1. Introduction

We investigate how data representation influences the structure uncovered by unsupervised learning, using PCA and KMeans on two datasets with different dimensionalities (784-dim MNIST and 20-dim COVID).
This study aims to answer three questions:

- How does PCA improve KMeans performance on high-dimensional data?
- Is PCA still effective for low-dimensional data?
- What does the performance gap between supervised and unsupervised learning reveal?

2. Methods

We conduct three sets of experiments:

1) High-dimensional data (MNIST, 784-dim)

- Raw features are redundant and noisy, which may limit KMeans performance.
- After applying PCA, we:
    - examine whether cluster separability improves,
    - identify the effective number of principal components.

2) Low-dimensional data (COVID, 20–30 dim)

- Raw features are already compact and semantic, so KMeans may perform reasonably well.
- After PCA, we:
    - evaluate whether separability improves or degrades,
    - determine the efficient dimensionality.

3) Comparison with supervised learning (SVM)

- Compare clustering results with supervised classification to examine the performance gap.
- Investigate whether PCA benefits unsupervised learning in a manner similar to supervised methods.

2.1 Datasets

- **MNIST**: 70k samples, 784-dimensional raw pixel features with 10 digit classes.
- **COVID**: 100k samples, 20-dimensional semantic clinical features (age, sex, diabetes, etc.), binary labels (Alive vs. Death).

2.2 Representation Choice

We evaluate several data representations:

- Raw features.
- PCA-2D (for visualization).
- PCA-1D (extreme low-dimensional baseline).
- PCA dimensions determined by explained-variance thresholds (80%, 90%, 95%, 99%).

To select meaningful PCA subspaces, we computed the number of components required to reach different explained-variance thresholds (80%, 90%, 95%, 99%).
Let the eigenvalues of the covariance matrix be $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$.

The cumulative explained variance (CEV) for the first d components is

$$\mathrm{CEV}(d) = \frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{D} \lambda_i}. \tag{1}$$

These values define the PCA dimensionalities evaluated in later experiments.

2.3 K-Means Setup

- **K = 10** for MNIST and **K = 2** for COVID.
- Multiple random initializations to ensure stable clustering.
- Evaluation Metrics

We evaluate clustering performance using five standard metrics.
For each metric, we include both the formal definition and explanations of all variables.

Accuracy (ACC, Hungarian-matched)
ACC measures the proportion of samples whose predicted cluster label can be aligned with the correct class label.
Since cluster IDs are permutation-invariant, we use the Hungarian algorithm to find the optimal matching.

Let

- C_k: the set of samples assigned to cluster k
- L_j: the set of samples belonging to ground-truth class j
- n: total number of samples

Then,

$$\mathrm{ACC} = \frac{1}{n} \sum_{k=1}^{K} \max_{j} |C_k \cap L_j|. \tag{2}$$

Normalized Mutual Information (NMI)
NMI measures how much information is shared between predicted clusters and true class labels.

Let

- U: clustering assignment
- V: ground-truth labels
- $I(U, V)$: mutual information between U and V
- $H(U)$, $H(V)$: entropies of U and V

Then,

$$\mathrm{NMI}(U, V) = \frac{2I(U, V)}{H(U) + H(V)}. \tag{3}$$

A value of 1 indicates perfect agreement; 0 indicates independence.

Adjusted Rand Index (ARI)
ARI evaluates pairwise consistency between clustering assignments and ground truth, corrected for chance.

Let

- RI: Rand Index, the proportion of sample pairs on which clustering and labels agree
- $\mathbb{E}[RI]$: expected Rand Index under random labeling

Then,

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[RI]}{\max(\mathrm{RI}) - \mathbb{E}[RI]}. \tag{4}$$

ARI ranges from 0 (random clustering) to 1 (perfect pairwise consistency).

Inertia (Within-Cluster Sum of Squares)
Inertia measures cluster compactness and is the objective minimized by K-means.

Let

- x_i: sample i
- c_{z_i}: centroid of the cluster to which i belongs
- z_i: cluster assignment of sample i

Then,

$$\mathrm{Inertia} = \sum_{i=1}^{n} \|x_i - c_{z_i}\|^2. \tag{5}$$

Lower inertia indicates tighter clusters but is not directly comparable across dimensionalities.

Together, ACC, NMI, and ARI evaluate agreement with ground-truth class structure, while Inertia measures the geometric quality of clusters independently of labels.

2.4 Supervised Baseline (SVM)

- Linear SVM trained on both raw data and PCA-transformed features.
- Serves as a reference to quantify the performance gap between supervised and unsupervised learning.

3. Experiments

3.1 Data Preprocessing

MNIST:
Each 28×28 image was flattened into a 784-dimensional vector and normalized to [0, 1]. No additional preprocessing was applied.

COVID-19 Clinical Dataset:
Samples with missing outcome labels were removed. Continuous features were standardized, and categorical variables were one-hot encoded. The final input contained 20 clinical attributes, with the binary label (Alive=0, Death=1) derived from the DATE_DIED field.

3.2 Principle Component Analysis of Data Representations

3.2.1 PCA on MNIST

The cumulative explained variance curve (Fig. 1) provides the number of components required to preserve different proportions of the total variance. Based on the thresholds of 80%, 90%, 95%, and 99%, the corresponding PCA dimensionalities were:

- **150**, **238**, **332**, and **544** components, respectively.

To cover both extreme and mid-range representations, we additionally included:

- **1, 2, 3 dims** (strong compression and visualization),
- **20, 50 dims** (typical mid-level subspaces),
- **784 dims** (raw space).

Thus, the full set of PCA dimensions evaluated is:

$$d \in \{1,\ 2,\ 3,\ 20,\ 50,\ 150,\ 238,\ 332,\ 544,\ 784\}. \tag{6}$$

This range allows us to analyze clustering behavior across representations from highly compressed to high-variance subspaces and the original feature space.

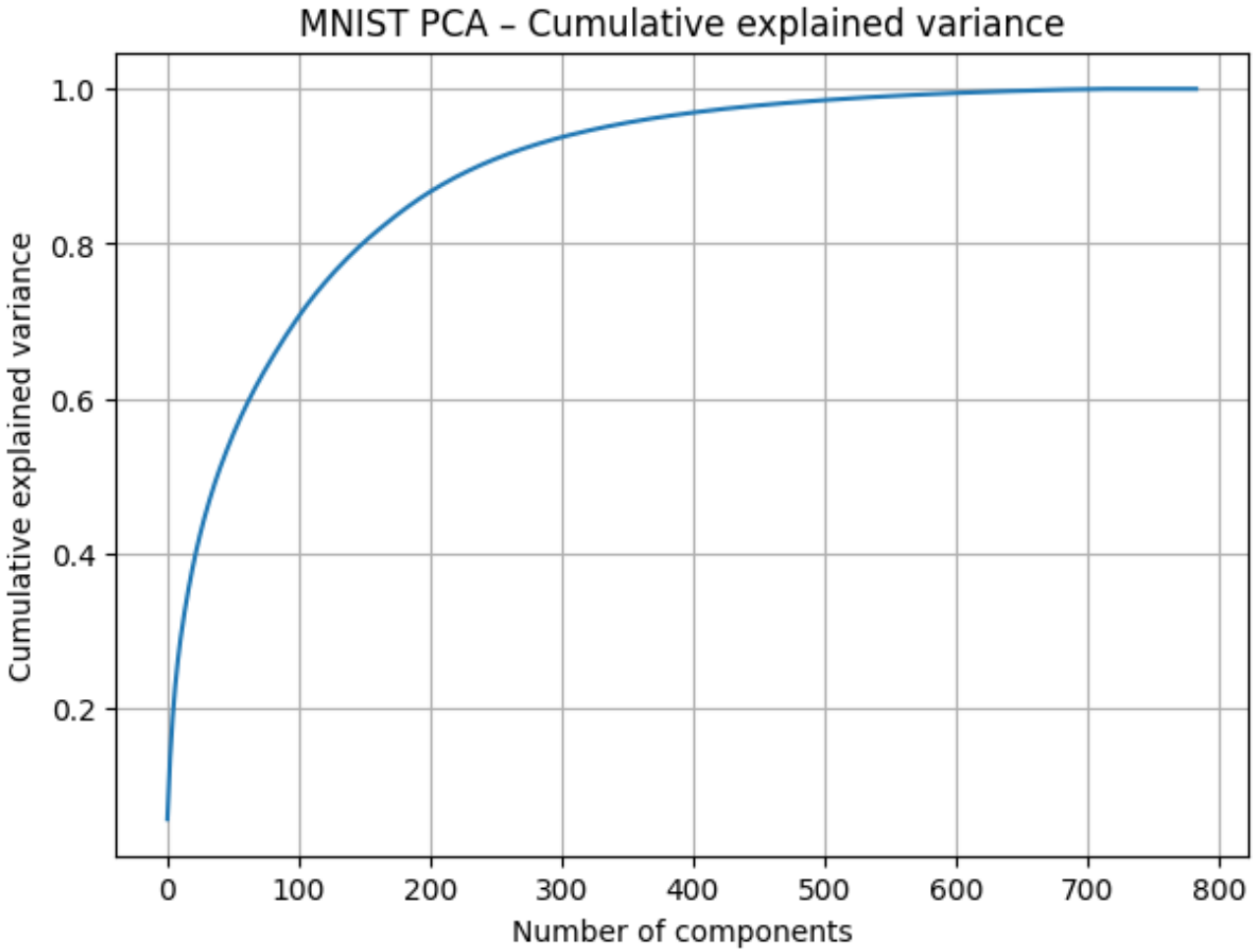

Figure 1. Cumulative explained variance curve of MNIST PCA components.
This curve shows how much variance is retained as the number of principal components increases.
The thresholds used in our experiments—80%, 90%, 95%, and 99%—correspond to approximately 150, 238, 332, and 544 components, respectively.

3.2.2 PCA on COVID-19 Clinical Data

PCA was applied to the standardized COVID-19 feature matrix to examine whether linear compression reveals additional clustering structure in this low-dimensional semantic dataset. The cumulative explained variance curve (Fig. 2) rises more gradually than in MNIST, reflecting the more evenly distributed information across features.

Based on the variance thresholds of 80%, 90%, 95%, and 99%, the corresponding PCA dimensionalities were:

- **5**, **7**, **9**, and **13** components, respectively.

To enable structured comparisons across representation granularities, we selected the following PCA dimensions:

$$d \in \{1,\ 2,\ 3,\ 5,\ 7,\ 9,\ 13,\ 20\}. \tag{7}$$

Here, 1–3 dimensions provide strong compression and visualization, 5–13 correspond to variance-based thresholds, and 20 approximates the original 20-dimensional space. These embeddings are used consistently in the clustering experiments (Section 3.4) and supervised baselines (Section 3.5).

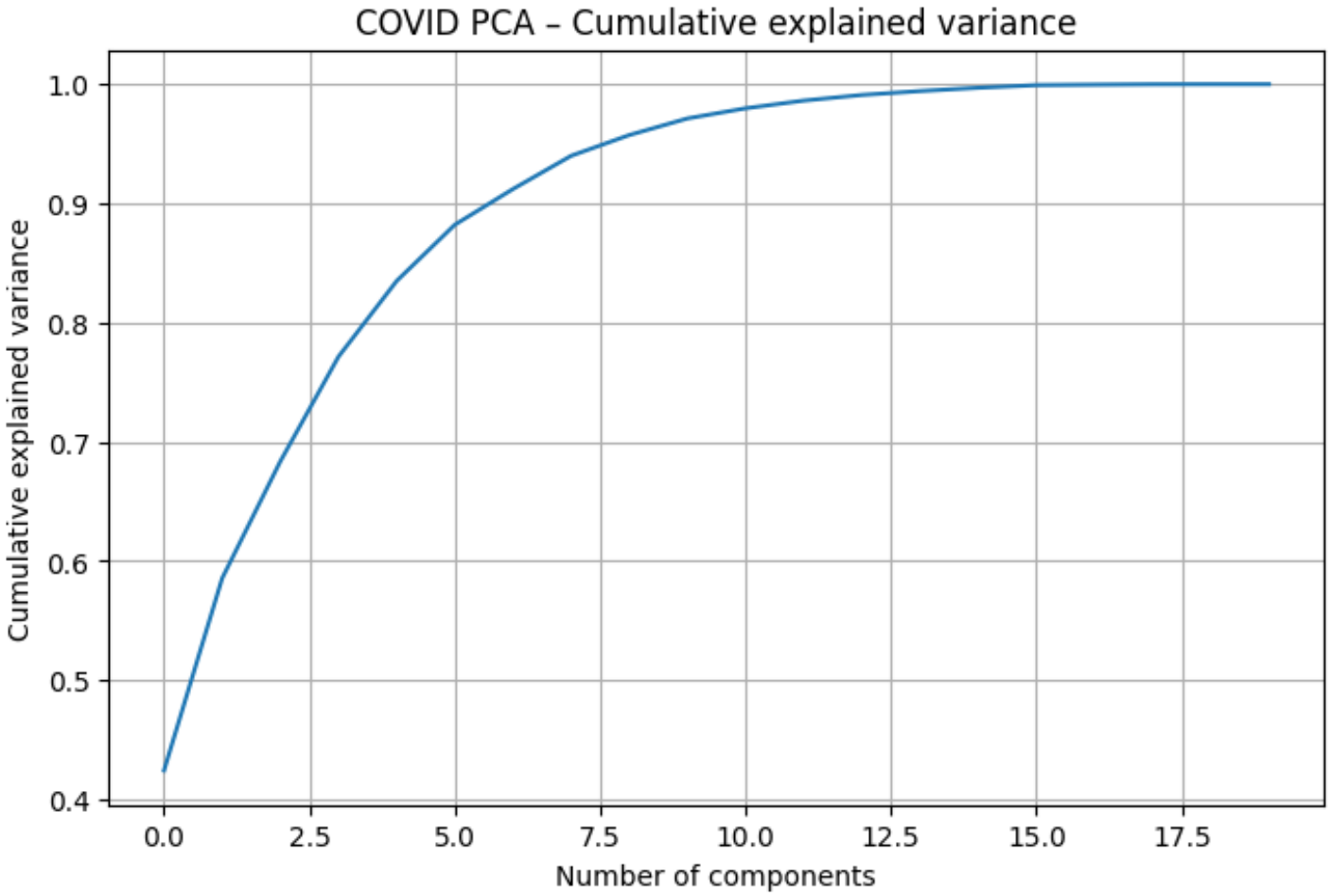

Figure 2. Cumulative explained variance curve of COVID-19 PCA components.
The curve shows that variance accumulates more gradually compared to MNIST, indicating lower redundancy in the clinical features.

3.3 K-means on Data Representations

3.3.1 K-means on MNIST

We evaluate K-means clustering on both the raw 784-dimensional pixel vectors and PCA-reduced embeddings. The complete results are provided in Table 1, covering all PCA dimensionalities.

Clustering on Raw Pixel Space

K-means on the raw space yields moderate performance (ACC = 0.4876, NMI = 0.4087, ARI = 0.2938).
These values indicate that although some alignment with digit labels exists, the raw pixel space does not naturally separate digits into distinct clusters.
This aligns with the well-known challenge of MNIST: intra-class handwriting variability often exceeds inter-class separability under Euclidean distance.

Effect of PCA Dimensionality

Low-dimensional PCA (1–3 dims).
Extreme compression severely degrades clustering performance (ACC ≤ 0.37, NMI ≤ 0.32, ARI ≤ 0.22).
These low-dimensional embeddings retain only coarse global variance and discard most digit-specific structure.

Intermediate PCA dimensions (20–50 dims).
This range produces the strongest clustering results (ACC ≈ 0.50, NMI ≈ 0.43, ARI ≈ 0.32).
These dimensions preserve essential digit-discriminative variance while reducing pixel-level noise, making them the most effective representations for K-means.

High-dimensional PCA (150–544 dims).
Beyond approximately 150 components, ACC, NMI, and ARI plateau and no longer improve.
This reflects diminishing returns once most variance has been captured, and adding additional components does not enhance cluster separability.

Overall.
PCA improves K-means performance relative to the raw pixel space, with the 20–50 dimensional range performing the best.
However, even the best PCA configurations achieve only moderate clustering quality, confirming that MNIST digits do not form well-separated Euclidean clusters.

2D PCA Visualization

Fig. 3 shows the MNIST PCA-2D projection, colored by ground-truth labels (left) and K-means assignments (right).
The ground-truth plot reveals substantial overlap among digit classes, forming intertwined manifolds with no clear boundaries.
The K-means partition instead aligns with global variance directions and produces regions that do not correspond to true digit categories.
These visualizations further demonstrate MNIST's weak geometric separability.

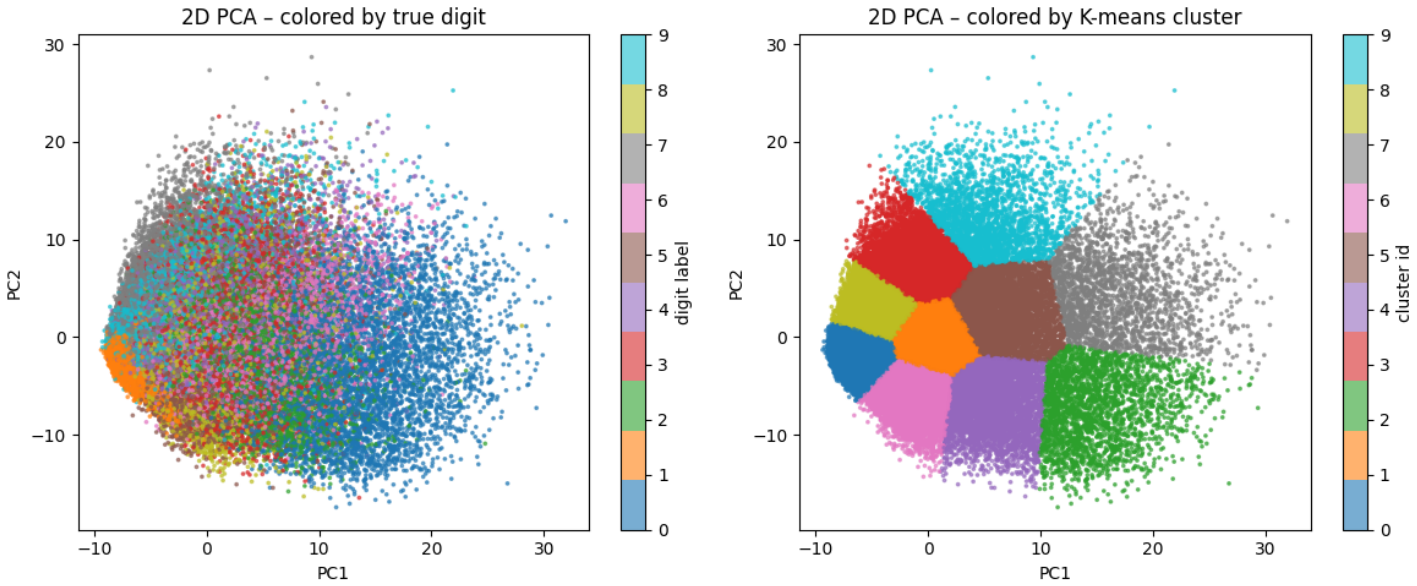

Figure 3. MNIST projected onto the first two principal components.
Left: colored by true digit labels.
Right: colored by K-means cluster assignments.

Table 1. K-means clustering performance on MNIST under different PCA dimensionalities.

PCA Dim	ACC	NMI	ARI	Inertia / point
1	0.2599	0.1949	0.1192	0.6739
2	0.2991	0.2141	0.1415	9.3170
3	0.3711	0.3201	0.2166	22.3625
20	**0.4959**	**0.4266**	**0.3171**	164.8088
50	**0.5000**	**0.4321**	**0.3227**	284.5029
150	0.4956	0.4314	0.3219	463.3275
238	0.5026	0.4346	0.3241	540.3907
332	**0.5026**	**0.4347**	**0.3242**	571.8054
544	0.5026	0.4345	0.3240	601.6856
784	**0.4876**	**0.4087**	**0.2938**	**608.1520**

3.3.2 K-means on COVID

We next evaluate K-means clustering on the COVID-19 clinical dataset. Unlike MNIST, this dataset is low-dimensional and composed of semantically meaningful attributes, making it a useful contrastive case. The full results across PCA dimensions are summarized in Table 2.

Clustering Across PCA Dimensions

Stable but limited accuracy.
ACC remains between 0.66 and 0.69 across all PCA dimensions, indicating weak cluster structure with respect to the Alive/Death outcome. Increasing dimensionality does not yield improvements, which stands in contrast to the behavior observed in MNIST.

Extremely low NMI and ARI.
NMI (≈0.10) and ARI (≈0.09) remain consistently near zero, demonstrating that the cluster assignments share very little information with the true labels. The two outcome groups are highly intertwined in the feature space, making them unsuitable for unsupervised separation.

Inertia increases with dimensionality.
Inertia grows monotonically (0.38 → 1.85) as PCA dimensionality increases, reflecting the geometric expansion of embedding space. However, this increase does not correspond to improved clustering quality.

Minimal effect of PCA.
Overall, PCA has almost no influence on clustering behavior. None of the PCA configurations result in meaningful performance differences, indicating that the Alive vs. Death labels do not correspond to natural Euclidean clusters.

2D PCA Visualization

Fig. 4 shows the COVID dataset projected onto the first two principal components.
The left panel shows substantial overlap between the Alive and Death classes, with no discernible boundary. The right panel illustrates that K-means imposes a linear partition unrelated to the underlying outcomes. These visual patterns explain the consistently low NMI and ARI values: although K-means produces a deterministic split, the data itself lacks geometric separability.

Overall, K-means achieves only limited performance on this dataset. PCA dimensionality does not meaningfully alter clustering behavior, underscoring the fundamental difficulty of applying unsupervised clustering to low-dimensional clinical outcome prediction.

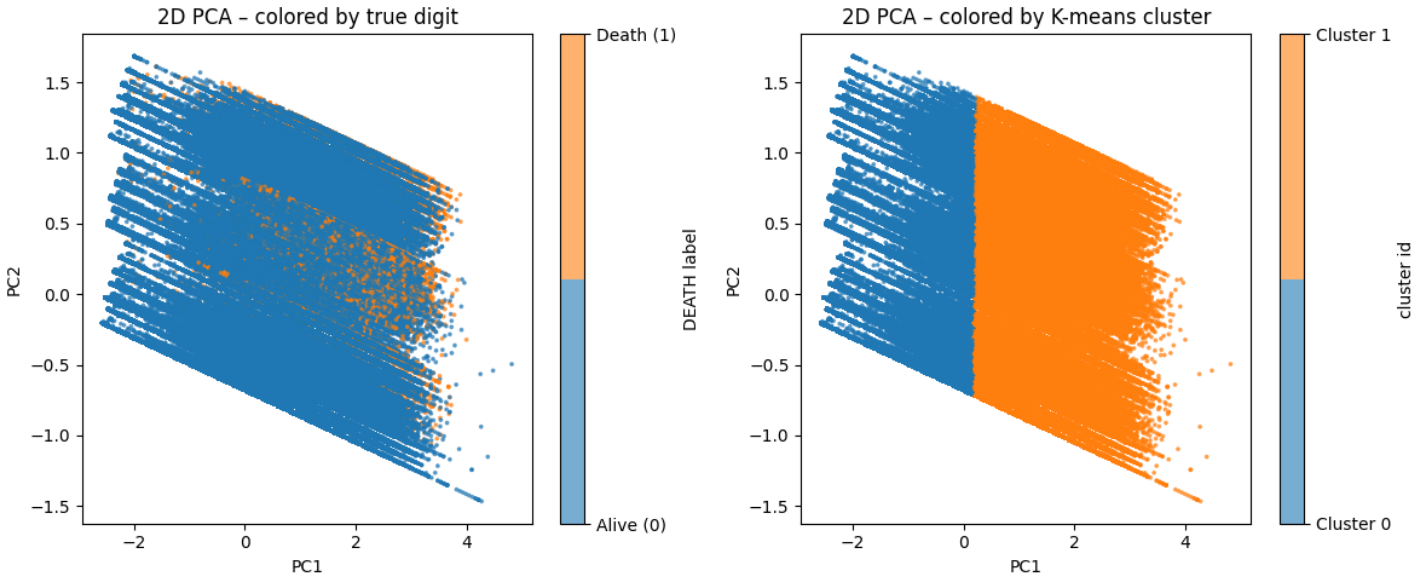

Figure 4. COVID-19 dataset projected onto the first two principal components.
Left: true outcomes.
Right: K-means cluster assignments.

Table 2. K-means clustering performance on the COVID dataset under different PCA dimensionalities.

PCA Dim	ACC	NMI	ARI	Inertia / point
1	0.6793	0.1057	0.0908	0.3802
2	0.6690	0.1018	0.0825	0.7925
3	**0.6837**	0.1067	0.0943	1.0438
5	0.6691	0.1028	0.0828	1.4307
7	0.6824	0.1076	0.0935	1.6282
9	0.6828	0.1078	0.0939	1.7441
13	0.6823	0.1076	0.0934	1.8303
20	**0.6691**	**0.1028**	**0.0828**	**1.8533**

3.4 Comparison Between MNIST and COVID Clustering

To understand how data characteristics influence the behavior of K-means, we compare clustering outcomes between MNIST and the COVID clinical dataset. Although both experiments apply the same pipeline—PCA dimensionality reduction followed by K-means clustering—their results differ sharply, revealing key differences in structural complexity and clusterability.

Differences in Data Structure

MNIST is a high-dimensional image dataset (784-D) with strong nonlinear structure. Although digit classes overlap heavily in pixel space, they lie on curved submanifolds shaped by writing styles and stroke patterns. These manifolds preserve class-related variance that PCA can partially uncover.

In contrast, the COVID dataset contains only 20 low-dimensional tabular clinical features. These features exhibit substantial overlap between Alive and Death outcomes and do not form separable manifolds or clusters. As a result, the dataset lacks the geometric structure required for effective unsupervised clustering.

Effect of PCA Dimensionality

MNIST.
Clustering performance improves noticeably when PCA reduces the 784-D pixel space to the 20–50 dimensional range.
ACC increases to ≈0.50, and NMI/ARI also rise, indicating that PCA retains discriminative structure while suppressing pixel-level noise.

COVID.
PCA dimensionality has almost no effect:
ACC stays in the narrow range of 0.66–0.69 across all PCA dimensions, and NMI/ARI remain extremely low (≈0.10 and ≈0.09).
This stability suggests that no PCA subspace—even extremely compressed or lightly compressed ones—reveals more separable structure.

Geometric Separability

PCA visualizations clearly illustrate the difference:

- In **MNIST**, the PCA-2D projection reveals broad but somewhat organized regions corresponding to different digits. Although the classes overlap significantly, there is still partial structure.
- In **COVID**, the Alive and Death samples almost completely overlap in PCA space. K-means imposes a partition, but the resulting boundary bears no relationship to the true labels.

These visual findings are consistent with the quantitative results: MNIST contains weak but recoverable cluster structure, while COVID contains almost none.

Interpretation

The comparison highlights a fundamental contrast:

1. **MNIST is weakly clusterable.**
   Its high-dimensional manifold structure contains digit-related variance that PCA can capture, enabling moderate clustering recovery.
2. **COVID is essentially unclusterable.**
   The Alive/Death label does not correspond to any geometric separation in the feature space, making K-

means ineffective regardless of PCA representation.

This analysis emphasizes that the success of unsupervised clustering depends heavily on the inherent geometry of the data and that PCA benefits only datasets with meaningful underlying structure.

4. Supervised Baseline

4.1 Supervised Baseline on MNIST

To evaluate the linear separability of MNIST digit images, we train a linear SVM classifier on the same PCA representations used in the clustering experiments. The classifier is implemented using an SGD-based linear SVM (hinge loss) without additional regularization tuning, ensuring a fair and direct comparison with the unsupervised results.

Table. 3 reports the test accuracy across all PCA dimensions:

PCA Dimension d	Test Accuracy
1	0.2037
2	0.3306
3	0.4259
20	0.8027
50	0.8421
150	0.8783
238	0.8961
332	**0.8991**
544	0.8917
784 (raw)	0.8925

Several observations emerge:

1. Supervised learning dramatically outperforms K-means.
Even at very low dimensions ($d \leq 3$), linear SVM achieves 0.20–0.43 accuracy—already comparable to or better than K-means on high-dimensional representations. This highlights the strong role of supervision in revealing discriminative structure.

2. PCA enables highly compact yet discriminative representations.
Accuracy rises above 0.80 with only 20 principal components and exceeds 0.89 with 150–332 components. The best performance occurs at $d \approx 332$ (0.8991), slightly higher than the raw 784-D feature baseline (0.8925).

3. MNIST is highly linearly separable under supervision.
While K-means struggles to recover cluster structure (Section 3.3), linear SVM trained with labels achieves near-90% accuracy, showing that discriminative information exists but does not manifest as clear Euclidean clusters.

These results demonstrate that MNIST contains strong class-discriminative variance that PCA preserves, and that supervised linear models can fully exploit this structure despite the weak clusterability observed in unsupervised settings.

4.2 Supervised Baseline on COVID

We evaluate a linear SVM classifier on the COVID clinical dataset using the same PCA representations as in the clustering experiments. This allows us to assess the linear separability of the Alive/Death labels and compare supervised performance against the K-means results in Section 3.3.2.

Table. 4 reports the SVM test accuracy across PCA dimensions:

PCA Dimension d	Test Accuracy
1	0.7922
2	0.8313
3	0.8703
5	**0.8871**
7	0.8870
9	0.8741
13	0.8741
20	0.8741

Several patterns emerge:

1. Rapid improvement in low dimensions.
Accuracy increases sharply from 1 to 5 components, peaking at **0.8871**. This suggests that most discriminative variance lies within the first few principal components.

2. Performance plateaus for $d \geq 7$.
Beyond 7 dimensions, accuracy remains stable (0.87–0.89), indicating limited additional discriminative structure in higher PCA components.

3. Supervised learning substantially outperforms K-means.
Compared to the K-means accuracy of 0.66–0.69 (Section 3.3.2), linear SVM achieves far higher performance, demonstrating that the Alive/Death labels possess mild linear separability even though they do not correspond to natural geometric clusters.

Overall, these results align with the clustering analysis: the COVID dataset contains **weak but nonzero** discriminative structure that can be exploited by supervised learning, but the features do not form well-separated groups in any PCA subspace. As a result, supervised accuracy remains moderate (<0.90) and PCA has only limited influence on performance.

5. Discussion

The results across MNIST and the COVID clinical dataset highlight how data geometry fundamentally determines the effectiveness of PCA, K-means, and supervised linear models.

PCA and Data Structure

PCA benefits MNIST because its 784-dimensional pixel space is noisy and highly redundant. Individual pixels carry little meaning, and digit information emerges only through correlated patterns across many dimensions. PCA isolates these coherent variance directions while suppressing pixel-level noise, which explains the improved K-means performance in the 20–50 dimensional range.

In contrast, the COVID dataset consists of 20 semantic clinical variables (e.g., age, comorbidities) that already exist in a low-dimensional, weakly correlated space. The data does not form a manifold-like structure with redundant variance, leaving little for PCA to uncover. As a result, dimensionality reduction neither clarifies structure nor improves clustering.

Fig. 5 illustrates the first ten PCA components of MNIST, each resembling digit-like stroke patterns. These eigenvectors confirm that MNIST contains meaningful variance structure that aligns with semantically relevant directions—precisely the setting where PCA enhances clustering.

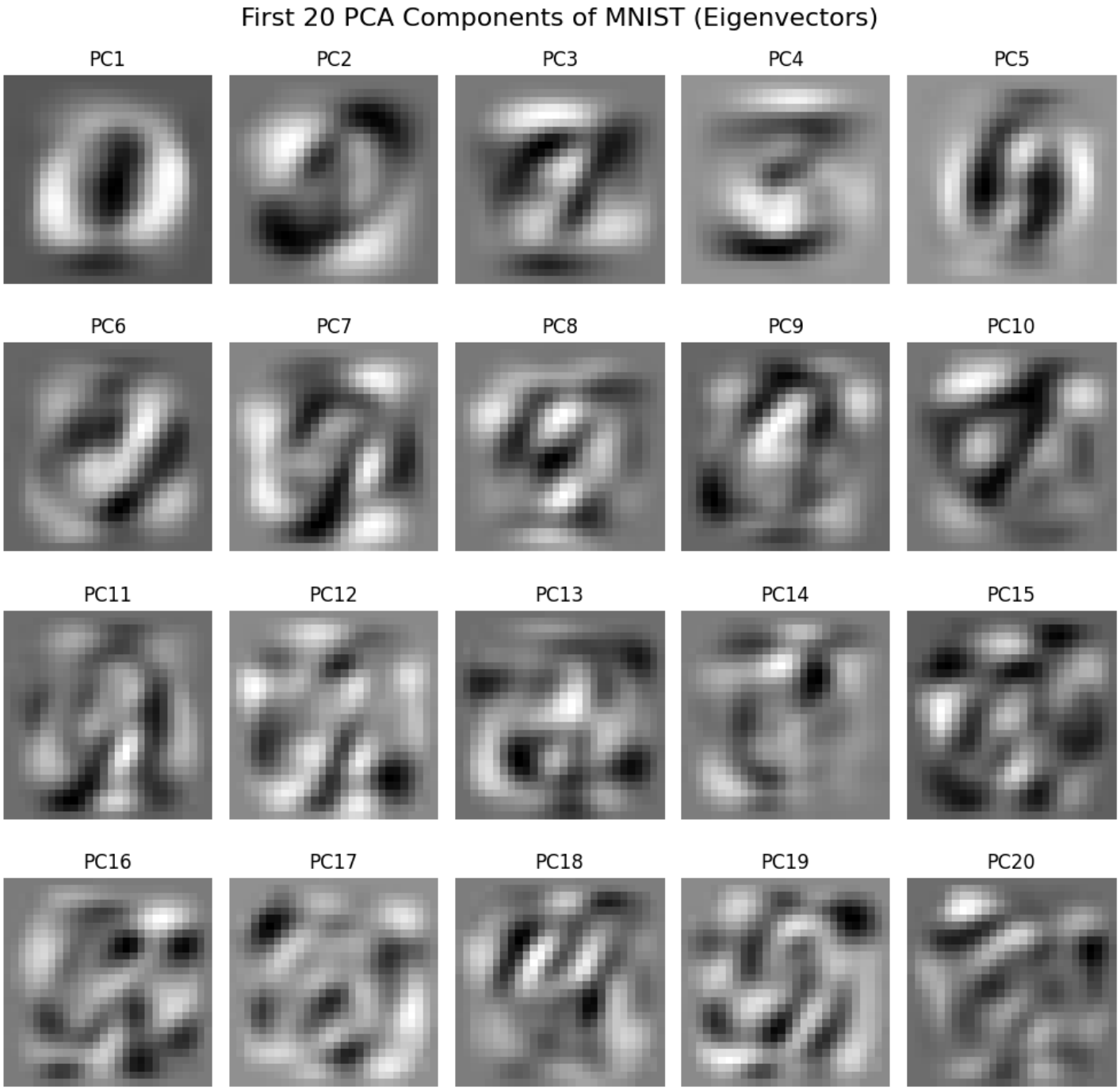

Figure 5. First ten PCA components of MNIST, visualized as 28×28 patterns.
The components capture digit-like strokes, demonstrating that PCA extracts meaningful low-frequency structure rather than noise.

Clusterability

MNIST is **weakly clusterable**: although digit classes overlap, they exhibit partial manifold organization, allowing K-means to achieve moderate alignment with ground-truth labels.
COVID is **essentially unclusterable**: Alive and Death samples occupy nearly identical regions in all PCA subspaces, and K-means boundaries do not correspond to clinical outcomes.

Supervised vs. Unsupervised Performance

Linear SVM outperforms K-means on both datasets.
On MNIST, SVM reaches ~90% accuracy, revealing strong supervised separability that does not manifest as geometric clusters—consistent with the low silhouette scores.
On COVID, SVM reaches ~88%, but this performance largely reflects class imbalance and the presence of a few highly predictive clinical variables rather than a coherent geometric boundary. This aligns with the near-zero NMI and ARI values observed in clustering.

Overall Insight

These results show that:

1. **PCA improves clustering only when meaningful variance structure exists.**
2. **Clustering succeeds only when class labels align with geometric separability.**
3. **Supervised models can exploit discriminative directions that are absent from unsupervised geometry.**

Ultimately, dataset structure—rather than algorithm choice—most strongly determines the effectiveness of PCA and the viability of unsupervised clustering.

6. Conclusion

This study examined how PCA dimensionality reduction influences K-means clustering across two contrasting datasets—MNIST and a COVID clinical dataset—and compared these results with linear SVM baselines to assess intrinsic separability.

For MNIST, PCA enhances clustering performance by suppressing pixel-level noise and exposing low-frequency, digit-related structure. Although K-means reaches only moderate accuracy, the linear SVM achieves nearly 90%, indicating that MNIST is highly separable under supervision but does not form distinct Euclidean clusters.

For the COVID dataset, the linear SVM achieves reasonably high accuracy (~0.88), but this performance is largely driven by class imbalance and a few highly predictive clinical variables rather than a meaningful geometric separation. This is consistent with the near-zero NMI and ARI values and the uniformly poor K-means results, indicating that the Alive/Death labels do not form separable clusters in feature space.

Overall, our findings highlight three key points:

1. **PCA improves clustering only when its principal components align with class-discriminative variance.**
2. **Clustering performance is governed more by the data's intrinsic geometry than by dimensionality choices.**
3. **Supervised baselines are essential for interpreting clustering results**, revealing whether limitations arise from the algorithm or from the data itself.

In summary, effective unsupervised learning depends critically on the presence of naturally separable structures in the data. When such structure is absent—as in the COVID dataset—neither PCA nor clustering algorithms can compensate for the lack of intrinsic separability.