is based on the idea that a data set can be modeled as a combination of several Gaussian distributions, where each distribution represents a cluster. The model's parameters include the mean, covariance, and the weight of each distribution in the mixture[20].

Mathematically, GMM assumes that the probability of observing a data point x is given by a mixture of K Gaussian distributions, and the probability that the data point x belongs to the k-th component of the GMM is calculated using the following formula:

$$p(x)=\sum_{k=1}^{K}\pi_k N(x\mid\mu_k,\sum_k) \tag{6}$$

Where $p(x)$ is the probability that the data point x belongs to one of the components of the mixture, K is the number of components (clusters), π_k is the weight of the K-th component in the mixture, and $N(x|\mu_k, \sum_k)$ represents the Gaussian distribution with mean μ_k and covariance matrix $\sum_k$ That describes the K-th component.

GMM uses the following parameters: π_k (the weight of the k-th component), μ_k (the mean of the k-th component), and $\sum_k$ (the covariance matrix of the k-th component). Each component is μ_k represented as a Gaussian distribution with parameters μ_k and $\sum_k$ .In a simple case, when it is assumed that the components are independent and identically distributed, the covariance may be a simple scalar value (diagonal covariance). Here's the translation: In Figure 3, yellow color represents the number of patients where no mutation was identified. The number 1 represents the samples in which a mutation was identified on both alleles for most types of thrombophilia (dominant homozygote), marked in pink. The number 0, indicated by the blue color, represents the number of samples in which a mutation was identified on only one allele (heterozygote) for most types of thrombophilia.

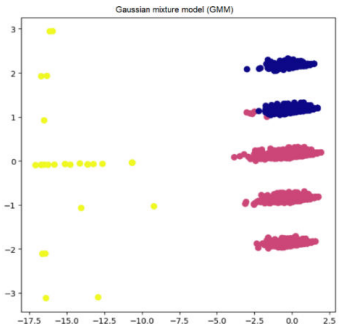

Figure 3 GMM

D. DBSCAN

Density-based clustering, such as DBSCAN, identifies clusters based on the density of points in space. DBSCAN does not require a predefined number of clusters (Fig 4), but instead defines clusters as sets of connected points where each point has at least minPts neighbors within a radius ϵ, which is formally represented by the density connectivity $N_\epsilon(x)$ (7). This set of points is defined as:

$$N_\epsilon(x) = \{y \in X : \|x - y\| \leq \epsilon\} \quad (7)$$

Where x is a point in the dataset x, and y are the neighbors within the radius ∈. A point x is considered a core point if the number of points in its neighborhood $N_\epsilon(x)$ is greater than or equal to the minimum number of neighbors **minPts**, as expressed by:

$$|N_\epsilon(x)| \geq \min Pts \quad (8)$$

DBSCAN classifies points into three categories: core points (which have enough neighbors), border points (which are not core points but are connected to core points), and noise (points that are not connected to any cluster)[21]. The advantage of DBSCAN is that it can detect clusters of any shape, unlike K-means, which is based on spherical clusters. DBSCAN can also identify and discard noise in the data, making it suitable for analyses with errors or irregularities. However, the parameters ϵ and minPts must be carefully chosen, as selecting incorrect values can lead to poor results. DBSCAN is particularly useful in situations with clusters of unequal densities, but it may have difficulty detecting clusters when the density varies within the dataset. In Figure 4, based on the available data, the data is classified into clusters based on data without mutations, while the dominant cluster is the one that includes both homozygous and heterozygous.

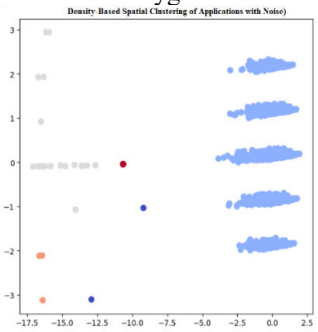

Figure 4 Density – Based Spatial Clustering for Applications with Noise

E. HIERARCHICAL CLUSTERING

Hierarchical clustering is a technique that creates a hierarchical structure of clusters, organized in the form of a tree (dendrogram). The clustering can be agglomerative (bottom-up), where each data point starts as a separate cluster, and then the closest clusters are iteratively merged, or divisive (top-down), where all data points start in a single cluster, which is then split into smaller clusters (Fig 5)[22].

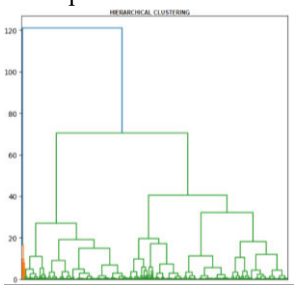

Figure 5 Hierarchical clustering

Mathematically, the distance between clusters can be defined in various ways: single linkage uses the minimum distance between points in two clusters, calculated as

$$d(C_1, C_2) = \min_{x \in C_1, y \in C_2} d(x, y) \quad (9)$$

Average linkage calculates the average distance between all pairs of points from two clusters, as

$$d(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{x \in C_1} \sum_{y \in C_2} d(x, y) \quad (10)$$

complete linkage uses the maximum distance,

$$d(C_1, C_2) = \max_{x \in C_1, y \in C_2} d(x, y) \quad (11)$$

while centroid linkage calculates the distance between the centroids of clusters:

$$d(C_1, C_2) = d(\mu C_1, \mu C_2) \tag{12}$$

where μC is the centroid of the cluster [23]. Agglomerative clustering begins with each data point as a separate cluster, and at each step, the closest clusters are merged until there is only one cluster containing all the points. The dendrogram is a graphical representation of this process, where each point or cluster gradually merges to form a hierarchical structure. At the end of the process, the clustering results can be cut at a certain height of the tree, determining the number of clusters. This method is very flexible, as it can detect clusters of various shapes and structures, and the choice of the distance method between clusters is crucial for the quality of the results.

V. EVALUATION MODEL

The code provided includes evaluations of different clustering algorithms (K-means, Gaussian Mixture Models, DBSCAN, and hierarchical clustering) using three important clustering metrics: Silhouette Score, Davies-Bouldin Index and Calinski-Harabasz Index. These metrics help in assessing the quality and effectiveness of the clustering models [24]. The Silhouette Score measures how similar each point is to its own cluster compared to other clusters. It ranges from -1 to +1-A value near +1 indicates that the points are well-clustered. A value near 0 suggests that the points are on or very close to the decision boundary between clusters.

A value near -1 indicates that the points may have been incorrectly clustered. Mathematically, the silhouette score for a point is calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{13}$$

Where $a(i)$ represents the average distance between point i and all other points in the same cluster (cohesion), and $b(i)$ is the average distance between point i and all points in the nearest cluster to which iii does not belong (separation). The Davies-Bouldin Index (DBI) is another measure of cluster quality. The goal is to minimize this index. It considers the average similarity ratio of each cluster with the cluster that is most similar to it. A lower value of DBI indicates better clustering performance. Mathematically, the Davies-Bouldin Index is defined as:

$$DB = \frac{1}{N} \sum_{i=1}^{N} \max_{i \neq j} \left(\frac{s_i + s_j}{d(c_i, c_j)} \right) \tag{14}$$

Where S_i is the average distance of all points in cluster i to its centroid $d(c_i, c_j)$ is the distance between the centroids of clusters i and j. N is the number of clusters. The Calinski-Harabasz Index (Variance Ratio Criterion) is used to assess the compactne ss and separation of clusters. The higher the value, the better the clusters are formed. It is also known as the Variance Ratio Criterion (VRC). The formula is:

$$CH = \frac{Tr(B_k)}{Tr(W_k)} \times \frac{N - k}{k - 1} \tag{15}$$

The Calinski-Harabasz index measures how well-separated the clusters are and how compact they are within themselves. In the code, this index is computed for all the clustering algorithms, with higher values suggesting better clustering. In formula $Tr(B_k)$ is the trace of the between-cluster dispersion matrix, $Tr(W_k)$ is the trace of the within-cluster dispersion matrix. N is the number of points in the dataset, and k is the number of clusters. The evaluation of clustering algorithms shows that K-means, GMM, DBSCAN, and hierarchical clustering exhibit different performances based on key metrics**.** K-means achieved an inertia of 3479.20, a Silhouette Score of 0.42, a Davies-Bouldin Index of 0.69, and a Calinski-Harabasz Index of 3836.63, proving to be a balanced and reliable model with well-defined clusters. Inertia (also referred to as WCSS – Within-Cluster Sum of Squares) is a metric used to evaluate the quality of clusters in K-means clustering (see Table 1).

TABLE I. RESULTS OF EVALUATION MODEL

Method	Inertia (WCSS)	Silhouette Score	DBI	CHI
K-means	3479.20	0.42	0.69	3836.63
GMM	-	0.42	0.69	3816.38
Hierarchical	-	0.41	0.71	4011.23

It measures the sum of squared distances between each data point and the centroid of the cluster to which it belongs. Lower inertia values generally indicate that the points are closer to their respective centroids, implying better-defined clusters. Mathematically, inertia is calculated as (16):

$$Inertia = \sum_{i=1}^{n} \sum_{k=1}^{K} I(x_i \in C_k) \cdot \|x_i - \mu_k\|^2 \tag{16}$$

where x_i is the data point, μ_k is the centroid of cluster C_k, and $\|x_i - \mu_k\|^2$ is the squared distance between the data point and the centroid [25]. GMM produced similar results, with a Silhouette Score of 0.42, a Davies-Bouldin Index of 0.69, and a slightly lower Calinski-Harabasz Index of 3816.38, but it did not offer additional advantages over K-means. DBSCAN failed to identify a sufficient number of valid clusters for evaluation, highlighting its sensitivity to hyperparameters and data structure. Hierarchical clustering achieved the highest Calinski-Harabasz Index of 4011.23, but a slightly lower Silhouette Score of 0.41 and a higher Davies-Bouldin Index of 0.71, indicating weaker cluster compactness. Considering the balanced results and reliability, K-means stands out as the best model for this dataset. The results of the clustering indicate three identified clusters in the dataset. Cluster 1 contains the dominant number of data points, specifically 1946 instances, making it the largest cluster in the analyzed set. Cluster 0 is smaller, with 778 data points, while Cluster 2 represents the smallest cluster with only 35 instances. These results suggest a significant imbalance in the

distribution of data among the clusters, with Cluster 1 having a much larger number of instances compared to the remaining clusters **(**Fig 6**).** This distribution may be a result of the specific structure of the dataset or the complexity of the clusters themselves, which can be further explored in subsequent analyses.

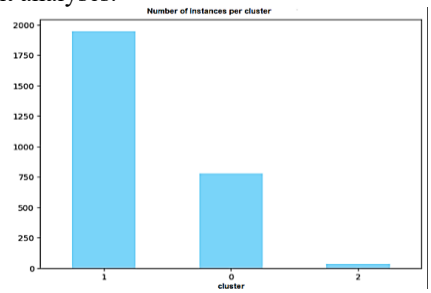

Figure 6 Number of instances per cluster

VI. Conclusion

The clustering analysis revealed three distinct clusters, with Cluster 1 being the most prominent, containing the largest number of instances (1946). This clustering pattern, along with the unsupervised machine learning model, has provided valuable insights into the distribution and characteristics of different sub-groups in relation to hereditary thrombophilia. The model performance evaluation indicated that K-means clustering outperformed methods such as hierarchical clustering, DBSCAN, and GMM. This was confirmed by its higher accuracy in identifying clusters with the highest frequency of specific genotypes, highlighting the relevance of K-means in this context. Furthermore, the analysis suggested that homozygotes are more likely to experience severe complications, while heterozygotes face relatively milder outcomes, which is consistent with previous findings in the literature. This clustering approach enables the prediction of thrombophilia-related risks, particularly for pregnancy-related complications, and lays the groundwork for future AI-based predictive modeling. The ability to identify risk sub-groups with high accuracy can aid in the early detection and better management othrombophilia, improving clinical outcomes for pregnant women.

References

[1] A. Hatzaki et al., "The impact of heterozygosity for the factor V Leiden and factor II G20210A mutations on the risk of thrombosis in Greek patients," International Angiology: A Journal of the International Union of Angiology, vol. 22, no. 1, pp. 79-82, 2003.

[2] J. Kvasnicka et al., "Prevalence trombofilních mutací FV Leiden, protrombinu G20210A a PAI-1 4G/5G a jejich vzájemných kombinací v souboru 1450 zdravých osob stredního veku v regionu Praha a strední Cechy (výsledky real-time PCR analýzy FRET)," Casopis lekaru ceskych, vol. 151, no. 2, pp. 76–82, 2012.

[3] P. Hundsdoerfer et al., "Homozygous and double heterozygous Factor V Leiden and Factor II G20210A genotypes predispose infants to thromboembolism but are not associated with an increase of foetal loss," Thrombosis and Haemostasis, vol. 90, no. 4, pp. 628–635, 2003, doi: 10.1160/TH03-02-0096.

[4] J. L. Kujovich, "Prothrombin-related thrombophilia," GeneReviews, University of Washington, Seattle, WA, 2006.

[5] A. Dautaj et al., "Hereditary thrombophilia," Acta Biomedica, vol. 90, Suppl. 10, pp. 44–46, 2019.

[6] M. M. Patnaik and S. Moll, "Inherited antithrombin deficiency: A review," Haemophilia, vol. 14, no. 6, pp. 1229–1239, 2008.

[7] N. Aračić et al., "The impact of inherited thrombophilia types and low molecular weight heparin treatment on pregnancy complications in women with previous adverse outcomes," Yonsei Medical Journal, vol. 57, no. 5, pp. 1230–1236, 2016.

[8] A. M. Pritchard, P. W. Hendrix, and M. J. Paidas, "Hereditary thrombophilia and recurrent pregnancy loss," Clinical Obstetrics and Gynecology, vol. 59, no. 3, pp. 487–497, 2016.

[9] N. Wawrusiewicz-Kurylonek, A. J. Krętowski, and R. Posmyk, "Frequency of thrombophilia-associated gene variants: Population-based study," BMC Medical Genetics, vol. 21, no. 1, p. 198, 2020.

[10] S. Ota et al., "Contribution of fetal ANXA5 gene promoter polymorphisms to the onset of preeclampsia," Placenta, vol. 34, pp. 1202–1210, 2013.

[11] S. V. Dugalić, "Complications of pregnancy in patients with hereditary thrombophilia and the effects of therapy on pregnancy outcomes," Doctoral dissertation, Belgrade, 2020.

[12] P. G. Lindqvist, M. Procházka, R. Laurini, and K. Maršál, "Umbilical artery Doppler in relation to placental pathology and FV Leiden in pregnant women and their offspring," Journal of Maternal-Fetal & Neonatal Medicine, vol. 26, no. 14, pp. 1394–1398, 2013.

[13] L. Coriu et al., "Inherited thrombophilia in pregnant women with intrauterine growth restriction," Maedica (Buchar), vol. 9, no. 4, pp. 351–355, 2014.

[14] M. A. Rodger et al., "Meta-analysis of low-molecular-weight heparin to prevent recurrent placenta-mediated pregnancy complications," Blood, vol. 123, no. 6, pp. 822–828, 2014.

[15] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," Information Sciences, vol. 622, pp. 178–210, 2023.

[16] J. Sun et al., "Inter-cluster and intra-cluster joint optimization for unsupervised cross-domain person re-identification," Knowledge-Based Systems, vol. 251, p. 109162, 2022.

[17] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," Electronics, vol. 9, no. 8, p. 1295, 2020.

[18] P. Bholowalia and A. Kumar, "EBK-means: A clustering technique based on elbow method and k-means in WSN," International Journal of Computer Applications, vol. 105, no. 9, 2014.

[19] M. Cui, "Introduction to the k-means clustering algorithm based on the elbow method," Accounting, Auditing and Finance, vol. 1, no. 1, pp. 5-8, 2020.

[20] D. A. Reynolds, "Gaussian mixture models," Encyclopedia of Biometrics, vol. 741, pp. 659–663, 2009.

[21] J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Density-based clustering in spatial databases: The algorithm gdbscan and its applications," Data Mining and Knowledge Discovery, vol. 2, pp. 169–194, 1998.

[22] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview, II," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 7, no. 6, e1219, 2017.

[23] A. M. Jarman, "Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method," Georgia Southern University, 2020.

[24] R. Baraku, I. N. Yanda, and R. Liwardana, "Analysis of elbow, silhouette, Davies-Bouldin, Calinski-Harabasz, and rand-index evaluation on k-means algorithm for classifying flood-affected areas in Jakarta," Journal of Applied Informatics and Computing, vol. 7, no. 1, pp. 95–1, 2023.

[25] M. Omran, A. Salman, and A. Engelbrecht, "Dynamic clustering using particle swarm optimization with application in unsupervised image classification," Fifth World Enformatika Conference (ICCI 2005), Prague, Czech Republic, pp. 199–204, Nov. 2005.