VERGLEICH VON DATENGETRIEBENEN CLUSTERINGMETHODEN IM CIFAR10- UND MNIST-DATENSATZ

von Emre Kaplan, Bünyamin Budak und Maxim Speczyk

Motivation

1 ,5 BILLIONEN DOLLAR
VERLUST DURCH STILLSTÄNDE

2,3 MILLIONEN DOLLER
PRO STUNDE KOSTET STILLSTAND

EXTREME MENGE AN DATEN DIE
DIES VERHINDERN KÖNNTEN

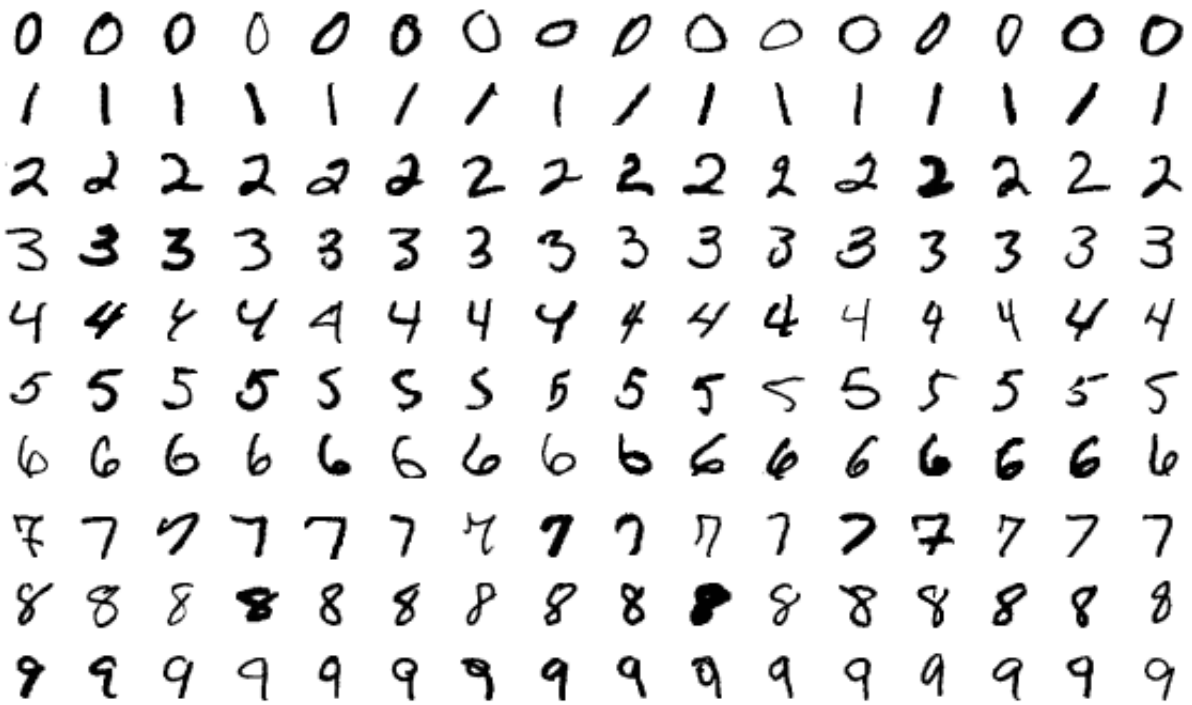
Motivation & Zielsetzung

Supervised Learing reicht nicht aus

Wir benötigen Unsupervised Learning

Wie gut sind moderne Clustering-Methoden?

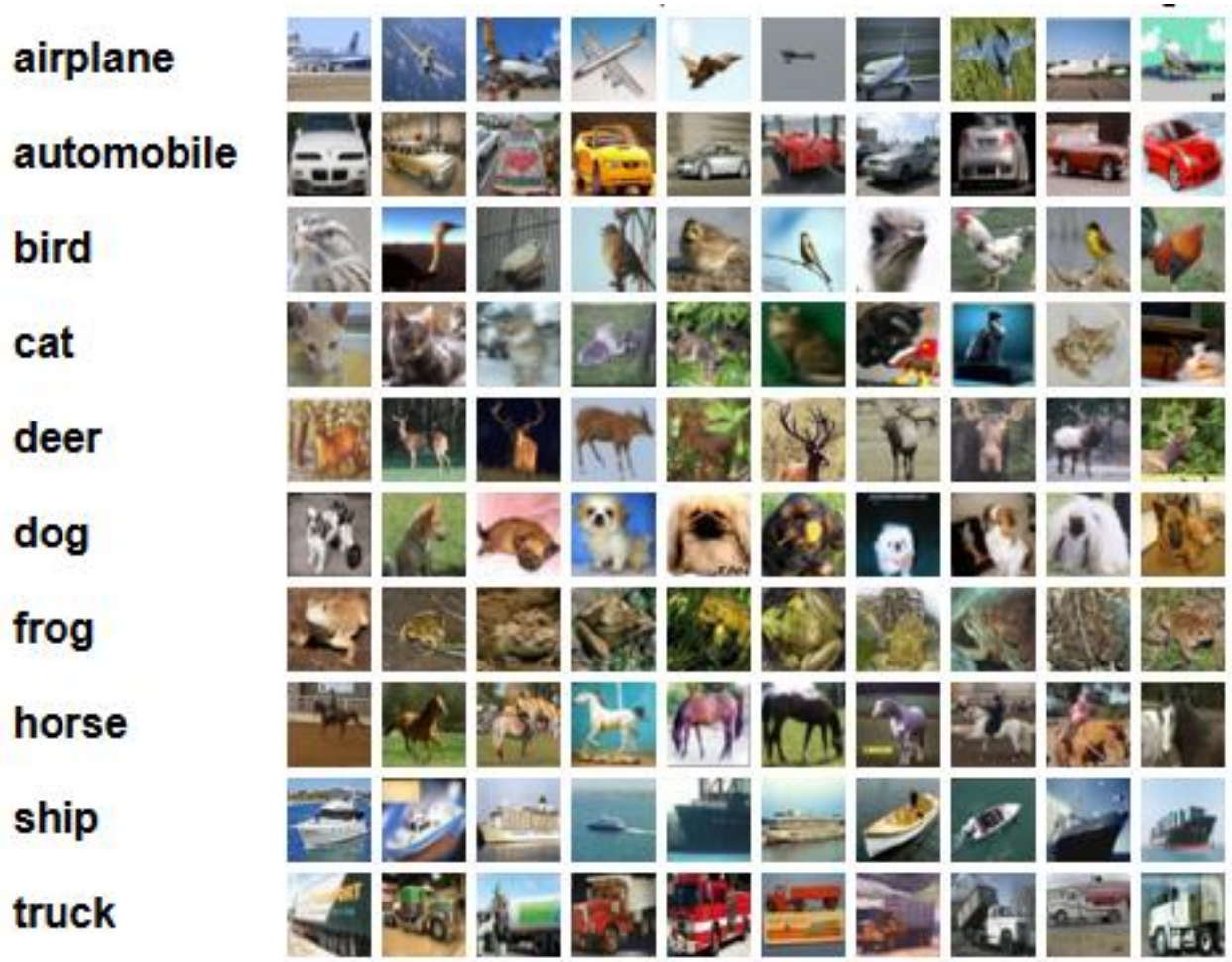
Datensätze - MNIST



https://de.wikipedia.org/wiki/MNIST-Datenbank#/media/Datei:MnistExamples.png

- Handgeschrieben Ziffern
- 28 x 28 Pixel
- Graustufen

-> Geringe Komplexität

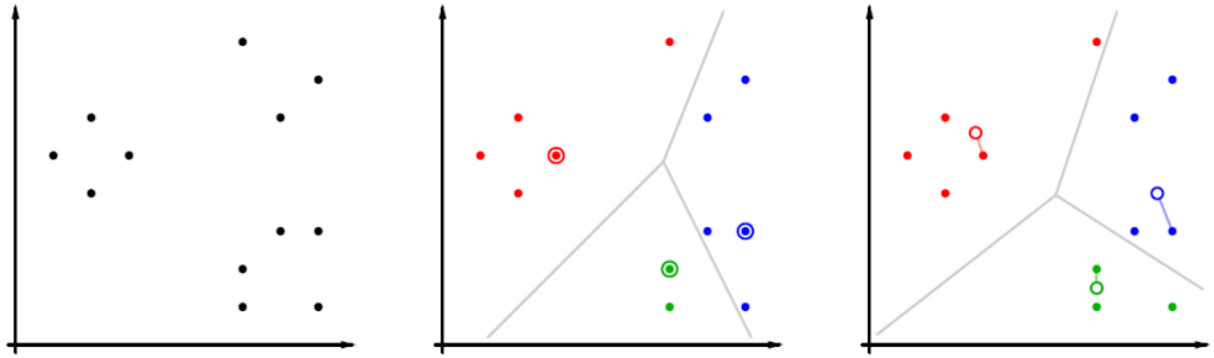


https://www.cs.toronto.edu/~kriz/cifar.html

Datensätze – CIFAR10

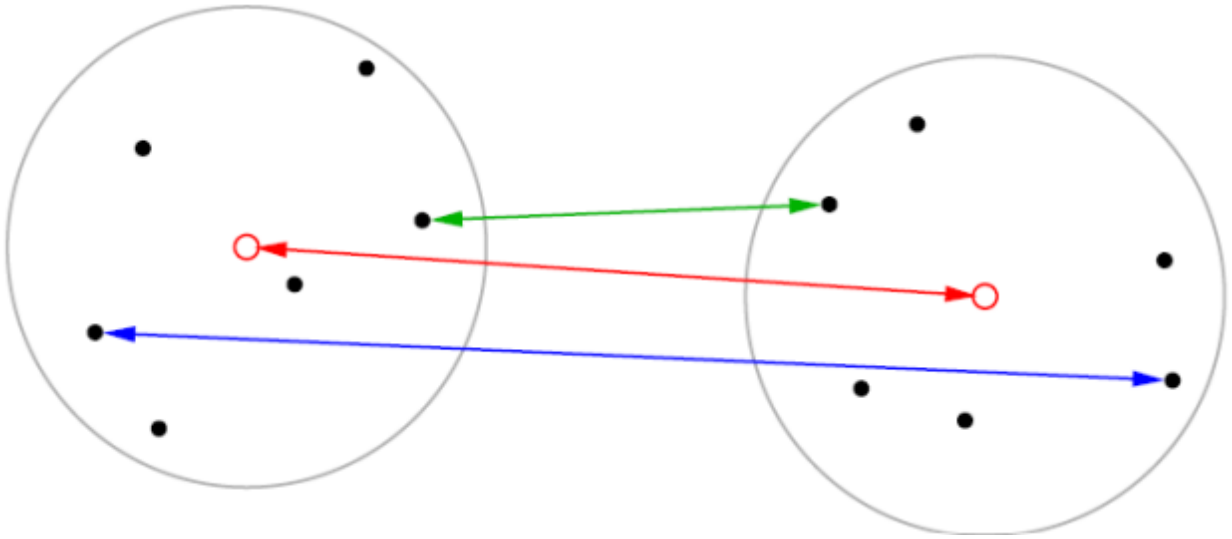
- Realistische Objekte
- 32 x 32 Pixel
- RGB

-> Hohe Komplexität & Varianz



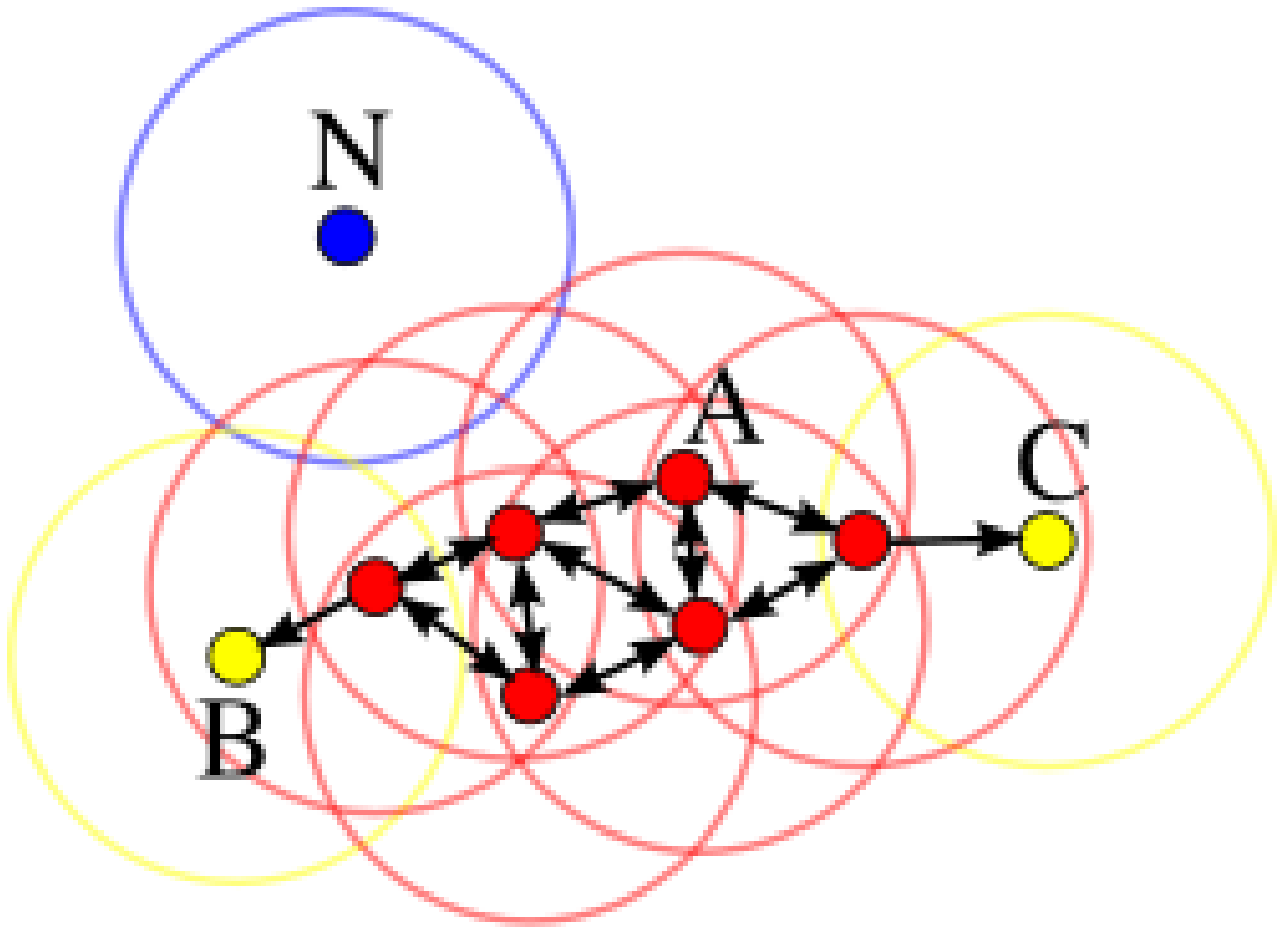
K-means - Clustering Verfahren

- iteratives, zentroid-basiertes Verfahren.
- Minimierung der quadratischen Abstände zum Cluster-Mittelpunkt.
- Wechselspiel: Datenpunkt-Zuweisung ↔ Update der Zentren.
- Vorteile: Effizienz & Einfachheit
- Nachteile: Ergebnisse hängen stark vom Startpunkt ab, ***k*** muss festgelegt sein



Hierarchisch - Clustering Verfahren

- Bottom-Up-Ansatz: Jeder Punkt startet als eigenes Cluster.
- Schrittweise Verschmelzung der ähnlichsten Cluster.
- Vorteil: Deterministisch (liefert immer das gleiche Ergebnis, kein Zufall).
- Nachteil: Hoher Rechenaufwand

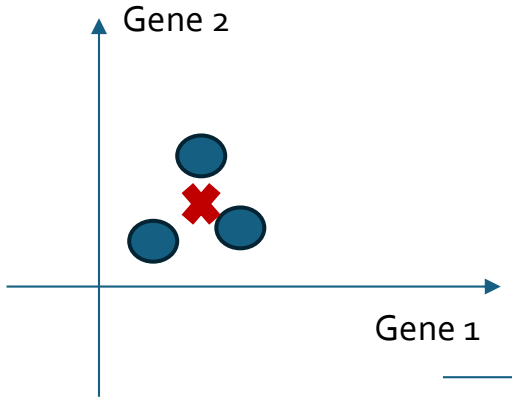


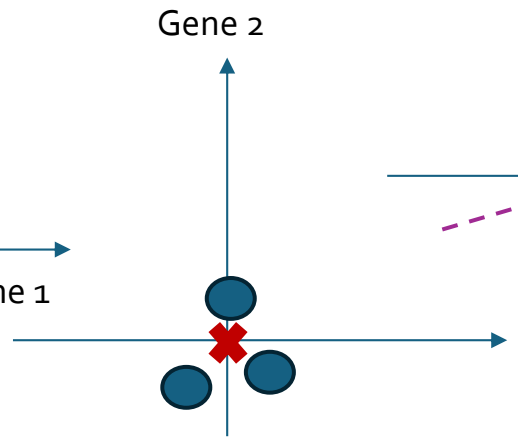
Dichtebasiertes Clustering - Clustering Verfahren

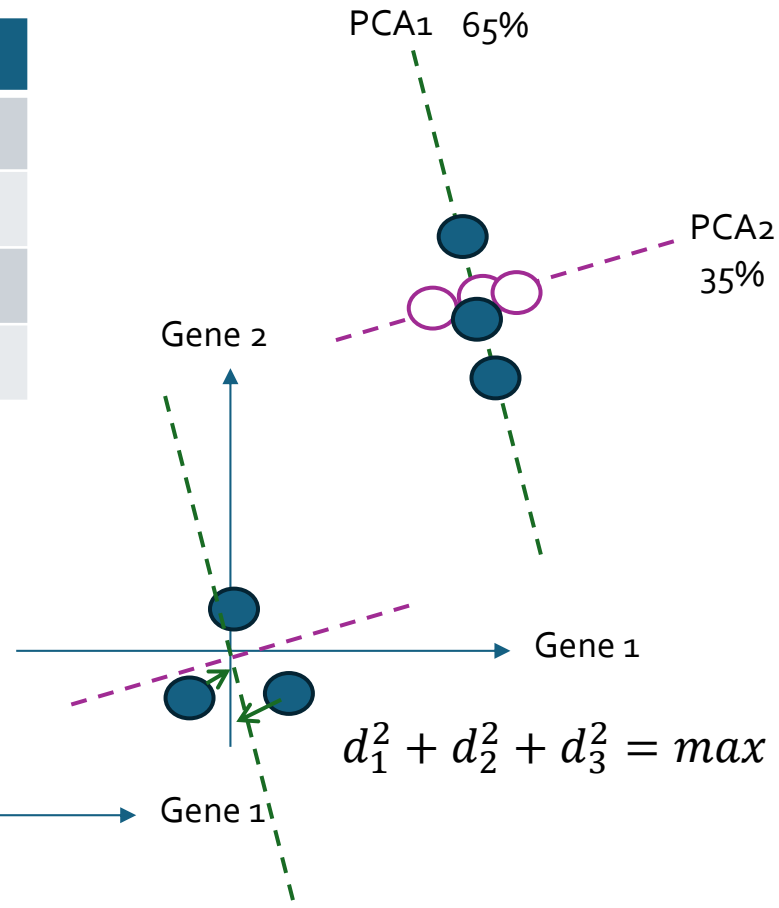
- Definition von Cluster durch Punktdichte (statt Abstand zum Zentrum).
- Unterscheidung in Kernpunkte, Randpunkte und Rauschen.
- Vorteil: Erkennt Cluster beliebiger Form (nicht nur Kugeln).
- Feature: Kann Noise (Rauschen) identifizieren und aussortieren (wichtig bei chaotischen Daten wie CIFAR-10).

PCA

	Gene 1	Gene 2	…
Maus 1	1	2	…
Maus 2	3	4	…
Maus 3	2	3	…
…	…	…	…







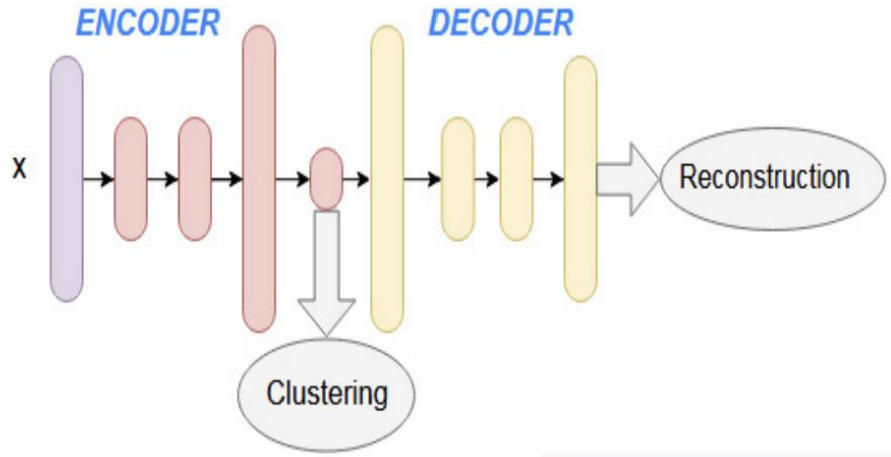
$$d_1^2 + d_2^2 + d_3^2 = max$$

- Mehrdimensionalität schwer im Koordinatensystem darzustellen
- Man nimmt zwei Spalten und analysiert sie
- Man sucht Schwerpunkt der Daten
- Man verschiebt die Punkte (erlaubt, da Relativität nicht verloren geht)
- Man zieht Gerade durch den Ursprung (wie sie liegt, hängt von Abständen hab)
- Man führt die Punkte direkt (im 90°) zur Gerade
- Steigung der Gerade = Verhältnis (-x/y)
- Oft Länge der Steigung = 1 für Eigenvektor und Eigenwert
- Eigenwert (Varianz) beschreibt Einfluss

Autoencoder

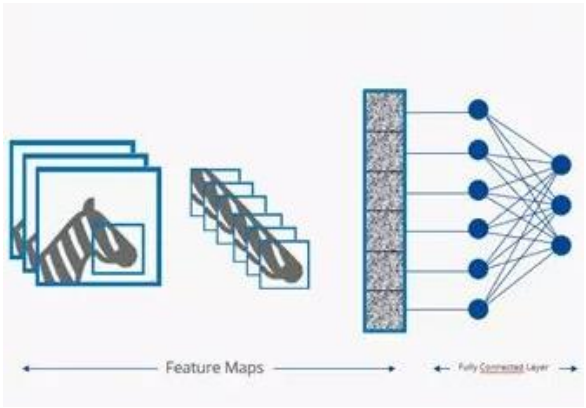
$\mathbb{R}^D \leftrightarrow \mathbb{R}^d$ (ideal)

- Statt zu Projizieren komprimiert der Encoder die Daten durch Aktivierungsfunktion
- Zwischenraum für Datenverarbeitung
- Verliert Interpretierbarkeit
- Decoder zum Prüfen der komprimierten Darstellung



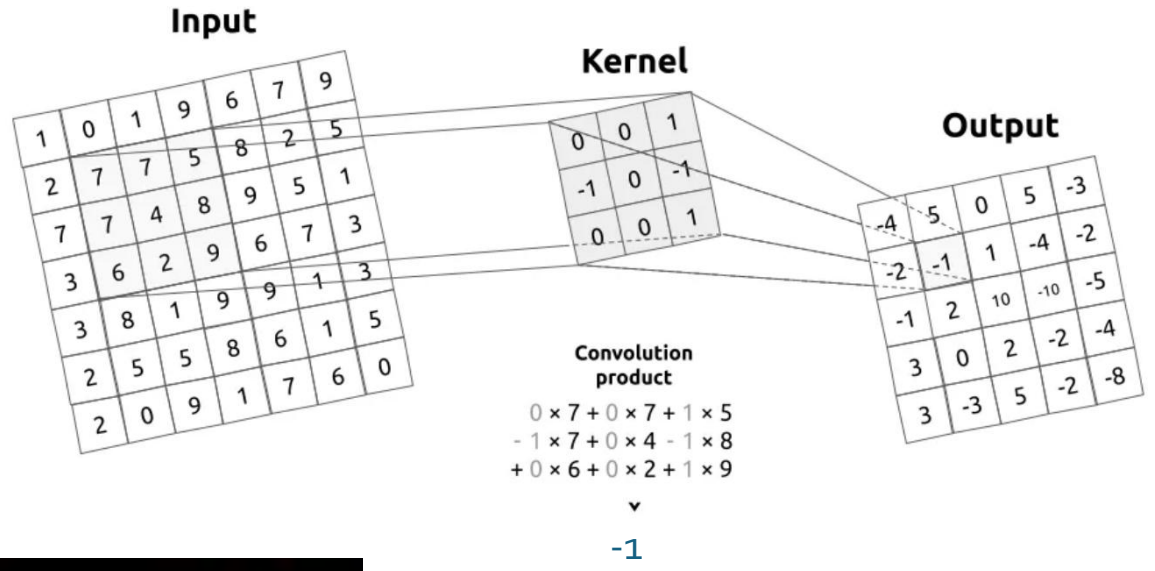
$$z = h(W \cdot x + b)$$

- Linearer Autoencoder (LAE) mit linearer Aktivierungsfunktion (vollständig verbundene Schicht) für MNIST
- Convolutional Autoencoder (CAE) mit Faltungsschichten für CIFAR10



Convolution (Faltung)

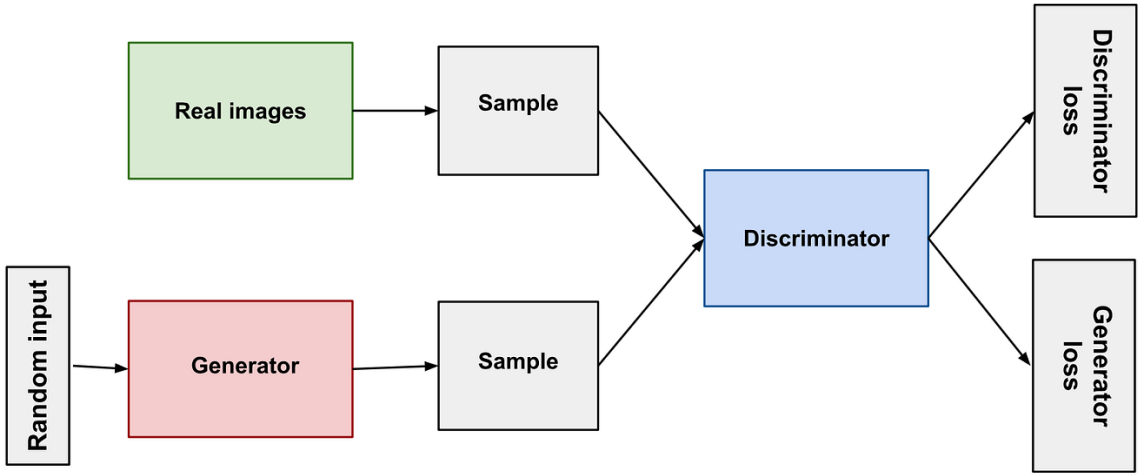
- Input mit Filter multipliziert
- Output nennt man „Feature Map“
- 7X7 und 3x3 ergibt 5x5 Feature Map
- OH = IH – FH + 1
- OB = IB – FB + 1

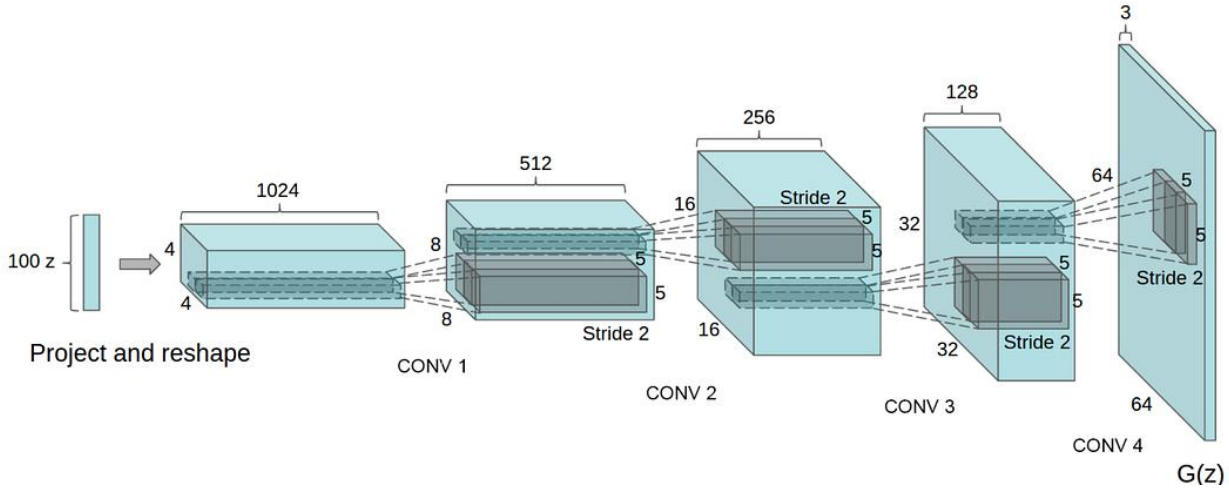


Beispiel: Darstellung von Figuren

DCGAN (generativ)

- DCGAN ist ein GAN mit veränderter Schichtstruktur für die Implementierung des Generators G und Diskriminators

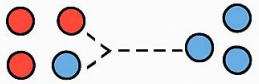
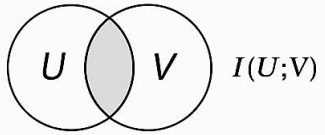
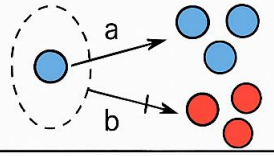


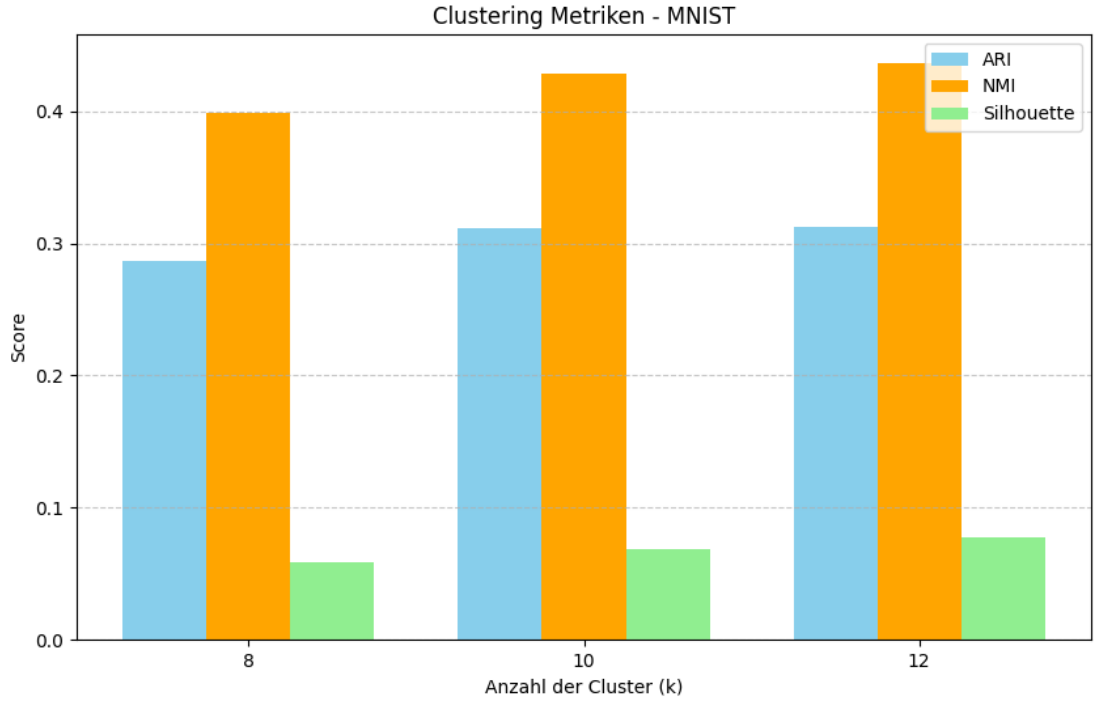


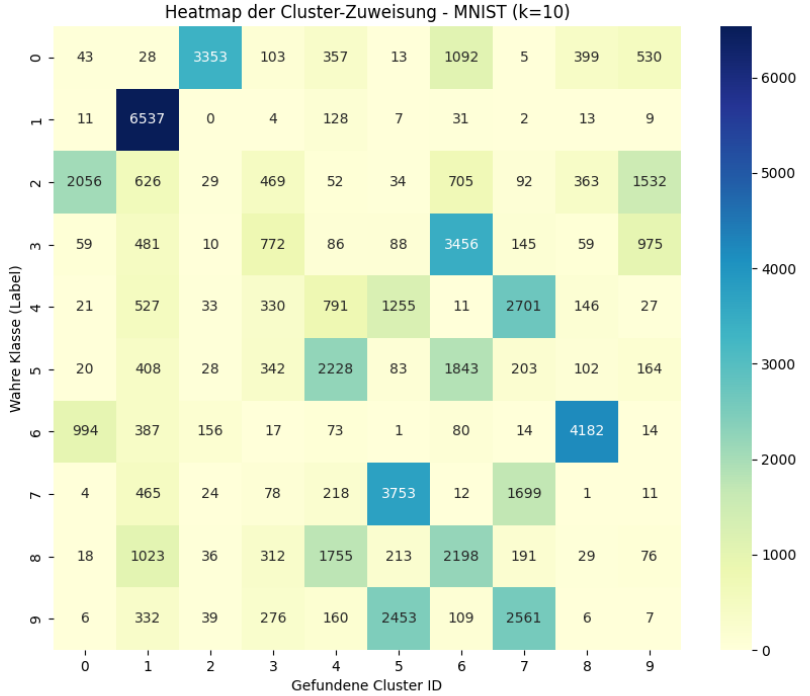
- Generator generiert Fake-Daten
- Diskriminator erkennt Daten
- Fake- und Real-Daten werden gemischt
- Entweder täuscht Generator oder Diskriminator erkennt (nicht beides!)

Metriken (Maß)

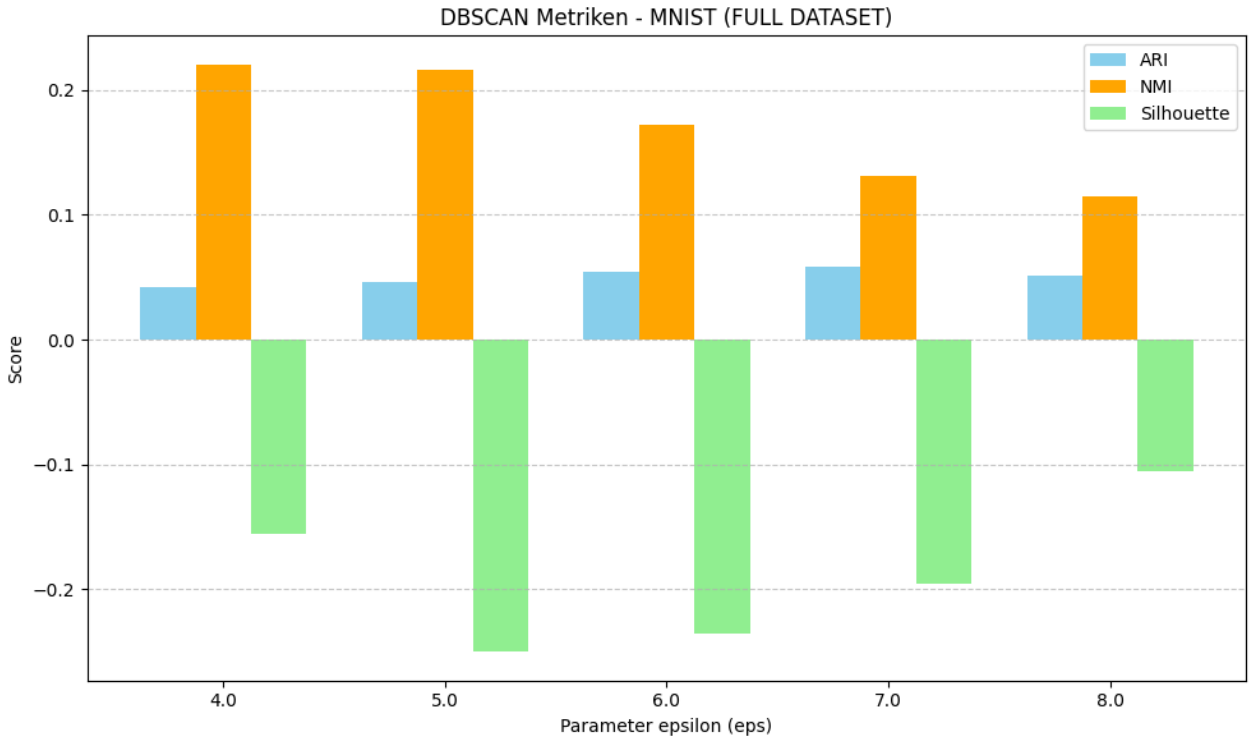
- ARI – Übereinstimmung zwischen zwei Cluster
- NMI – informationsbasierte Ähnlichkeit
- Silhouette score – Distanzvergleich: eigener Cluster vs. nächster

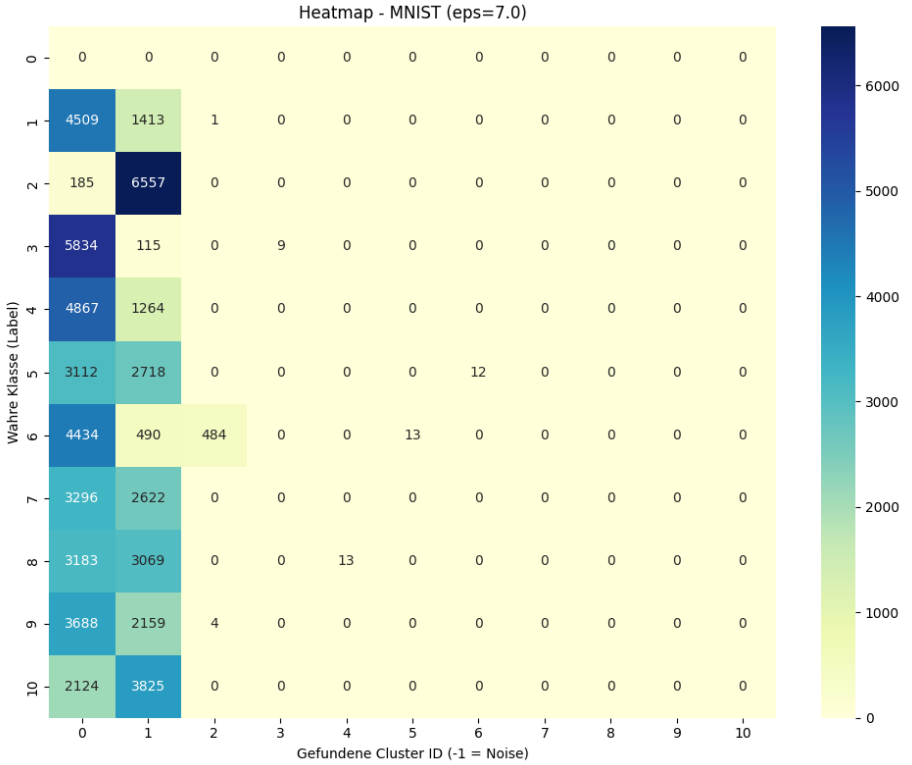
ARI	NMI	Silhouette
Agreement between two clusterings	Information-based similarity	Cluster structure quality
Range −1 1	**Range** 0 1	**Range** −1 1
Calculation Basis 		
Interpretation 1 = perfect 0 = random <0 = worse than random	**Interprettation** 1 = perfect 0 = no shared information	**Interpretation** 1 = well placed 0 = on boundary <0 = misassigned



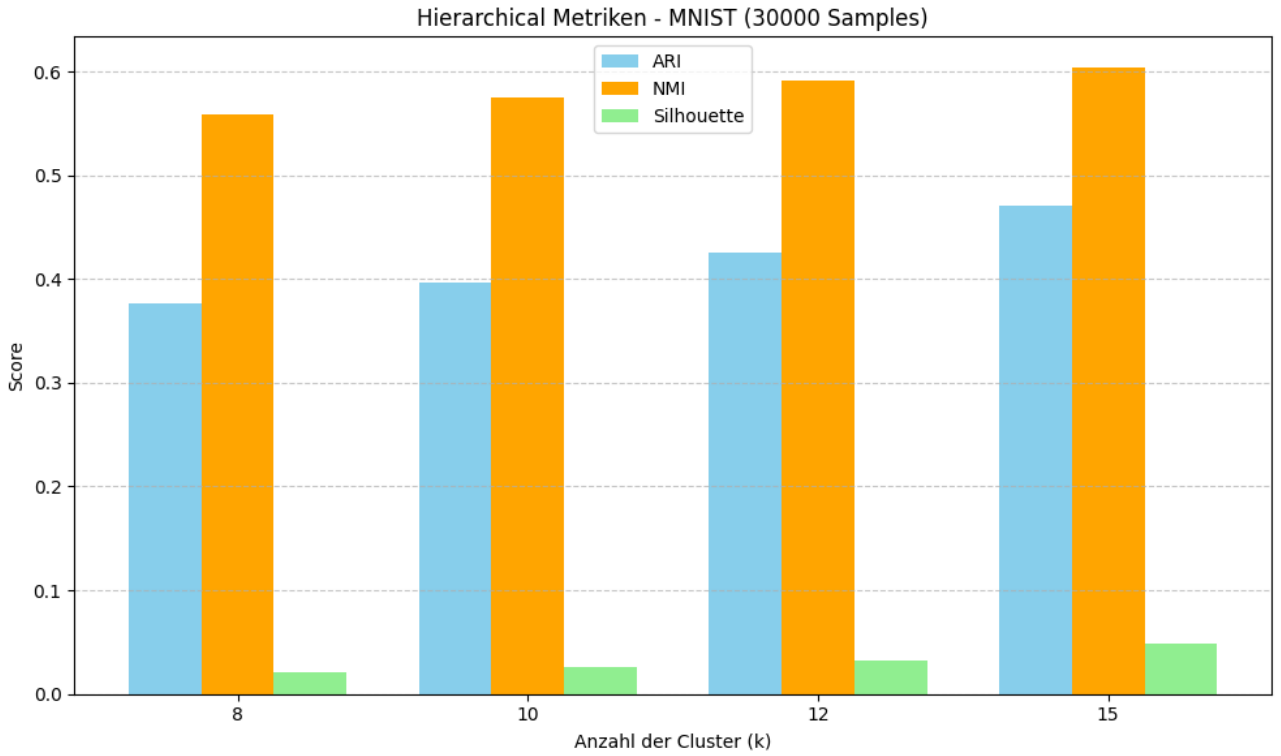


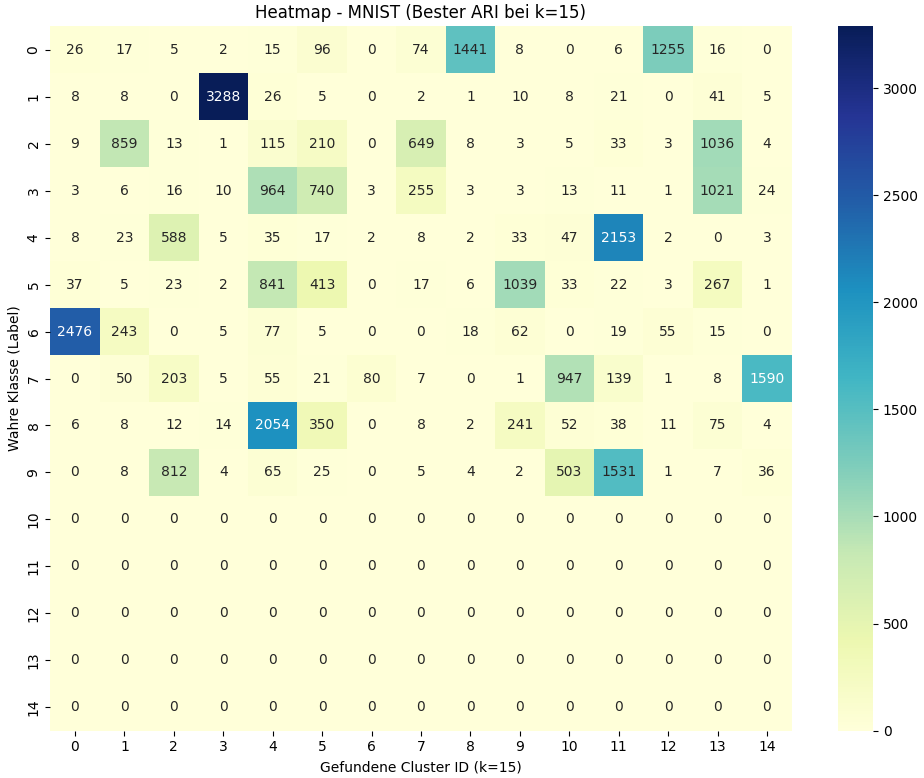
KMEANS BASELINE MIT PCA (MNIST)



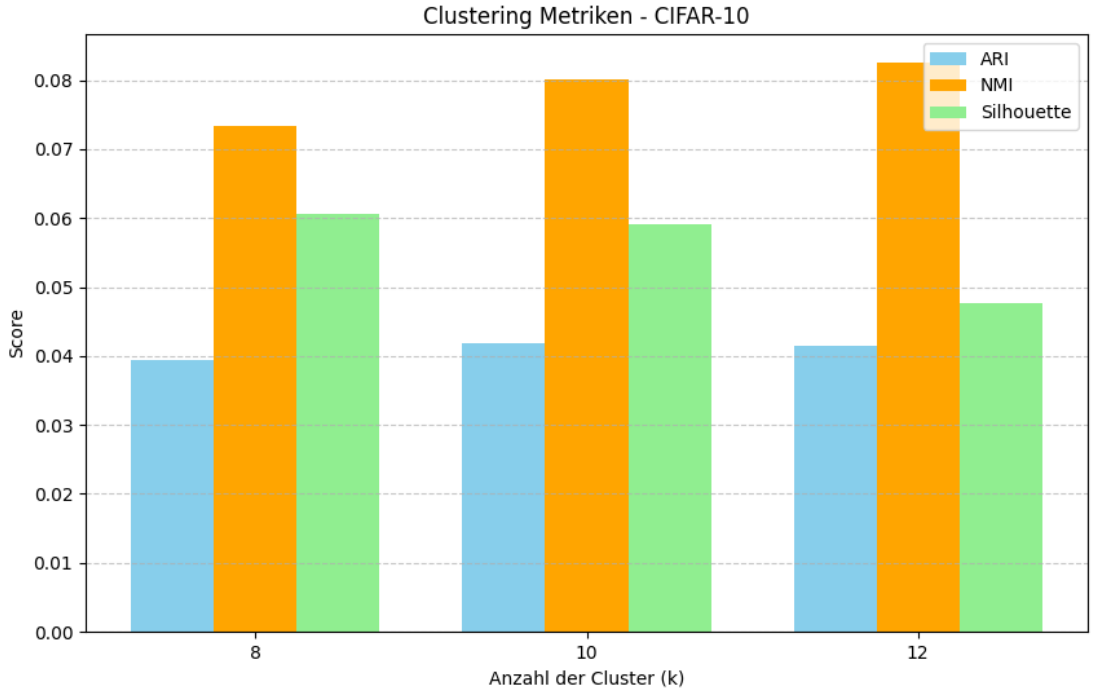


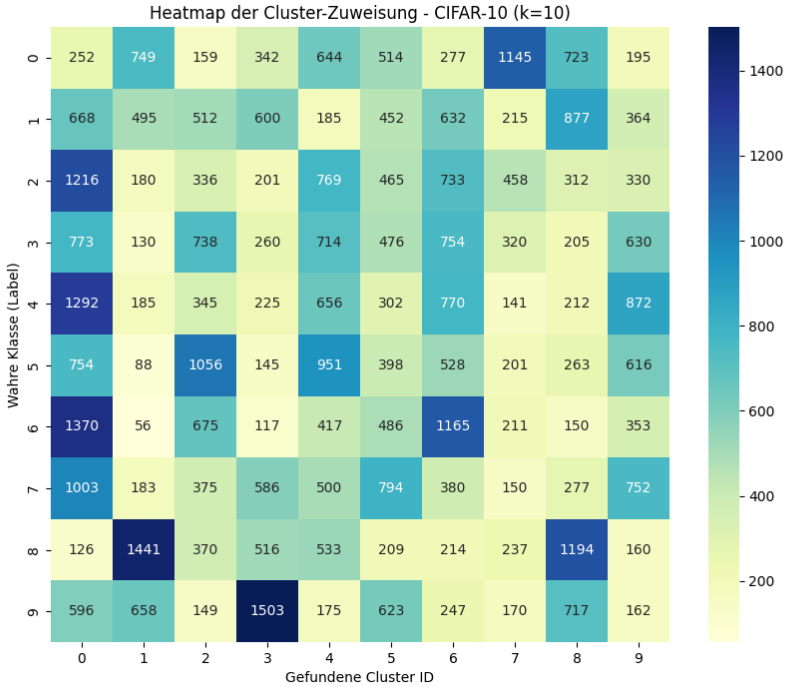
DBSCAN BASELINE MIT PCA (MNIST)



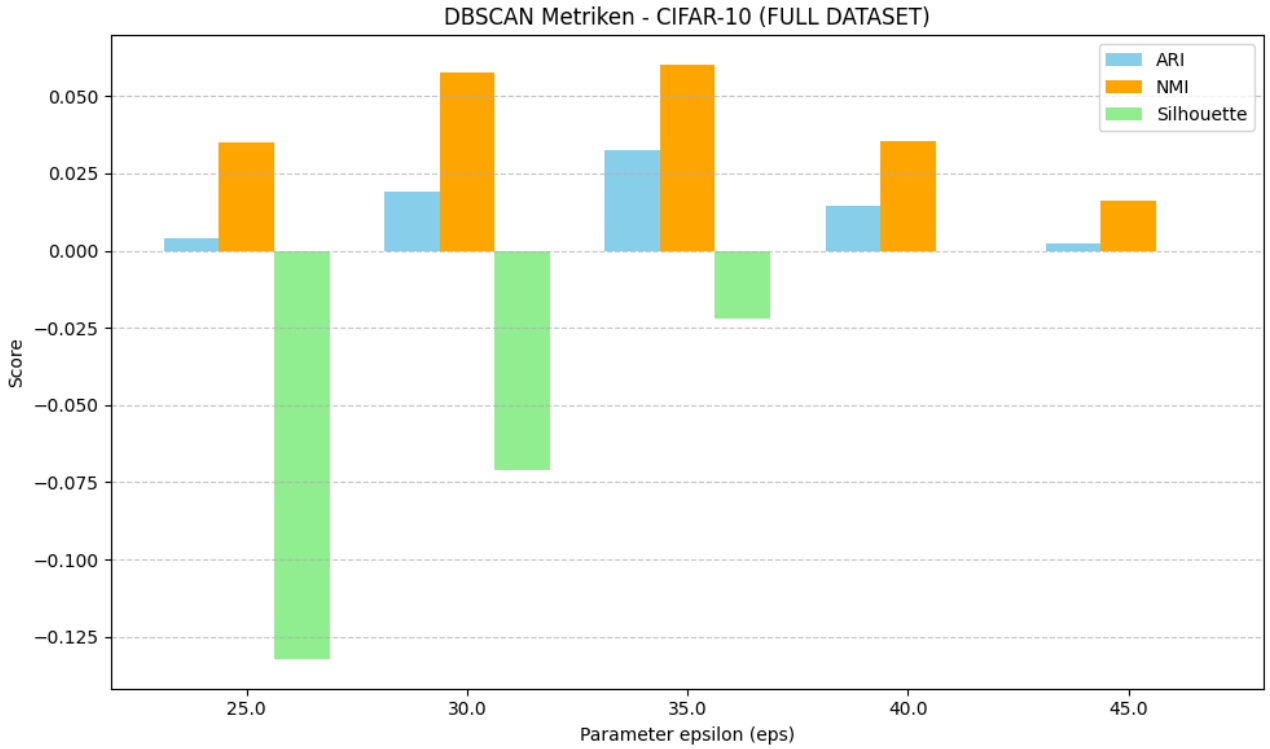


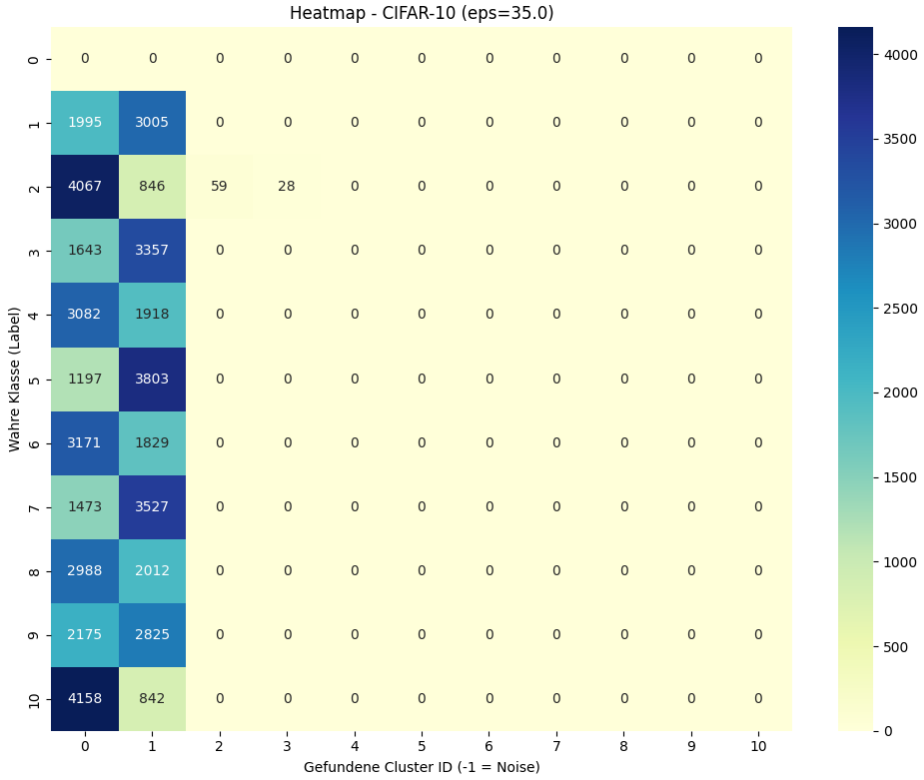
HIERARCHISCHES BASELINE MIT PCA (MNIST)



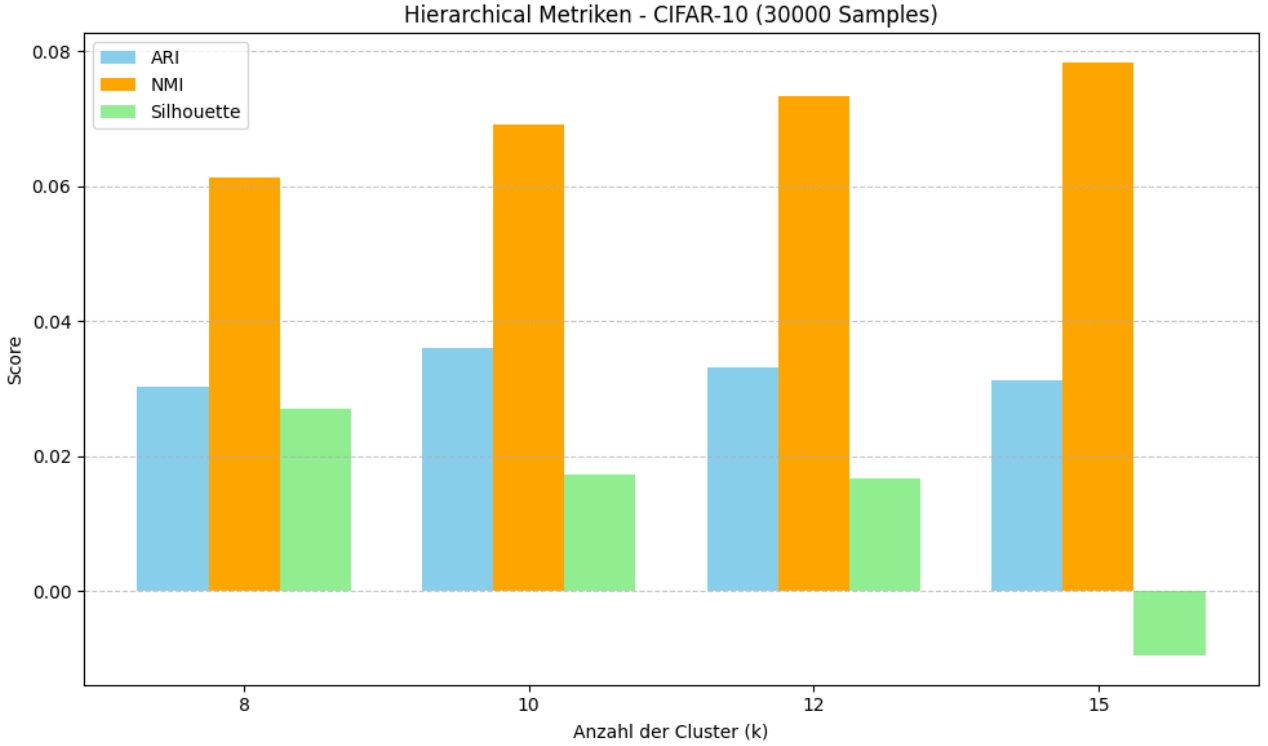


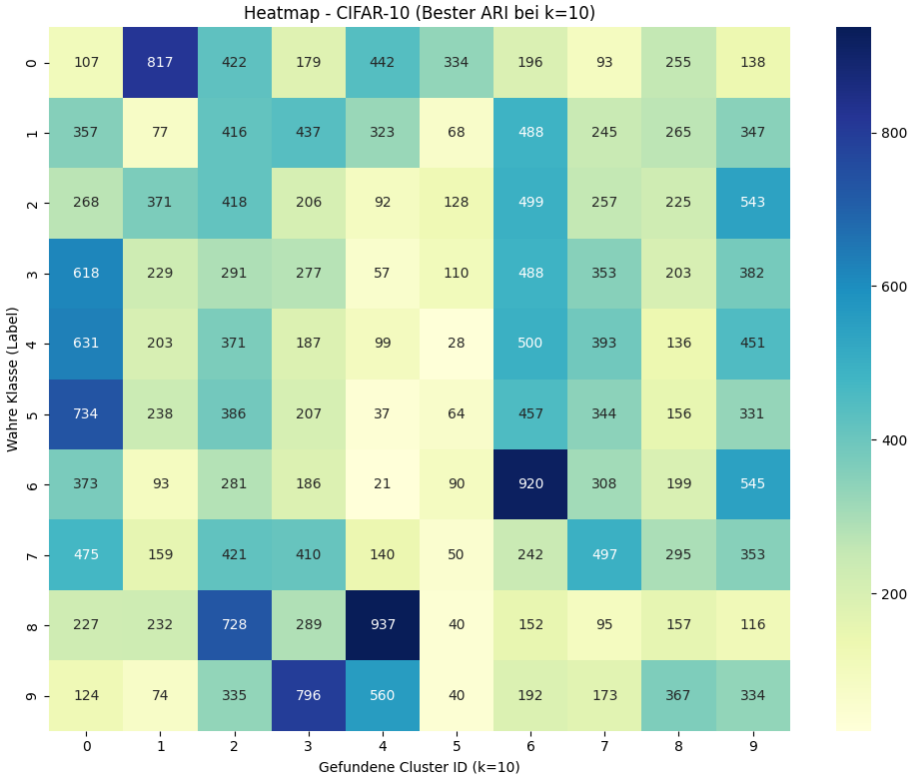
KMEANS BASELINE MIT PCA (CIFAR-10)





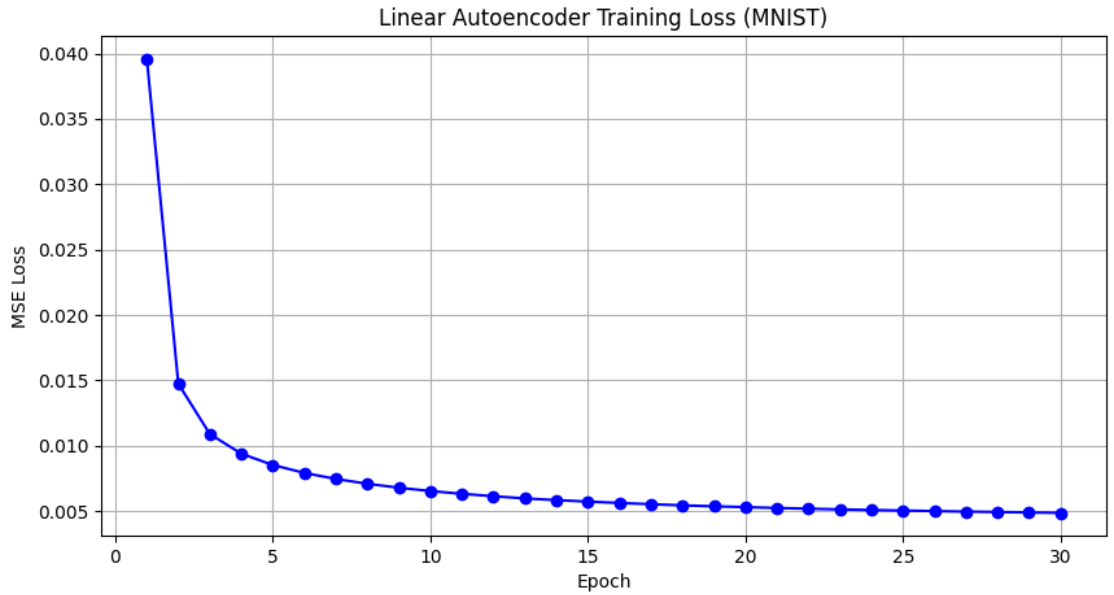
DBSCAN BASELINE MIT PCA (CIFAR-10)

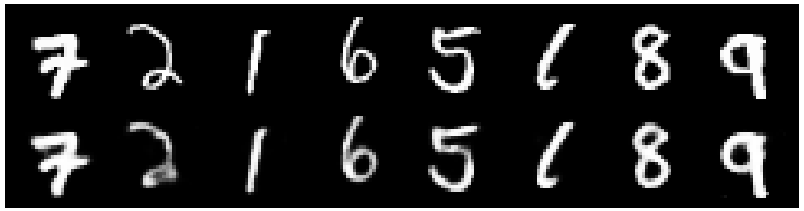




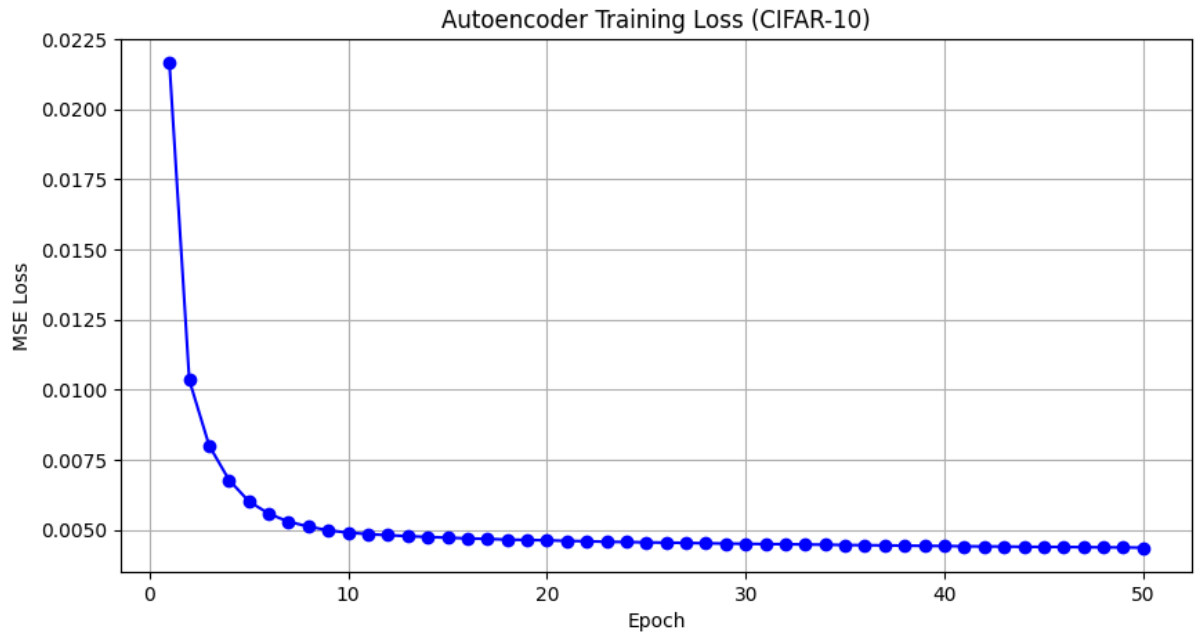
HIERARCHISCHES BASELINE MIT PCA (CIFAR-10)

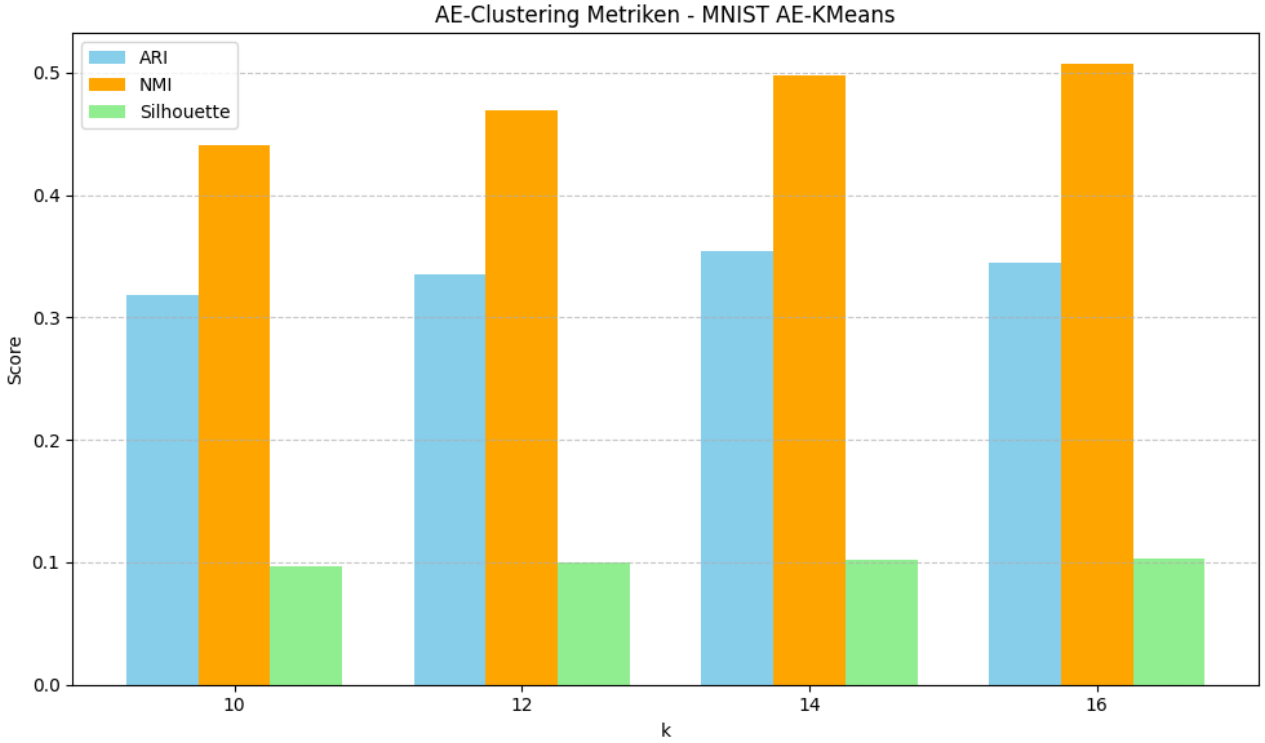
Linear Autoencoder Training

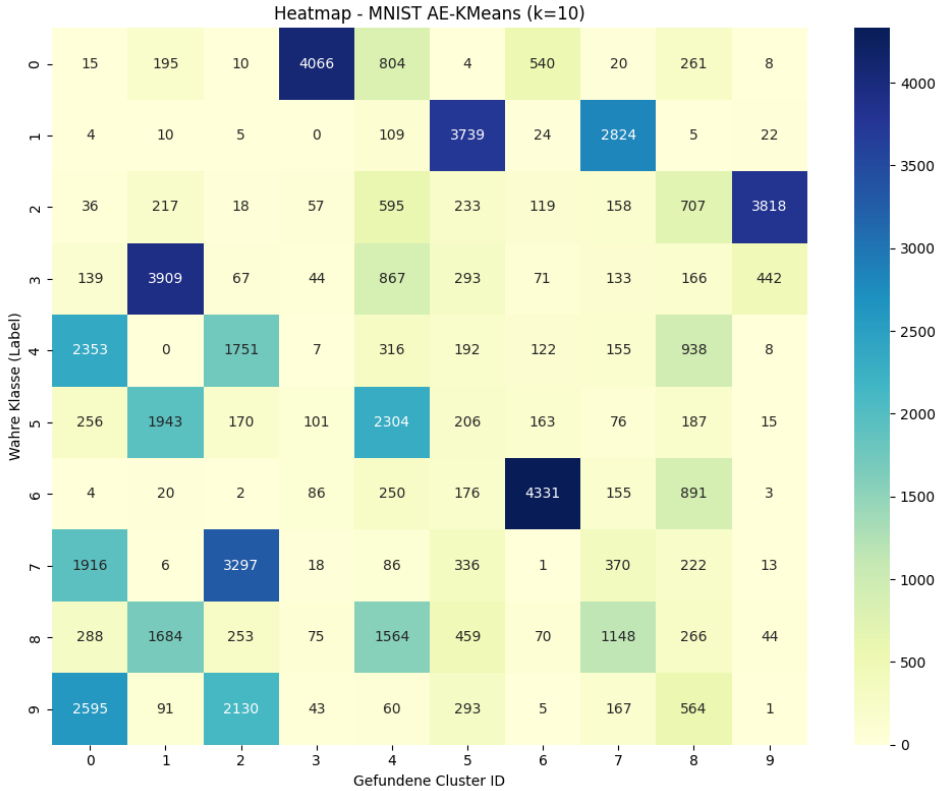




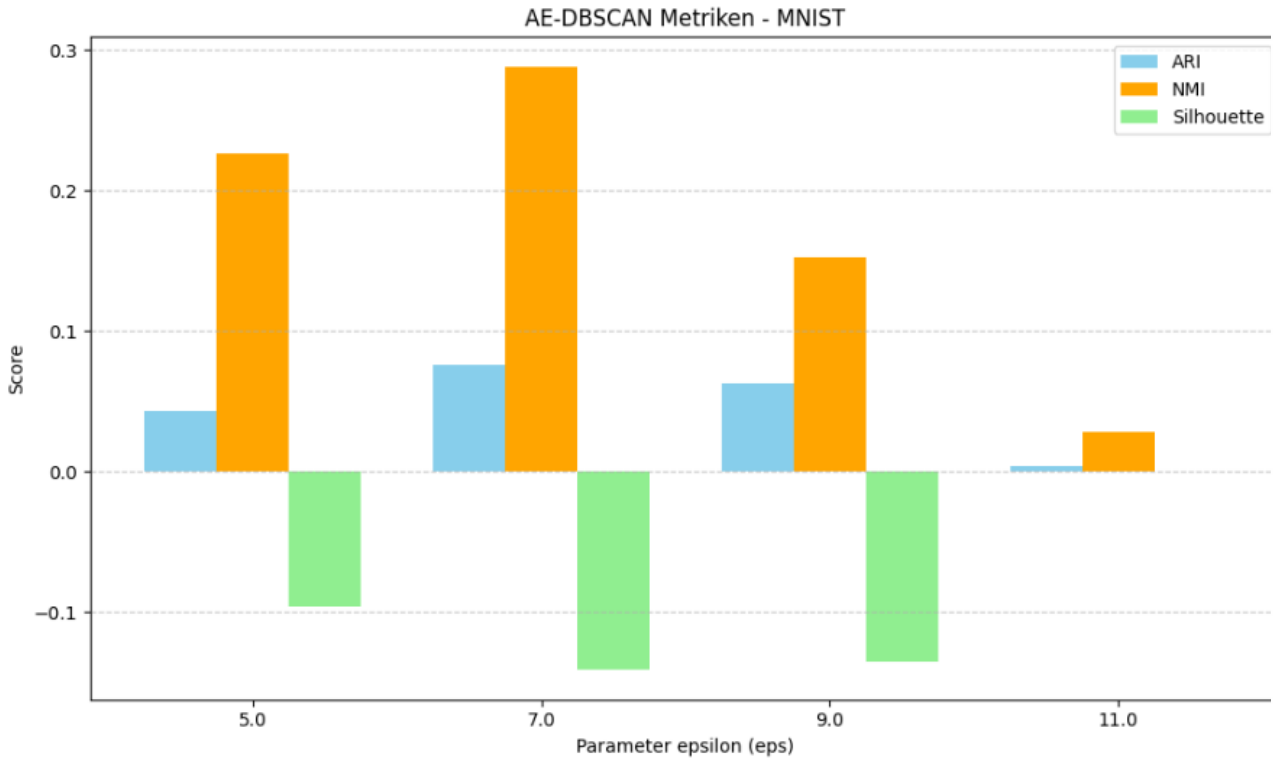
Convolutional Autoencoder Training

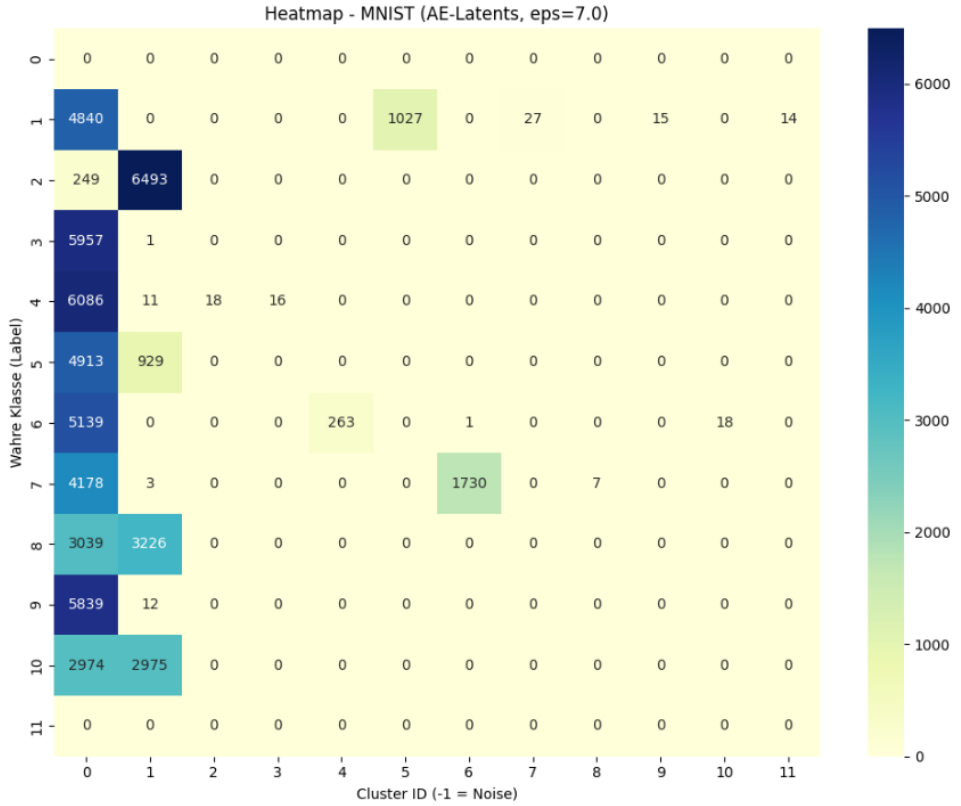




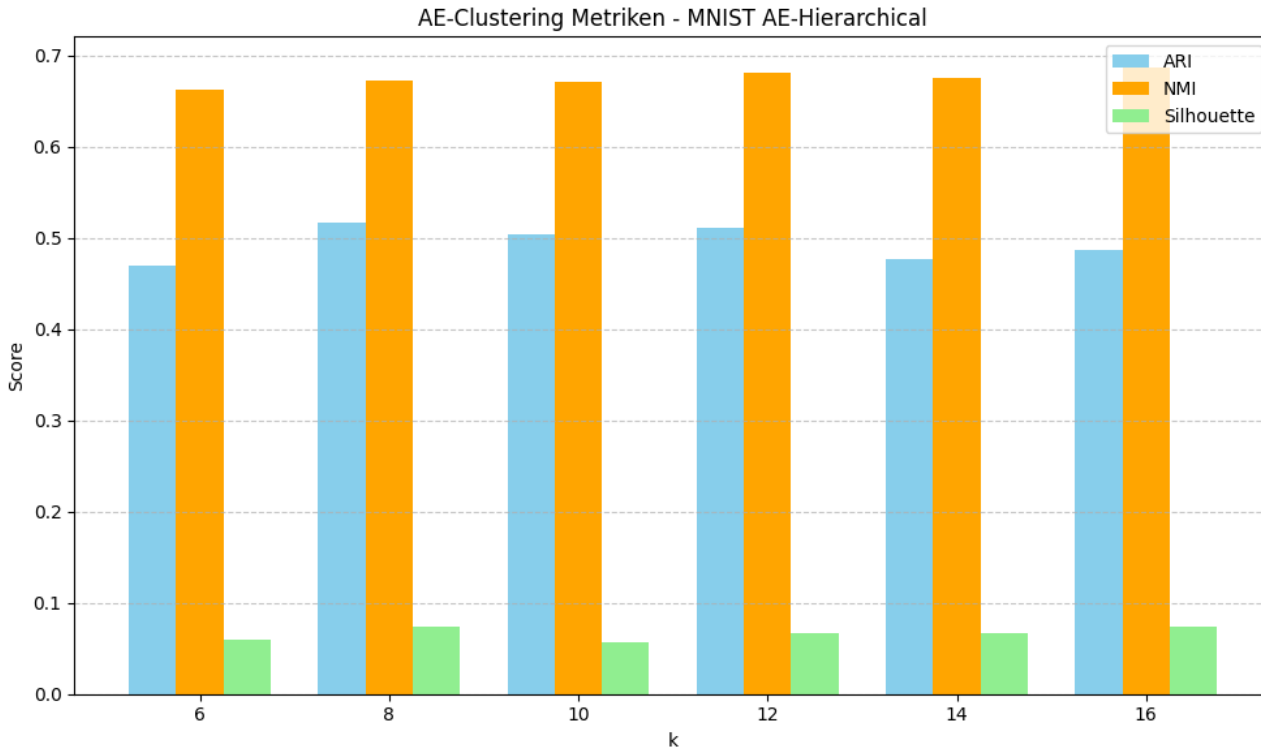


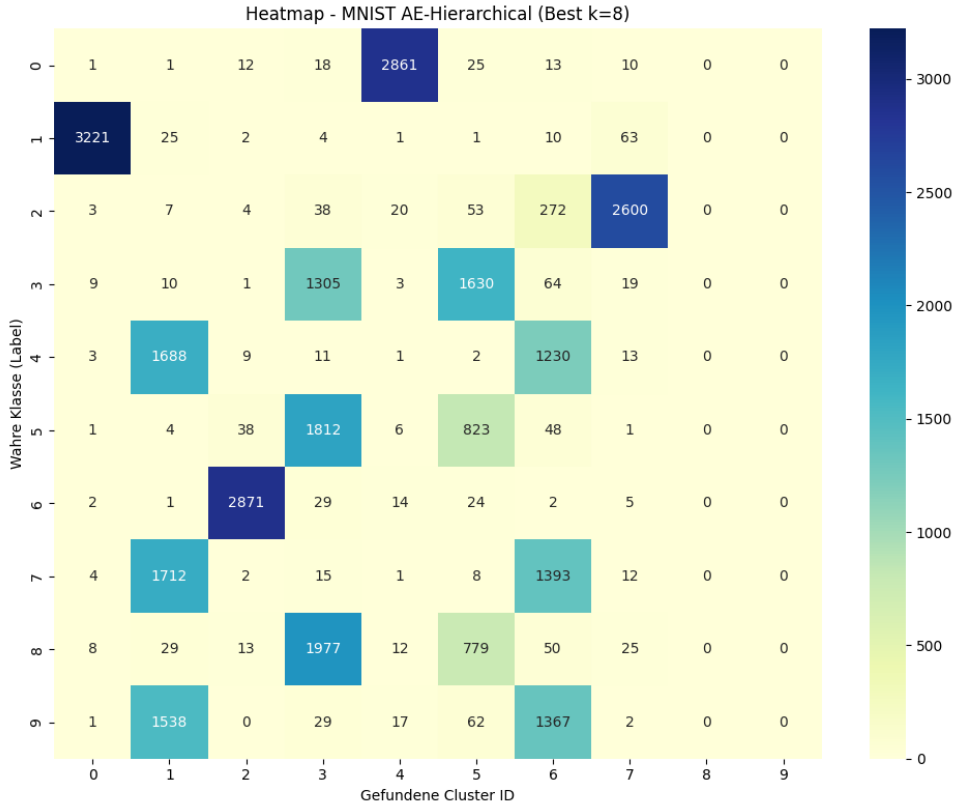
KMEANS AUTOENCODER (MNIST)



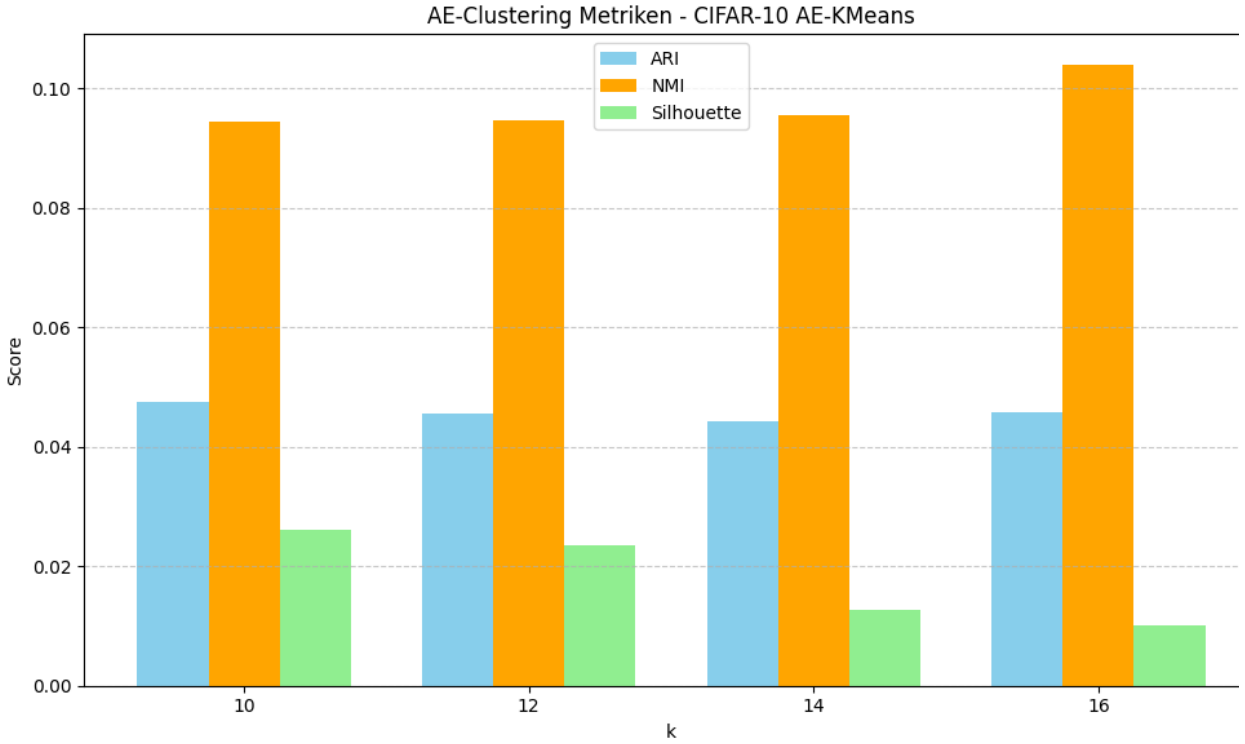


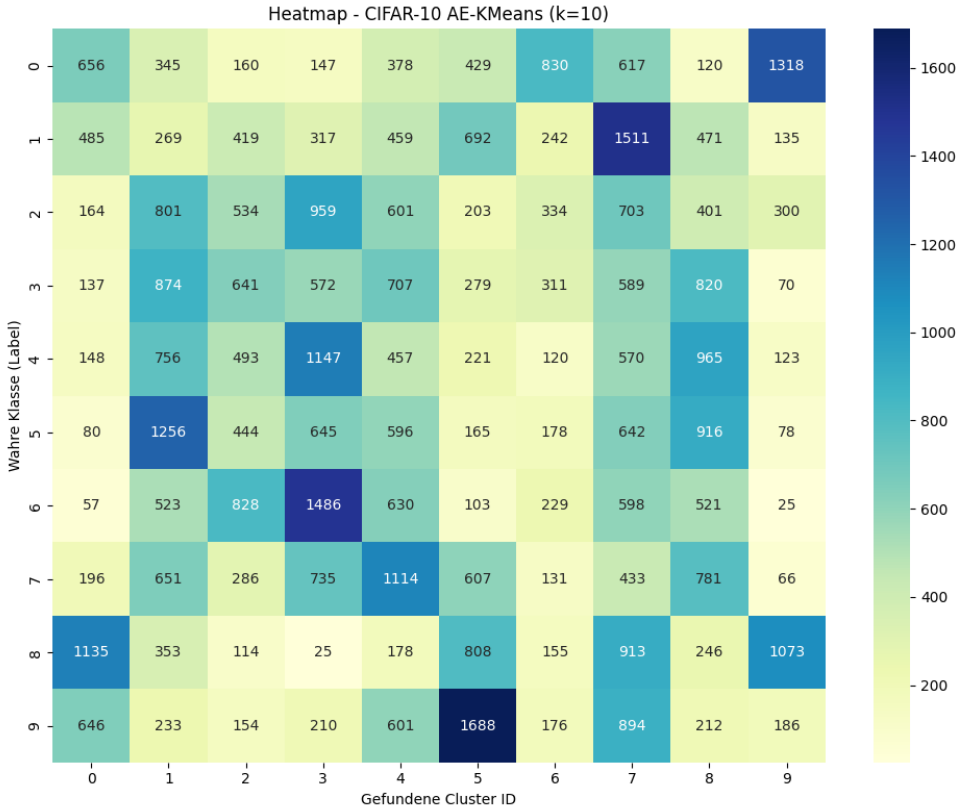
DBSCAN AUTOENCODER (MNIST)



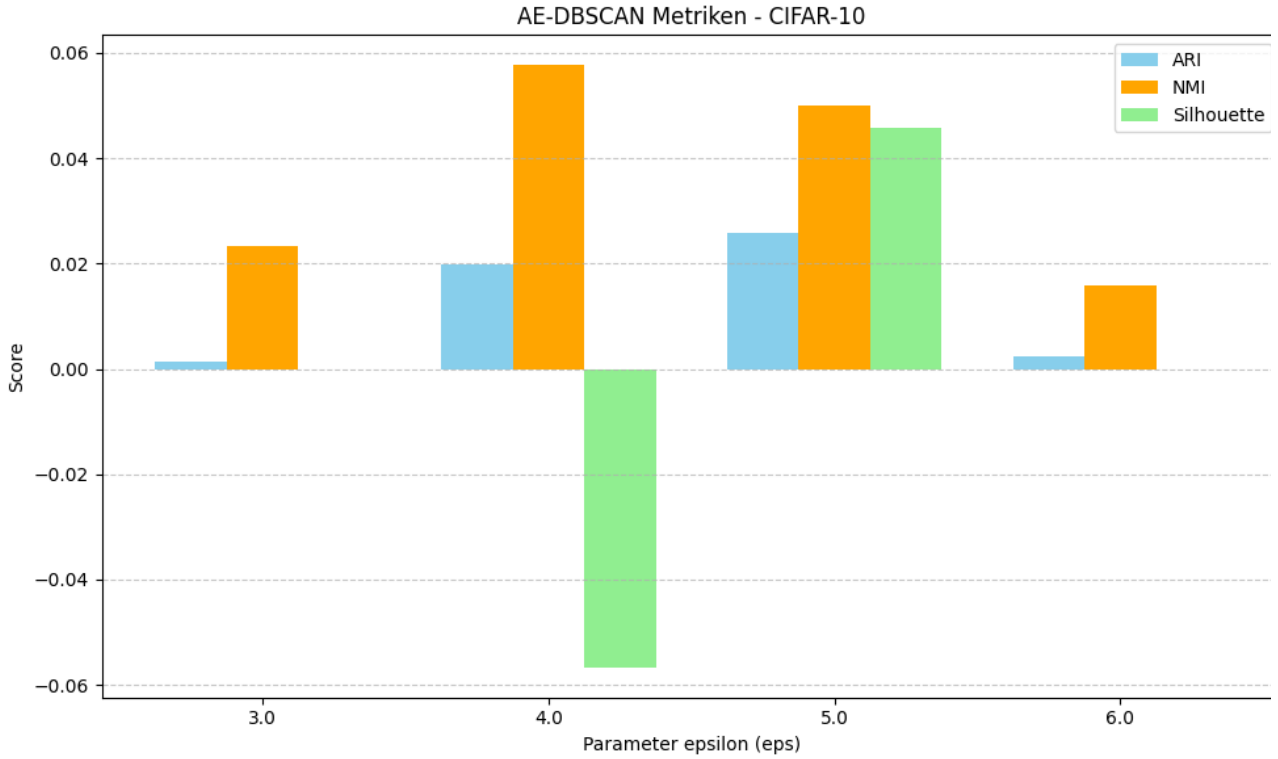


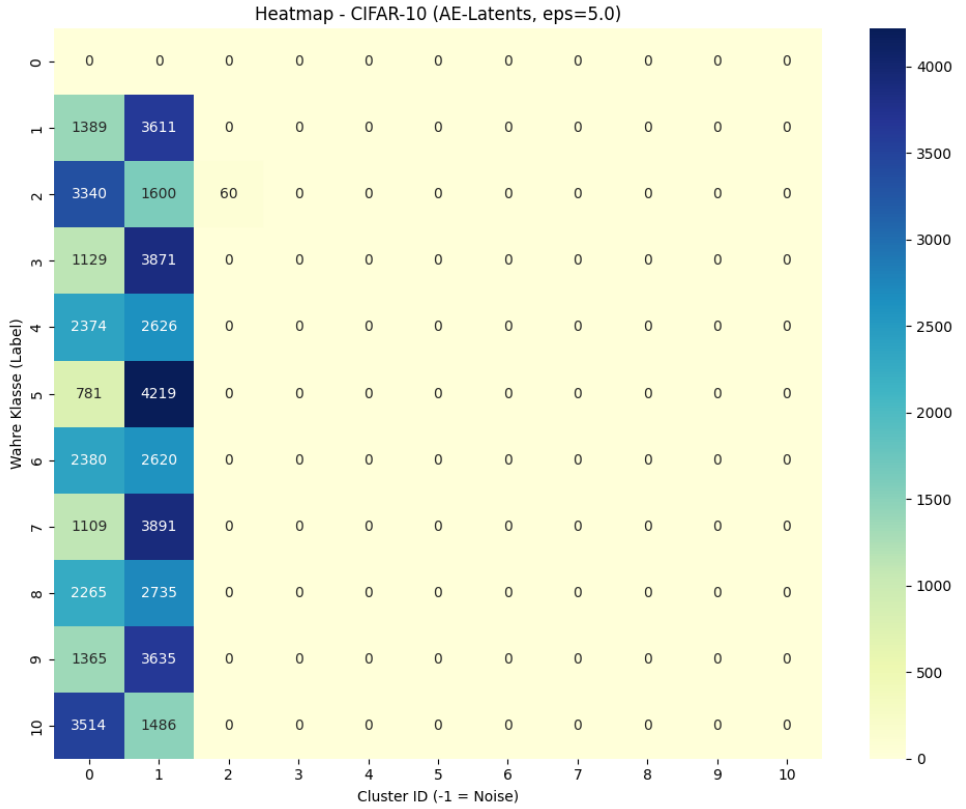
HIERARCHISCHES MIT AUTOENCODER (MNIST)



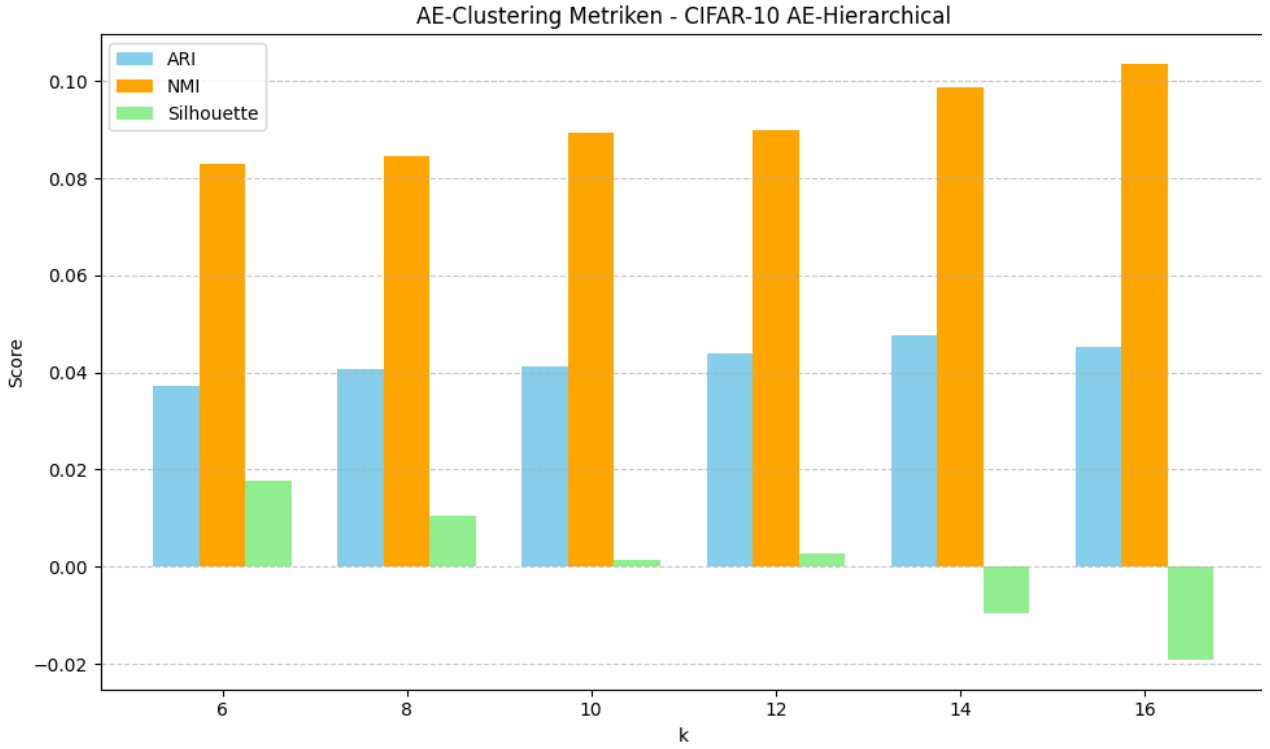


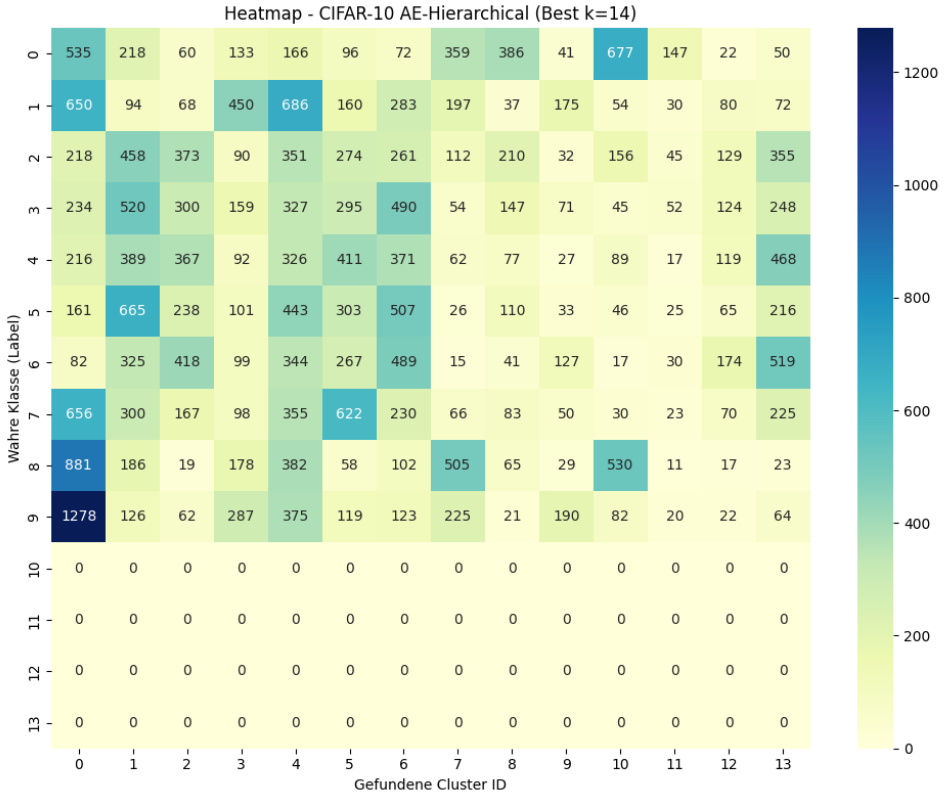
KMEANS AUTOENCODER (CIFAR-10)





DBSCAN AUTOENCODER (CIFAR-10)





HIERARCHISCHES AUTOENCODER (CIFAR-10)

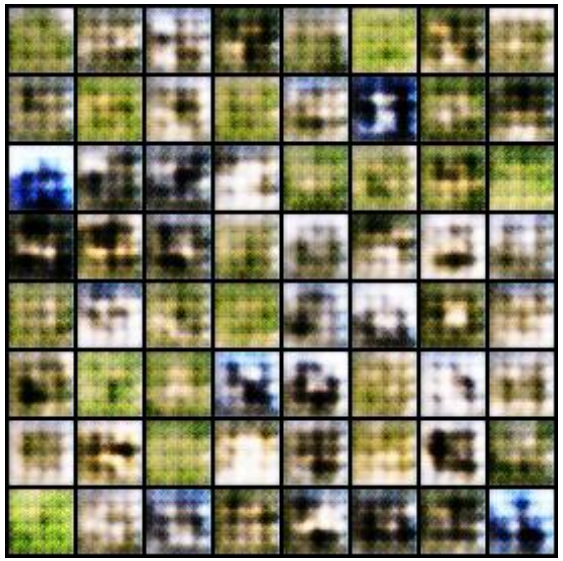
Epoche 1

Epoche 25

Epoche 50

DCGAN TRAINING (MNIST)

DCGAN Training (CIFAR-10)

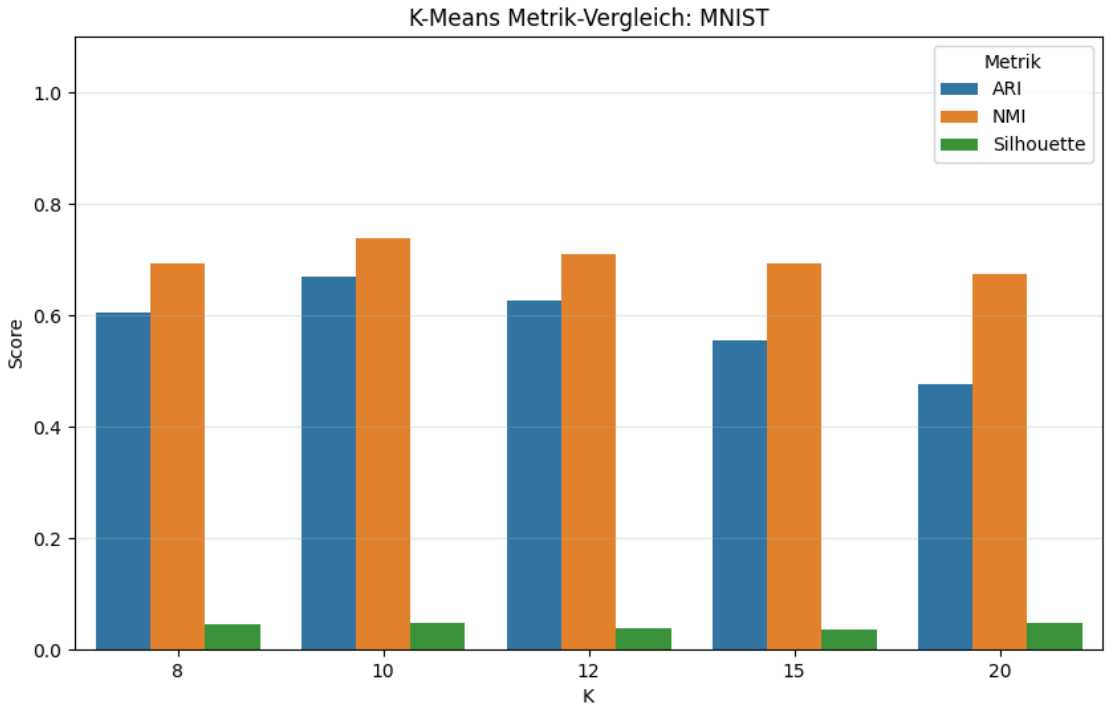


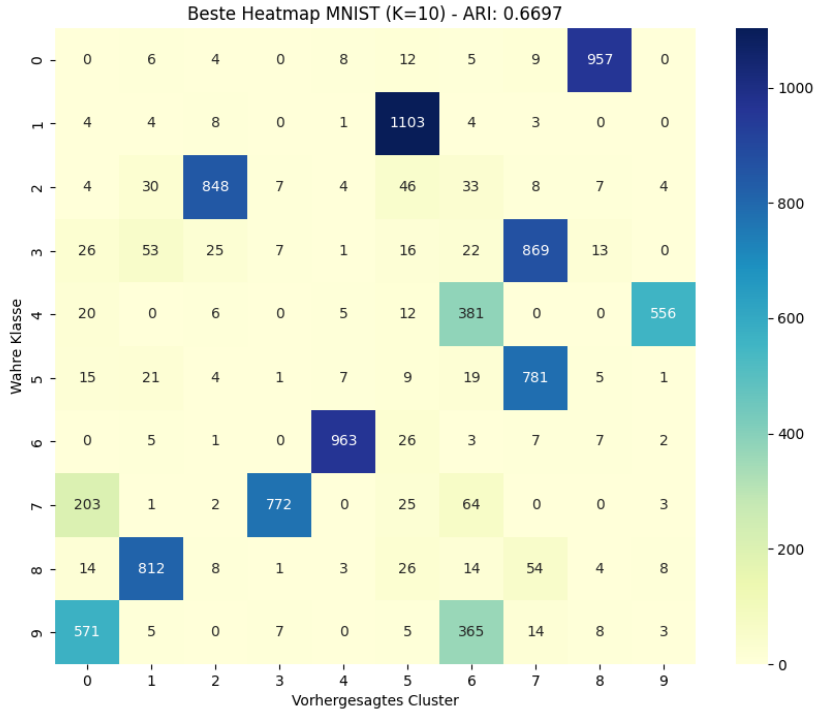
Epoche 1

Epoche 35

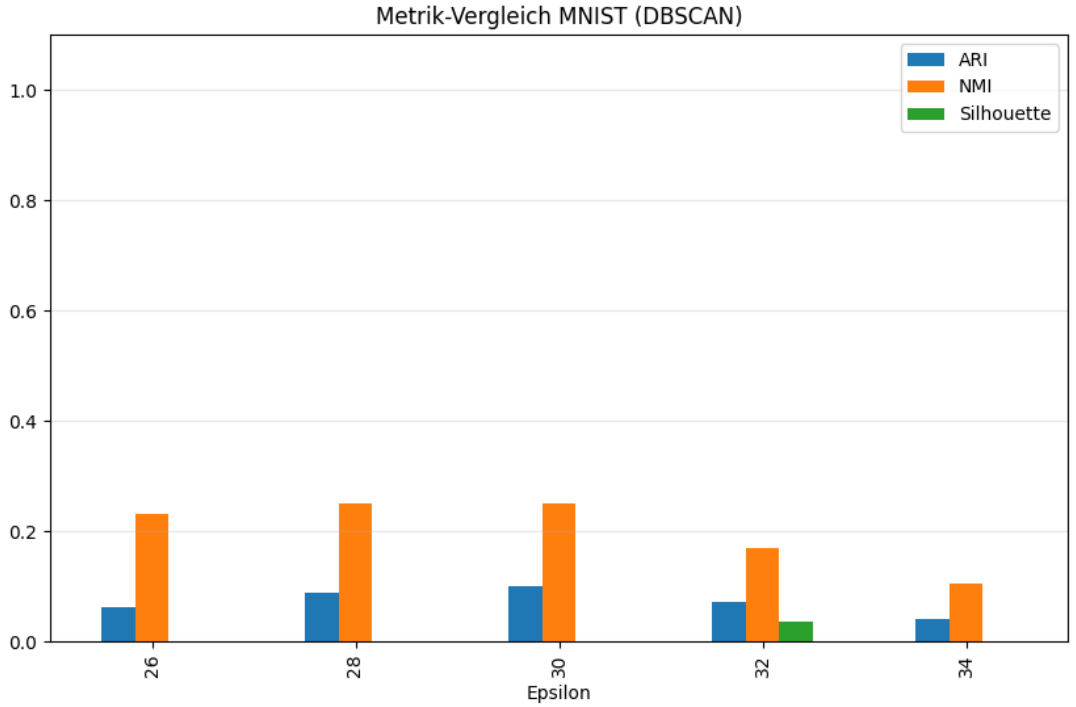
Epoche 75

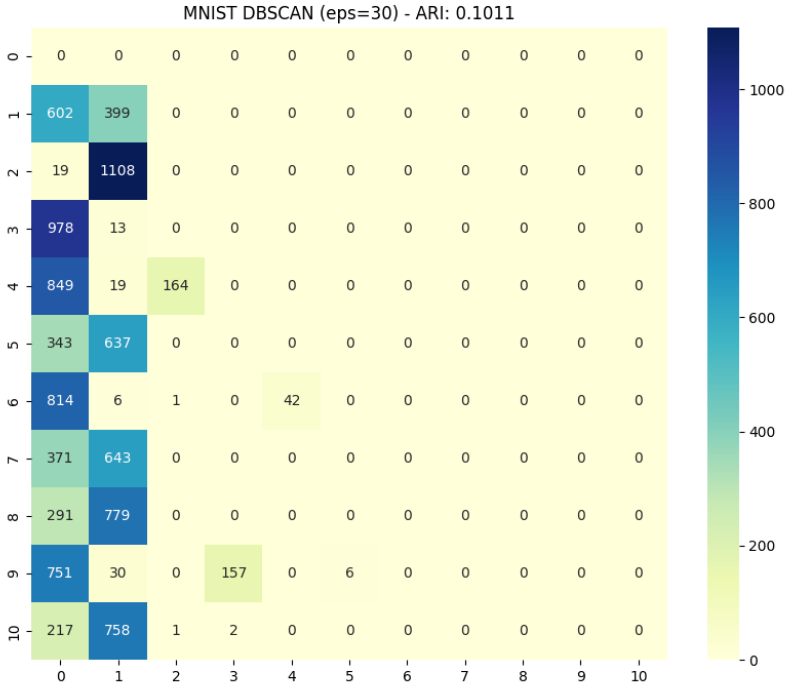
Epoche 100



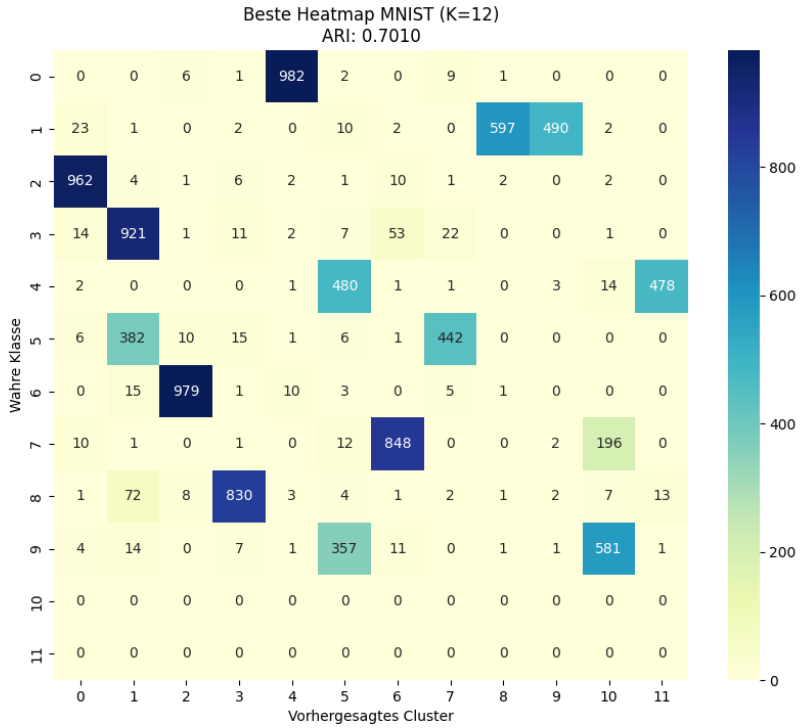


KMEANS MIT DCGAN (MNIST)

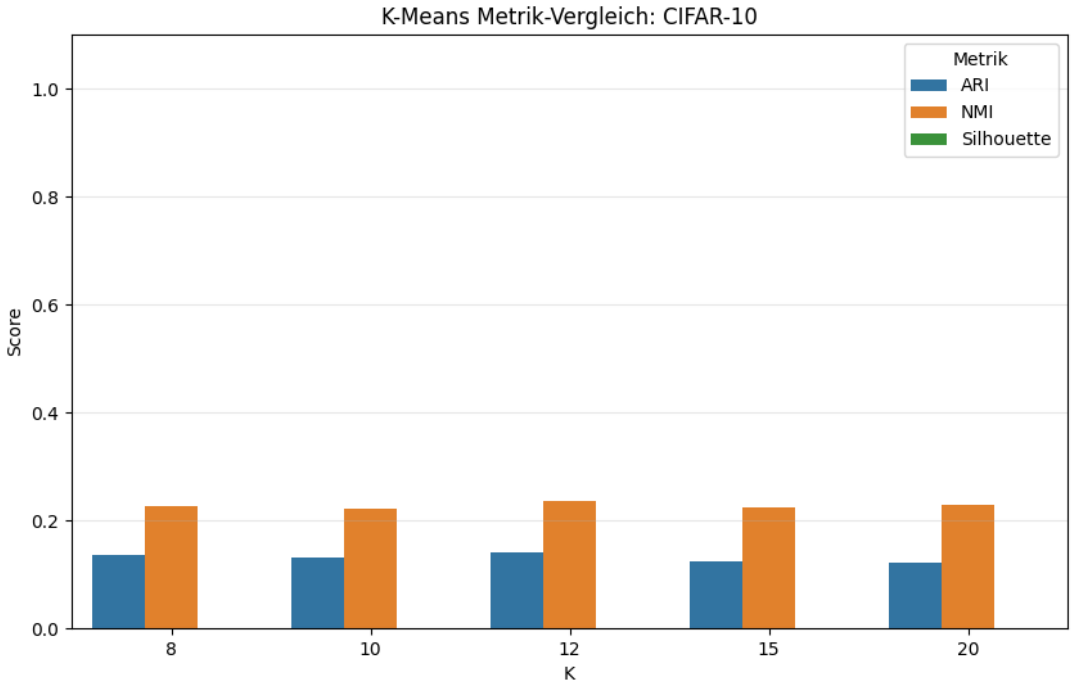


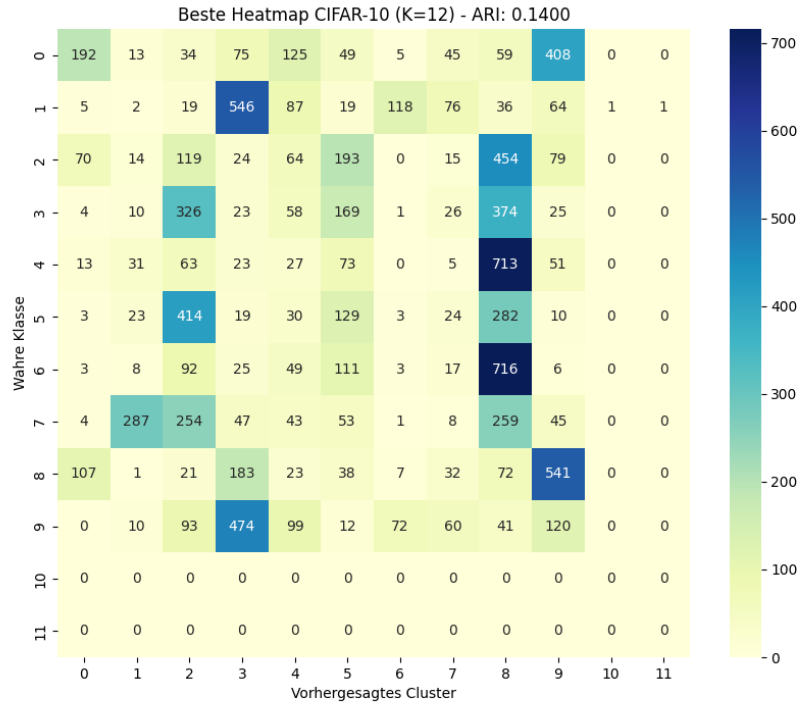


DBSCAN MIT DCGAN (MNIST)

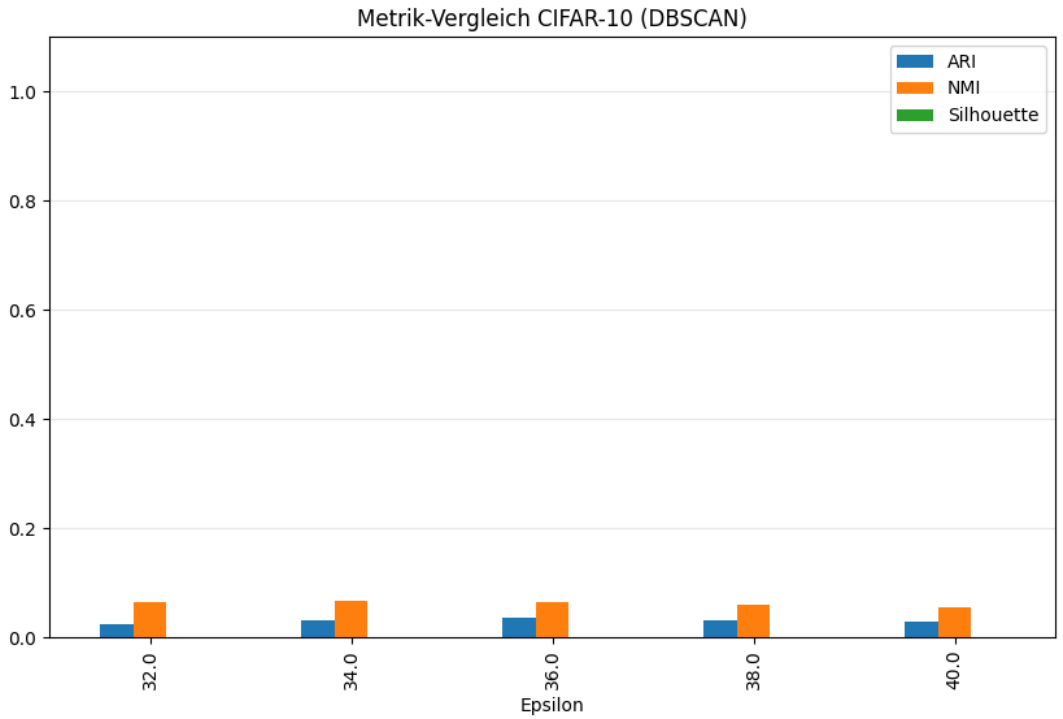


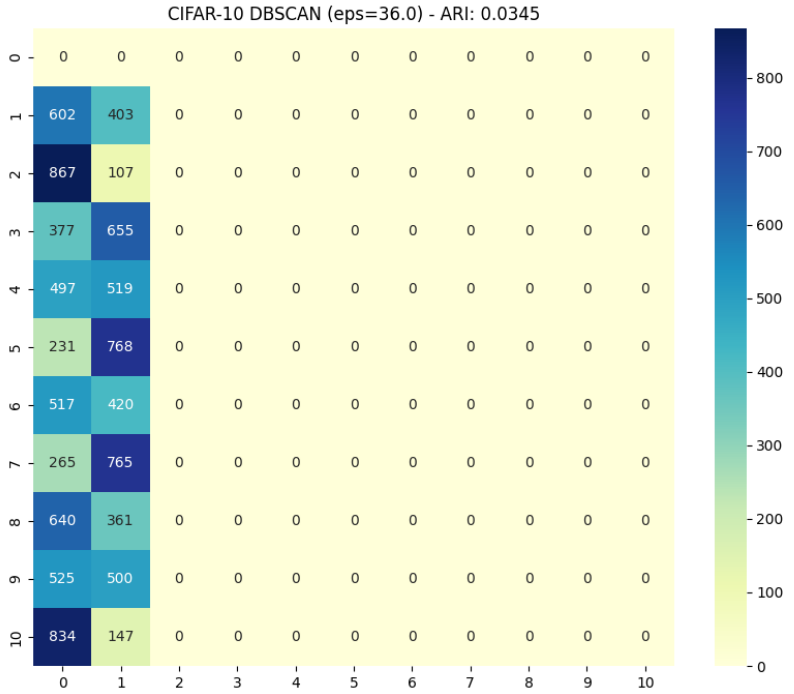
HIERARCHISCHES MIT DCGAN (MNIST)



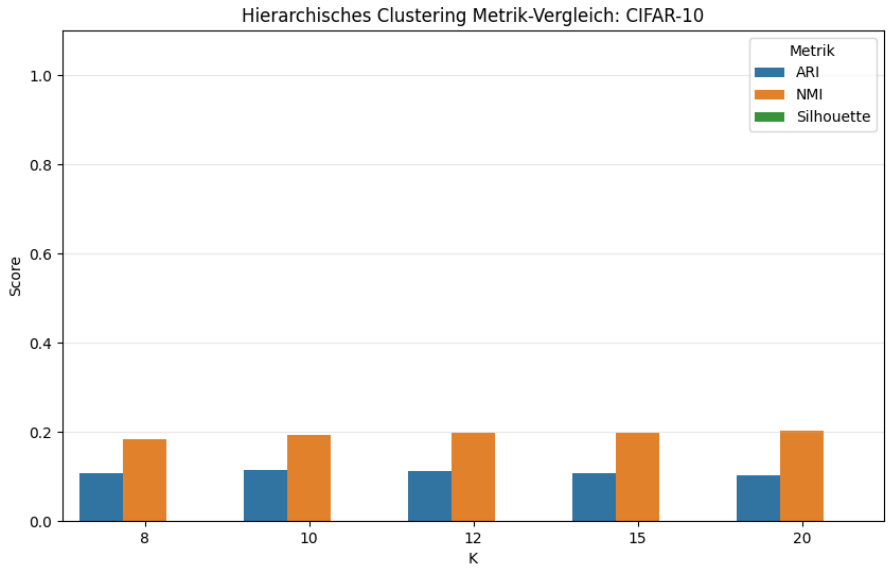


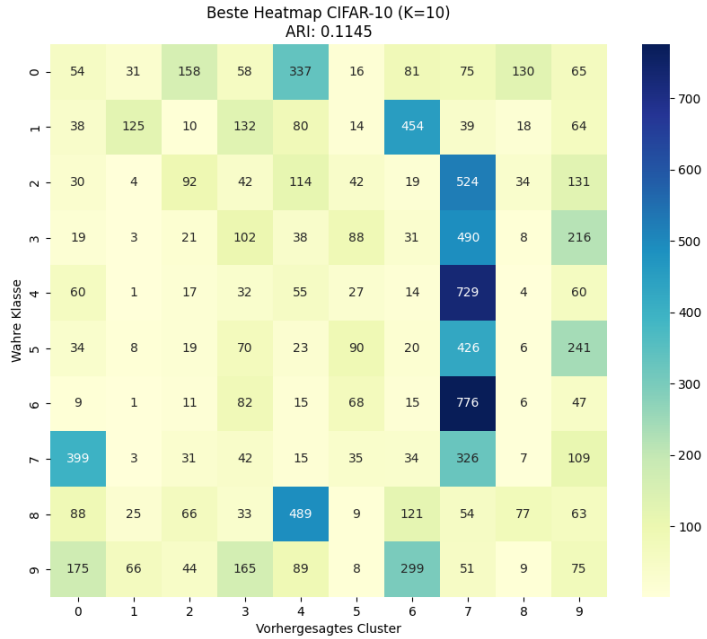
KMEANS MIT DCGAN (CIFAR-10)





DBSCAN MIT DCGAN (CIFAR-10)





HIERARCHISCHES MIT DCGAN (CIFAR-10)

Fazit

- Clustering auf Rohdaten ist bei komplexen Datensätzen wie CIFAR-10 weitgehend ineffektiv.
- Autoencoder verbessern die Strukturierung, bleiben jedoch durch rekonstruktive Features begrenzt.
- DCGAN-basierte Repräsentationen liefern die besten Ergebnisse durch semantisch diskriminative Merkmale.
- Die Kombination aus DCGAN-Features und hierarchischem Clustering erzielt die höchste Clustering-Qualität.