CS 7641 Machine Learning:
Unsupervised Learning and Dimensionality Reduction

Jilong Cui

jcui@gatech.edu

1 OVERVIEW

This report presents an analysis of two clustering algorithms, KMeans and Expectation Maximization (EM), combined with four dimensionality reduction techniques: Principal Component Analysis (PCA), Independent Component Analysis (ICA), Randomized Projections, and Lasso Feature Selection. These methodologies are systematically applied and fine-tuned on two distinct classification datasets, each with its own unique characteristics. Subsequently, the datasets processed through these dimensionality reduction and clustering pipelines are employed to train Neural Network classifiers. The performance outcomes, as well as the behavior and implications of each method on the datasets, are examined and detailed in the analysis.

2 DATA SET

The datasets chosen for this investigation – UFC Fight and Wine – present distinct challenges. The **UFC Fight Dataset** provides the nuances of mixed martial arts, capturing 156 features that represent a fighter's performance and characteristics in a binary classification problem—predicting the winner from either the red or blue corner. The dataset, comprising 3,592 entries, reflects the complexities of competitive sports, where outcomes are not solely determined by quantitative measures but are also influenced by the unpredictable nature of human competition. The data reveals a pronounced imbalance, with the red corner—typically the favorite—emerging as the victor in 66.26% of the cases. This imbalance underscores the significance of accurately predicting blue corner victories. The dataset's attributes, a mixture of continuous and binary data, reflect the diverse elements of a fighter's skill set and strategic approach, indicating that clustering could provide valuable insights into different fighting styles and techniques, beyond the binary outcome of wins and losses. Using f1_score (macro) as a metric allows for a more balanced consideration of predictive performance across both classes, particularly valuing the less represented blue corner.

On the other hand, the **Wine Dataset**, derived from a chemical analysis of wines from three cultivars in Italy, offers a completely different landscape. With 13 continuous features and 5,000 entries (selected for computational efficiency), the dataset invites a three-class classification to discern the cultivar origin of each wine. The features here are objective, quantifiable, and predictable, making for a less complex classification task than the UFC data, as evidenced by higher achievable accuracies (over 90%). The balanced nature of the dataset, in conjunction with the use of f1_score (weighted) as an optimization metric, provides a fair representation of each class in the predictive modeling process. Clustering within the Wine Dataset could be particularly effective in identifying inherent groupings based on chemical compositions, potentially revealing subtler distinctions between the cultivars that might not be immediately apparent. This approach contrasts with the UFC dataset, where clustering could unravel the layered strategies and combat styles of fighters, which are not directly linked to the fight's outcome.

In summary, while the UFC Fight Dataset offers a fertile ground for clustering analyses to understand fighter profiles and tactics, the Wine Dataset is well-suited for clustering based on chemical properties.

3 CLUSTERING ON ORIGINAL DATA SET

KMeans and Expectation Maximization (EM) are implemented by python sklearn KMeans and GaussianMixture.

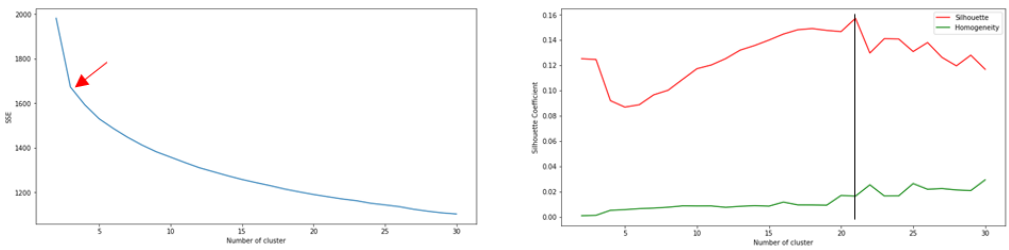
The elbow method (Figure 1 a - Wine dataset as example) is used to determine the optimal number of clusters by finding the "elbow" point on a plot of explained variation versus the number of clusters. For the Wine dataset, the elbow was identified at three clusters, which coincides with the actual number of cultivars. To assess clustering effectiveness, the Silhouette coefficient and homogeneity score were employed, with the optimal clusters chosen based on the highest Silhouette coefficient and a high homogeneity score. The UFC dataset's optimal number of clusters was identified as 22 (Figure 1 b, UFC dataset as example), while for the Wine dataset it remained 3.

The matching score between the cluster labels and the original dataset labels is also considered to evaluate accuracy. The UFC data, with its high-dimensionality and class imbalance, posed a challenge to KMeans, resulting in a modest match score of 0.429. This suggests that the straightforward distance-based partitioning of KMeans couldn't effectively tackle the dataset's intricacies. However, the match score improved significantly with EM, reaching 0.663, likely due to its probabilistic framework which accommodates the UFC dataset's subtleties better. The Wine dataset, with distinct and well-separated clusters, aligned neatly with KMeans, achieving a high match score of 0.903. This indicates that the simpler structure of the Wine data is well-suited for KMeans. EM's performance, while slightly lower at 0.87, still suggests a good fit, albeit with less precision than KMeans, possibly due to EM's sensitivity to initialization and convergence nuances.

The clusters obtained from KMeans on the UFC dataset didn't align as closely with the labels, implying that the algorithm might benefit from adjustments like feature scaling or kernel methods to handle non-linear separations. For EM, incorporating constraints or priors might help in guiding the clustering to a more accurate alignment with the labels. The Wine dataset's clusters corresponded well with the labels, affirming that the chemical attributes form natural groupings reflective of the cultivars. This natural alignment bolsters the interpretability and justifies the cluster choices.

Comparing the algorithms, KMeans' performance hinges on the dataset's linearity and cluster balance, while EM's flexibility offers better handling of complex data structures at the cost of potentially overfitting. Enhancements like dimensionality reduction or advanced initialization techniques might bolster their respective performances. The selection of datasets undoubtedly influenced the outcomes. The UFC dataset's high-dimensional space and inherent unpredictability of fight outcomes introduced complexities that skewed the clustering results. Conversely, the Wine dataset's clarity and lower dimensionality facilitated a more straightforward clustering process.

In conclusion, the performance of both clustering algorithms is not only algorithm-dependent but also heavily influenced by the characteristics of the datasets. To improve performance, fine-tuning the algorithms to align with the datasets' properties, such as feature selection or parameter optimization, would be essential.



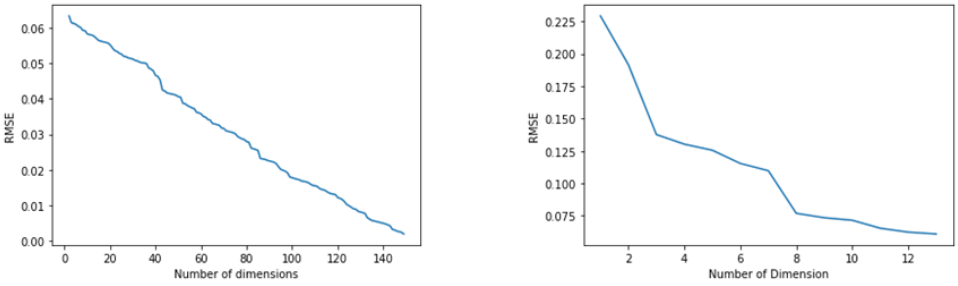
(a) Wine Dataset – Elbow (b) UFC Dateset – KMean Coef.

Figure 1. Clustering Analysis – KMean & Expectation Maximization (EM)

4 DIMENSIONALITY REDUCTION ON DATA SET

In the dimensionality reduction phase of the analysis, four algorithms were employed: Principal Component Analysis (PCA), Independent Component Analysis (ICA), Gaussian Random Projection (GRP), and Lasso feature selection.

Reconstruction error was the main metric for assessing dimensionality reduction, reflecting information loss when data is compressed and then restored. This error was charted against retained dimensions, with Figure 2 illustrating this for Gaussian Random Projection (GRP). The reduction extent was decided by balancing this error with dataset specifics and expert insights, such as key chemical attributes in the Wine dataset. Table 1 compiles the final dimension count and corresponding errors for each reduction technique used.



(a) UFC – Randomized Projections (d) Wine – Randomized Projections

Figure 2. Reconstruction Error vs. Dimensionality (Randomized Projections as Example)

Table 1. Summary of Dimensionality Reduction

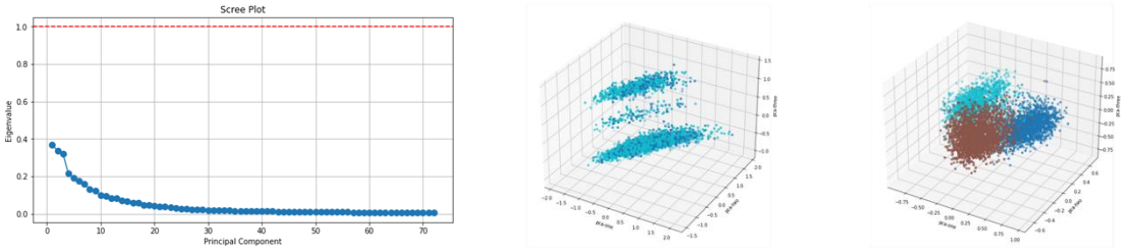
		Original	PCA	ICA	RP	Lasso
UFC	**Features**	156	72	80	40	39
	Reconstruction Error	-	0.04	0.027	0.042	-
Wine	**Features**	13	8	3	9	4
	Reconstruction Error	-	0.19	0.163	0.075	-

In newly crafted spaces through dimensionality reduction, the data assumes different geometries.

PCA transforms data into a space where eigenvalues gauge the variance each principal component holds. Here, the UFC dataset's eigenvalues diminish significantly after the 20th component (Figure 3 a), corresponding to its cluster analysis, suggesting substantial variance is captured within the first few components. The Wine dataset's reduction to 8 dimensions, while incurring a higher reconstruction error, still reflects its three cultivars' defining traits, with the first few components being paramount. 3D visualizations using the top three components illustrate the data's new structure. The UFC dataset (Figure 3 b) forms three broad groups, which, intriguingly, don't align with match winners or losers, suggesting PCA's strength lies in uncovering fighters' tactics rather than forecasting results. Conversely, the Wine dataset's 3D plot (Figure 3 c) shows distinct separation among its classes, affirming PCA's ability to retain essential data characteristics that match the original class labels.

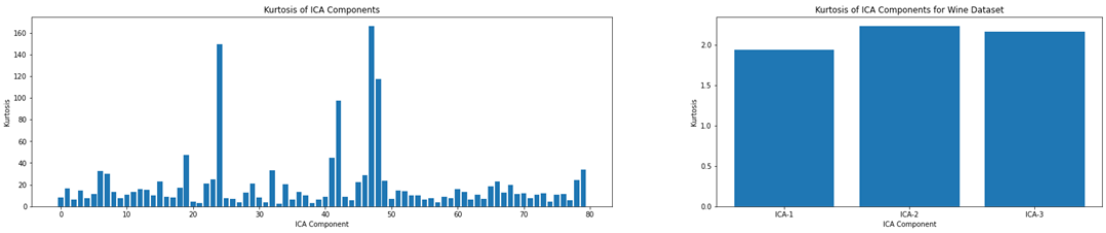
ICA distills the UFC dataset to 80 features, a reduction accompanied by a modest reconstruction error, hinting at the presence of statistically significant independent components. This efficiency is likely due to ICA's unique capability to identify data points that are statistically independent rather than merely uncorrelated as in PCA. For the Wine dataset, ICA compresses it to 3 components with a

relatively minor increase in reconstruction error. This underscores the potential redundancy in the original features and ICA's effectiveness in isolating the core informational essence of the data. The kurtosis analysis of ICA, illustrated for the UFC dataset (Figure 4 a), revealed that only a handful of components carry the bulk of the dataset's information richness. The corresponding 3D component distribution plot, although not explicitly separating winners from losers, uncovers stylistic nuances between fighters. In the case of the Wine dataset, the near-equal kurtosis of the reduced features underscores a harmonious contribution to class differentiation, with the 3D plot likely reinforcing the clear class demarcation seen in PCA's output. Thus, ICA excels in reducing dimensions while preserving enough information for accurate classification, particularly evident in the Wine dataset's clear class delineation.



(a) UFC – PCA eigenvalue (b) UFC – PCA first three PC (c) Wine – PCA first three PC

Figure 3. PCA Eigenvalue and Principal Component Distribution



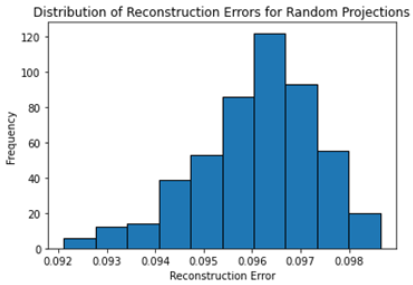
(a) UFC – ICA kurtosis (b) Wine – ICA Kurtosis

Figure 4. ICA Kurtosis and Independent Component Distribution

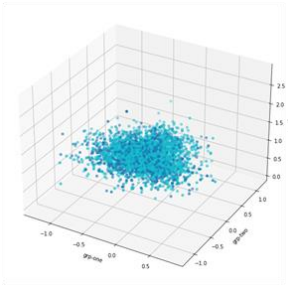
GRP's reduction of the UFC dataset to 40 dimensions with a slight error increase points to a potential loss of complex information vital for outcome prediction. Yet, for the Wine dataset, a reduction to 9 features with moderate error suggests the core class structure remains intact despite dimensionality reduction. This resilience in the Wine dataset could be due to the pronounced differences between classes being well-preserved by GRP's stochastic nature, which generalizes data structure without explicitly maintaining all pairwise distances. Random projections over multiple trials resulted in a Gaussian error distribution (Figure 5 a), likely influenced by the Central Limit Theorem, where aggregating numerous random variables (the projections) trends towards a Gaussian. Visualizations of the UFC dataset with principal components (Figure 5 b) fail to distinctly categorize winners, indicating GRP's limitations in refining features for predictive tasks. Conversely, the Wine dataset's visualization (Figure 5 c) showcases well-defined class clusters, suggesting that GRP, while simplifying the space, does not compromise the data's intrinsic categorical divisions, beneficial for clustering and classification efforts.

Lasso employs regularization to streamline feature selection by shrinking some regression coefficients to zero, effectively discarding less influential variables. It isolated 39 features for the UFC and 4 for the Wine dataset, reflecting its capacity to prioritize the most impactful predictors. This process doesn't aim to reconstruct the original dataset but to optimize prediction, striking a balance between error reduction and model simplicity. The predictive power of Lasso-selected features is pivotal, particularly their relevance to the target variable,

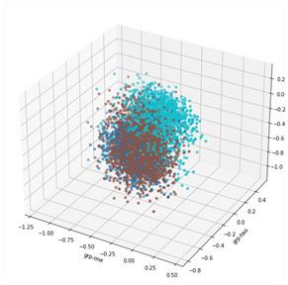
which sometimes enhances model performance. In the clustering analysis, these selected features will be scrutinized to determine if their predictive value translates to meaningful clusters. While a strong predictor-feature correlation is essential, it doesn't guarantee effective clustering, as good predictive features may not always delineate clear cluster boundaries. The subsequent clustering will validate the efficacy of Lasso's feature selection in capturing the underlying data structure necessary for robust clustering.



(a) UFC – GRP error distribution



(b) UFC – GRP first three PC



(c) Wine – GRP first three PC

Figure 5. Randomized Projection Reconstruction Error and Principal Component Distribution

Overall, noise impacts algorithms differently; PCA and ICA are sensitive to noise due to their reliance on data variance and independence, while GRP and Lasso may be more robust due to their probabilistic and penalization methods, respectively. The rank of the data, indicating the dimensionality of its feature space, is fully realized in the number of non-zero eigenvalues in PCA and components in ICA that significantly contribute to the variance or independence. Colinearity, the degree to which features correlate linearly, is reduced post-dimensionality reduction, with PCA aiming to eliminate it, while ICA seeks to capitalize on non-Gaussian distributions, and Lasso on predictive power. Specific properties, like the inherent structure or noise level of the datasets, influence algorithm outputs. For UFC data, the intrinsic dimensionality appears lower than the feature count, beneficial for PCA and ICA, while the class distinctions in the Wine dataset are well-preserved across methods. The reconstruction accuracy and robustness against noise vary by technique, with PCA and ICA offering nuanced insights into data variance and independence, GRP providing a probabilistic simplification, and Lasso targeting predictive relevance.

5 CLUSTERING ON REDUCED DIMENSION DATA SET

The clustering analysis on dimensionality-reduced datasets reveals nuanced insights. Table 2 summarizes the clustering analysis results for the 16 cases with the matching score to the original label. For the UFC dataset, PCA and Lasso, when combined with KMeans, significantly boosted matching scores by around 30%, indicating these methods excel at distilling the dataset to its most salient features for clustering. This enhanced matching score suggests not just the same clusters as before, but more refined ones that better align with the original labels. In the Wine dataset, PCA and ICA incrementally improved matching scores, implying their effectiveness in isolating essential clustering traits. However, the consistency of EM's performance across various models underscores its capacity to adapt to changes in the feature space, owing to its probabilistic approach that accounts for the likelihood of each data point's cluster membership.

Following the initial analysis, t-SNE was also employed as a metric to assess clustering performance. The visualization it provides complements the quantitative measures by demonstrating how dimensionality reduction techniques like ICA and Lasso, especially when used with KMeans, impact the clustering structure. In the case studies presented, ICA and Lasso's integration with KMeans serves as a

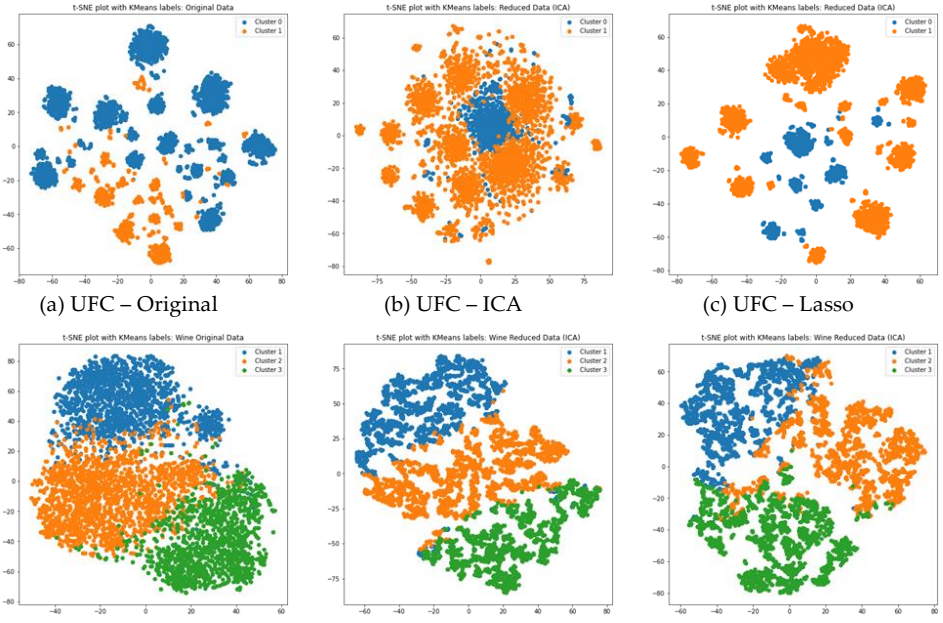
demonstrative result, where t-SNE visualizations offer additional insight into the clustering's effectiveness and the dimensional reduction's influence on the data's local structure.

t-SNE's role in the UFC dataset (Figure 6 a b c) is particularly interesting as it compacts the data under a single label, hinting that the independent components from ICA might not adequately represent the actual clusters due to the UFC dataset's complex structure. Conversely, Lasso's approach to feature selection retains and even clarifies the inherent clustering pattern within the UFC data, likely due to its focus on variables with strong predictive power. For the Wine dataset (Figure 6 d e f), ICA's effectiveness is underscored by its clear demarcation of the expected tripartite groupings, likely because the distinct factors separating wine varieties align well with the independent components ICA identifies. Lasso's performance suggests that even with a reduced feature set, the critical aspects that define the Wine dataset's classes are captured and emphasized, facilitating improved cluster definition.

In summary, the clusters post-dimensionality reduction are not identical to the original; they've been optimized. PCA and Lasso have sharpened the UFC dataset's clusters, aligning them more closely with inherent categories, whereas ICA and Lasso have solidified the Wine dataset's clusters, suggesting that the dimensionality reduction techniques are not just preserving but enhancing the data's intrinsic clustering characteristics. This enhancement may be due to the removal of noise and redundant information, allowing the algorithms to focus on the most defining features of the data.

Table 2. Summary of Clustering on Reduced Dimension Data Set (Matching Score with Original Label)

		Original	PCA	ICA	RP	Lasso
UFC	**Kmean**	0.429	0.571	0.604	0.403	0.597
	EM	0.663	0.663	0.663	0.663	0.663
Wine	**Kmean**	0.903	0.902	0.911	0.799	0.813
	EM	0.87	0.897	0.907	0.829	0.798



(a) UFC – Original (b) UFC – ICA (c) UFC – Lasso

(d) Wine – Original (e) Wine – ICA (f) Wine – Lasso

Figure 6. t-SNE Plots with KMeans Clustering with Reduced Dimension Data Set (ICA & Lasso)

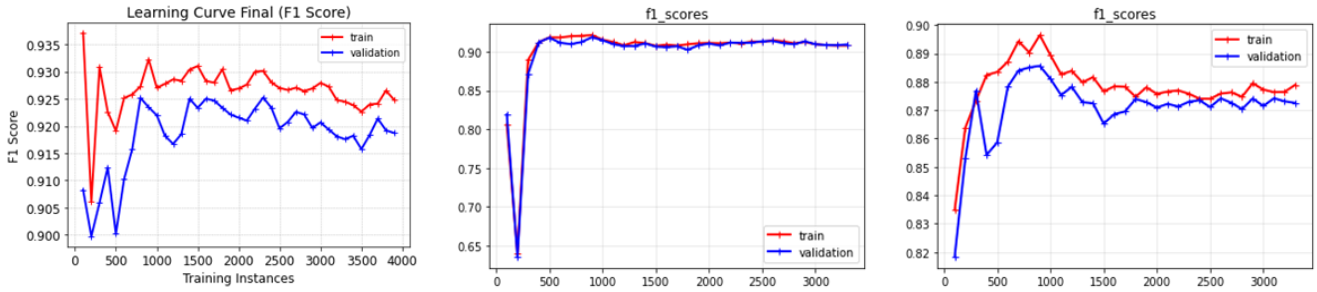
6 NUERUAL NETWORK WITH ICA AND LASSO REDUCED FEATURES (WINE DATA SET)

When re-running the neural network algorithms with reduced features from ICA and Lasso, differences in performance and computation time were observed. Figure 7 depicts the learning curve for the original, ICA, and Lasso reduced feature dataset. With fewer features to process, both ICA and Lasso led to faster training and prediction times, which is advantageous for real-time applications.

Specifically, ICA's reduction resulted in a diminished variance without notably impacting the bias. This outcome is consistent with ICA's ability to remove noise and irrelevant information, which likely contributed to a more consistent model performance across various runs and datasets. The reduced variance suggests that the network became more robust against overfitting, which is a desirable trait indicating that the model generalizes well to new data. Lasso's performance maintained a level comparable to the full dataset. The accuracy only modestly decreased from 0.92 to 0.88, which is impressive given the substantial reduction in feature set size. This points to Lasso's efficiency in selecting highly predictive features that retain most of the necessary information for the neural network to function effectively.

To evaluate the differences in performance, one would consider metrics like training speed, prediction time, accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC). Additionally, learning curves, like the ones attached for the original, ICA-reduced, and Lasso-selected feature datasets, provide visual insights into the model's learning process over time. From the learning curves, it can be deduced that the neural network with the original dataset may have higher accuracy but at the cost of longer training times and potential overfitting, as indicated by the larger gap between the training and validation scores. For the ICA and Lasso curves, the gap between the training and validation scores is narrower, particularly for Lasso, indicating a better generalization to unseen data. The Lasso curve suggests that while there is a slight compromise in maximum attainable accuracy, the trade-off for computational efficiency and model robustness could be well worth it, especially in time-sensitive applications.

In summary, both ICA and Lasso optimizations contribute to more efficient neural network operations with slightly different performance trade-offs: ICA for consistency and Lasso for preserving accuracy. These findings would be critical when deciding on the dimensionality reduction technique to apply, depending on the specific needs of the application, whether it's speed, accuracy, or a balance of both.



(a) NN on Wine Original (b) NN on Wine ICA Reduced (c) NN on Wine Lasso Selected

Figure 7. Neural Network Learning Curve on Wine Original, ICA Feature Reduced, Lasso Selected Feature Dataset

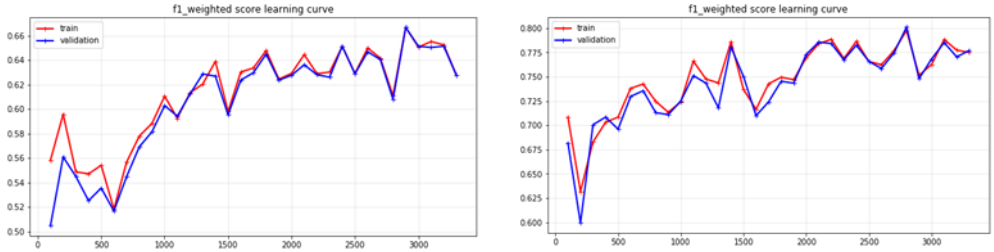
7 NUERUAL NETWORK WITH CLUSTERING LABELS (WINE DATA SET)

Incorporating KMeans and EM clustering labels as features offers a distilled feature space, potentially streamlining the neural network's training and leading to quicker convergence. These clustering labels, serving as a form of unsupervised feature engineering, seem to stabilize the feature set, evidenced by the consistent performance regardless of the clustering algorithm used. Adding the clustering labels to the feature set resulted in a marginal improvement in performance, indicating that while the labels do contribute useful information, they do not drastically alter the learning dynamics; the learning curves maintain their overall trajectory. This is consistent with a slight increase in computational time, likely due to the neural network processing the additional clustering label features.

An intriguing experiment was conducting neural network training exclusively with clustering labels, which still yielded acceptable accuracy levels, as shown in Figure 8. This suggests that the labels themselves encapsulate enough data structure to inform predictions. Training solely on clustering labels revealed a high bias but low variance scenario in the learning curves, indicating that while the model may not capture all the complexities of the data (high bias), its performance is consistent across different datasets (low variance). This high bias yet low variance phenomenon when using only clustering labels illuminates the clustering algorithms' effectiveness in extracting the datasets' fundamental characteristics. Such characteristics, when used alongside the original dataset, can enhance the neural network's predictive power by reinforcing the data's structural patterns.

In evaluating the performance differences upon re-running the neural networks, it's essential to consider not only accuracy but also how the model's learning curve changes with the integration of clustering labels. The shape of the learning curve can reveal much about the model's learning process — whether it's improving or plateauing over time. Furthermore, computational time is a critical factor, especially in operational environments where speed is of the essence.

The findings suggest that while clustering labels as standalone features can lead to a simpler, faster-converging model with acceptable accuracy, their true value is realized when combined with the original dataset. This combination can provide a robust feature set that balances complexity and performance, potentially offering a sweet spot for neural network training. The dual-layer approach, where clustering algorithms precede neural network classification, acts as a form of ensemble learning that can effectively capture and utilize complex patterns within the data for improved prediction outcomes.



(a) NN on Wine KMean Cluster (b) NN on Wine EM Cluster

Figure 8. Neural Network Learning Curve on Wine Dataset Cluster Label