mathematics

MDPI

Article

Exploring the Molecular Interaction of PCOS and Endometrial Carcinoma through Novel Hyperparameter-Optimized Ensemble Clustering Approaches

Pınar Karadayı Ataş

Department of Computer Engineering, Faculty of Engineering, Istanbul Arel University, 34537 Istanbul, Turkey; pinaratas@arel.edu.tr

Abstract: Polycystic ovary syndrome (PCOS) and endometrial carcinoma (EC) are gynecological conditions that have attracted significant attention due to the higher prevalence of EC in patients with PCOS. Even with this proven association, little is known about the complex molecular pathways that connect PCOS to an increased risk of EC. In order to address this, our study presents two main innovations. To provide a solid basis for our analysis, we have first created a dataset of genes linked to EC and PCOS. Second, we start by building fixed-size ensembles, and then we refine the configuration of a single clustering algorithm within the ensemble at each step of the hyperparameter optimization process. This optimization evaluates the potential performance of the ensemble as a whole, taking into consideration the interactions between each algorithm. All the models in the ensemble are individually optimized with the suitable hyperparameter optimization method, which allows us to tailor the strategy to the model's needs. Our approach aims to improve the ensemble's performance, significantly enhancing the accuracy and robustness of clustering outcomes. Through this approach, we aim to enhance our understanding of PCOS and EC, potentially leading to diagnostic and treatment breakthroughs.

Keywords: machine learning; molecular biology; mathematical modeling; bioinformatics; PCOS; endometrial cancer

MSC: 62H30; 65K10; 92D10; 46N60; 62P10

Citation: Karadayı Ataş, P. Exploring the Molecular Interaction of PCOS and Endometrial Carcinoma through Novel Hyperparameter-Optimized Ensemble Clustering Approaches. *Mathematics* **2024**, *12*, 295. https://doi.org/10.3390/math12020295

Academic Editors: Md Masud Rana, Duc Duy Nguyen and Jin Wang

Received: 30 November 2023
Revised: 26 December 2023
Accepted: 11 January 2024
Published: 16 January 2024

1. Introduction

Polycystic ovary syndrome (PCOS) is a common endocrine disorder that affects 4–12% of females who are of reproductive age, characterized by features such as oligomenorrhea, hyperandrogenism, and polycystic ovaries [1,2]. Despite its widespread occurrence, the association between PCOS and endometrial carcinoma (EC), the most common gynecological cancer in North American and European women, has long been a subject of medical concern [3,4]. Although PCOS and EC may appear to be two separate illnesses at first glance, an intriguing correlation has evolved that has received much attention in the medical field. According to recent research, the chance of women with PCOS developing EC was three times higher than women without PCOS. This correlation has not only aroused interest but also raised critical questions regarding the precise molecular mechanisms that underlie this increased risk. A thorough examination of the underlying genetic and proteomic makeup that links PCOS and EC is necessary due to the complex web of relationships between these two disorders [5–8].

Machine learning is a multifaceted discipline that applies predefined model assumptions to address research issues [9–11]. It makes use of computational power to derive model parameters from training data, enabling it to make predictions and perform data analysis. Notably, machine learning is used in many fields, such as molecular biology and genomics, where it is essential for identifying complex relationships and patterns in genetic

and molecular databases [8,12,13]. Various machine learning modes, such as supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning, are used, depending on the particular research goals and datasets, to drive groundbreaking discoveries and extract valuable insights in the field of genetics and in molecular research [14–16]. The literature has featured multiple research studies that highlight the synergy between machine learning and genetic data, especially in the context of EC [17,18] and PCOS [19,20]. In a recent study [21] focusing on cardiovascular diseases (CVDs), including heart failure (HF) and atrial fibrillation (AF), the application of artificial intelligence (AI) and machine learning (ML) techniques to RNA-seq-driven gene expression data demonstrated the potential for personalized treatments and predictive analysis. In [22], driver genes linked to pathogenic survival that were correlated with patient prognosis were identified using machine learning (ML) analysis. Following the identification of RABGAP1L, MYH9, and DRD4 as candidate genes through the integration of copy number variation and gene expression data, these genes were utilized in conjunction with tumor stages to generate predictive survival models.

The researchers in [23] used an open-source dataset of 541 patients from Kerala, India, to train heterogeneous machine learning and deep learning classifiers to detect PCOS among fertile patients. The objective of their project is to precisely identify PCOS and to provide medical practitioners with an automated screening framework that includes interpretable machine learning technologies. In study [24], a multi-center retrospective analysis at European gynecologic cancer centers was carried out by researchers in order to create a personalized predictive model for EC based on patient and disease characteristics. The primary outcomes of the trial were disease-free survival (DFS) and cancer-specific survival (CSS) at three and five years. Two models were developed using machine learning algorithms: one for pre-treatment and another that combined characteristics related to therapy, perioperative care, and postoperative recovery. The purpose of the other study [25] was to evaluate the diagnostic and prognostic utility of atrial fibrillation (AF)-related gene expression in EC. After examining gene expression data from EC tissues and nearby control tissue, the researchers integrated more genes from earlier research and chose noteworthy genes. In 36 EC patients, they used qPCR to confirm these genes. In addition, a machine learning model was created using these gene expressions to forecast EC grade.

Clustering is an important machine learning technique in the fields of biomedical and molecular biology. It is a fundamental tool that helps researchers identify complex relationships, recognize sophisticated patterns, and make well-informed decisions. It is crucial to recognize that a multitude of thorough studies utilizing clustering approaches across a range of crucial applications have been conducted in this specialized subject. The authors of study [26] introduce a method for clustering DNA sequences without the need for sequence alignment, sequence homology, or taxonomic identifiers. This method is called DeLUCS (Deep Learning method for the Unsupervised Clustering of DNA Sequences). They utilize DNA sequences' frequency chaos game representations (FCGRs) and mimic FCGRs to allow genomic signatures to self-learn via multiple neural networks. In study [27], they mention a procedure that used agglomerative hierarchical clustering in R to classify proteins from the KEGG database based on sequence similarities and Gene Ontology annotations. PPI network analysis and the DAVID Program were also used in their strategy, which provided additional insights. Researchers in [28] thoroughly analyzed more than a million severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike protein sequences in a related study. Their research, to which we make reference in our work, centered on figuring out how these sequences evolved and how they interacted with one another in different global variants. Through the use of methods such as clustering and network analysis, they were able to demonstrate that certain mutations occurred at the same time and had a major impact on important characteristics like the virus's ability to evade immune responses, bind to receptors, and be transmissible. An enhanced network clustering technique called FCAN-MOPSO is provided in [29]. This technique considerably improves convergence rates over the state-of-the-art FCAN algorithm by combining multi-

objective particle swarm optimization. In order to solve imbalances in fuzzy memberships, FCAN-MOPSO adds instance-frequency-weighted regularization to the original FCAN optimization model. The optimization is then broken down into smaller issues, resulting in a balanced local exploitation and global exploration. Compared to other cutting-edge algorithms, FCAN-MOPSO exhibits better accuracy and convergence in complex networks and has been shown to converge globally.

The effectiveness of clustering algorithms is largely dependent on hyperparameter tuning. These algorithms, which include well-known techniques like Hierarchical Clustering, DBSCAN, and K-Means, heavily rely on the proper setting of hyperparameters. The quality and result of the clustering can be significantly changed by adjusting parameters like the epsilon distance in DBSCAN or the number of clusters in K-Means.

Grid Search, Random Search, and Bayesian Optimization are the methods for hyperparameter tuning that are most frequently used in the literature when it comes to clustering algorithms [30]. Depending on the computational power and data complexity, each of these approaches has a unique set of uses and benefits [31]. Grid Search is a systematic and exhaustive method that evaluates every combination of hyperparameters within a predefined grid [32,33]. This technique is particularly advantageous for simpler or smaller datasets, where the hyperparameter space is relatively constrained. Its comprehensive nature ensures that all potential combinations are explored, making it a reliable method for thorough exploration. However, Grid Search's efficiency decreases significantly with the increase in the dimensionality of the hyperparameter space, leading to substantial computational demands. On the other hand, Random Search provides a more effective method for handling bigger datasets. By sampling the hyperparameter space at random, this method allows for a more thorough investigation of possible solutions. Because Random Search can find efficient hyperparameter combinations that more structured methods might miss, it is especially helpful in situations where the optimal hyperparameter range is not well defined [34]. However, given its stochastic nature, it is possible that it will not always find the ideal hyperparameters. Among the more sophisticated methods, Bayesian Optimization stands out as being especially useful for high-dimensional spaces. By building a probabilistic model of the objective function, this approach determines the most promising hyperparameters for further analysis in an astute manner. A balance is struck between discovering new hyperparameters and honing in on those that exhibit potential in Bayesian Optimization [35]. Compared to Grid or Random Search, it is typically more efficient in high-dimensional settings, but its implementation and conceptual understanding are more difficult to fully comprehend.

Furthermore, ensemble learning is one of the most significant machine learning tools [36]. In order to increase prediction accuracy and performance, ensemble learning combines the outputs of several machine learning models. This method works especially well in the fields of genetics and molecular biology, where complex and multidimensional datasets call for sophisticated comprehension [37,38]. Researchers can efficiently capture the subtleties and complexities of genetic and molecular data by utilizing ensemble learning techniques. This can lead to discoveries and provide a more thorough understanding of these complex systems. The existing literature encompasses numerous studies on this topic. The study in [39] presents SELPPI, a stacking ensemble framework for machine learning-based protein–protein interaction modulator prediction. Combining different tree-based techniques with a genetic algorithm takes several chemical descriptors as input parameters. The framework provides a dependable method for finding novel modulators that target protein-protein interactions by achieving predictions using primary and meta-learners. A novel deep ensemble learning-based framework for retinal vascular segmentation was presented in a recent study [40]. In benchmarking comparisons, their model outperformed current techniques, demonstrating its robustness and efficacy across a variety of datasets. Especially useful for our work is this approach's ability to integrate different deep learning models, such as the FCN-Transformer and Pyramid Vision Transformer, to capture discriminative feature representations.

Due to the importance of clustering algorithms in uncovering the underlying genetic and molecular interaction underpinnings of EC and PCOS diseases, increasing the performance of such algorithms is of high importance, and this work is motivated to pursue such enhancements, as there is a noticeable gap that exists in the literature: no model has been found that concentrates on optimizing ensemble weights and hyperparameters at the same time, especially for clustering tasks. This improvement is required due to the complexity of the genetic and molecular information associated with conditions such as EC and PCOS. These disorders are distinguished by complex biological processes and a variety of genetic markers, necessitating the use of sophisticated analytical techniques to precisely determine their molecular causes. Furthermore, since a more accurate and nuanced understanding of these diseases can result in more precisely targeted therapeutic strategies and personalized medicine approaches, improving clustering algorithms has a direct impact on disease characterization and treatment. Thus, our work aims to close this crucial gap by creating an advanced ensemble clustering methodology that addresses the particular difficulties posed by the molecular data of PCOS and EC while simultaneously optimizing weights and hyperparameters. It is possible that this observation will lead to novel approaches in the optimization of machine learning models for clustering in gene-based datasets, providing a special chance for research and generation in the field. Our study is primarily motivated by this gap in the ensemble model generation step. To improve ensemble learning and attain better predictive performance, we seek to investigate and tackle these neglected facets, especially the inclusion of hyperparameter tuning in the ensemble-building process.

In this paper, we present a novel framework to generate an optimal ensemble, a methodology that, for the clustering problem, uniquely combines the simultaneous weighting of models and the tuning of hyperparameters—an approach not previously investigated in previous studies. We have devised a nested algorithm in our design that is specifically tailored to tackle the two problems of tuning hyperparameters and optimizing ensemble weights in the ensemble generation step, where there are gaps in the literature. Instead of using more thorough methods like Grid Search, we use a heuristic approach based on Bayesian search, Hyberband, and Sequential Model-Based Algorithm Configuration (SMAC) to improve the effectiveness of our learning and optimization processes. Conventional weighted ensemble approaches usually adjust hyperparameters separately and treat model weighting and hyperparameter tuning as two distinct processes. On the other hand, our methodology combines these two processes and can choose optimal hyperparameters for each one in order to create the best ensemble overall. This integrated approach presents a novel perspective in ensemble model optimization and represents a significant departure from traditional approaches. This study represents a dual novelty in the field: first, we have compiled gene data specifically for endometrial carcinoma and PCOS, which we have carefully gathered; second, we have introduced a new methodological approach that adds a new perspective to the body of existing research.

2. Materials and Methods

2.1. Dataset

The integrity of the dataset performs a critical role in the building of a predictive model that explains the relationship between PCOS and EC, as it serves as the foundation for the model's eventual predictive power. In order to fully identify the genes linked to both PCOS and EC, a thorough review of the literature was carried out. Our dataset's genes were carefully chosen according to their frequency and significance in the identified literature. Genes such as CYP11a, CYP21, CYP17, and CYP19 for PCOS and MUTYH, CHEK2, TP53, and MLH1 for EC are included in the final list. Numerous studies have demonstrated the established roles that these genes play in the corresponding diseases, which led to their selection. Our manuscript has these genes listed in a comprehensive table with the corresponding references (Table 1). Our search approach comprised a set of keywords including "genetic link", "molecular interaction", "endometrial cancer", and "polycystic ovary syndrome". To guarantee a comprehensive review of the body of research,

we made use of multiple NCBI databases, such as PubMed, Gene, and Protein. We were able to compile a broad range of studies with this method, including research identifying intersecting genes from different sources or experiments as well as experimental data directly connecting PCOS and EC. This custom dataset helped us achieve our goal of creating a model that both complements and adds to the body of current knowledge. Our analysis is built on a foundation that is both scientifically sound and specifically adapted to the subtleties of our research inquiry because we have assembled data from multiple scholarly sources on a personalized basis. The amino acid sequences of these genes and the corresponding proteins were then obtained. Furthermore, we expanded our genomic landscape by including genes that are first- and second-degree relational to our initial gene set using the GeneMANIA software developed by [41]. Understanding the limitations of depending just on genes that are explicitly cited in the literature, we enriched our dataset in our analysis by including both second and third-degree linked genes from the identified studies. Using the research in Table 1, we assembled 192 gene candidates for PCOS, including those first- and second-degree relational to our initial gene set. As with EC, we used 177 gene candidates, such as MUTYH, CHEK2, TP53, POLD1, MLH1, MSH2, MSH6, PMS2, EPCAM, as well as Phosphatase and Tensin Homolog. This painstaking compilation procedure has produced a solid and extensive dataset that will enable our predictive engine to make more accurate predictions. We recognize that, although our dataset is extensive, it comes from both direct and indirect sources. We have therefore been cautious in extrapolating conclusions from our research. Motivated by this dataset, the algorithmic part of our work seeks to identify putative molecular interactions and associations. We are aware, nevertheless, that these results are preliminary and should be confirmed by experiments.

Table 1. Genes associated with PCOS and endometrial cancer.

Disease	Genes	References
PCOS	T2DM, CYP21, CYP17, CYP19, SHBG, AMH, INSR, Calpain10, FTO, CYP11B2, CYP17A1, CYP19A1, CYP1A1, CYP21A2, CYP3A7, Kir6.2, KCNJ11, PPARG, CYP11A, H6PD, Follistatin, LH β-subunit, FSH β-subunit, Dopamine D3 receptor, FSH receptor, Insulin, Insulin receptor, Microsatellite D19S884, IRS1 and IRS2, CAPN10, Resistin, IGF2, PPP1R3, PC-1, Paraoxonase, PAI-1, IL-6, Adiponectin, IL-6 receptor complex, EPHX, Aldosterone synthatase, Tumor necrosis factor receptor-2, Matrix metalloproteinase-1, Factor V, AR	[42–61]
Endometrial Carcinoma	HNF1B, CYP19A1, SH2B3, SOX4, KLF5, AKT1, EIF2AK4, HEY2/NCOA7, MLH1, MSH2, MSH6, PMS2, EPCAM, PTEN, BRCA1, BRCA2, MUTYH, CHEK2, TP53, POLD1, PALB2, BRIP1, RAD51C, RAD51D, STK11, SDHB, SDHC, SDHD, AKT1, PIK3CA, KLLN, SEC23B, NTLH1, RINT1, FAN1, NBN, APC, ATM, FANCO, FANCJ, FANCC, MMR	[62–82]

2.2. Feature Encoding

Protein feature extraction is a more complicated problem than DNA and RNA sequencing because of the wide variety of amino acids and the unique structures and activities of proteins. Many different feature extraction methods have been proposed over time to deal with this complexity [83].

We have prepared the protein sequence data for computer analysis by encoding it using the Composition of k-Spaced Amino Acid Pairs (CKSAAP). This encoding technique captures the essence of short-range interactions between amino acid residues inside a protein sequence or its fragments, effectively representing the sequence context surrounding ubiquitination sites. By utilizing the CKSAAP technique for feature extraction, we highlight the local interactions seen in k-spaced amino acid pairs by analyzing their composition. This method, which is based on the ideas presented by [84], computes the frequency of each k-spaced amino acid pair, where k is the number of intervening residues, and enables us to measure the structural and functional subtleties of protein sequences. This approach im-

proves our analysis and offers a comprehensive perspective on the characteristics connected to proteins that are essential to comprehending these illnesses.

Given a protein sequence, the Composition of k-Spaced Amino Acid Pairs (CKSAAP) can be calculated using the following formula:

$$F_{(AA1,AA2,k)} = \frac{N_{(AA1,AA2,k)}}{N_{\text{total}}} \tag{1}$$

where

- $F_{(AA1,AA2,k)}$ denotes the frequency of the amino acid pair $(AA1, AA2)$ with exactly k amino acids between them in the sequence.
- $N_{(AA1,AA2,k)}$ is the count of the occurrences of the pair $(AA1, AA2)$ separated by k amino acids within the sequence.
- N_{total} is the total number of k-spaced amino acid pair possibilities in the sequence, which serves as a normalization factor for the frequency calculation.

This quantitative measure allows us to encode protein sequences into a numeric format suitable for machine learning algorithms, thereby facilitating the prediction and analysis of protein characteristics and functions.

2.3. Dimensionality Reduction

Principal component analysis (PCA) is used to minimize the feature space and extract the important information from high-dimensional datasets, thus reducing the problem of dimensionality [85]. PCA operates by identifying the eigenvectors $\mathbf{v}_i$ of the data's covariance matrix $\mathbf{C}$, which align with the largest eigenvalues λ_i:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^\top\mathbf{X}, \quad \mathbf{C}\mathbf{v}_i = \lambda_i\mathbf{v}_i. \tag{2}$$

We utilize singular value decomposition (SVD) to decompose the centered data matrix $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top. \tag{3}$$

Through PCA, the dimensionality of our protein feature vector space is reduced from 24,000 to 200:

$$\mathbf{X}_{\text{reduced}} = \mathbf{X}\mathbf{V}_{\text{reduced}}. \tag{4}$$

Here, $\mathbf{V}_{\text{reduced}}$ comprises the leading eigenvectors that capture the bulk of the variance within the dataset, effectively preserving the essence of the data while facilitating more efficient computational analyses. Principal component analysis (PCA) was employed as the dimensionality reduction technique in this study due to its effectiveness in capturing the greatest variety of features and maintaining important information while decreasing noise [86–88]. PCA is especially well-suited for protein data dimensionality reduction. Furthermore—and this is crucial considering the complex patterns of protein sequences—it finds and eliminates connections between amino acid characteristics. With PCA, the data are transformed into a set of orthogonal components, preserving the essential structural and functional properties of the reduced dataset that are required for precise machine learning predictions.

2.4. Clustering Methods

In the field of protein sequence analysis, selecting and optimizing clustering algorithms is crucial to obtaining pertinent insights. This work starts a detailed exploration of several clustering strategies, each unique in how it puts protein sequences together based on shared traits or patterns. We focus on four popular methods: K-Means, Gaussian Mixture Model (GMM), Hierarchical Clustering, and the density-based spatial clustering method (DBSCAN). Knowing the intricacies of every one of these methods is crucial to their

effective use since they all have special advantages and challenges when applied to protein sequence analysis.

The K-Means algorithm is a popular clustering technique that identifies 'k' prototypes as the cluster centroids [89]. K-Means' principal goal is to reduce the sum of squared errors, which can be expressed mathematically as:

$$\min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2, \tag{5}$$

where x indicates a data point, C_i is the i-th cluster in C, μ_i is the centroid of C_i, and C represents the set of clusters. The mean of the points within each cluster is used to compute the centroids. Most of the work in fine-tuning the K-Means algorithm is done on the number of clusters, 'n_clusters'. Additionally, 'k-means++', 'random', or a predefined array can be set for the centroid initialization method, 'init'. Further parameters that affect the algorithm's performance are 'max_iter', the maximum number of iterations per run, and 'n_init', which indicates the number of times the algorithm is run with different centroid seeds.

A probabilistic model known as the GMM makes the assumption that every data point is a combination of multiple Gaussian distributions with unknown parameters [90]. In the realm of clustering, GMMs are widely utilized and can be expressed as follows.

The GMM assumes that each observation x_i is derived from one of K Gaussian distributions given a series of observations $\{x_1, x_2, \ldots, x_N\}$, where each observation is a d-dimensional real vector. The expression π_k, where π_k is the mixing coefficient, represents the likelihood that an observation x_i will be produced by a Gaussian distribution k.

The probability density function of a GMM is a weighted sum of K Gaussian component densities, given by:

$$p(x|\Theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k), \tag{6}$$

where $\mathcal{N}(x|\mu_k, \Sigma_k)$ is the k-th Gaussian density function, μ_k is the mean of the k-th Gaussian, Σ_k is the covariance matrix of the k-th Gaussian, and $\Theta = \{\pi_k, \mu_k, \Sigma_k\}$ represents the parameters of the mixture model.

The hierarchical clustering methods implemented by the sklearn library repeatedly join or split preexisting clusters to form a hierarchy of clusters [91]. The process is represented by a tree structure, in which the leaves represent clusters made up of individual samples, and the root represents a single cluster containing all the samples. "AgglomerativeClustering" is a function in sklearn that is widely used for hierarchical clustering. Because it regulates how sets of observations are separated from one another, the linkage parameter in hierarchical clustering algorithms is significant in agglomerative clustering. There are the following options for this parameter: the smallest distance between two clusters' observations is considered in `single`; the maximum distance between two clusters' observations is used in `complete`; the average distance between two clusters' observations is used in `average`; and `ward` minimizes the variance within the clusters. The linkage criterion selected has a significant impact on the properties and structure of the resulting cluster hierarchy.

The number of clusters, `n_clusters`, is another important hyperparameter in agglomerative clustering. On the other hand, the number of clusters can be automatically calculated by setting the `distance_threshold` parameter, which defines the linkage distance threshold for merging clusters. The relationship can be expressed mathematically as follows:

$$n\_clusters = f(distance\_threshold, linkage), \tag{7}$$

where f is a function determining the number of clusters based on the distance threshold and linkage criteria.

The popular clustering technique DBSCAN does not require pre-specifying the number of clusters [92]. The DBSCAN algorithm relies on two hyperparameters in particular: `eps` (epsilon), the first hyperparameter, specifies the scan radius that is used to find a point's

neighborhood. When evaluating the local point density, this radius is essential. In order to consider a region dense enough to form a cluster, a minimum number of points must be within the `eps` radius. This is specified by the second hyperparameter, `min_samples`. The density threshold for cluster formation in DBSCAN is determined by these parameters taken together.

The clustering process can be mathematically described as follows:

$$\forall p \in \text{Dataset}, \text{Cluster}(p) = \begin{cases} \text{Core Point}, & \text{if } |N_{eps}(p)| \geq \text{min\_samples} \\ \text{Border Point}, & \text{if } |N_{eps}(p)| < \text{min\_samples but } p \in N_{eps}(q), \\ \text{Noise}, & \text{otherwise} \end{cases} \tag{8}$$

where the set of points inside point p's 'eps' radius is denoted by $N_{eps}(p)$. If a point's neighborhood has at least 'min_samples' points, then that point is considered a core point. All other points are categorized as noise, while points near a core point but with fewer neighbors than 'min_samples' are called border points. DBSCAN examines every point inside the 'eps' distance after beginning at an unexplored point. A cluster is identified at any point within this radius that equals or surpasses 'min_samples'. By going through this process recursively for every point, dense regions are recognized as clusters.

2.5. Hyperparameter Optimization

One important part of machine learning is hyperparameter optimization, which is fine-tuning external parameters, $\gamma = \{\gamma_1, \gamma_2, \ldots\} \in \Gamma$, that are set a priori and not learned during training. This procedure is a component of a bi-level optimization problem, wherein optimizing performance with respect to γ is the secondary objective, and optimizing the model's parameter θ is the primary goal. Two datasets are used in the optimization: one for training (X_T) and another for hyperparameter tuning (X_V). They are both sampled from a distribution D independently and identically (i.i.d.). The goal is to minimize the empirical generalization error on X_V. Along with other methods, this is usually accomplished by using a zero–one loss function. In order to represent the generalization error as a function of hyperparameters and to provide guidance for the probabilistic modeling process when choosing new hyperparameters, Bayesian Optimization is frequently utilized. In this iterative process, models are trained using chosen hyperparameters and assessed using validation data. More recently, new approaches such as warm starting and enhanced acquisition functions have been introduced with the goal of improving the speed, accuracy, and applicability of hyperparameter optimization. Throughout this investigation, we utilize an array of sophisticated optimization methodologies, such as Bayesian Optimization (BO), Sequential Model-based Algorithm Configuration (SMAC), and Hyperband, to meticulously adjust and assess our models.

2.5.1. Bayesian Optimization (BO)

Bayesian Optimization (BO) is a well-liked iterative method for resolving hyperparameter optimization (HPO) problems [93]. The foundation of Bayesian Optimization (BO) for modeling objective functions is a Gaussian process (GP). In the configuration space D of hyperparameters, the predictions of outputs $y = f(x)$ for any input x follow a normal distribution, assuming that a function f with mean ℓ and covariance σ^2 is a realization of a GP. The expression for this relationship is:

$$p(y|x; D) = \mathcal{N}(y|\hat{\ell}, \hat{\sigma}^2), \tag{9}$$

where the configuration space is denoted by D, and the evaluation result for each hyperparameter value x is given by $y = f(x)$.

The GP model's confidence intervals are used in the BO–GP framework to determine which points to evaluate next, once a set of predictions has been obtained. The sample records are updated with each newly tested point, allowing the model to be continuously improved with fresh data. Until a predetermined termination criterion is satisfied, this iterative process is repeated.

The application of BO–GP to a dataset of size n incurs a time complexity of $O(n^3)$ and a space complexity of $O(n^2)$. However, one primary limitation of BO–GP is its cubic time complexity with respect to the number of instances, which constrains its capacity for parallelization. Moreover, BO–GP is predominantly used for optimizing continuous variables, limiting its applicability in certain scenarios. There is a time complexity of $O(n^3)$ and a space complexity of $O(n^2)$ when using BO–GP on a dataset of size n. The cubic time complexity of BO–GP with respect to the number of instances, however, is one of its main drawbacks and limits its parallelization potential. Furthermore, BO–GP's applicability in some scenarios is limited because it is primarily used to optimize continuous variables.

2.5.2. Sequential Model-Based Algorithm Configuration (SMAC)

Bayesian Optimization effectively utilizes Sequential Model-based Algorithm Configuration (SMAC) with Random Forest (RF) as a surrogate model for optimizing hyperparameters [94]. The core of SMAC lies in its ensemble of regression trees, B, which collectively model the objective function. The mathematical foundation of this approach is outlined by the mean $\hat{l}$ and variance $\hat{r}^2$ calculations for the regression function $r(x)$ within a Gaussian model $N(\hat{y}|\hat{l}, \hat{r}^2)$:

$$\hat{l} = \frac{1}{|B|} \sum_{r \in B} r(x) \quad \text{(Mean Estimate)} \tag{10}$$

$$\hat{r}^2 = \frac{1}{|B| - 1} \sum_{r \in B} (r(x) - \hat{l})^2 \quad \text{(Variance Estimate).} \tag{11}$$

The procedure for implementing SMAC is as follows. Firstly, B regression trees are constructed by sampling instances from the training set. Each tree is then developed by selecting a split node from a subset of the hyperparameters, with the computational cost controlled by predefined parameters such as the minimum number of instances for further splits and the total number of trees. The mean and variance for each new configuration are estimated using the RF model.

SMAC stands out for its support of various types of variables, including continuous, discrete, categorical, and conditional hyperparameters. The time complexities for fitting and predicting variances with SMAC are efficient, being $O(n \log n)$ and $O(\log n)$, respectively, thereby enhancing the overall efficiency of the BO process.

2.5.3. Hyperband

Developed to efficiently allocate computational resources for hyperparameter tuning, Hyperband is an advanced optimization algorithm [95]. It uses the Successive Halving algorithm to dynamically balance the number of hyperparameter configurations (n) and the budgets allotted to them. The primary concept involves partitioning the entire budget (B) into n components and allocating each component ($b = \frac{B}{n}$) to various configurations, gradually removing the less efficient ones.

Given the budget constraints $b_{\max}$ and $b_{\min}$, the algorithm operates in the following manner:

1. Set $s_{\max} = \log(\frac{b_{\max}}{b_{\min}})$.
2. Iterate over s from $b_{\max}$ to $b_{\min}$:
   (a) Determine the number of configurations, $n = \text{DetermineBudget}(s)$.
   (b) Sample n configurations, $c = \text{SampleConfigurations}(n)$.
   (c) Apply Successive Halving to c.
3. Return the best configuration found so far.

The total number of data points, the minimum number of instances needed to train a model, and the available budgets are used to determine the initial values of the budget constraints, $b_{\min}$ and $b_{\max}$. Next, as indicated in steps 2–3, the algorithm determines n and the budget size for every configuration. Samples of the configurations are run through the successive halving process, which iteratively promotes the more successful configurations while systematically removing the under-performing ones.

By incorporating the Successive Halving method, the Hyperband algorithm's computational complexity is reduced to $O(n \log n)$, which improves its ability to find the ideal hyperparameter configuration.

Table 2 provides a detailed comparison of Hyperband, BO–GP, and SMAC, as well as an explanation of their distinct qualities and HPO efficiency levels.

Table 2. Comparison of Hyperband, BO–GP, and SMAC in HPO.

HPO Method	Strengths	Limitations	Time Complexity
Hyperband	Enables parallelization.	Ineffective with conditional HPs. Makes it necessary for subsets with limited resources to be representative.	$O(n \log n)$
BO–GP	Fast convergence speed for continuous HPs.	Poor capacity for parallelization. Ineffective when using conditional HPs.	$O(n^3)$
SMAC	Efficient with all types of HPs.	Poor capacity for parallelization.	$O(n \log n)$

2.6. Ensemble Generation Through Hyperparameter Optimization

The notion of producing ensembles through hyperparameter optimization has attracted significant interest. Researchers in [96] illustrated this by using a multi-stage boosting-like technique for hyperparameter optimization to produce better image representations. The Sequential Model-Based Ensemble Optimization (SMBEO) was first presented in study [97]. It simulates multiple independent optimization processes using bootstrapped validation datasets, which are then integrated using the agnostic Bayesian combination method.

Many trained models are usually obtained from hyperparameter optimization, and one model is frequently selected based on its generalization error, $\gamma^* = \arg\min_\gamma L(h_\gamma | X_V)$. However, this method, which is similar to a point estimate, runs the risk of overfitting. Choosing multiple models can be an effective countermeasure to reduce overfitting and improve generalization performance.

As proposed by [98], a simple approach to creating an ensemble is to hold onto the models created during the optimization procedure. This 'post hoc' ensemble generation involves building a model pool for potential future fusion. Pruning this pool is accomplished well by the forward greedy selection, as described by [99]. A model is added to the ensemble at each iteration in order to reduce the empirical error on the validation dataset:

$$h_t = \arg\min_{h \in H} L(E \cup \{h\} | X_V) \tag{12}$$

$$L(E \cup \{h\} | X_V) = \sum_{i=0}^{|X_V|} l_{0-1}(g(x_i, E \cup \{h\}), y_i), \tag{13}$$

where the combination function for the model in the ensemble E on sample x_i is $g(x_i, E)$. Majority voting is the preferred combination rule because it has a lower tendency to overfit. Agnostic Bayesian combination, stacking, and weighted voting are additional combination techniques.

Empirically, it has been demonstrated that this ensemble approach performs better than the best single model from hyperparameter optimization, mainly because of the combination's ability to reduce variance.

3. Clustering Performance Metrics

Various clustering performance measures were employed in our study to properly comprehend and implement the clustering results [100]. When evaluating each clustering

model's performance on our dataset, these metrics were quite important. As an illustration of the suitability of cluster assignments, the Silhouette score allowed us to assess how compact and well-separated the clusters were. The separation and cohesiveness of the clusters were measured using the Calinski–Harabasz Index, which indicated the overall quality of clustering. With lower values indicating better grouping, the Davies–Bouldin Index allowed us to assess the dispersion both within and between clusters. Finally, the degree of agreement between the clustering results and any pre-defined categories was understood with the use of Normalized Mutual Information (NMI). By utilizing these criteria, we made sure that our clustering strategy was thoroughly and impartially analyzed, which resulted in more accurate and trustworthy conclusions. These metrics are explained in detail below.

3.1. Silhouette Score

The effectiveness of the clustering in terms of density and separation is gauged by the Silhouette score. It is described as:

$$S = \frac{b - a}{\max(a, b)}, \tag{14}$$

where a is the mean distance to the other elements in the same cluster, and b is the mean nearest-cluster distance. Higher values, closer to 1, indicate well-separated and densely packed clusters.

3.2. Calinski–Harabasz Index

The variance ratio criterion, or Calinski–Harabasz Index, is a tool for assessing the quality of clustering. It has the following definition:

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

where the number of clusters is k, the number of data points is N, and the between-group and within-group dispersion matrices are B_k and W_k, respectively.

3.3. Davies–Bouldin Index

Lower values of the Davies–Bouldin Index, a measure of clustering algorithm quality, correspond to better clustering. It has the following definition:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right), \tag{16}$$

where the centroid of cluster i is denoted by c_i, the average distance to it is σ_i, and the distance between the centroids of cluster (c_i, c_j) is $d(c_i, c_j)$.

3.4. Normalized Mutual Information (NMI)

The Mutual Information (MI) score is normalized to scale the results between 0 (no mutual information) and 1 (perfect correlation), which is known as Normalized Mutual Information. It has the following definition:

$$NMI(U, V) = \frac{2 \cdot MI(U, V)}{H(U) + H(V)}, \tag{17}$$

where H is the entropy, U and V are two sets of clusters, and MI is the mutual information between U and V.

3.5. Proposed Method

In this study, our proposed methodology is centered around the exploration of molecular interactions between PCOS and EC. The core objective is to harness advanced ensemble clustering techniques with novel hyperparameter optimization to unveil the intricate molecular networks that may link these two conditions. PCOS, a complex endocrine disorder, has been increasingly recognized for its broader implications, notably its potential connection to EC. The molecular cross-talk between these conditions, however, remains poorly understood. Our methodological approach is designed to dissect these complex interactions, offering new insights into the shared molecular pathways and potential points of convergence between PCOS and EC.

Using our proposed method requires consideration of the ensemble performance at each optimization iteration. A thorough approach is necessary to ensure that the optimization of each individual model has a positive impact on the overall effectiveness of the ensemble. Such a strategy not only increases the overall predictive power but also fosters robustness and generalization capabilities in a variety of application scenarios.

In the existing literature, studies have predominantly focused on applying a single hyperparameter optimization algorithm across all models within an ensemble. However, it has been observed that not all hyperparameter optimization algorithms yield the best results for every model; different models can often benefit from different optimization strategies [101,102]. Our study introduces innovations on two fronts. First, it explores new ground by applying hyperparameter optimization to the domain of ensemble generation for clustering problems, an area in which this technique has not been applied before. Second, we adopt a customized strategy in which, during the generation stage, each model is optimized using the hyperparameter optimization method best suited to its unique features. This methodology aims to maximize the overall performance by selectively applying the most effective optimization strategy for each individual model in the ensemble, thereby potentially enhancing the accuracy and robustness of the resulting clustered outcomes.

Selecting the elements that best fit the unique requirements of clustering tasks is a crucial step in adapting the concurrent ensemble construction and hyperparameter optimization processes for clustering. Given its ability to evaluate clustering quality quantitatively, the choice of NMI as the loss function in this case is strategic. The NMI value is a reliable metric for assessing clustering performance because a higher value denotes a higher degree of similarity between the true and predicted clustering. NMI performs exceptionally well in assessing the degree of agreement between the true labels and the clustering results, producing a normalized score that takes into account the mutual information between the two datasets. For this reason, NMI is particularly suitable when the cluster assignments need to be precise and the underlying structure of the data is complex.

The decision to use majority voting for the integration function was also carefully considered. By allocating each data point to the cluster that the majority of models concur with, this technique aggregates the clustering decisions made by individual models within the ensemble. In mathematical terms, the final cluster label $L(x)$ for a set of models $\{M_1, M_2, \ldots, M_n\}$ and a data point x can be found as follows:

$$L(x) = \arg\max_{c} \sum_{i=1}^{n} \mathbb{1}\{M_i(x) = c\}, \tag{18}$$

where the indicator function is $\mathbb{1}\{\}$, and the cluster label assigned to x by model M_i is $M_i(x)$. By reducing the impact of individual model biases and errors, this method may produce clustering results that are more accurate and stable. By lowering the noise and variances of individual models, majority voting produces clustering results that are more reliable and stable. By utilizing the combined knowledge of several models, this method frequently produces better performance than any one model working alone in the ensemble.

Our clustering ensemble uses a variety of algorithms, including Gaussian Mixture, DBSCAN, K-Means, and Hierarchical Clustering. To maximize each method's perfor-

mance both individually and collectively within the ensemble, particular hyperparameter optimization techniques are applied.

K-Means and Hierarchical Clustering Optimization: Hyperband, an advanced hyperparameter optimization technique, is what our ensemble has decided to use to optimize the K-Means and Hierarchical Clustering algorithms. Hyperband's ability to effectively explore the parameter space of these clustering techniques is what motivated this decision. Fast and efficient, Hyperband works especially well for adjusting important parameters like the number of clusters in K-Means and the linkage criteria in Hierarchical Clustering. It is the perfect option for these algorithms because of its speedy evaluation and iteration over a broad range of parameter configurations, which guarantees a more efficient and effective optimization process.

DBSCAN Optimization: SMAC is used to optimize the DBSCAN algorithm, which is well-known for its density-based clustering methodology. SMAC is the perfect optimizer for DBSCAN because of its versatility and ability to handle complex parameter spaces, especially when it comes to adjusting the parameters for epsilon and minimum points.

Gaussian Mixture Model Optimization: BO–GP is applied to Gaussian Mixture Models. Gaussian Mixture Models are well-suited to BO–GP's efficiency in handling continuous and probabilistic parameter spaces, as they optimize parameters such as component count and covariance type.

A round-robin technique is used to manage the ensemble and update its members. Every optimization iteration aims to improve a particular model within the ensemble. A preset order is followed when selecting the model to optimize in a particular iteration, guaranteeing that every model is updated on a regular basis.

The ensemble is updated in a greedy manner at the conclusion of every optimization iteration. The better-performing model is kept after the recently optimized model is compared to its predecessor in the ensemble. With this strategy, the ensemble is guaranteed to continuously evolve by incorporating the best iterations of each clustering algorithm.

Since the selected datasets are genetic sequences, such biological data should be preprocessed in order to be used as input to the clustering algorithms. We have applied the CKSAAP encoding scheme to transform the sequences into frequency encoding more suitable for machine learning operations.

This algorithm, called "Ensemble Clustering Optimization" (Algorithm 1), describes a systematic way to build an optimized clustering ensemble by using different clustering algorithms and optimization techniques accordingly. The algorithm is made to evaluate and incorporate optimized models methodically, thereby iteratively improving the ensemble. Following is a thorough explanation of each step:

1. **Initialization:** Three sets are initialized at the initial phase of the algorithm—H for the model history, G for the hyperparameter collection, and E for the ensemble itself. These sets are initially empty.
2. **Iterative Process:** An iterative process that completes N iterations constitutes the algorithm's core. By going through this process, the ensemble can be gradually improved.
3. **Model Replacement and Updating:** Every iteration, represented by i, finds a replacement model in the ensemble using the algorithm. The variable j, which is computed as $i \mod m$, indicates this, with m denoting the ensemble size. The jth model is eliminated if it is present in the ensemble E.
4. **Hyperparameter Optimization for Each Algorithm:** Every clustering algorithm in the set A carries out a specific optimization process, including GMM, DBSCAN, K-Means, and Hierarchical Clustering. This entails implementing KMeansHierarchicalOptimize for both K-Means and Hierarchical Clustering, applying DBSCANOptimize for the DBSCAN algorithm, and using GMMOptimize for Gaussian Mixture Models. Using the training data XT, validation data XV, and the hyperparameter space Γ, these optimization steps are specifically designed to determine the optimal hyperparameters γ_i for each model.

5. **Model Training and History Tracking:** For each clustering algorithm, a new model h_i is trained using the optimized hyperparameters γ_i. The hyperparameters of this model are then saved in G, and it is subsequently added to the history set H.
6. **Selecting and Updating the Best Model:** The algorithm then uses the loss function L, which is intended to assess clustering performance, to choose the model that performs the best out of H. This best-performing model h_j is added to the ensemble E, thereby replacing the model that was previously removed.
7. **Final Ensemble:** After the iterations are finished, the ensemble E is made up of a number of optimized clustering models, each of which adds to an overall improvement in the clustering performance.

Algorithm 1 Ensemble Clustering Optimization

Require: XT—Training data, XV—Validation data, N—Number of iterations, s—Ensemble size, A—Set of clustering algorithms {GMM, DBSCAN, K-Means, Hierarchical}, Γ—Hyperparameter space for all algorithms, L—Loss function for clustering performance
Ensure: H—History of models, E—The final ensemble of optimized clustering models

```
Biological data preprocessing
Initialize H, G, E to empty sets
for i = 1 to N do
    j ← i mod s
    if E contains jth model then
        Remove the jth model from E
    end if
    for each clustering algorithm α in A do
        if α is GMM then
            γ_i = GMMOptimize(XT, XV, Γ)
        else if α is DBSCAN then
            γ_i = DBSCANOptimize(XT, XV, Γ)
        else if α is K-Means or Hierarchical then
            γ_i = KMeansHierarchicalOptimize(XT, XV, Γ)
        end if
        h_i = TrainModel(α, γ_i)
        Add γ_i to G and h_i to H
    end for
    h_j = SelectBestModel(H, L)
    Update E by adding h_j
end for
```

The basic flowchart in Figure 1 illustrates an algorithmic procedure for optimizing clustering models in an ensemble framework. Various hyperparameter tuning strategies, such as SMAC and BO–GP, are applied to particular algorithms, such as GMM, DBSCAN, K-Means, and Hierarchical Clustering; these are then iterated through a round-robin method based on a history of performance metrics, and the final ensemble set is chosen when a predetermined condition is satisfied.

One notable feature of this method is its systematic approach to improving ensemble clustering. It guarantees that the final ensemble is not only diversified but also optimized and integrated across multiple models iteratively for optimal performance. The ensemble is strong and efficient in managing challenging clustering tasks because each clustering algorithm's unique optimization strategies are used to further guarantee that each model performs at its peak.

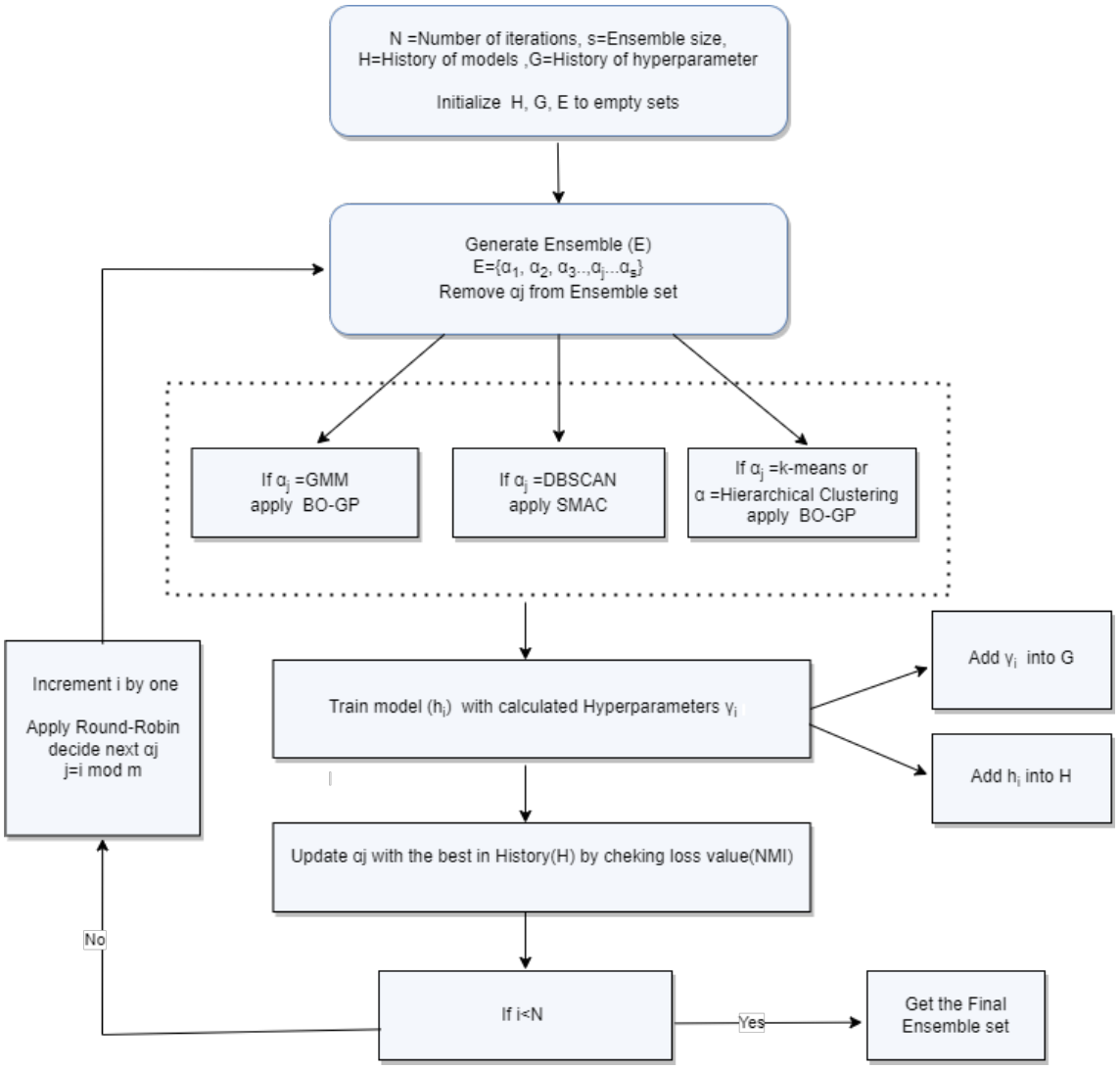

Figure 1. Flowchart of ensemble clustering optimization utilizing algorithm-specific hyperparameter tuning and iterative model enhancement.

4. Results

The outcomes of our investigation, which used PCOS and EC datasets, highlight how well hyperparameter optimization and ensemble learning can be combined.

Using the CKSAAP technique, feature encoding was performed as part of the preprocessing steps before implementing PCA on the PCOS and EC datasets. In order to convert the protein sequences into a numerical format that machine learning algorithms could use, this encoding was essential. Following the standardization of these encoded features, PCA was used to lower the data's dimensionality. An illustrated representation of the explained variance ratio served as a guide for determining the number of principal components (k) to keep. Each principal component's individual and cumulative explained variance is depicted in the plot in Figure 2. Fourteen principal components were selected as a result of this analysis since they accounted for 90% of the variance in the dataset. This decision ensures that the most pertinent features are included while preserving computational efficiency by striking a balance between keeping a sizable amount of information and simplifying the dataset. The study now turns its attention to the next phase, which combines optimization techniques with ensemble clustering selection, after determining the ideal number of principal components. This following stage is essential to improving our strategy even more and raising the methodology's overall efficacy.

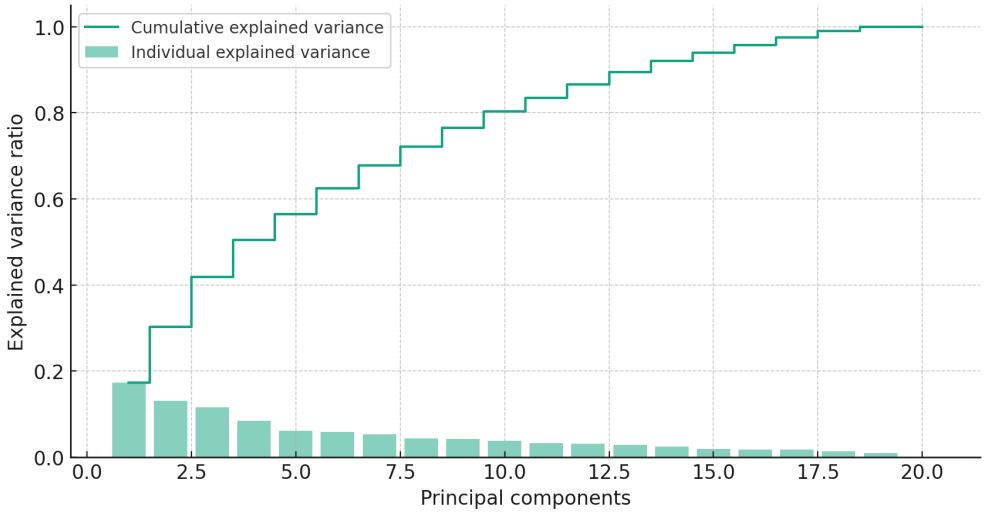

Figure 2. Explained variance by each principal component and the cumulative explained variance.

The Gaussian Mixture Model and K-Means, which were optimized with Hyperband and showed notable increases in cluster definition and accuracy, were two of the models where we saw the most gains in clustering performance. SMAC-optimized DBSCAN produced inconsistent results, whereas Hyperband-optimized Hierarchical Clustering performed better in creating distinct cluster hierarchies. Our ensemble's use of majority voting as the integration function significantly improved the accuracy and stability of the clustering findings by efficiently combining the judgments made by each individual model and reducing biases. Overall, the performance of individual algorithms as well as the ensemble's overall predictive capacity were improved by our customized method for ensemble creation and hyperparameter optimization, which has tremendous potential to further genomic research in PCOS and EC.

The dataset was subjected to an exhaustive 5-fold cross-validation procedure, and for each strategy, we set the number of optimization iterations to 100, guaranteeing a thorough assessment. The hyperparameter optimization techniques applied to various clustering algorithms are shown in Table 3. Every algorithm in our group used a different optimization technique: Hyperband for K-Means and Hierarchical Clustering, SMAC for DBSCAN, and Bayesian Optimization with Gaussian Processes (BO–GP) for GMM. These techniques were specially selected to match the properties and optimization requirements of every algorithm, leading to increased efficiency and accuracy. The algorithms' performance was greatly enhanced by the application of these various optimization strategies, proving the benefit of customized hyperparameter tuning for clustering tasks. For every clustering technique, we produced a different number of models: the ensemble sizes varied from 10, 20, 30, 40, and 50 models for every technique.

Table 3. Machine learning algorithms and their hyperparameter optimization.

ML Algorithm	Main HPs	Optional HPs	HPO Methods	Libraries
K-Means	n_clusters	init, n_init, max_iter	Hyperband	Hyperopt, Hyperband
Hierarchical Clustering	n_clusters, distance_threshold	linkage	Hyperband	Hyperopt, Hyperband
DBSCAN	eps, min_samples	-	SMAC	SMAC
Gaussian Mixture Model (GMM)	n_components	covariance_type, max_iter, tol	BO–GP	Skpot

The performance metrics of different clustering algorithms after the hyperparameter optimization procedures are shown in Table 4. In particular, the Silhouette score and

Calinski–Harabasz Index, which indicate improved cluster separation and cohesion, significantly improve when the K-Means algorithm is optimized using Hyperband. Furthermore, Table 4 with default hyperparameter performance metrics offers a baseline comparison that emphasizes the effectiveness of the optimization procedure. Comparably, the Gaussian Mixture Model (GMM) exhibits a significant rise in Silhouette and NMI scores, indicating enhanced clustering accuracy after being adjusted using Bayesian Optimization using Gaussian Processes (BO–GP). We also observe notable improvements in the metrics for Hierarchical Clustering and DBSCAN, which were optimized through Hyperband and SMAC, respectively. DBSCAN shows a notable increase in the NMI score, indicating improved alignment with the true class labels. Taken as a whole, these findings highlight how well-customized hyperparameter optimization methods work with various clustering algorithms individually.

Table 4. Comparison of default and improved metrics for clustering algorithms.

Clustering Algorithm	Metric	Default HPs	Optimized	Improvement
K-Means	Silhouette	0.15	0.21	+0.06
	Calinski–Harabasz	50.00	69.64	+19.64
	Davies–Bouldin	2.50	2.03	−0.47
	NMI	0.25	0.31	+0.06
GMM	Silhouette	0.13	0.21	+0.08
	Calinski–Harabasz	45.00	52.83	+7.83
	Davies–Bouldin	2.20	1.72	−0.48
	NMI	0.28	0.34	+0.06
Hierarchical Clustering	Silhouette	0.10	0.18	+0.08
	Calinski–Harabasz	40.00	53.92	+13.92
	Davies–Bouldin	2.30	1.88	−0.42
	NMI	0.20	0.32	+0.12
DBSCAN	Silhouette	0.08	0.12	+0.04
	Calinski–Harabasz	30.00	48.29	+18.29
	Davies–Bouldin	1.10	0.87	−0.23
	NMI	0.18	0.31	+0.13

A thorough examination of ensemble performance metrics for various optimization techniques is shown in Table 5, which also illustrates how effective each technique is for a range of ensemble sizes. The table presents the results in terms of Silhouette, Calinski–Harabasz, Davies–Bouldin, and NMI scores for the following ensemble size categories: 4×10, 4×20, 4×30, 4×40, and 4×50. The ensemble sizes were chosen based on a trade-off between attained performance and efficiency of computation. The process involves creating an ensemble using m models from each of the n different clustering algorithms in the ensemble configuration denoted as "nxm" in Table 5. With this method, the ensemble is a well-designed aggregation rather than just a collection of models, with m distinct models contributed by each of the n-selected clustering techniques. For instance, Table 5's first row presents the outcomes of an ensemble clustering approach. This approach generated an ensemble of 40 models, with 10 models from each of the four clustering algorithms. When the Bayesian Optimization with Gaussian Processes (BO–GP) method is the only one used for hyperparameter optimization during the ensemble generation step, it displays the performance metrics. NMI, Davies–Bouldin Index, Calinski–Harabasz Index, and Silhouette score are among the metrics. The methods include BO–GP, SMAC, Hyperband, and the proposed method. The rows for BO–GP, SMAC, and Hyperband in Table 5 show the results of applying these hyperparameter optimization techniques consistently to all clustering algorithms in the ensemble generation stage.

Table 5. Ensemble performance metrics for different optimization methods.

Method/Ensemble Size	Silhouette	Calinski–Harabasz	Davies–Bouldin	NMI
Ensemble of 4 × 10 Models				
Post Hoc Ensemble Result	0.14	44	1.9	0.56
BO–GP	0.16	47	1.7	0.58
SMAC	0.15	46	1.8	0.57
Hyperband	0.17	50	1.6	0.59
Proposed Method	0.19	54	1.4	0.71
Ensemble of 4 × 20 Models				
Post Hoc Ensemble Result	0.15	45	1.8	0.57
BO–GP	0.17	48	1.6	0.59
SMAC	0.16	47	1.7	0.58
Hyperband	0.18	51	1.5	0.60
Proposed Method	0.20	55	1.3	0.72
Ensemble of 4 × 30 Models				
Post Hoc Ensemble Result	0.16	46	1.7	0.58
BO–GP	0.18	49	1.5	0.60
SMAC	0.17	48	1.6	0.59
Hyperband	0.19	52	1.4	0.61
Proposed Method	0.21	56	1.2	0.73
Ensemble of 4 × 40 Models				
Post Hoc Ensemble Result	0.17	47	1.6	0.59
BO–GP	0.19	50	1.4	0.61
SMAC	0.18	49	1.5	0.60
Hyperband	0.20	53	1.3	0.62
Proposed Method	0.22	57	1.1	0.74
Ensemble of 4 × 50 Models				
Post Hoc Ensemble Result	0.18	48	1.5	0.60
BO–GP	0.20	51	1.3	0.62
SMAC	0.19	50	1.4	0.61
Hyperband	0.21	54	1.2	0.63
Proposed Method	0.23	58	1.0	0.75

A wide range of metrics, including Davies–Bouldin, NMI, Silhouette, and Calinski–Harabasz, are used in Table 5 to assess the performance of different ensemble sizes that have been optimized using different techniques, including BO–GP, SMAC, Hyperband, and our proposed method. All of these metrics together provide information about how well the algorithms replicate the real data structure, how cohesively they cluster, how well they separate, and more. As the ensemble size grows, all methods show a consistent improvement in the Silhouette score, which measures the cohesion and separation of the clusters. Significantly, our proposed method performs better than other methods in larger ensembles (4×40 and 4×50 models), indicating that it can produce clusters that are more distinct and well-separated.

A similar trend is noted in terms of the Calinski–Harabasz Index, which assesses the cluster validity. When compared to other methods, the proposed method consistently shows higher values, especially in larger ensembles, suggesting that it forms clusters with better-defined structures and higher separation between them. There is also a pattern of improvement in the Davies–Bouldin Index, where lower values indicate better clustering quality. The proposed method's superior ability to form compact and well-separated clusters is demonstrated by its lowest Davies–Bouldin scores in larger ensembles.

Most importantly, our proposed method exhibits a significant upward trend in the NMI (Normalized Mutual Information) metric, which evaluates the clustering quality relative to the true class labels. Particularly in the 4×50 model ensemble, it attains the highest NMI scores of all ensemble sizes, demonstrating its superior performance in faithfully capturing the underlying data structure. It is clear from the table that our suggested approach produces higher performance metrics on average than the widely used post hoc ensemble method. Lower Silhouette, Calinski–Harabasz, and NMI scores as

well as higher Davies–Bouldin scores for the post hoc ensemble results are indicative of this. The superior effectiveness of our proposed approach is highlighted by the post hoc ensemble method's comparative under-performance, highlighting its potential advantages in ensemble modeling contexts.

Reflecting on the biological implications of the obtained results, we found a correlation between multiple gene pairs from the utilized datasets. This correlation provides important information about how metabolic control and genomic stability interact. For instance, insulin resistance is a typical characteristic of PCOS, and it is also known that the function of CAPN10 in glucose metabolism and insulin signaling is very important. Insulin resistance-related elevated insulin levels can have a significant impact on ovarian function and cause disturbances in the synthesis of steroid hormones. The pathophysiology of PCOS is mostly attributed to these hormonal abnormalities, which have an impact on a variety of body functions, including metabolism and fertility.

This is further complicated by MUTYH's involvement in DNA repair. Its main purpose of repairing oxidative DNA damage raises the possibility of a defense mechanism against genomic instability, which is frequently linked to the emergence of cancer. The interplay of these two genes suggests the following: metabolic abnormalities associated with PCOS, mediated by CAPN10, may intensify genomic instability via the MUTYH pathway. This association may play a part in the oncogenic processes seen in EC, where genomic instability is a major issue. This link emphasizes the necessity of investigating focused therapies that address DNA repair mechanisms in addition to metabolic dysregulation as a combined therapy approach.

Another noteworthy association found between CYP17 and MSH2, MLH1, and BRIP1 offer important new understandings of the interplay between hormonal control and DNA repair processes. Important for steroidogenesis, CYP17 controls key hormonal pathways by affecting levels of estrogen and androgen. The disruption of these pathways in PCOS results in diseases like hyperandrogenism, which exacerbates the disorder's hallmark symptoms like irregular menstrual cycles and infertility. These hormone abnormalities may interfere with the regular endometrial cycle, raising the possibility of pathological illnesses like EC.

Particularly instructive are the associations found between CYP17 and DNA repair genes, including MSH2, MLH1, and BRIP1. In order to stop mutations that can cause cancer, MSH2 plays a crucial function in preserving DNA integrity throughout cell division. Similarly, mutations in the genes MLH1 and BRIP1 have been connected to an increased risk of cancer. These genes are essential parts of the machinery that repairs DNA. The relationship that CYP17 has with these genes points to a possible convergence of the pathways involved in DNA repair and steroid production. This convergence may be crucial to comprehend the complex relationship between CYP17-mediated hormonal dysregulation in PCOS and DNA repair integrity, which in turn affects EC formation. This realization is crucial because it provides fresh perspectives on how hormone pathway modulation may affect DNA repair mechanisms and vice versa, potentially providing treatment targets for EC and PCOS. Our study's findings paint a complete picture of the molecular interaction between EC and PCOS. We learn more about the biological mechanisms causing these circumstances by examining these interactions by using our proposed method. Our results point to a complex network that may be involved in the transition from PCOS to EC. This network crosses metabolic pathways, hormone regulation, and genomic stability.

In summary, all approaches perform better as the size of the ensemble increases, but our proposed method performs better across the board, especially in larger ensembles, according to all measured metrics. This indicates that it can handle challenging clustering tasks with resilience and flexibility, which makes it a good option for producing high-quality clustering ensembles.

5. Discussion

Our study has shown the efficacy of ensemble learning methodology and hyperparameter optimization applied to a dataset containing genes associated with EC and

PCOS. Silhouette and Calinski–Harabasz scores were significantly improved by employing Hyperband and BO–GP to optimize algorithms like K-Means and GMM. This suggests that the algorithms created more recognizable and distinct clusters. More specifically, the higher NMI score indicates that the algorithms generated clustering outcomes that were more in line with the true class labels. These results have great promise for deciphering the intricate relationships between diseases and for better comprehending the intricate structures of genetic datasets. In particular, this method might help comprehend molecular pathways and genetic variations. For example, applying this method could enable a deeper examination of the molecular causes of diseases in cancer genetics research or the study of hereditary diseases.

After the analysis of the internal mechanisms within the proposed ensemble clustering algorithms, we identified several genetic co-occurrences in the results with significant molecular interaction implications. The interplay of these gene pairs highlights a complex network where metabolic, hormonal, and genomic stability pathways intersect, particularly relevant in understanding the progression from PCOS to EC:

- ***CAPN10* and *MUTYH***
  - *CAPN10* affects ovarian function and the generation of steroid hormones in PCOS via being involved in insulin signaling and glucose metabolism.
  - *MUTYH* is essential for DNA repair, indicating a connection between genomic stability and metabolic dysregulation in PCOS.
  - **Interaction Implication:** By means of *CAPN10*, metabolic disturbances in PCOS may intensify genomic instability through *MUTYH*, potentially playing a role in carcinogenic processes in EC.
- ***CYP17* and *MSH2***
  - Involved in steroidogenesis, *CYP17* affects both estrogen and testosterone levels and is linked to hormonal imbalances in PCOS.
  - Defects raise the risk of EC, while DNA fidelity is maintained by *MSH2*.
  - **Interaction Implication:** A possible explanation for the increased risk of EC in PCOS patients with *CYP17–MSH2* correlation is a potential cross-talk between hormonal imbalance and DNA repair mechanisms.
- ***CYP17*, *CYP21*, and *RAD51C***
  - *RAD51C* is essential for DNA repair, particularly for repairing double-strand breaks.
  - PCOS pathology is impacted by the involvement of *CYP17* and *CYP21* in the biosynthesis of steroid hormones.
  - **Interaction Implication:** There is a connection between the dysregulation of steroid hormones and DNA repair processes, indicating that hormonal imbalances in PCOS may affect DNA repair pathways and accelerate the development of EC.
- ***CYP17* and *MLH1*, *CYP17*, and *BRIP1***
  - *BRIP1* and *MLH1* are linked to increased cancer risk and are involved in DNA repair.
  - DNA repair and steroid biosynthesis pathways are converging, as evidenced by their interaction with *CYP17*.
  - **Interaction Implication:** Understanding EC development in the context of PCOS necessitates an understanding of the potential connection between DNA repair integrity and hormonal irregularities in PCOS, as demonstrated by *CYP17*.

The results of this study imply that comparable techniques may find wider use in the analysis of genetic and biological data. This approach could be used in future studies to more precisely define the connections between various disease types and genetic traits. Furthermore, the methodology's potential benefits could be investigated in applications related to personalized medicine, particularly in the development of customized treatment plans based on genetic data.

6. Conclusions

Our methodology consists of building an ensemble of a given size and methodically optimizing the configuration of a particular clustering algorithm within this ensemble at every phase of the hyperparameter optimization procedure. We meticulously assess each algorithm's potential performance throughout this process, taking into account its interactions with the other models in the ensemble. By fine-tuning each element in relation to the others, this iterative process of adjustment seeks to improve the ensemble's overall efficacy. This approach enabled the ensemble, comprising a combination of different models, to achieve a level of performance and accuracy higher than what each model could have provided individually. This innovation is particularly important in the context of genetic data analysis, considering the high dimensionality and complexity inherent in such datasets.

In conclusion, the interplay of these gene pairs highlights a complex network where metabolic, hormonal, and genomic stability pathways intersect. This network might be particularly relevant in understanding the progression from PCOS, a condition characterized by metabolic and hormonal imbalances, to EC, where genomic instability plays a crucial role. These insights could be vital for developing targeted interventions and personalized management strategies for individuals with PCOS who are at a heightened risk of developing EC.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available upon reasonable request. Interested researchers may contact the corresponding author to gain access to the data.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PCOS	Polycystic ovary Syndrome
EC	Endometrial Carcinoma
CVDs	Cardiovascular Diseases
HF	Heart Failure
AF	Atrial Fibrillation
AI	Artificial Intelligence
ML	Machine Learning
MS	Metastatic Score
DFS	Disease-Free Survival
CSS	Cancer-Specific Survival
RSF	Random Forest Tree
FCGRs	Frequency Chaos Game Representations
SMAC	Sequential Model-Based Algorithm Configuration
NCBI	National Center for Biotechnology Information
CKSAAP	Composition of k-Spaced Amino Acid Pairs
PCA	Principal Component Analysis
SVD	Singular Value Decomposition
GMM	Gaussian Mixture Model
DBSCAN	Density-Based Spatial Clustering Method
HPO	Hyperparameter Optimization
BO	Bayesian Optimization
NMI	Normalized Mutual Information
BO–GP	Bayesian Optimization with Gaussian Processes

References

1. Okamura, Y.; Saito, F.; Takaishi, K.; Motohara, T.; Honda, R.; Ohba, T.; Katabuchi, H. Polycystic ovary syndrome: Early diagnosis and intervention are necessary for fertility preservation in young women with endometrial cancer under 35 years of age. *Reprod. Med. Biol.* **2017**, *16*, 67–71. [CrossRef] [PubMed]
2. Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertil. Steril.* **2004**, *81*, 19–25. [CrossRef] [PubMed]
3. Markowska, A.; Chudecka-Głaz, A.; Pityński, K.; Baranowski, W.; Markowska, J.; Sawicki, W. Endometrial Cancer Management in Young Women. *Cancers* **2022**, *14*, 1922. [CrossRef] [PubMed]
4. Zhang, Y.; Hu, Y.; Yu, J.; Xie, X.; Jiang, F.; Wu, C. Landscape of PCOS co-expression gene and its role in predicting prognosis and assisting immunotherapy in endometrial cancer. *J. Ovarian Res.* **2023**, *16*, 129. [CrossRef] [PubMed]
5. Prakash, A.; Nourianpour, M.; Senok, A.; Atiomo, W. Polycystic Ovary Syndrome and Endometrial Cancer: A Scoping Review of the Literature on Gut Microbiota. *Cells* **2022**, *11*, 3038. [CrossRef] [PubMed]
6. Johnson, J.E.; Daley, D.; Tarta, C.; Stanciu, P.I. Risk of endometrial cancer in patients with polycystic ovarian syndrome: A meta-analysis. *Oncol. Lett.* **2023**, *25*, 1–9. [CrossRef]
7. Tanos, P.; Dimitriou, S.; Gullo, G.; Tanos, V. Biomolecular and genetic prognostic factors that can facilitate fertility-sparing treatment (FST) decision making in early stage endometrial cancer (ES-EC): A systematic review. *Int. J. Mol. Sci.* **2022**, *23*, 2653. [CrossRef]
8. Kumari, J.; Kumar, E.; Kumar, D. A Structured Analysis to Study the Role of Machine Learning and Deep Learning in the Healthcare Sector with Big Data Analytics. *Arch. Comput. Methods Eng.* **2023**, *30*, 1–29. [CrossRef]
9. Neijzen, D.; Lunter, G. Unsupervised learning for medical data: A review of probabilistic factorization methods. *Stat. Med.* **2023**, *42*, 5541–5554. [CrossRef]
10. Topuz, B.; Alp, N.Ç. Machine learning in architecture. *Autom. Constr.* **2023**, *154*, 105012. [CrossRef]
11. Ooi, K.B.; Tan, G.W.H.; Al-Emran, M.; Al-Sharafi, M.A.; Capatina, A.; Chakraborty, A.; Dwivedi, Y.K.; Huang, T.L.; Kar, A.K.; Lee, V.H.; et al. The potential of Generative Artificial Intelligence across disciplines: Perspectives and future directions. *J. Comput. Inf. Syst.* **2023**, *2023*, 1–32. [CrossRef]
12. Singh, A.V.; Varma, M.; Laux, P.; Choudhary, S.; Datusalia, A.K.; Gupta, N.; Luch, A.; Gandhi, A.; Kulkarni, P.; Nath, B. Artificial intelligence and machine learning disciplines with the potential to improve the nanotoxicology and nanomedicine fields: A comprehensive review. *Arch. Toxicol.* **2023**, *97*, 963–979. [CrossRef] [PubMed]
13. Sharifani, K.; Amini, M. Machine Learning and Deep Learning: A Review of Methods and Applications. *World Inf. Technol. Eng. J.* **2023**, *10*, 3897–3904.
14. Mazalan, M.; Do, T.D.; Zaman, W.S.W.K.; Ramlan, E.I. Machine Learning Approaches for Stem Cells. *Curr. Stem Cell Rep.* **2023**, *2023*, 1–14. [CrossRef]
15. Arjmand, B.; Hamidpour, S.K.; Tayanloo-Beik, A.; Goodarzi, P.; Aghayan, H.R.; Adibi, H.; Larijani, B. Machine learning: A new prospect in multi-omics data analysis of cancer. *Front. Genet.* **2022**, *13*, 824451. [CrossRef]
16. Liu, G.; Stokes, J.M. A brief guide to machine learning for antibiotic discovery. *Curr. Opin. Microbiol.* **2022**, *69*, 102190. [CrossRef]
17. Bhardwaj, V.; Sharma, A.; Parambath, S.V.; Gul, I.; Zhang, X.; Lobie, P.E.; Qin, P.; Pandey, V. Machine learning for endometrial cancer prediction and prognostication. *Front. Oncol.* **2022**, *12*, 852746. [CrossRef] [PubMed]
18. Naqvi, N.Z.; Kaur, K.; Khanna, S.; Singh, S. An Overview of Machine Learning Techniques Focusing on the Diagnosis of Endometriosis. In *Machine Vision and Augmented Intelligence: Select Proceedings of MAI 2022*; Springer: Berlin, Germany, 2023; pp. 61–84.
19. Vaswania, J.; Mulchandanib, H.; Vaghelac, R.; Pateld, R. A Systematic literature review on diagnosis of PCOS using machine learning algorithms. *GIT J. Eng. Technol.* **2022**, *14*, 5.
20. Ahmed, S.; Rahman, M.S.; Jahan, I.; Kaiser, M.S.; Hosen, A.S.; Ghirime, D.; Kim, S.H. A Review on the Detection Techniques of Polycystic Ovary Syndrome Using Machine Learning. *IEEE Access* **2023**, *11*, 86522–86543. [CrossRef]
21. Venkat, V.; Abdelhalim, H.; DeGroat, W.; Zeeshan, S.; Ahmed, Z. Investigating genes associated with heart failure, atrial fibrillation, and other cardiovascular diseases, and predicting disease using machine learning techniques for translational research and precision medicine. *Genomics* **2023**, *115*, 110584. [CrossRef]
22. Lee, C.J.; Baek, B.; Cho, S.H.; Jang, T.Y.; Jeon, S.E.; Lee, S.; Lee, H.; Nam, J.S. Machine learning with in silico analysis markedly improves survival prediction modeling in colon cancer patients. *Cancer Med.* **2023**, *12*, 7603–7615. [CrossRef] [PubMed]
23. Khanna, V.V.; Chadaga, K.; Sampathila, N.; Prabhu, S.; Bhandage, V.; Hegde, G.K. A distinctive explainable machine learning framework for detection of polycystic ovary syndrome. *Appl. Syst. Innov.* **2023**, *6*, 32. [CrossRef]
24. Shazly, S.A.; Coronado, P.J.; Yılmaz, E.; Melekoglu, R.; Sahin, H.; Giannella, L.; Ciavattini, A.; Carpini, G.D.; Di Giuseppe, J.; Yordanov, A.; et al. Endometrial Cancer Individualized Scoring System (ECISS): A machine learning-based prediction model of endometrial cancer prognosis. *Int. J. Gynecol. Obstet.* **2023**, *161*, 760–768. [CrossRef] [PubMed]
25. Roškar, L.; Kokol, M.; Pavlič, R.; Roškar, I.; Smrkolj, Š.; Rižner, T.L. Decreased Gene Expression of Antiangiogenic Factors in Endometrial Cancer: qPCR Analysis and Machine Learning Modelling. *Cancers* **2023**, *15*, 3661. [CrossRef] [PubMed]

26. Millán Arias, P.; Alipour, F.; Hill, K.A.; Kari, L. DeLUCS: Deep learning for unsupervised clustering of DNA sequences. *PLoS ONE* **2022**, *17*, e0261531. [CrossRef] [PubMed]
27. Rezaei-Tavirani, M.; Zamanian-Azodi, M.; Rajabi, S.; Masoudi-Nejad, A.; Rostami-Nejad, M.; Rahmatirad, S. Protein clustering and interactome analysis in Parkinson and Alzheimer's diseases. *Arch. Iran. Med.* **2016**, *19*, 101–109.
28. Negi, S.S.; Schein, C.H.; Braun, W. Regional and temporal coordinated mutation patterns in SARS-CoV-2 spike protein revealed by a clustering and network analysis. *Sci. Rep.* **2022**, *12*, 1128. [CrossRef]
29. Hu, L.; Yang, Y.; Tang, Z.; He, Y.; Luo, X. FCAN-MOPSO: An Improved Fuzzy-based Graph Clustering Algorithm for Complex Networks with Multi-objective Particle Swarm Optimization. *IEEE Trans. Fuzzy Syst.* **2023**, *31*, 3470–3484. [CrossRef]
30. Bischl, B.; Binder, M.; Lang, M.; Pielok, T.; Richter, J.; Coors, S.; Thomas, J.; Ullmann, T.; Becker, M.; Boulesteix, A.L.; et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2023**, *13*, e1484. [CrossRef]
31. Ali, Y.A.; Awwad, E.M.; Al-Razgan, M.; Maarouf, A. Hyperparameter search for machine learning algorithms for optimizing the computational complexity. *Processes* **2023**, *11*, 349. [CrossRef]
32. Prabu, S.; Thiyaneswaran, B.; Sujatha, M.; Nalini, C.; Rajkumar, S. Grid Search for Predicting Coronary Heart Disease by Tuning Hyper-Parameters. *Comput. Syst. Sci. Eng.* **2022**, *43*, 737–749. [CrossRef]
33. Belete, D.M.; Huchaiah, M.D. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *Int. J. Comput. Appl.* **2022**, *44*, 875–886. [CrossRef]
34. Anh, D.T.; Pandey, M.; Mishra, V.N.; Singh, K.K.; Ahmadi, K.; Janizadeh, S.; Tran, T.T.; Linh, N.T.T.; Dang, N.M. Assessment of groundwater potential modeling using support vector machine optimization based on Bayesian multi-objective hyperparameter algorithm. *Appl. Soft Comput.* **2023**, *132*, 109848. [CrossRef]
35. Rusch, T.; Mair, P.; Hornik, K. Structure-based hyperparameter selection with Bayesian optimization in multidimensional scaling. *Stat. Comput.* **2023**, *33*, 28. [CrossRef]
36. Yang, Y.; Lv, H.; Chen, N. A survey on ensemble learning under the era of deep learning. *Artif. Intell. Rev.* **2023**, *56*, 5545–5589. [CrossRef]
37. Zhu, X.; Li, J.; Ren, J.; Wang, J.; Wang, G. Dynamic ensemble learning for multi-label classification. *Inf. Sci.* **2023**, *623*, 94–111. [CrossRef]
38. Charoenkwan, P.; Schaduangrat, N.; Moni, M.A.; Manavalan, B.; Shoombuatong, W. SAPPHIRE: A stacking-based ensemble learning framework for accurate prediction of thermophilic proteins. *Comput. Biol. Med.* **2022**, *146*, 105704. [CrossRef]
39. Gao, M.; Zhao, L.; Zhang, Z.; Wang, J.; Wang, C. Using a stacked ensemble learning framework to predict modulators of protein–protein interactions. *Comput. Biol. Med.* **2023**, *161*, 107032. [CrossRef]
40. Du, L.; Liu, H.; Zhang, L.; Lu, Y.; Li, M.; Hu, Y.; Zhang, Y. Deep ensemble learning for accurate retinal vessel segmentation. *Comput. Biol. Med.* **2023**, *158*, 106829. [CrossRef]
41. Warde-Farley, D.; Donaldson, S.L.; Comes, O.; Zuberi, K.; Badrawi, R.; Chao, P.; Franz, M.; Grouios, C.; Kazi, F.; Lopes, C.T.; et al. The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **2010**, *38*, W214–W220. [CrossRef]
42. Khan, M.J.; Ullah, A.; Basit, S. Genetic basis of polycystic ovary syndrome (PCOS): Current perspectives. *Appl. Clin. Genet.* **2019**, *2019*, 249–260. [CrossRef] [PubMed]
43. Diamanti-Kandarakis, E.; Bartzis, M.I.; Bergiele, A.T.; Tsianateli, T.C.; Kouli, C.R. Microsatellite polymorphism (tttta) n at- 528 base pairs of gene CYP11α influences hyperandrogenemia in patients with polycystic ovary syndrome. *Fertil. Steril.* **2000**, *73*, 735–741. [CrossRef] [PubMed]
44. Wang, Y.; Wu, X.; Cao, Y.; Yi, L.; Chen, J. A microsatellite polymorphism (tttta) n in the promoter of the CYP11a gene in Chinese women with polycystic ovary syndrome. *Fertil. Steril.* **2006**, *86*, 223–226. [CrossRef] [PubMed]
45. Witchel, S.; Aston, C. The role of heterozygosity for CYP21 in the polycystic ovary syndrome. *J. Pediatr. Endocrinol. Metab. JPEM* **2000**, *13*, 1315–1317. [PubMed]
46. Takayama, K.; Suzuki, T.; Bulun, S.E.; Sasano, H.; Yilmaz, B.; Sebastian, S. Organization of the human aromatase p450 (CYP19) gene. *Proc. Semin. Reprod. Med.* **2004**, *22*, 5–9.
47. Wickham III, E.P.; Ewens, K.G.; Legro, R.S.; Dunaif, A.; Nestler, J.E.; Strauss, J.F., III. Polymorphisms in the SHBG gene influence serum SHBG levels in women with polycystic ovary syndrome. *J. Clin. Endocrinol. Metab.* **2011**, *96*, E719–E727. [CrossRef]
48. Gorsic, L.K.; Kosova, G.; Werstein, B.; Sisk, R.; Legro, R.S.; Hayes, M.G.; Teixeira, J.M.; Dunaif, A.; Urbanek, M. Pathogenic anti-Müllerian hormone variants in polycystic ovary syndrome. *J. Clin. Endocrinol. Metab.* **2017**, *102*, 2862–2872. [CrossRef] [PubMed]
49. Baban, A.S.S.; Korsheed, S.H.; Al Hayawi, A.Y. The FSHR polymorphisms association with polycystic ovary syndrome in women of Erbil, Kurdistan in North of Iraq. *Ibn Al Haitham J. Pure Appl. Sci.* **2018**, *2018*, 257–272. [CrossRef]
50. Nardo, L.; Patchava, S.; Laing, I. Polycystic ovary syndrome: Pathophysiology, molecular aspects and clinical implications. *Panminerva Medica* **2008**, *50*, 267–278.
51. Sir-Petermann, T.; Perez-Bravo, F.; Angel, B.; Maliqueo, M.; Calvillan, M.; Palomino, A. G972R polymorphism of IRS-1 in women with polycystic ovary syndrome. *Diabetologia* **2001**, *44*, 1200–1201.

52. Ajmal, N.; Khan, S.Z.; Shaikh, R. Polycystic ovary syndrome (PCOS) and genetic predisposition: A review article. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **2019**, *3*, 100060. [CrossRef] [PubMed]
53. Wojciechowski, P.; Lipowska, A.; Rys, P.; Ewens, K.G.; Franks, S.; Tan, S.; Lerchbaum, E.; Vcelak, J.; Attaoua, R.; Straczkowski, M.; et al. Impact of FTO genotypes on BMI and weight in polycystic ovary syndrome: A systematic review and meta-analysis. *Diabetologia* **2012**, *55*, 2636–2645. [CrossRef] [PubMed]
54. Urbanek, M. The genetics of the polycystic ovary syndrome. *Nat. Clin. Pract. Endocrinol. Metab.* **2007**, *3*, 103–111. [CrossRef]
55. Joseph, S.; Barai, R.S.; Bhujbalrao, R.; Idicula-Thomas, S. PCOSKB: A KnowledgeBase on genes, diseases, ontology terms and biochemical pathways associated with PolyCystic Ovary Syndrome. *Nucleic Acids Res.* **2016**, *44*, D1032–D1035. [CrossRef]
56. Babu, K.A.; Rao, K.L.; Kanakavalli, M.; Suryanarayana, V.; Deenadayal, M.; Singh, L. CYP1A1, GSTM1 and GSTT1 genetic polymorphism is associated with susceptibility to polycystic ovaries in South Indian women. *Reprod. Biomed. Online* **2004**, *9*, 194–200. [CrossRef]
57. Zhang, C.w.; Zhang, X.l.; Xia, Y.j.; Cao, Y.x.; Wang, W.j.; Xu, P.; Che, Y.n.; Wu, X.k.; Yi, L.; Gao, Q.; et al. Association between polymorphisms of the CYP11A1 gene and polycystic ovary syndrome in Chinese women. *Mol. Biol. Rep.* **2012**, *39*, 8379–8385. [CrossRef] [PubMed]
58. Zhao, S.; Tang, X.; Shao, D.; Dai, H.; Dai, S. Association study between a polymorphism of aldosterone synthetase gene and the pathogenesis of polycystic ovary syndrome. *Zhonghua Fu Chan Ke Za Zhi* **2003**, *38*, 94–97.
59. Li, L.; Gu, Z.P.; Bo, Q.M.; Wang, D.; Yang, X.S.; Cai, G.H. Association of CYP17A1 gene-34T/C polymorphism with polycystic ovary syndrome in Han Chinese population. *Gynecol. Endocrinol.* **2015**, *31*, 40–43. [CrossRef]
60. Goodarzi, M.O.; Xu, N.; Azziz, R. Association of CYP3A7* 1C and serum dehydroepiandrosterone sulfate levels in women with polycystic ovary syndrome. *J. Clin. Endocrinol. Metab.* **2008**, *93*, 2909–2912. [CrossRef]
61. Goodarzi, M.O. The genetic basis of the polycystic ovary syndrome. In *Androgen Excess Disorders in Women: Polycystic Ovary Syndrome and Other Disorders*; Springer: Berlin, Germany, 2007; pp. 223–233.
62. Spurdle, A.B.; Bowman, M.A.; Shamsani, J.; Kirk, J. Endometrial cancer gene panels: Clinical diagnostic vs research germline DNA testing. *Mod. Pathol.* **2017**, *30*, 1048–1068. [CrossRef]
63. Spurdle, A.B.; Thompson, D.J.; Ahmed, S.; Ferguson, K.; Healey, C.S.; O'mara, T.; Walker, L.C.; Montgomery, S.B.; Dermitzakis, E.T.; Group, A.N.E.C.S.; et al. Genome-wide association study identifies a common variant associated with risk of endometrial cancer. *Nat. Genet.* **2011**, *43*, 451–454. [CrossRef]
64. Painter, J.N.; O'mara, T.A.; Batra, J.; Cheng, T.; Lose, F.A.; Dennis, J.; Michailidou, K.; Tyrer, J.P.; Ahmed, S.; Ferguson, K.; et al. Fine-mapping of the HNF1B multicancer locus identifies candidate variants that mediate endometrial cancer risk. *Hum. Mol. Genet.* **2015**, *24*, 1478–1492. [CrossRef] [PubMed]
65. Setiawan, V.W.; Doherty, J.A.; Shu, X.o.; Akbari, M.R.; Chen, C.; De Vivo, I.; DeMichele, A.; Garcia-Closas, M.; Goodman, M.T.; Haiman, C.A.; et al. Two estrogen-related variants in CYP19A1 and endometrial cancer risk: a pooled analysis in the Epidemiology of Endometrial Cancer Consortium. *Cancer Epidemiol. Biomark. Prev.* **2009**, *18*, 242–247. [CrossRef]
66. O'Mara, T.A.; Glubb, D.M.; Painter, J.N.; Cheng, T.; Dennis, J.; Attia, J.; Holliday, E.G.; McEvoy, M.; Scott, R.J.; Ashton, K.; et al. Comprehensive genetic assessment of the ESR1 locus identifies a risk region for endometrial cancer. *Endocr. Relat. Cancer* **2015**, *22*, 851. [CrossRef] [PubMed]
67. Cheng, T.H.; Thompson, D.; Painter, J.; O'Mara, T.; Gorman, M.; Martin, L.; Palles, C.; Jones, A.; Buchanan, D.D.; Win, A.K.; et al. Meta-analysis of genome-wide association studies identifies common susceptibility polymorphisms for colorectal and endometrial cancer near SH2B3 and TSHZ1. *Sci. Rep.* **2015**, *5*, 17369. [CrossRef]
68. Chen, M.M.; O'Mara, T.A.; Thompson, D.J.; Painter, J.N.; (ANECS), A.N.E.C.S.G.; Attia, J.; Black, A.; Brinton, L.; Chanock, S.; Chen, C.; et al. GWAS meta-analysis of 16 852 women identifies new susceptibility locus for endometrial cancer. *Hum. Mol. Genet.* **2016**, *25*, 2612–2620. [CrossRef] [PubMed]
69. Cheng, T.H.; Thompson, D.J.; O'Mara, T.A.; Painter, J.N.; Glubb, D.M.; Flach, S.; Lewis, A.; French, J.D.; Freeman-Mills, L.; Church, D.; et al. Five endometrial cancer risk loci identified through genome-wide association analysis. *Nat. Genet.* **2016**, *48*, 667–674. [CrossRef]
70. Ligtenberg, M.J.; Kuiper, R.P.; Chan, T.L.; Goossens, M.; Hebeda, K.M.; Voorendt, M.; Lee, T.Y.; Bodmer, D.; Hoenselaar, E.; Hendriks-Cornelissen, S.J.; et al. Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3′ exons of TACSTD1. *Nat. Genet.* **2009**, *41*, 112–117. [CrossRef]
71. Haraldsdottir, S.; Hampel, H.; Tomsic, J.; Frankel, W.L.; Pearlman, R.; De La Chapelle, A.; Pritchard, C.C. Colon and endometrial cancers with mismatch repair deficiency can arise from somatic, rather than germline, mutations. *Gastroenterology* **2014**, *147*, 1308–1316. [CrossRef]
72. Mensenkamp, A.R.; Vogelaar, I.P.; van Zelst-Stams, W.A.; Goossens, M.; Ouchene, H.; Hendriks-Cornelissen, S.J.; Kwint, M.P.; Hoogerbrugge, N.; Nagtegaal, I.D.; Ligtenberg, M.J. Somatic mutations in MLH1 and MSH2 are a frequent cause of mismatch-repair deficiency in Lynch syndrome-like tumors. *Gastroenterology* **2014**, *146*, 643–646. [CrossRef]
73. Buchanan, D.D.; Tan, Y.Y.; Walsh, M.D.; Clendenning, M.; Metcalf, A.M.; Ferguson, K.; Arnold, S.T.; Thompson, B.A.; Lose, F.A.; Parsons, M.T.; et al. Reply to J. Moline et al. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **2014**, *32*, 2278–2279. [CrossRef] [PubMed]

74. Dowty, J.G.; Win, A.K.; Buchanan, D.D.; Lindor, N.M.; Macrae, F.A.; Clendenning, M.; Antill, Y.C.; Thibodeau, S.N.; Casey, G.; Gallinger, S.; et al. Cancer risks for MLH 1 and MSH 2 mutation carriers. *Hum. Mutat.* **2013**, *34*, 490–497. [CrossRef] [PubMed]
75. Senter, L.; Clendenning, M.; Sotamaa, K.; Hampel, H.; Green, J.; Potter, J.D.; Lindblom, A.; Lagerstedt, K.; Thibodeau, S.N.; Lindor, N.M.; et al. The clinical phenotype of Lynch syndrome due to germ-line PMS2 mutations. *Gastroenterology* **2008**, *135*, 419–428. [CrossRef] [PubMed]
76. Kempers, M.J.; Kuiper, R.P.; Ockeloen, C.W.; Chappuis, P.O.; Hutter, P.; Rahner, N.; Schackert, H.K.; Steinke, V.; Holinski-Feder, E.; Morak, M.; et al. Risk of colorectal and endometrial cancers in EPCAM deletion-positive Lynch syndrome: A cohort study. *Lancet Oncol.* **2011**, *12*, 49–55. [CrossRef] [PubMed]
77. Palles, C.; Cazier, J.B.; Howarth, K.M.; Domingo, E.; Jones, A.M.; Broderick, P.; Kemp, Z.; Spain, S.L.; Guarino, E.; Salguero, I.; et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat. Genet.* **2013**, *45*, 136–144. [CrossRef] [PubMed]
78. Valle, L.; Hernández-Illán, E.; Bellido, F.; Aiza, G.; Castillejo, A.; Castillejo, M.I.; Navarro, M.; Seguí, N.; Vargas, G.; Guarinos, C.; et al. New insights into POLE and POLD1 germline mutations in familial colorectal cancer and polyposis. *Hum. Mol. Genet.* **2014**, *23*, 3506–3512. [CrossRef] [PubMed]
79. Rohlin, A.; Zagoras, T.; Nilsson, S.; Lundstam, U.; Wahlström, J.; Hultén, L.; Martinsson, T.; Karlsson, G.B.; Nordling, M. A mutation in POLE predisposing to a multi-tumour phenotype. *Int. J. Oncol.* **2014**, *45*, 77–81. [CrossRef]
80. Elsayed, F.A.; Kets, C.M.; Ruano, D.; Van Den Akker, B.; Mensenkamp, A.R.; Schrumpf, M.; Nielsen, M.; Wijnen, J.T.; Tops, C.M.; Ligtenberg, M.J.; et al. Germline variants in POLE are associated with early onset mismatch repair deficient colorectal cancer. *Eur. J. Hum. Genet.* **2015**, *23*, 1080–1084. [CrossRef]
81. Billingsley, C.C.; Cohn, D.E.; Mutch, D.G.; Stephens, J.A.; Suarez, A.A.; Goodfellow, P.J. Polymerase ε (POLE) mutations in endometrial cancer: Clinical outcomes and implications for L ynch syndrome testing. *Cancer* **2015**, *121*, 386–394. [CrossRef]
82. Mahdi, H.; Mester, J.L.; Nizialek, E.A.; Ngeow, J.; Michener, C.; Eng, C. Germline PTEN, SDHB-D, and KLLN alterations in endometrial cancer patients with Cowden and Cowden-like syndromes: An international, multicenter, prospective study. *Cancer* **2015**, *121*, 688–696. [CrossRef]
83. Zhang, J.; Liu, B. A review on the recent developments of sequence-based protein feature extraction methods. *Curr. Bioinform.* **2019**, *14*, 190–199. [CrossRef]
84. Chen, Z.; Chen, Y.Z.; Wang, X.F.; Wang, C.; Yan, R.X.; Zhang, Z. Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PLoS ONE* **2011**, *6*, e22930. [CrossRef] [PubMed]
85. Labrín, C.; Urdinez, F. Principal component analysis. In *R for Political Data Science*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2020; pp. 375–393.
86. Yao, F.; Coquery, J.; Lê Cao, K.A. Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. *BMC Bioinform.* **2012**, *13*, 1–15. [CrossRef] [PubMed]
87. Ernst, M.; Sittel, F.; Stock, G. Contact-and distance-based principal component analysis of protein dynamics. *J. Chem. Phys.* **2015**, *143*, 244114. [CrossRef]
88. You, Z.H.; Lei, Y.K.; Zhu, L.; Xia, J.; Wang, B. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinform.* **2013**, *14*, 1–11. [CrossRef] [PubMed]
89. Ikotun, A.M.; Ezugwu, A.E.; Abualigah, L.; Abuhaija, B.; Heming, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci.* **2022**, *622*, 178–210. [CrossRef]
90. Reynolds, D.A.; et al. Gaussian mixture models. *Encycl. Biom.* **2009**, *741*, 10.
91. Nielsen, F.; Nielsen, F. Hierarchical clustering. *Introduction to HPC with MPI for Data Science*; Springer: Berlin, Germany, 2016; pp. 195–211.
92. Sahu, R.T.; Verma, M.K.; Ahmad, I. Density-based spatial clustering of application with noise approach for regionalisation and its effect on hierarchical clustering. *Int. J. Hydrol. Sci. Technol.* **2023**, *16*, 240–269. [CrossRef]
93. Wang, X.; Jin, Y.; Schmitt, S.; Olhofer, M. Recent advances in Bayesian optimization. *ACM Comput. Surv.* **2023**, *55*, 1–36. [CrossRef]
94. Li, H.; Liang, Q.; Chen, M.; Dai, Z.; Li, H.; Zhu, M. Pruning SMAC search space based on key hyperparameters. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e5805. [CrossRef]
95. Alkaff, A.K.; Prasetiyo, B. Hyperparameter Optimization on CNN Using Hyperband on Tomato Leaf Disease Classification. In Proceedings of the 2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), Malang, Indonesia, 16–18 June 2022; pp. 479–483.
96. Nguyen, B.; Morell, C.; De Baets, B. Scalable large-margin distance metric learning using stochastic gradient descent. *IEEE Trans. Cybern.* **2018**, *50*, 1072–1083. [CrossRef] [PubMed]
97. Lacoste, A.; Larochelle, H.; Laviolette, F.; Marchand, M. Sequential model-based ensemble optimization. *arXiv* **2014**, arXiv:1402.0796.
98. Feurer, M.; Springenberg, J.; Hutter, F. Initializing bayesian hyperparameter optimization via meta-learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 29.
99. Dai, Q.; Ye, R.; Liu, Z. Considering diversity and accuracy simultaneously for ensemble pruning. *Appl. Soft Comput.* **2017**, *58*, 75–91. [CrossRef]
100. Kumar, V.; Chhabra, J.K.; Kumar, D. Performance evaluation of distance metrics in the clustering algorithms. *INFOCOMP J. Comput. Sci.* **2014**, *13*, 38–52.

101. Feurer, M.; Hutter, F. Hyperparameter optimization. In *Automated Machine Learning: Methods, Systems, Challenges*; Springer: Berlin, Germany, 2019; pp. 3–33.
102. Yang, L.; Shami, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* **2020**, *415*, 295–316. [CrossRef]