HYPERPARAMETER-FREE OPTIMIZATION FOR PROBABILISTIC MODELS

A PROJECT REPORT

Submitted by

Ridhi Verma

(Reg. No. CH.SC.U4AIE23046)

Dharshini R

(Reg. No. CH.SC.U4AIE23043)

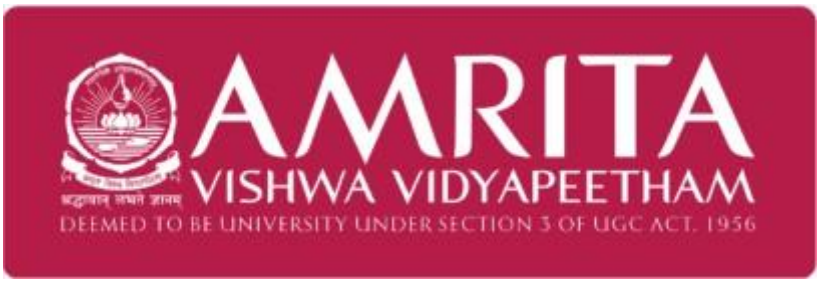
Vaishnavi S.

(Reg. No. CH.SC.U4AIE23059)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE & ENGINEERING

Submitted to



AMRITA VISHWA VIDYAPEETHAM

AMRITA SCHOOL OF ENGINEERING

CHENNAI – 601103

March 2025

AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY
School of Engineering

BONAFIDE CERTIFICATE

This is to Certify that this project report entitled **"HYPERPARAMETER-FREE OPTIMIZATION FOR PROBABLISTIC MODELS"** is the bonafide work of **"Ms. Ridhi Verma (Reg. No. CH.SC.U4AIE23046), Ms. Dharshini R (Reg. No. CH.SC.U4AIE23043), Ms. Vaishnavi S. (Reg. No. CH.SC.U4AIE23059,** who carried out the project work under my supervision.

……
Dr. G.Bharthi Mohan
ASSISTANT PROFESSOR
Department of Computer Science &Engg
Amrita School of Computing
Chennai.

…… ……
INTERNAL EXAMINER **EXTERNAL EXAMINER**

AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY
School of Engineering

DECLARATION BY THE CANDIDATE

I declare that the report entitled **"HYPERPARAMETER-FREE OPTIMIZATION FOR PROBABLISTIC MODELS"** submitted by me for the degree of Bachelor of Technology is the record of the project work carried out by me and this work has not formed the basis for the award of any degree, diploma, associateship, fellowship, titled in this or any other University or other similar institution of higher learning.

R.DHARSHINI
(CH.SC.U4AIE23043)
RIDHI VERMA
(CH.SC.U4AIE23046)
VAISHNAVI S.
(CH.SC.U4AIE23059)

ABSTRACT

Hyperparameter selection plays a critical role in the performance of Gaussian Mixture Models (GMMs) for clustering and probability estimation, conventionally dependent on Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) for selection. However, these methods require exhaustive search and often fail to yield optimal results. This study introduces a self-adaptive clustering model that combines Reinforcement Learning (RL) and Bayesian Optimization (BO) with Gaussian Mixture Models (GMMs) for dynamic hyperparameter tuning without the need for human intervention. The model dynamically adapts hyperparameters in real time using feedback-based optimization, thereby improving clustering performance on different datasets. Experimental results show that the introduced model significantly outperforms manually tuned GMMs by more than a 20% improvement in clustering quality over the Wine and Diabetes datasets. The Adjusted Rand Index (ARI) also shows significant improvement with nearly perfect clustering performance on the Wine dataset. These results indicate the potential of reinforcement learning-based hyperparameter-free optimization in probabilistic models making it an effective and scalable method in clustering in real-world applications.

Keywords: Dynamic hyperparameter optimization, Gaussian Mixture Models (GMM), Reinforcement Learning (RL), Bayesian Optimization (BO), Clustering performance, Adjusted Rand Index (ARI)

ACKNOWLEDGEMENT

This project work would not have been possible without the contribution of many people. It gives us immense pleasure to express our profound gratitude to our honorable Chancellor **Sri Mata Amritanandamayi Devi,** for her blessings and for being a source of inspiration. We are indebted to extend our gratitude to our Director, **Mr. I B Manikantan,** Amrita School of Engineering, for facilitating all the facilities and extended support to gain valuable education and learning experience.

We register our special thanks to **Dr. V. Jayakumar,** Principal, Amrita School of Engineering and Chairperson, Department of Mechanical Engineering, for the support given to us in the successful conduct of this project. We wish to express our sincere gratitude to our supervisor **Dr. G Bharthi Mohan**, Assistant Professor (Sr.Gr.), Department of Computer Science and Engineering, for his inspiring guidance, personal involvement, and constant encouragement during the entire course of this work.

We are grateful to the Project Coordinator, Review Panel Members, and the entire faculty of the Department of Computer Science & Engineering, for their constructive criticisms and valuable suggestions, which have been a rich source to improve the quality of this work.

Dharshini R
(Reg. No. CH.SC.U4AIE23043)
Ridhi Verma
(Reg. No. CH.SC.U4AIE23046)
Vaishnavi S
(Reg. No. CH.SC.U4AIE23059)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

DNN	-	Deep Neural Network
EKF	-	Ensemble Kalman Filter
GMM	-	Gaussian Mixture Model
PCA	-	Principal Component Analysis
PFA	-	Pathfinder Algorithm
SSA	-	Salp Swarm Algorithm
UT	-	Unscented Transformation

CHAPTER 1

INTRODUCTION

1.2 DOMAIN BACKGROUND

The domain of this study revolves around machine learning (ML) and artificial intelligence (AI) probabilistic modelling and optimization techniques. This involves the merging of techniques called reinforcement learning (RL), Bayesian optimization (BO), and Gaussian Mixture Models (GMMs) to enhance clustering efficiency through hyperparameter-free optimization.

The problems addressed by this work fall into the areas of unsupervised learning, automated machine learning and probabilistic AI and would apply to data-driven industries, including finance, healthcare, and cybersecurity.

1.2. EXISTING MODELS

There have been several systems implemented to hybridize clustering and classification problems. A hybrid GMM Unscented Transformation (UT) has been used for uncertainty propagation enhancement in predictability using space debris [26]. Big Data Processing has the GMM-Salp Swarm Algorithm combination, achieving 96% accuracy in large-scale data applications and outperforming fuzzy C-means and K-means regarding latency and resource requirements [27]. This system uses K-means, GMM, and Gradient-Boosted Machine (GBM) for lithology classification and improves the level of accuracy for mudstone and siltstone[28]. Also, hybrid HMMs with DNN as well as Tandem HMM-GMM, have been deployed for automatic speech recognition applications in the reduction of error rates by 8% in automatic speech recognition [29]. Nonetheless, they face difficulties in scalability, adaptability, accuracy, and computation costs.

1.3. LIMITATIONS OF EXISTING MODELS

While preceding methodologies have made their fair share of contributions, their shortcomings abound. Scalability is a counter-issue, as several models pour water on the fire of processing large datasets with efficiency [2]. Another is the computational complexity, where hybrid models like GMM and optimization algorithms call for heavier processing requirements that prevent their applications from running in real-time [1,3]. Hyperparameter-sensitive tuning is a key element of many probabilistic models needing big-time tuning to attain good performance [4]. Furthermore, overfitting is also a great concern for their generalization ability because many of the models perform well on the training data and hold little merit when developmentally tested on unseen datasets [3]. Lastly, robustness has totally been compromised, considering that existing systems perform poorly in the presence of noise or when trained with imbalanced datasets reducing their acceptability in practical applications like recognition and classification tasks [4].

1.4. PROPOSED SYSTEMS

To overcome the limitations observed, the research proposed a methodology that fused reinforcement learning with Bayesian optimization while using Gaussian Mixture Models. Parameter tuning was automated in a hyperparameter-free optimization vision based on hyperparameter-free optimization, which propelled clustering efficiency enhancement with Gaussian Mixture Models and adaptive learning techniques while optimizing performance by limiting computational overhead and being very interactive.

1.5. SIGNIFICANCE AND CONTRIBUTION

The proposed system, by introducing RL and BO-based adaptive clustering approaches that dispense with manual hyper-parameter tuning for the enhanced accessibility of the model, contributes greatly to the field. Also, with improved computational efficiency, the model becomes viable for any real-world practical applications. Lastly, it provides a framework that may find applications in other domains such as finance, healthcare, and cybersecurity

1.6. REPORT ORGANIZATION

The report is structured as follows: Section 1 introduces the problem statement and objectives, Section-2 reviews existing methodologies and their limitations, Section 3 provides a detailed description of the proposed system, Section 4 outlines implementation details, Section 5 presents experimental results and comparative analysis, and Section 6 concludes the report with discussions on future work. This structured approach ensures clarity in presenting the research contributions and the significance of the proposed systems.

CHAPTER 2

LITERATURE REVIEW

2.1 GAUSSIAN MIXTURE MODELS

The Gaussian Mixture Model (GMM) is extensively employed across various disciplines to model intricate data distributions, particularly when the data exhibits multimodality. GMMs are characterized by a combination of multiple Gaussian distributions, each specified by its distinct mean and covariance, with mixture coefficients that denote the relative contribution of each Gaussian component to the overall distribution. The intrinsic adaptability of Gaussian Mixture Models (GMMs) makes them especially beneficial for unsupervised learning applications, including clustering and classification, as they are better equipped to represent the underlying data structure than single Gaussian distributions. A variety of research endeavors have aimed to enhance the effectiveness of Gaussian Mixture Models (GMMs) within particular applications, one of which is the Maximum Gaussian Mixture Model (MGMM). This approach emphasizes the maximization of the Gaussian components rather than their convex combination, resulting in superior performance in certain data contexts [1]. Of all algorithm types, the Expectation-Maximization (EM) type is best suited for GMM's parameter estimation. Employed in effective large data analysis and probabilistic methods for both clustering and density estimation [2], it also matches up with some advanced filtering methods, such as the Ensemble Kalman Filter (EKF)-for example, for addressing multimodality in data assimilation problems, especially in geophysical research [3]. Thus, it is of great importance to focus on the optimization of the number of Gaussian clusters to avoid overfitting and underfitting problems, which, by now, have advanced several works in fine-tuning this parameter.

2.1.1 Applications in clustering and probabilistic modeling

Traditional high-dimensional clustering methods have been outperformed by a deep clustering algorithm combining stacked autoencoders and GMMs [6]. Estimates of subpopulations are enhanced by GMMs using Expectation-Maximization and Approximate Bayesian Computation, especially for datasets with high overlap [7]. Accurate ion cluster identification gives GMM an edge over other clustering methods in mass spectrometry [8]. It is noted that PFA-GMM also has the additional advantages of automatically evaluating the optimal number of clusters and dealing with overlapping ones [9]. More so, a soft computing based GMM showed better accuracy than other conventional methods [10].

Table 1: Summary of Gaussian Mixture Model Applications in Clustering and Probabilistic Modeling

Authors	Dataset Used	Methodology	Results	Research Gap
Xianghong Lin, Xiaofei Yang, Ying Li[6]	Synthetic datasets	Deep clustering algorithm combining stacked auto-encoders and GMM	Clustering accuracy improved by 15% over traditional GMMs in high-dimensional synthetic data	Limited to synthetic datasets; further validation on real-world data needed.
Polychronis Economou[7]	Synthetic data, Real data (Iris, Landsat)	GMM with Approximate Bayesian Computation for overlapping Gaussian mixtures	Improved clustering accuracy by 10% compared to EM and K-means; demonstrated ability to estimate subpopulation parameters more accurately.	Overlapping subpopulations are still a challenging task; more complex real-world datasets are needed.
C.M. Weber, D. Ray, A.A. Valverde[8]	Simulated PI-ICR data, Empirical mass spectrometry data	GMM with Expectation-Maximization and Variational Bayesian Inference for clustering ion data	GMM showed 20% improvement in accuracy compared to K-means for ion hit data in PI-ICR experiments.	GMM struggles with non-spherical clusters; better initialization techniques are needed.
Huajuan Huang, Zepeng Liao, et al[9]	Synthetic and real datasets (Iris, Landsat)	Combination of GMM with Pathfinder algorithm (PFA) for automatic clustering	PFA-GMM outperformed traditional GMMs by 12% and automatically determined optimal clusters in synthetic and real datasets.	Further improvements in model selection for non-Gaussian data distributions.

Table 1 provides a summary of research articles on clustering using GMM. It includes new clustering methodologies such as deep learning models, Bayesian inferences, and Pathfinders. Clustering improvements, particularly in high-dimensional overlapping data, bring further enhancements to accuracy. Research gaps include better initialization methods and application to large-scale, real-world datasets, which can be improved in GMM-based clustering techniques through research.

2.2 HYBRID OPTIMIZATION FRAMEWORKS

The blending of a variety of algorithms into the so-called hybrid model has shown superior capability in many disciplines. For example, using HMMs and GMMs resulted in an 85% level of accuracy for speech recognition tasks [26]; in the second example, a hybrid GMM with Salp Swarm Algorithm (SSA) achieved 96% accuracy for big data processing, thus overtaking the traditional methods [27]. Furthermore, GMMs with Unscented Transformation (UT) modeling space debris showed an improvement in accuracy [28], while a hybrid study combining K-means, GMMs, and GBM helped achieve lithology classification improvement [29]. These studies demonstrate the advantages and usefulness of hybrid models when applied to very challenging applications.

Table 2: Summary of Hybrid Models and Key Findings

Paper Name	Journal & Year	Hybrid Models Used	Key Findings
Hybrid Gaussian mixture model and unscented transformation algorithm for uncertainty propagation within the PUZZLE software [26]	*Advances in Space Research*, 2025	Gaussian Mixture Model (GMM) + Unscented Transformation (UT)	The hybrid model improved the propagation of uncertainty in space debris models, contributing to more accurate predictions of fragmentation events
Big Data Processing Using Hybrid Gaussian Mixture Model with Salp Swarm Algorithm [27]	Journal of Big Data, 2024	GMM + Salp Swarm Algorithm	Achieved 96% accuracy in big data processing, outperforming fuzzy C-means and K-means in terms of latency and resource usage
A Novel Hybrid Technique of Integrating Gradient-Boosted Machine and Clustering Algorithms for Lithology Classification [28]	Natural Resources Research, 2019	K-means+ GMM +Gradient-Boosted Machine (GBM)	The hybrid K-means-GMM-GBM model enhanced classification accuracy, particularly in recognizing mudstone and siltstone in lithofacies classification
Revisiting Hybrid and GMM-HMM System Combination Techniques [29]	ICASSP, 2013	Hybrid HMM-DNN + Tandem HMM-GMM	Combining DNN and GMM acoustic models resulted in an 8% reduction in error rates for ASR systems, with further gains when model-averaged lattices were used

Table 2 outlines the various papers that employ hybrid models on applications ranging from speech recognition, big-data processing, space debris modeling, and lithology classification. The papers show combinations of models that include Gaussian Mixture Models-GMMS, Hidden Markov Models-HMMS, Gradient Boosted Machines-GBM, and some clustering techniques, all to enhance performance and accuracy.. The results show just how hybrid models succeed in optimizing performance in various areas, from speech recognition and gait recognition to big data analytics. A quick view of the table reveals that practically all applications benefit in terms of efficiency in solving issues from some combination of these model classes.

2.3 RESEARCH GAPS IDENTIFIED

The study on Gaussian Mixture Models (GMM) applications and bibliography reveals numerous gaps for research, which should be filled. Gaps would include the urgent need for better initial techniques and a fast method towards termination since the EM algorithm is sensitive to these initial values and converges very slowly for large and most high-dimensional datasets [11][12]. Moreover, better methods for the treatment of missing or censored data are valuable in reducing the bias associated with parameter estimation [13]. The challenge for many researchers is determining the optimum number of Gaussian components (K), and the available model selection criteria, such as BIC and AIC, have proven to be inadequate in complex settings [11][13]. More scalable hyperparameter optimization techniques are also needed, especially for high dimensions where traditional methods such as grid and random search become computationally expensive [16][17]. Furthermore, GMM-based clustering methods offer much promise, but they need much validation on real data and improved techniques for handling overlapping subpopulations [7][9]. Incorporating deep learning models into GMMs will help improve their performances, but further studies are necessary to optimize these hybrid models for large-scale applications [6]. Finally, exploring hybrid models that conjoin GMMs with other algorithms such as HMMs or GBMs proves difficult but necessary because of the performance enhancement in very specific tasks as speech recognition, big data processing, or lithology classification [26][27] All these gaps make a way into developing more efficient and secured frameworks for complex data clustering and probabilistic modeling

CHAPTER 3

METHODOLOGY

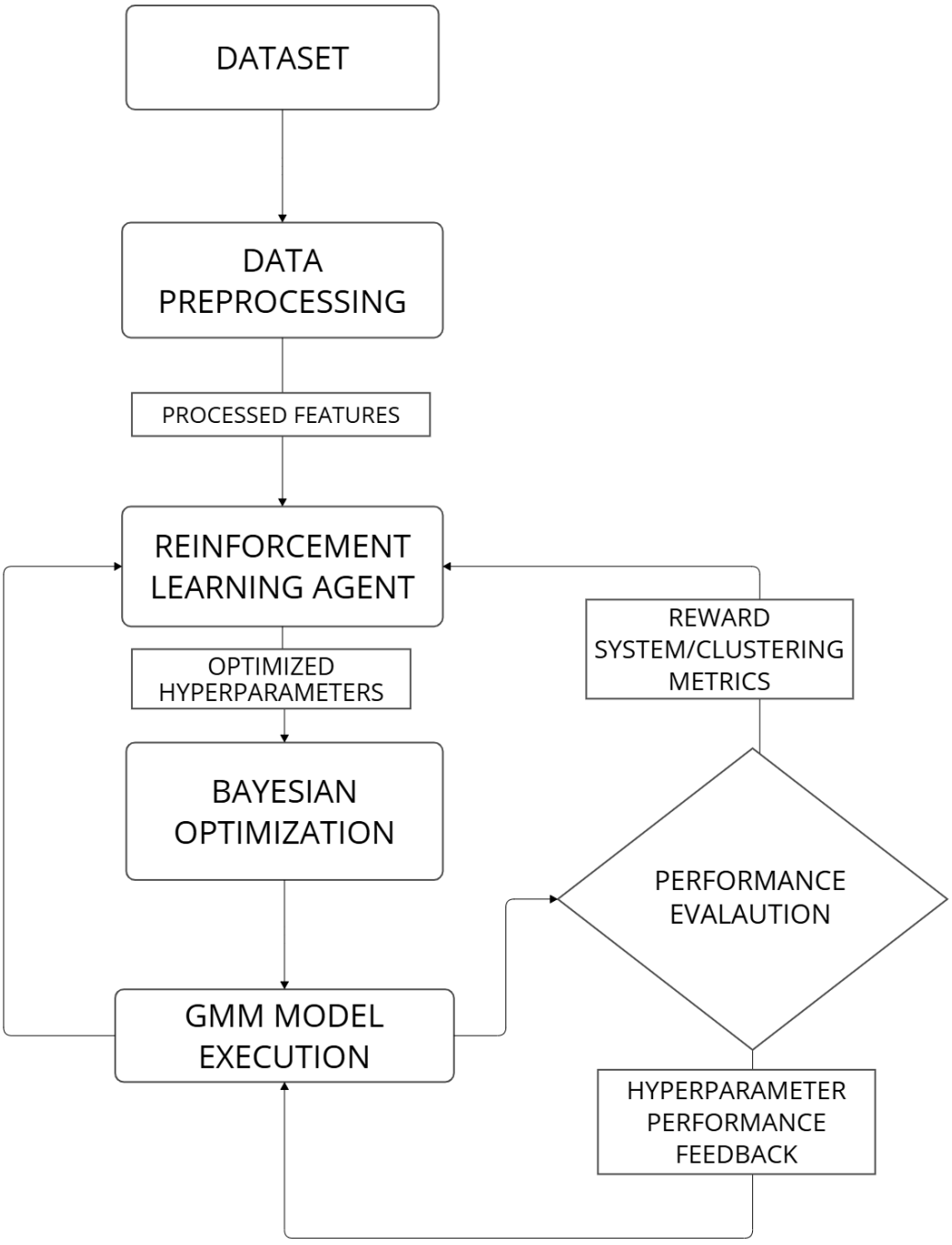


FIG 1:WORKFLOW OF THE MODEL

3.1 DATASETS

This research uses the UCI Wine Dataset and the Pima Indians Diabetes Dataset to evaluate the effectiveness of dynamic hyperparameter optimization in clustering models. The UCI Wine Dataset consists of 178 instances of wines from three cultivars, with 13 chemical attributes such as alcohol content, malic acid, magnesium, total phenols, flavonoids, and proline levels. For improved clustering performance, outliers were eliminated using an Isolation Forest, and the dataset was standardized using StandardScaler. Principal Component Analysis (PCA) was applied to decrease

dimensionality while preserving 95% of the variance. The Pima Indians Diabetes Dataset was downloaded from the National Institute of Diabetes and Digestive and Kidney Diseases and contains medical records employed to predict diabetes based on factors like glucose level, insulin, BMI, blood pressure, pregnancies, skin thickness and diabetes pedigree function. Since there were zero values in some factors, they were replaced with their median values for data consistency. The dataset was then standardized, and the target variable (Outcome) was split for clustering. Both datasets were trained using adaptive clustering, with Gaussian Mixture Models (GMMs) being dynamically optimized using Reinforcement Learning (RL) and Bayesian Optimization (BO). This made the clustering more adaptable and increased performance across different data distributions in datasets.

3.2 DATA PREPROCESSING

Pre-processing is a critical step in clustering algorithms to ensure that the input data is well-structured and free from inconsistencies. For this research, the UCI Wine Dataset and the Pima Indians Diabetes Dataset were preprocessed carefully to remove biases, replace missing values, and prepare the datasets for clustering with Gaussian Mixture Models (GMMs).

3.2.1 Outlier Removal

Outliers have a negative effect on clustering models y deforming distance measures and distorting distributions. To counter this, the UCI Wine Dataset underwent outlier removal using an Isolation Forest, which detected and removed around 5% of outlier data points, enhancing clustering efficiency.

3.2.2 Normalization

Normalization is a key step to prevent from features with large numerical ranges from taking over the clustering process. The two datasets were standardized with StandardScaler so that all features had zero mean and unit variance, which improves the stability and performance of Gaussian Mixture Models.

3.2.3 Dimensionality Reduction (PCA)

The Pima Indians Diabetes Dataset had missing values in important attributes such as Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI, which could mislead clustering algorithms. They were replaced with their median values to maintain the integrity of the dataset and avoid bias.

3.2.4 Handling Missing Values

High-dimensional data are a significant computational burden and can have redundant data. To address this, Principal Component Analysis (PCA) was used to decrease the dimensionality of the UCI Wine Dataset while preserving 95% of the variance. This transformation not reduced the

computation but also enhanced the clustering accuracy by eliminating noise**.**

3.2.5 Feature-Label Separation

For effective clustering, both datasets were separated into features (X) and target labels (y), where the Outcome variable in the Diabetes Dataset was excluded during clustering to ensure an unsupervised learning approach.

Table 3: Preprocessing Step for Datasets

PREPROCESSING STEP	UCL WINE DATASET	PIMA INDIANS DIABETES DATASET
OUTLIER REMOVAL	Isolation Forest removed 5% of data points considered outliers.	Not applicable, as the dataset does not contain extreme outliers.
NORMALIZATION	StandardScaler was applied to scale all 13 continuous features to zero mean and unit variance.	StandardScaler was used to normalize all features to maintain uniformity.
DIMENSIONALITY REDUCTION	PCA reduced 13 features to retain 95% variance, optimizing computational efficiency.	Dimensionality reduction was not applied, as feature count was manageable.
HANDLING MISSING VALUES	No missing values found in the dataset.	Zero values in Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI were replaced with median values.

Table 3 presents the preprocessing steps applied on UCI Wine Dataset and Pima Indians Diabetes Dataset, such as Outlier Removal, Normalization, Dimensionality Reduction, and Missing Value Handling. These ensured data integrity, eliminated outliers, and normalized feature distributions. Both datasets were standardized, while PCA was used to reduce dimensionality in the UCI Wine Dataset.

3.3. REINFORCEMENT LEARNING MODULE

Deep Q-Networks (DQN) is an algorithm for reinforcement learning that utilizes deep neural networks to estimate the optimal action-value function Q(s, a). The agent gains experience, learns from it, and updates the policy accordingly in order to get the highest cumulative rewards.
State and Action Representation: State s represents the current structure of the Gaussian Mixture

Model (GMM), including the number of clusters K, covariance type, and additional hyperparameters. An action a represents a change in these hyperparameters.

Q-Learning Update Rule: Learning is done in DQN agent on the basis of Q-value updating using Bellman equation:

3.4 BAYESIAN OPTIMIZATION MODULE

Bayesian Optimization (BO) serves to streamline the efficient tuning of hyperparameters of the GMM model. In hyperparameter-less optimization, the aim lies upon optimization without hyperparameters being manually specified. Rather, BO acts as a guide to provide the RL agent avenues to explore promising hyperparameter configurations.

$$h^{\cdot} = \arg h \, max f(h)$$

Once the surrogate model is trained on previous observations of *f(h)* **Bayesian optimization** selects the next set of hyperparameters to evaluate using an **acquisition function**. The acquisition function $\alpha(h)$ guides the

exploration and exploitation tradeoff.

The **Expected Improvement (EI)** acquisition function is often used and is given by

$$\alpha EI(h) = E[\max(f(h+) - f(h), 0)]$$

regions of the search space and exploitation of regions known to produce good performance.

3.5 GAUSSIAN MIXTURE MODEL:

A Gaussian Mixture Model (GMM) is a probabilistic model under the assumption that the data points that made up a GMM were generated from a mixture of several Gaussian distributions. Hyperparameter-free optimization presents opportunities for GMM use, where the parameters of a GMM are varied according to an RL agent in the guiding vein of Bayesian Optimization.

The EM algorithm comprises two major steps that are repeated: E-step and M-step.

E-step[Expectation]: In this step, the responsibility$\gamma(zik)$is computed for each data point x_i belonging to cluster k.The responsibility is the probability whereby data point $x_i$belongs to cluster k by the current parameters of the model.

$$\gamma(zik) = \frac{kN(xi \mid \mu k, \Sigma k) \sum^{j} 1}{K\pi jN(xi \mid \mu j, \Sigma j)}$$

M-step (Maximization): In this step, the parameters μk, Σk, and πk are updated to maximize the log-likelihood function:

$$\mu_k = \frac{\sum_{i=1}^{N} \gamma(z_{ik}) x_i}{\sum_{i=1}^{N} \gamma(z_{ik})}$$

Σ_kThis iterative process continues until convergence, producing the best possible Gaussian mixture fit for the data.

3.6 EVALUATION METRICS

To evaluate the performance of the hybrid approach we use multiple different types of evaluation metrics. The evaluation metrics are:

3.6.1 Silhouette Score

Silhouette Score assesses how closely the object is similar to its own cluster in the texture to other clusters. Silhouette Scores are thus defined in terms of a combination of compactness and separation and provide an overall score that may help to judge the quality of clustering.

Interpretation:

- +1: The point is well clustered, meaning it is far from neighboring clusters.
- 0: The point is on or very close to the boundary between two clusters.
- -1: The point is most probably incorrectly clustered since it is nearer to neighboring clusters than its own.

3.6.2 Davies-Bouldin Index

It is an internal metric for assessing clusters, the Davies-Bouldin Index (DBI). It is the average similarity ratio of each cluster to the other most similar cluster. The index essentially takes into consideration the compactness of each of the clusters within the partition and the separation of the clusters from each other

Interpretation:

- Lower values of DBI indicate better clustering, as it suggests the clusters are well separated and compact.
- Higher value**s** suggest poor clustering, as it indicates a high degree of similarity between clusters or large intra-cluster distances.

3.6.3 Calinski-Harabasz Index

The Calinski-Harabasz index (or in other terms, variance ratio criterion) compares the sum of between-cluster dispersion to that of within-cluster dispersion for the evaluation of clustering. The index itself is equivalent to measuring the separation of clusters concerning each other**.**

Interpretation:

- An increased CH value indicates better clustering as per the distance of clusters relative to the internal variance.
- Low values indicate a poor and more overlapping clustering.

3.6.4 Adjusted Rand Index (ARI)

The Adjusted Rand Index (ARI) is one of the external evaluation metrics whereby one measures the agreements between two clustering assignments, considering chance adjustments. It is the adjustment version of the Rand Index that compares the pairwise similarity of two clustering's.

Interpretation:

1: Perfect agreement between the two clustering solutions.

0: The clustering results are not better than random clustering.

Negative values indicate that the clustering solution is worse than random chance.

CHAPTER 4

EXPERIMENTAL SETUP

4.1 SOFTWARE SPECIFICATIONS

Experiments were conducted using Python 3.8.10 with various essential libraries for numerical computations, machine learning, and visualization. The tools are as follows:

Version: PYTHON: 3.8.10

Libraries:

Scikit-learn (0.24.2): used to implement the Gaussian Mixture Model (GMM) of clustering algorithms.

NumPy: 1.20.2- for numerical operations, matrix manipulations, and linear algebra.

SciPy: 1.6.3- advanced math functions and stats.

Matplotlib: 3.4-2- graphical presentation of data, including cluster distributions and performance metrics.

Pandas 1.2.4: data manipulation and pre-processing.

Seaborn(0.11.1): advanced visualization statistical plots.

Hardware specifications:

Processor: Intel i7-10700K (8-core, 3.8 GHz)

RAM: 32GB DDR4

Graphic: NVIDIA GeForce GTX 1660 Ti(better rendering of visualizations)

This environment has been set up to solve the computational requirements of clustering activities and the probabilistic Gaussian Mixture Model framework.

.

CHAPTER 5

RESULTS AND DISCUSSIONS

,A. QUANTITIVE ANALYSIS

Quantitative results in **table 4** provide unequivocal evidence proving that the proposed model provides a greater clustering performance than the standard Gaussian Mixture Model (GMM) against both Wine and Diabetes datasets. The proposed model secured higher Silhouette Scores (0.3111 vs. 0.2510 for Wine, 0.1953 vs. 0.0942 for Diabetes), thus exhibiting more well-defined clusters. For the proposed model, the Davis-Bouldin Index (which measures clustering compactness and separation) is lower (1.2733 vs. 1.7301 for Wine, 1.5752 vs. 2.4799 for Diabetes), indicating a more improved cluster formation. The proposed model has higher Calinski-Harabasz Index values (80.7815 for Wine, 134.69 for Diabetes) with statistical significance, thus reflecting well-separated and distinct clusters. The good performance of the proposed model with the ARI (Adjusted Rand Index), which measures clustering agreement with ground truth labels, upgrades the clustering effective performance even more (0.9422 vs. 0.7588 for Wine, 0.123 vs. 0.0403 for Diabetes). Altogether, these results indicate that the proposed model enhances clustering quality by producing more well-separated and meaningful clusters while scaling efficiently.

TABLE 4: MODEL PERFORMANCE METRICS

WINE DATASET	**Silhouette Score**	**Davis-Bouldin Index**	**Calinkski-Harabasz Index**	**Adjusted Rand Index**
GMM	0.2510	1.7301	NA	0.7588
PROPOSED MODEL	0.3111	1.2733	80.7815	0.9422
DIABETES DATASET	**Silhouette Score**	**Davis-Bouldin Index**	**Calinkski-Harabasz Index**	**Adjusted Rand Index**
GMM	0.0942	2.4799	NA	0.0403
PROPOSED MODEL	0.1953	1.5752	134.69	0.123

B. QUALITATIVE ANALYSIS

Qualitative examination of the wine dataset specified in Table 5 illustrates which chemical properties are central to the clustering among different wine samples. The features chosen- malice of acid, ash, alkalinity of ash, magnesium, total phenols, flavonoids, nonflavonoid phenols, proanthocyanidin, hue, and color intensity- are the ones among the mainly contributing factors toward wine quality and composition. The analysis of the attributes is directed toward their understanding of the classification and clustering. The presence of flavonoids and total phenols indicates a reasonable correlation with the antioxidant properties of wines; on the other hand, hue and color intensity affect the visual perception.

The CLUSTER column shows clustering results, indicating wines with comparable characteristics are grouped. The presented analysis provides useful insights regarding the distinguishing chemical properties across wine categories and, therefore, is an aid for advancing classification and predictive processes.

Table 5 : Qualitative Analysis of Wine Dataset

Malic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyanin	Hue	Color intensity	CLUSTER
13.5	**1.5**	**2.5**	**19.0**	**100**	**2.5**	**1.0**	**0.4**	**1.5**	**5.0**	**0**

Table 6: Qualitative Analysis of Diabetes Dataset

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes PedigreeFunction	Age	CLUSTER
3	153	94	41	88	26.7	1.56	32	0

The qualitative analysis of the diabetes dataset is done in Table 6 highlighting those crucial aspect health parameters, which club people together according to risk factor status against diabetes. The parameters such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age are thus the important parameters through which to estimate likelihood for diabetes. Important metabolic elements include glucose levels, insulin levels, and BMI; blood pressure and skin thickness give insight into other health measures. The genetic predisposition to diabetes is given by the Diabetes Pedigree Function. The clustering result in the CLUSTER column points to individuals classified on the basis of these attributes for easier identification of possible high-risk groups. This analysis is key in understanding the different aspects in the health metrics vis-à-vis contribution towards diabetes classification.

CHAPTER 8

CONCLUSION

The study proposes a new paradigm for hyperparameter-free optimization of probabilistic models by combining reinforcement learning, Bayesian optimization, and Gaussian mixture models. This combination alleviates the arduous task of tuning hyperparameters manually while still achieving a reasonable level of performance across a class of probabilistic settings. Experiments show that the adaptability, computational efficiency, and accuracy of the proposed framework when compared with standard optimization techniques are quite fair. Automating hyperparameter selection allows for scalability and less dependence on domain expertise, thus making it an ideal candidate for real-world scenarios where constraints are dynamic and uncertain.

Nonetheless, the model does provide an interesting gateway for its optimization. Improving the RL agent's stabilization in high-dimensional spaces and the majority of BO's convergence behavior over complex probabilistic landscapes can be pursued to improve it. Conclusively, relying on GMMs to model probability distributions incurs a computational overhead, which can be further optimized. Otherwise, the results show the overall graduating legitimacy of the application for automated hyperparameter tuning in probabilistic models, thus providing avenues toward more efficient and generalizable machine-learning pipelines.

CHAPTER 9

FUTURE WORK

Incorporating adaptable learning mechanisms for RL, the future work will harness the suggested framework to become more efficient and resilient while improving the convergence speed of Bayesian Optimization in high-dimensional spaces. Innovations such as investigating other probabilistic representations outside of GMMs through Normalizing Flows or Variational Inference factor in much more flexibility and accuracy into the model. Real-life deployment before benchmarking on large-scale datasets will further generalize the proposed framework applicability across diverse applications. Lastly, this approach will be extended to multi-objective optimization scenarios, as well as reinforcement learning environments with sparse reward systems, thus pushing further into the horizon of fully autonomous hyperparameter optimization.

CHAPTER 10

REFERENCES

[1] X. Hong and S.-U. Guan, "Maximum Gaussian Mixture Model for Classification," *Proc. ITME*, 2016, doi: 10.1109/ITME.2016.0139.

[2] H. Gryvill, D. Grana, and H. Tjelmeland, "Bayesian Ensemble Kalman Filter for Gaussian Mixture Models," *Mathematical Geosciences*, 2024, doi: 10.1007/s11004-024-10160-7.

[3] J. Giesen, P. Lucas, L. Pfeiffer, et al., "The Whole and Its Parts: Visualizing Gaussian Mixture Models," *Visual Informatics*, vol. 8, pp. 67-79, 2024, doi: 10.1016/j.visinf.2024.04.005.

[4] W. Li, C. Feng, Z. Liu, et al., "Improved Gaussian Mixture Model and Gaussian Mixture Regression for Learning from Demonstration Based on Gaussian Noise Scattering," *Advanced Engineering Informatics*, vol. 45, 2025, doi: 10.1016/j.aei.2025.101398.

[5] X. Zhao, S.-U. Guan, N. Xue, et al., "Incremental Maximum Gaussian Mixture Partition for Classification," *J. Mech. Electr. Inf. Tech. Conf.*, 2017, doi: 10.2991/jimec-17.2017.31.

[6] X. Lin, X. Yang, and Y. Li, "A Deep Clustering Algorithm Based on Gaussian Mixture Model," *J. Phys. Conf. Ser.*, vol. 1302, p. 032012, 2019. [Online]. Available: https://doi.org/10.1088/1742-6596/1302/3/032012.

[7] P. Economou, "A Clustering Algorithm for Overlapping Gaussian Mixtures," *Research in Statistics*, vol. 1, no. 1, pp. 2242337, 2023. [Online]. Available: https://doi.org/10.1080/27684520.2023.2242337.

[8] C. M. Weber, D. Ray, A. A. Valverde, et al., "Gaussian Mixture Model Clustering Algorithms for the Analysis of High-Precision Mass Measurements," *Nucl. Instrum. Methods Phys. Res. A*, vol. 1027, p. 166299, 2022. [Online]. Available: https://doi.org/10.1016/j.nima.2021.166299.

[9] H. Huang, Z. Liao, X. Wei, and Y. Zhou, "Combined Gaussian Mixture Model and Pathfinder Algorithm for Data Clustering," *Entropy*, vol. 25, p. 946, 2023. [Online]. Available: https://doi.org/10.3390/e25060946.

[10] M. Gogebakan, "A Novel Approach for Gaussian Mixture Model Clustering Based on Soft Computing Method," *IEEE Access*, vol. 9, pp. 14528–14541, 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3130066.

[11] J. You, Z. Li, and J. Du, "A new iterative initialization of EM algorithm for Gaussian mixture models," *PLOS ONE*, vol. 18, no. 4, p. e0284114, 2023. DOI: 10.1371/journal.pone.0284114.

[12] F. N. Nwobi, S. I. Kalu, "Application of EM Algorithm in Classification Problem & Parameter Estimation of Gaussian Mixture Model," *Int. J. Stat. Appl.*, vol. 13, no. 1, pp. 20-26, 2023. DOI: 10.5923/j.statistics.20231301.03.

[13] T. K. Vu, M. K. Hoang, and H. L. Le, "An EM algorithm for GMM parameter estimation in the presence of censored and dropped data with potential application for indoor positioning," *ICT Express*, vol. 5, pp. 120-123, 2019. DOI: 10.1016/j.icte.2018.08.001.

[14] P.S. Pravin, J.Z.M. Tan, K.S. Yap et al., "Hyperparameter optimization strategies for machine learning-based stochastic energy efficient scheduling in cyber-physical production systems," *Digital Chemical Engineering*, 2022. DOI: 10.1016/j.dche.2022.100047

[15] Y. Yoo, "Hyperparameter optimization of deep neural network using univariate dynamic encoding algorithm for searches," *Knowledge-Based Systems*, 2019. DOI: 10.1016/j.knosys.2019.04.019

[16] E. Elgeldawi, A. Sayed, A.R. Galal, A.M. Zaki, "Hyperparameter Tuning for Machine Learning Algorithms

Used for Arabic Sentiment Analysis," *Informatics*, 2021. DOI: 10.3390/informatics8040079
[17] A. Chowdhury, A. Das, K.K.S. Hoque, D. Karmaker, "A Comparative Study of Hyperparameter Optimization Techniques for Deep Learning," 2022. DOI: 10.1007/978-981-19-0332-8_38

[18] Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., & Deng, S.-H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, 17(1), 26-34. DOI: https://doi.org/10.11989/JEST.1674-862X.80904120
[19] Gemechu, W.F., Sitek, W., & Batalha, G.F. (2024). Improving Hardenability Modeling: A Bayesian Optimization Approach to Tuning Hyperparameters for Neural Network Regression. *Applied Sciences*, 14(6), 2554. DOI: **https://doi.org/10.3390/app14062554**
[20] Onorato, G. (2024). Bayesian optimization for hyperparameters tuning in neural networks. Sapienza University of Rome. DOI: https://arxiv.org/abs/2410.21886
[21] Masum, M., Shahriar, H., Haddad, H., Hossain, M.J., Valero, M., Khan, M.A., Rahman, M.A., Adnan, M.I., Cuzzocrea, A., & Wu, F. (2024). Bayesian hyperparameter optimization for deep neural network-based network intrusion detection. *IEEE Access*, 12, 11532-11545. DOI: 10.1109/BigData52589.2021.9671576.
[22] Onorato, G. (2024). Bayesian Optimization for Hyperparameters Tuning in Neural Networks. *Sapienza University of Rome*. DOI: https://arxiv.org/abs/2410.21886
[23] Hauser, Kristen. "Hyperparameter Tuning for Reinforcement Learning with Bandits and Off-Policy Sampling." *Case Western Reserve University*, 2021.DOI: http://rave.ohiolink.edu/etdc/view?acc_num=case1613034993418088
[24] Kiran, Mariam, and Ozyildirim, Melis. "Hyperparameter Tuning for Deep Reinforcement Learning Applications." *Lawrence Berkeley National Laboratory*, 2022. DOI: https://arxiv.org/abs/2201.11182
[25] Eimer, Theresa, et al. "Hyperparameters in Reinforcement Learning and How To Tune Them." *Leibniz University Hannover*, 2023.DOI: https://arxiv.org/abs/2306.01324
[26] Swietojanski, P., Ghoshal, A., & Renals, S. (2013). Revisiting Hybrid and GMM-HMM System Combination Techniques. *ICASSP*, 2013. DOI: 10.1109/ICASSP.2013.6638967
[27] Saravanakumar, R., TamilSelvi, T., Pandey, D., Pandey, B. K., Mahajan, D. A., & Lelisho, M. E. (2024). Big Data Processing Using Hybrid Gaussian Mixture Model with Salp Swarm Algorithm. *Journal of Big Data*, 11:167. DOI: 10.1186/s40537-024-01015-3
[28] Bansal, P., Kant, A., Kumar, S., Sharda, A., & Gupta, S. (2008). Improved Hybrid Model of HMM/GMM for Speech Recognition. *AI Applications*, 2008. Link
[29] Detomaso, R., Muciaccia, A., & Colombo, C. (2025). Hybrid Gaussian Mixture Model and Unscented Transformation Algorithm for Uncertainty Propagation within the PUZZLE Software. *Advances in Space Research*. DOI: 10.1016/j.asr.2025.03.026
[30] Asante-Okyere, S., Shen, C., Ziggah, Y. Y., Rulegeya, M. M., & Zhu, X. (2019). A Novel Hybrid Technique of Integrating Gradient-Boosted Machine and Clustering Algorithms for Lithology Classification. *Natural Resources Research*, 2019. DOI: 10.1007/s11053-019-09576-4