**Πανεπιστήμιο Μακεδονίας**
Πρόγραμμα Μεταπτυχιακών Σπουδών

**Μάθημα:**
Μέθοδοι και Εργαλεία Τεχνητής Νοημοσύνης

# Αξιολόγηση Συνδυαστικών Μοντέλων Μείωσης Διάστασης και Συσταδοποίησης στο Fashion-MNIST Dataset

ΣΥΓΓΡΑΦΕΙΣ:
Κωνσταντίνος Θωμασιάδης, mai25016
Ευστάθιος Ιωσηφίδης, mai25017
Αικατερίνη Κρότκα, mai25031

Θεσσαλονίκη, Μάιος 2025

# Περιεχόμενα



# Κατάλογος σχημάτων



# Κατάλογος Πινάκων

# 1. Εισαγωγή

Η παρούσα εργασία εξετάζει την αξιολόγηση διαφορετικών συνδυαστικών μοντέλων μείωσης διάστασης χώρου και συσταδοποίησης επί εικόνων του fashion-mnist dataset. Η μείωση διάστασης είναι μια σημαντική τεχνική για την επεξεργασία δεδομένων υψηλής διάστασης, ενώ η συσταδοποίηση επιτρέπει την ανακάλυψη εγγενών δομών στα δεδομένα. Σε αυτήν την εργασία, υλοποιήθηκαν και συγκρίθηκαν οι τεχνικές Principal Component Analysis (PCA) και Stacked Autoencoder (SAE) για τη μείωση διάστασης, καθώς και οι αλγόριθμοι Mini-Batch K-means και Gaussian Mixture Model (GMM) για τη συσταδοποίηση. Στόχος της εργασίας είναι η αξιολόγηση της επίδρασης των τεχνικών μείωσης διάστασης στην απόδοση των αλγορίθμων συσταδοποίησης και η σύγκριση των αποτελεσμάτων που προκύπτουν από την εφαρμογή των αλγορίθμων συσταδοποίησης απευθείας στα raw δεδομένα με τα αποτελέσματα που προκύπτουν από την εφαρμογή τους στα δεδομένα μετά από μείωση διάστασης. Επιπλέον, αξιολογείται η καταλληλότητα των clusters με τη χρήση των μετρικών Calinski-Harabasz, Davies-Bouldin και Silhouette score.
Η υπόλοιπη αναφορά είναι δομημένη ως εξής: Στο Κεφάλαιο 2 παρουσιάζεται το υπόβαθρο και η σχετική βιβλιογραφία. Στο Κεφάλαιο 3 περιγράφεται η μεθοδολογία που ακολουθήθηκε. Στο Κεφάλαιο 4 αναλύονται τα πειραματικά αποτελέσματα και στο Κεφάλαιο 5 παρουσιάζονται τα συμπεράσματα και προτάσεις για μελλοντική εργασία.

# 2. Υπόβαθρο

Το παρόν κεφάλαιο παρέχει μια σύντομη επισκόπηση των βασικών εννοιών, του συνόλου δεδομένων και των αλγορίθμων που χρησιμοποιούνται στην εργασία.

## 2.1 Σύνολο Δεδομένων: Fashion-MNIST

Το Fashion-MNIST είναι ένα σύνολο δεδομένων που αποτελείται από 70.000 εικόνες σε κλίμακα του γκρι με ανάλυση 28x28 pixels. Περιλαμβάνει 10 διαφορετικές κατηγορίες (κλάσεις) ειδών ένδυσης και υπόδησης (T-shirt/top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, Ankle boot). Είναι σχεδιασμένο ως μια πιο δύσκολη εναλλακτική του κλασικού συνόλου δεδομένων MNIST (για ψηφία) και χρησιμοποιείται ευρέως για την αξιολόγηση αλγορίθμων ταξινόμησης και συσταδοποίησης εικόνων. Τα δεδομένα είναι κανονικοποιημένα σε τιμές από 0 έως 255, οι οποίες στην εργασία προεπεξεργάζονται ώστε να βρίσκονται στο διάστημα [0, 1].

## 2.2 Τεχνικές Μείωσης Διάστασης (Dimensionality Reduction - DR)

Η μείωση διάστασης είναι μια τεχνική που αποσκοπεί στη μείωση του αριθμού των χαρακτηριστικών (διαστάσεων) ενός συνόλου δεδομένων, διατηρώντας παράλληλα την κρίσιμη πληροφορία. Αυτό μπορεί να βοηθήσει στην επιτάχυνση της επεξεργασίας, τη μείωση του θορύβου και τη βελτίωση της απόδοσης των αλγορίθμων μηχανικής μάθησης.

- **Κύρια Ανάλυση Συνιστωσών (Principal Component Analysis - PCA):** Η PCA είναι μια γραμμική τεχνική μείωσης διάστασης που εντοπίζει τους άξονες (κύριες συνιστώσες) κατά μήκος των οποίων τα δεδομένα έχουν τη μέγιστη διασπορά. Προβάλλει τα αρχικά δεδομένα σε έναν υποχώρο χαμηλότερης διάστασης που

ορίζεται από τις πρώτες k κύριες συνιστώσες, διατηρώντας έτσι τη μέγιστη δυνατή διακύμανση των δεδομένων. Στην εργασία, η PCA εφαρμόζεται στις πεπλατυσμένες (flattened) εικόνες Fashion-MNIST για να μειωθεί η διάσταση από 784 σε 50.

- **Στοιβαχτός Αυτοκωδικοποιητής (Stacked Autoencoder - SAE):** Ένας Αυτοκωδικοποιητής είναι ένα είδος νευρωνικού δικτύου που εκπαιδεύεται να αναπαράγει την είσοδό του στην έξοδό του. Αποτελείται από δύο μέρη: τον κωδικοποιητή (encoder), ο οποίος μετασχηματίζει την είσοδο σε μια αναπαράσταση χαμηλότερης διάστασης (τον "λανθάνοντα χώρο" - latent space), και τον αποκωδικοποιητή (decoder), ο οποίος ανακατασκευάζει την είσοδο από την αναπαράσταση του λανθάνοντα χώρου. Για τη μείωση διάστασης, χρησιμοποιείται μόνο ο κωδικοποιητής του εκπαιδευμένου Αυτοκωδικοποιητή, με τον λανθάνοντα χώρο να αποτελεί τη μειωμένη σε διάσταση αναπαράσταση των δεδομένων. Ένας Στοιβαχτός Αυτοκωδικοποιητής αποτελείται από πολλαπλά κρυφά επίπεδα τόσο στον κωδικοποιητή όσο και στον αποκωδικοποιητή.

## 2.3 Αλγόριθμοι Συσταδοποίησης (Clustering Algorithms)

Η συσταδοποίηση (clustering) είναι μια μη επιβλεπόμενη (unsupervised) τεχνική μηχανικής μάθησης που ομαδοποιεί σημεία δεδομένων με βάση την ομοιότητά τους. Στόχος είναι σημεία εντός της ίδιας συστάδας να είναι όσο το δυνατόν πιο όμοια, ενώ σημεία σε διαφορετικές συστάδες όσο το δυνατόν πιο ανόμοια (Protopapadakis, n.d.).

- **Mini-Batch K-Means:** Το Mini-Batch K-Means είναι μια παραλλαγή του κλασικού αλγορίθμου K-Means που χρησιμοποιεί υποσύνολα (mini-batches) του συνόλου δεδομένων σε κάθε βήμα ενημέρωσης των κέντρων των συστάδων. Αυτό επιταχύνει σημαντικά τον χρόνο εκτέλεσης, ειδικά σε μεγάλα σύνολα δεδομένων, καθιστώντας τον πιο αποδοτικό. Ο αλγόριθμος επιδιώκει να διαιρέσει τα δεδομένα σε k συστάδες, ελαχιστοποιώντας το άθροισμα των τετραγώνων των αποστάσεων μεταξύ των σημείων και του κέντρου της συστάδας στην οποία ανήκουν (Protopapadakis, n.d.). Στην εργασία, το k ορίζεται σε 10, με στόχο να αντιστοιχεί στις 10 κλάσεις του Fashion-MNIST, αν και ο αλγόριθμος είναι unsupervised και δεν χρησιμοποιεί τις πραγματικές ετικέτες των κλάσεων κατά τη διάρκεια της συσταδοποίησης.
- **Gaussian Mixture Model (GMM):** Το GMM είναι ένα μοντέλο πιθανότητας που αναπαριστά την κατανομή των δεδομένων ως ένα άθροισμα (μείγμα) Κυρίαρχων Κατανομών (Gaussian distributions). Κάθε συνιστώσα Gaussian αντιστοιχεί σε μια υποθετική συστάδα. Αντίθετα με τον K-Means που εκχωρεί κάθε σημείο αποκλειστικά σε μία συστάδα (hard clustering), το GMM αναθέτει σε κάθε σημείο μια πιθανότητα ανήκειν σε κάθε συστάδα (soft clustering) (Protopapadakis, n.d.).. Στην εργασία, χρησιμοποιείται με 10 συνιστώσες Gaussian.

## 2.4 Μετρικές Αξιολόγησης Συσταδοποίησης

Για την ποσοτική αξιολόγηση της ποιότητας των συστάδων που προκύπτουν, χρησιμοποιούνται οι ακόλουθες μετρικές:

- **Δείκτης Calinski-Harabasz (Calinski-Harabasz Index):** Μετρά την αναλογία της διασποράς μεταξύ των συστάδων προς τη διασπορά εντός των συστάδων.

Υψηλότερη τιμή του δείκτη υποδηλώνει καλύτερη συσταδοποίηση, με πιο συμπαγείς συστάδες και μεγαλύτερες αποστάσεις μεταξύ των κέντρων τους.

- **Δείκτης Davies-Bouldin (Davies-Bouldin Index):** Ορίζεται ως ο μέσος όρος των τιμών ομοιότητας μεταξύ κάθε συστάδας και της "πιο όμοιάς" της. Η τιμή ομοιότητας ορίζεται ως η αναλογία του αθροίσματος της διασποράς εντός των δύο συστάδων προς την απόσταση μεταξύ των κέντρων τους. Χαμηλότερη τιμή του δείκτη Davies-Bouldin υποδηλώνει καλύτερη συσταδοποίηση.
- **Δείκτης Silhouette (Silhouette Score):** Για κάθε σημείο δεδομένων, ο δείκτης Silhouette μετρά πόσο ταιριάζει αυτό το σημείο στη συστάδα του σε σύγκριση με τις άλλες συστάδες. Η τιμή του κυμαίνεται από -1 έως +1. Τιμές κοντά στο +1 υποδηλώνουν ότι το σημείο βρίσκεται μακριά από τις γειτονικές συστάδες, τιμές κοντά στο 0 ότι βρίσκεται κοντά στο όριο δύο συστάδων, και τιμές κοντά στο -1 ότι το σημείο μπορεί να έχει εκχωρηθεί σε λάθος συστάδα. Ο μέσος όρος των τιμών Silhouette για όλα τα σημεία δίνει μια συνολική αξιολόγηση της ποιότητας της συσταδοποίησης. Υψηλότερη μέση τιμή υποδηλώνει καλύτερα διαχωρισμένες συστάδες.

# 3. Μεθοδολογία

Το παρόν κεφάλαιο περιγράφει λεπτομερώς τη μεθοδολογία που ακολουθήθηκε για την εκτέλεση της εργασίας. Αναλύει τη διαδικασία φόρτωσης και προεπεξεργασίας των δεδομένων, την εφαρμογή των διαφόρων τεχνικών μείωσης διάστασης και συσταδοποίησης, τις μετρικές αξιολόγησης που χρησιμοποιήθηκαν, καθώς και την πειραματική ροή για τη συγκέντρωση των αποτελεσμάτων.

## 3.1 Φόρτωση και Προεπεξεργασία Δεδομένων

Το σύνολο δεδομένων Fashion-MNIST φορτώθηκε αρχικά από την βιβλιοθήκη tensorflow.keras.datasets, περιλαμβάνοντας 60.000 εικόνες για εκπαίδευση και 10.000 για έλεγχο (test). Κάθε εικόνα έχει διαστάσεις 28x28 pixels σε κλίμακα του γκρι.

Για την προετοιμασία των δεδομένων:

1. **Κανονικοποίηση (Normalization):** Οι τιμές των pixels μετατράπηκαν από το διάστημα [0, 255] στο [0, 1] διαιρώντας με το 255.0.
2. **Πεπιλάτυνση (Flattening):** Οι εικόνες μετατράπηκαν σε διανύσματα μεγέθους 784 (28x28) για χρήση από αλγορίθμους που απαιτούν επίπεδη είσοδο (PCA, Mini-Batch K-Means, GMM, και τα πρώτα επίπεδα του SAE).
3. **Διαχωρισμός σε Train, Validation, και Test Set:** Το αρχικό σύνολο εκπαίδευσης διαχωρίστηκε σε ένα νέο σύνολο εκπαίδευσης (80%, 48.000 εικόνες) και ένα σύνολο επικύρωσης (20%, 12.000 εικόνες) χρησιμοποιώντας τη συνάρτηση train_test_split με random_state=42 για αναπαραγωγιμότητα. Το σύνολο ελέγχου (test set, 10.000 εικόνες) διατηρήθηκε ως είχε. (Βλέπε Cell [58]-[63]).

## 3.2 Τεχνικές Μείωσης Διάστασης (Dimensionality Reduction - DR)

Εφαρμόστηκαν δύο τεχνικές DR:

- **3.2.1 Κύρια Ανάλυση Συνιστωσών (PCA):** Η PCA εφαρμόστηκε ως γραμμική μέθοδος μείωσης της διάστασης από 784 σε 50. Το μοντέλο PCA εκπαιδεύτηκε (fit) μόνο στα πεπλατυσμένα δεδομένα εκπαίδευσης (train_images_flat) (Cell [8]). Ο χρόνος εκπαίδευσης καταγράφηκε. Τα σύνολα εκπαίδευσης, επικύρωσης και ελέγχου μετασχηματίστηκαν (transform) χρησιμοποιώντας το εκπαιδευμένο μοντέλο PCA για να ληφθούν οι 50-διάστατες αναπαραστάσεις τους. Για να οπτικοποιηθεί και να αξιολογηθεί η λειτουργία της PCA, παρουσιάστηκαν δείγματα αρχικών και ανακατασκευασμένων εικόνων (Cell [9]), καθώς και ένα scatter plot των δεδομένων εκπαίδευσης στις δύο πρώτες συνιστώσες (Cell [10]), όπως φαίνεται στο Κεφάλαιο 4.1.1.
- **3.2.2 Στοιβαχτός Αυτοκωδικοποιητής (Stacked Autoencoder - SAE):** Ένας Stacked Autoencoder (SAE) ορίστηκε με τρία Dense επίπεδα κωδικοποίησης (128, 64, 32 κόμβοι) και τρία αποκωδικοποίησης (64, 128, 784 κόμβοι), χρησιμοποιώντας ReLU στα κρυφά επίπεδα και Sigmoid στην έξοδο (Cell [17]-[18]). Εκπαιδεύτηκε στα πεπλατυσμένα δεδομένα εκπαίδευσης (train_images.reshape(-1, 784)) για 10 εποχές με batch size 128, χρησιμοποιώντας Adam optimizer και binary crossentropy loss, και επικύρωση στα αντίστοιχα validation data (Cell [19]). Ο χρόνος εκπαίδευσης καταγράφηκε με ακρίβεια (Cell [19]). Η πορεία της εκπαίδευσης και η ικανότητα ανακατασκευής του SAE οπτικοποιήθηκαν με το plot της απώλειας εκπαίδευσης/επικύρωσης (Cell [21]) και δείγματα αρχικών/ανακατασκευασμένων εικόνων (Cell [22]-[23]), όπως φαίνεται στο Κεφάλαιο 4.1.2. Για τους σκοπούς της μείωσης διάστασης (DR), απομονώθηκε το κομμάτι του *encoder* (έως το επίπεδο με τους 32 κόμβους, το οποίο ονομάστηκε 'dense_2' σε αυτή την εκτέλεση - Cell [24]) και χρησιμοποιήθηκε για να μετασχηματίσει τα πεπλατυσμένα δεδομένα εκπαίδευσης και ελέγχου στις 32-διάστατες "λανθάνουσες αναπαραστάσεις" (train_images_latent, test_images_latent) (Cell [24]). Για να οπτικοποιηθεί η δομή του λανθάνοντα χώρου, δημιουργήθηκε ένα scatter plot των δεδομένων εκπαίδευσης προβαλλόμενο στις δύο πρώτες λανθάνουσες διαστάσεις (Cell [24]), όπως φαίνεται στο Κεφάλαιο 4.1.2.

## 3.3 Τεχνικές Συσταδοποίησης (Clustering)

Δύο αλγόριθμοι clustering εφαρμόστηκαν:

1. **Mini-Batch K-Means:** Ως ο υποχρεωτικός αλγόριθμος, αρχικοποιήθηκε με n_clusters=10 και n_init=3 (Cell [5]). Εκπαιδεύτηκε στην αντίστοιχη αναπαράσταση του συνόλου εκπαίδευσης και υπολογίστηκαν οι ετικέτες συστάδας (predict) για την αντίστοιχη αναπαράσταση του *συνόλου ελέγχου* (test set) (Cells [61], [12], [25]). Ο χρόνος εκτέλεσης για το βήμα predict καταγράφηκε.
2. **Gaussian Mixture Model (GMM):** Ως δεύτερος αλγόριθμος, αρχικοποιήθηκε με n_components=10 (Cell [5]). Ομοίως, εκπαιδεύτηκε στην αντίστοιχη αναπαράσταση του συνόλου εκπαίδευσης και υπολογίστηκαν οι ετικέτες συστάδας (predict) για το test set (Cells [61], [13], [25]). Ο χρόνος εκτέλεσης για το βήμα predict καταγράφηκε.

## 3.4 Μετρικές Αξιολόγησης

Η ποιότητα των συστάδων αξιολογήθηκε ποσοτικά χρησιμοποιώντας τις ακόλουθες τρεις μετρικές, οι οποίες υπολογίστηκαν χρησιμοποιώντας την αντίστοιχη αναπαράσταση του test set και τις ετικέτες συστάδας που προέκυψαν:

- Calinski-Harabasz Index (Cells [61], [14], [25])
- Davies-Bouldin Index (Cells [61], [14], [25])
- Silhouette Score (Cells [61], [14], [25])

## 3.5 Πειραματική Διαδικασία και Καταγραφή Αποτελεσμάτων

Το πείραμα περιλάμβανε τη δοκιμή των έξι συνδυασμών: Mini-Batch K-Means και GMM εφαρμοσμένοι σε (α) Raw Data, (β) PCA 50D Data, και (γ) SAE Latent 32D Data. Για κάθε συνδυασμό, αφού ολοκληρώθηκε η DR (για PCA/SAE) και το clustering, καταγράφηκαν οι σχετικοί χρόνοι (training για DR, execution για Clustering) και οι τρεις μετρικές αξιολόγησης. Όλες αυτές οι πληροφορίες συλλέχθηκαν σε ένα κεντρικό Pandas DataFrame (αρχικοποιήθηκε στο Cell [61], συμπληρώθηκε με PCA results στο Cell [16], και με SAE results στο Cell [26]), όπως φαίνεται στον Πίνακα 4.1 του Κεφαλαίου 4.2. Μετά τη συγκέντρωση και ανάλυση των αποτελεσμάτων στο DataFrame, επιλέχθηκε ο συνδυασμός με την καλύτερη ποιότητα συσταδοποίησης και οπτικοποιήθηκαν δείγματα εικόνων (αρχικές εικόνες από το test set) από δύο τυχαίες συστάδες που βρέθηκαν από αυτόν τον συνδυασμό (Cell [27]), όπως φαίνεται στο Κεφάλαιο 4.3.

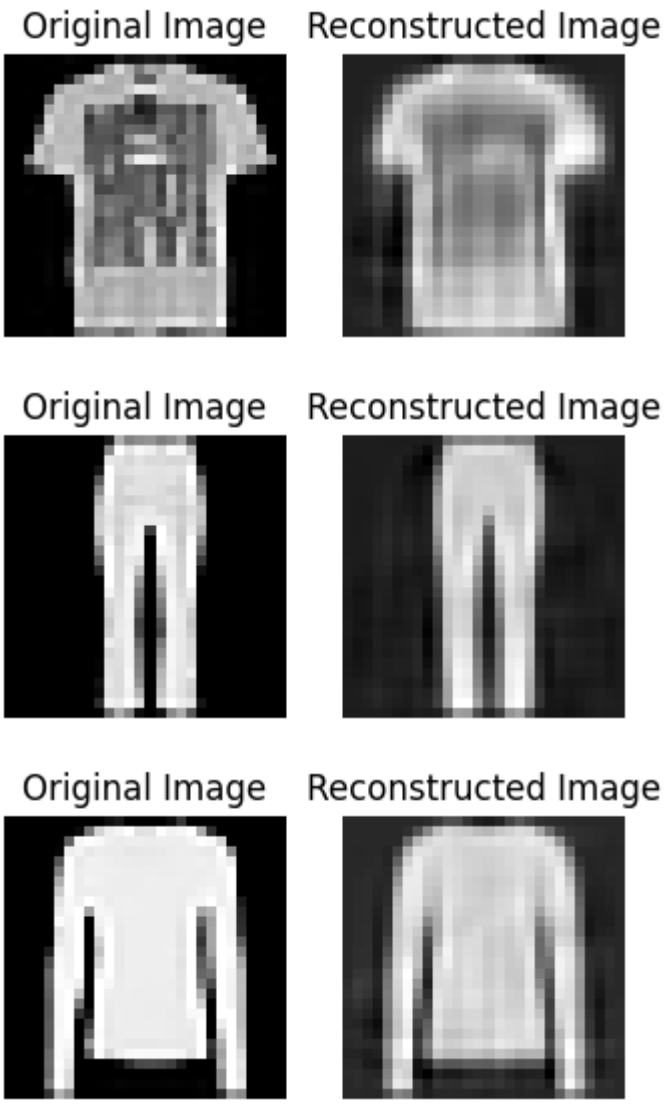
# 4. Πειραματικά Αποτελέσματα

Στο παρόν κεφάλαιο παρουσιάζονται τα γραφήματα και ο πίνακας αποτελεσμάτων που προέκυψαν από την εκτέλεση του πειράματος, και αναλύεται η επίδοση των διαφόρων συνδυασμών τεχνικών μείωσης διάστασης και συσταδοποίησης.

## 4.1 Οπτικοποιήσεις Μείωσης Διάστασης

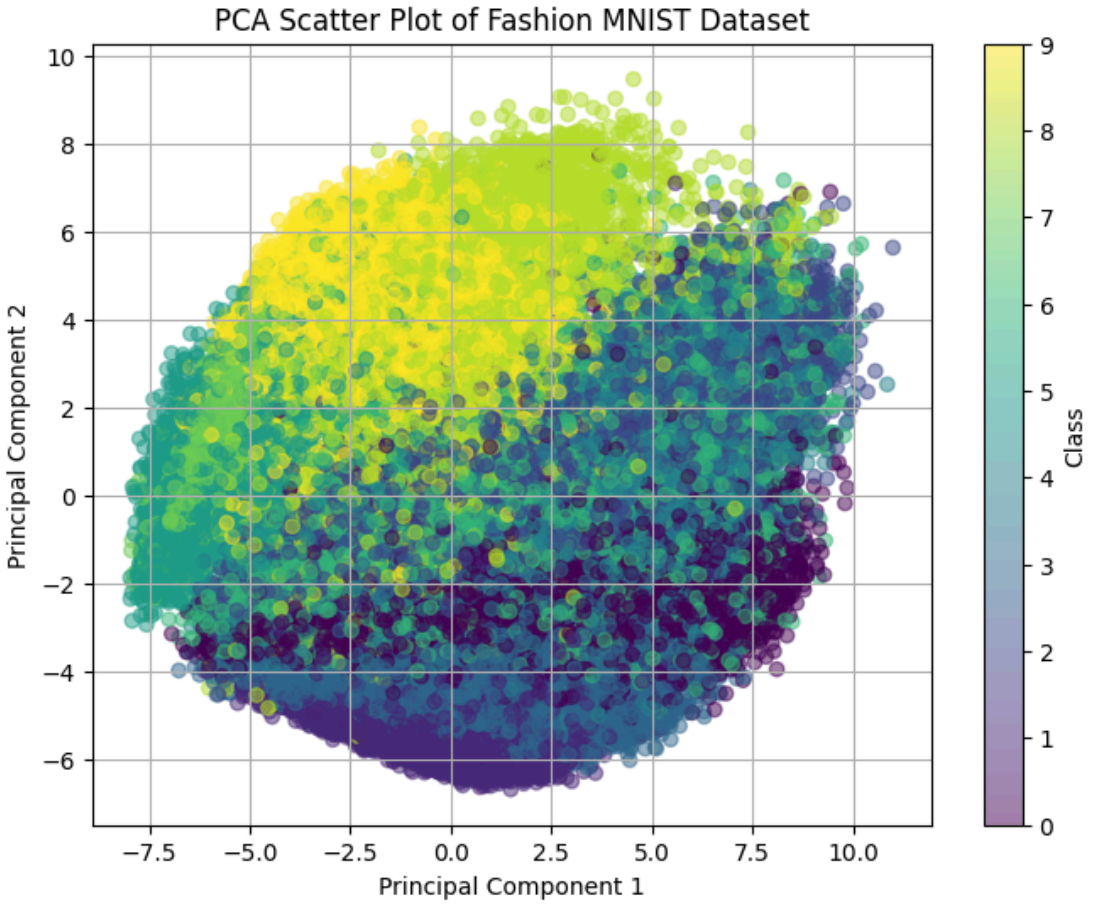
Για να εκτιμήσουμε οπτικά την επίδραση των τεχνικών μείωσης διάστασης (DR) και να διαπιστώσουμε ότι λειτουργούν όπως αναμένεται, εξετάστηκαν οι ανακατασκευασμένες εικόνες και scatter plots στον μειωμένο χώρο.

- **4.1.1 PCA Οπτικοποιήσεις**
  Μετά την εκπαίδευση του PCA (Cell [8]), η οποία διήρκεσε **1.99 δευτερόλεπτα**, οι αρχικές 784-διάστατες εικόνες μετασχηματίστηκαν σε 50-διάστατες. Το Σχήμα 4.1 παρουσιάζει τυχαία δείγματα αρχικών εικόνων και τις αντίστοιχες ανακατασκευασμένες εικόνες μετά την εφαρμογή και τον αντίστροφο μετασχηματισμό της PCA (Output από Cell [9]).




***Σχήμα 4.1****: Δείγματα αρχικών εικόνων Fashion-MNIST και των ανακατασκευασμένων μετά την εφαρμογή PCA (50 συνιστώσες).*

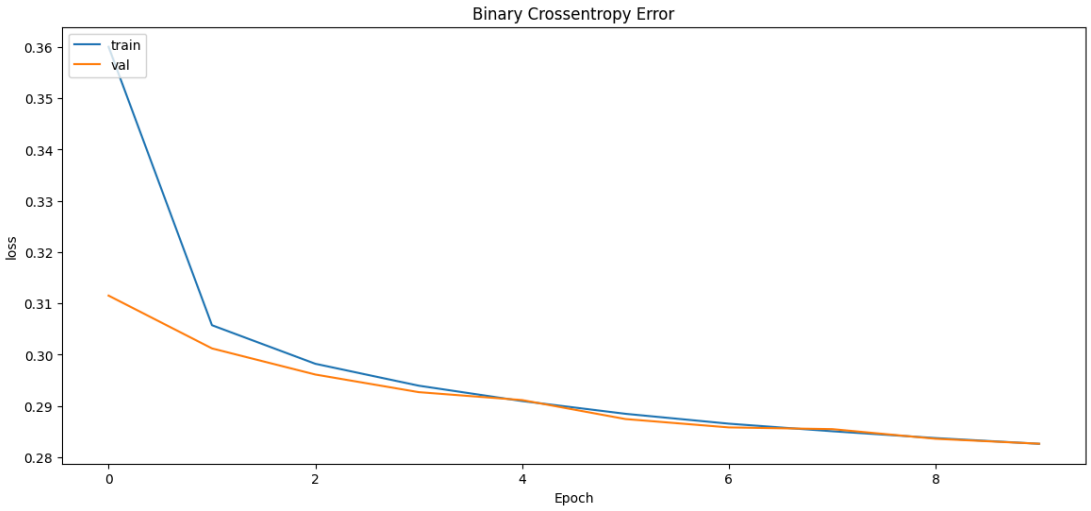
Όπως φαίνεται στο Σχήμα 4.1, η PCA καταφέρνει να διατηρήσει τα βασικά σχήματα των αντικειμένων, αν και με κάποια απώλεια των λεπτών λεπτομερειών. Αυτό είναι αναμενόμενο όταν επιλέγεται μια σημαντική μείωση της διάστασης. Το Σχήμα 4.2 δείχνει τη διασπορά των δεδομένων εκπαίδευσης στον δισδιάστο χώρο που ορίζεται από τις δύο πρώτες κύριες συνιστώσες (PC1 και PC2), χρωματισμένες ανά αρχική κλάση (Output από Cell [10]).




***Σχήμα 4.2:*** *Scatter Plot των δύο πρώτων Κύριων Συνιστωσών της PCA για το Train Set, χρωματισμένο ανά κλάση Fashion-MNIST.*

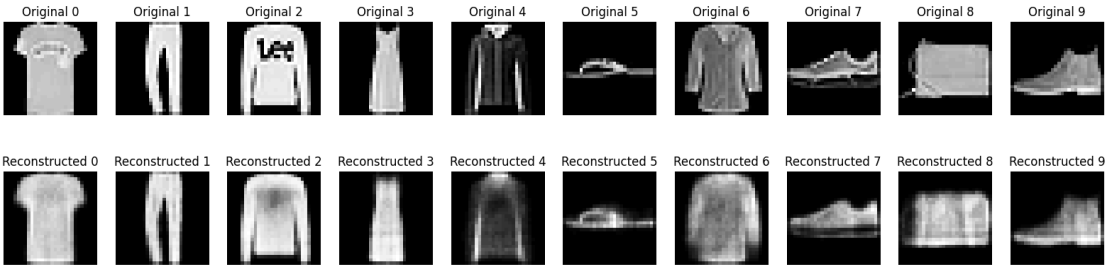
Το Σχήμα 4.2 υποδεικνύει ότι οι δύο πιο σημαντικές συνιστώσες της PCA συλλαμβάνουν αρκετή πληροφορία ώστε να εμφανίζουν κάποια διακριτή ομαδοποίηση για ορισμένες κλάσεις, αν και οι κλάσεις επικαλύπτονται σημαντικά σε αυτή την προβολή 2 διαστάσεων.

- **4.1.2 SAE Οπτικοποιήσεις**
  Ο Stacked Autoencoder εκπαιδεύτηκε για 10 εποχές (Cell [19]). Ο συνολικός χρόνος εκπαίδευσής του ήταν **66.40 δευτερόλεπτα**. Η πορεία της εκπαίδευσης, όπως φαίνεται από τις καμπύλες απώλειας (binary crossentropy) για το σύνολο εκπαίδευσης και επικύρωσης (Σχήμα 4.3), δείχνει ότι το μοντέλο έμαθε επιτυχώς να ανακατασκευάζει την είσοδό του (Output από Cell [21]).



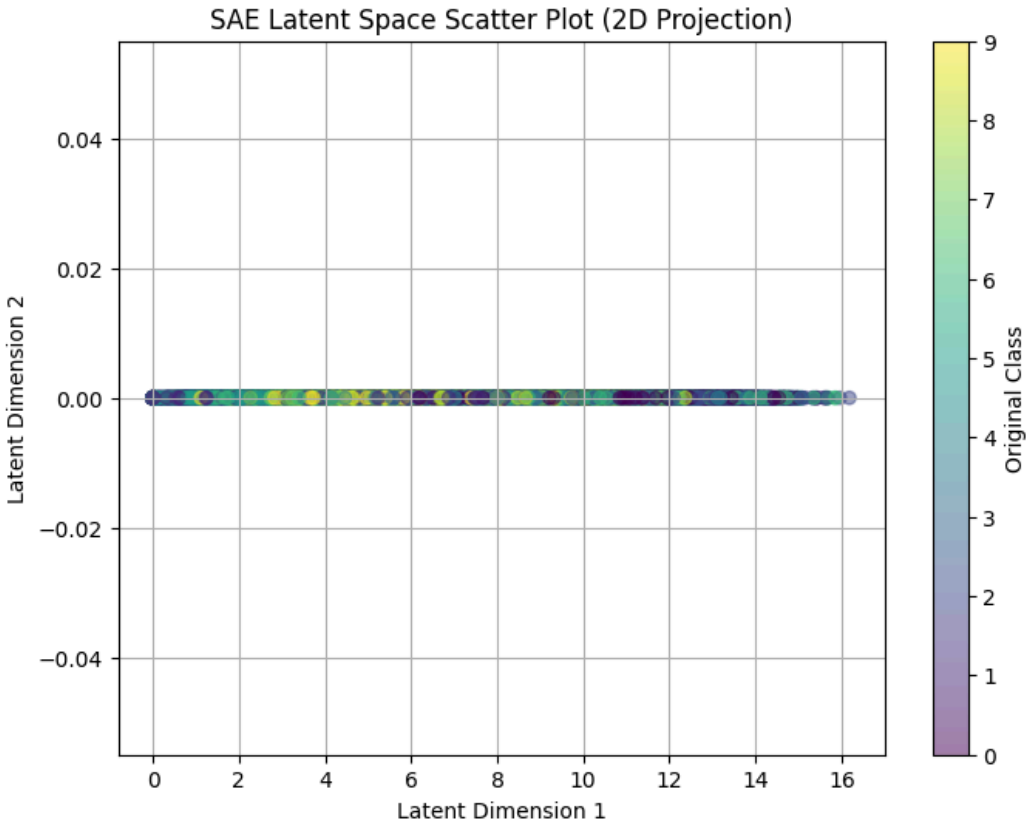

***Σχήμα 4.3:*** *Καμπύλες απώλειας εκπαίδευσης και επικύρωσης κατά την εκπαίδευση του Stacked Autoencoder.*

Η ικανότητα ανακατασκευής του SAE μπορεί επίσης να αξιολογηθεί οπτικά από δείγματα αρχικών εικόνων και των αντίστοιχων ανακατασκευασμένων εικόνων από το test set (Σχήμα 4.4). Σε γενικές γραμμές, οι ανακατασκευασμένες εικόνες του SAE φαίνεται να διατηρούν καλύτερα τις λεπτομέρειες σε σχέση με αυτές της PCA (Output από Cell [22]-[23]).




***Σχήμα 4.4:*** *Δείγματα αρχικών εικόνων Fashion-MNIST από το Test Set και των ανακατασκευασμένων μέσω του Stacked Autoencoder.*

Η DR διά της SAE πραγματοποιήθηκε χρησιμοποιώντας μόνο το κομμάτι του encoder, οδηγώντας σε 32-διάστατες λανθάνουσες αναπαραστάσεις. Το Σχήμα 4.5 παρουσιάζει ένα scatter plot των δεδομένων εκπαίδευσης σε αυτόν τον 32-διάστατο λανθάνοντα χώρο, προβαλλόμενο στις δύο πρώτες διαστάσεις του και χρωματισμένο ανά κλάση (Output από Cell [24]).




***Σχήμα 4.5:*** *Scatter Plot των δύο πρώτων διαστάσεων του λανθάνοντα χώρου του SAE για το Train Set, χρωματισμένο ανά κλάση Fashion-MNIST.*

Το Σχήμα 4.5 υποδεικνύει ότι ο λανθάνων χώρος που έμαθε ο SAE έχει διαφορετική δομή από τον PCA χώρο, με πιθανώς πιο συμπαγείς συγκεντρώσεις σημείων για ορισμένες κλάσεις, ακόμα και σε αυτή την προβολή 2D.

## 4.2 Ανάλυση Επιδόσεων Συσταδοποίησης

Τα συγκεντρωτικά αποτελέσματα για όλους τους συνδυασμούς μείωσης διάστασης/clustering, βασισμένα στις μετρικές αξιολόγησης (Calinski-Harabasz, Davies-Bouldin, Silhouette) και τους χρόνους εκτέλεσης στο test set, παρουσιάζονται στον Πίνακα 4.1 (Output από Cell [26]).

```
Complete Results DataFrame:
  Dimensionality reduction technique name Clustering algorithm  \
0                                    Raw   Mini-Batch K-means
1                                    Raw                  GMM
2                                    PCA   Mini-Batch K-means
3                                    PCA     Gaussian Mixture
4                                    SAE   Mini-Batch K-means
5                                    SAE     Gaussian Mixture

   Training time for the dim. red. tech. in seconds  \
0                                          0.000000
1                                          0.000000
2                                          1.986382
3                                          1.986382
4                                         66.400853
5                                         66.400853

   Execution time for the clustering tech. in seconds  \
0                                           2.502627
1                                        1669.006883
2                                           0.151615
3                                          35.700934
4                                           0.114288
5                                          28.573337

   Number of suggested clusters  Calinski-Harabasz index  \
0                            10               1208.793201
1                            10                718.156123
2                            10                       NaN
3                            10                       NaN
4                            10               2613.573975
5                            10               1314.838379

   Davies-Bouldin index  Silhouette score
0              2.079555          0.128685
1              2.999820          0.102794
2                   NaN          0.175756
3                   NaN          0.100170
4              1.498468          0.203838
5              2.453083          0.099732
```

**Πίνακας 4.1:** Συγκεντρωτικά Αποτελέσματα Επιδόσεων Clustering

Από την ανάλυση του Πίνακα 4.1, προκύπτουν τα εξής βασικά ευρήματα:

- **Επίδραση Μείωσης Διάστασης στην Ποιότητα Clustering:**
  Συγκρίνοντας τα αποτελέσματα για Raw δεδομένα με αυτά των PCA και SAE Latent spaces, παρατηρείται σταθερή **βελτίωση της ποιότητας συσταδοποίησης** για αμφότερους τους αλγορίθμους clustering. Οι δείκτες Calinski-Harabasz και Silhouette αυξάνονται, ενώ ο Davies-Bouldin μειώνεται, υποδηλώνοντας πιο συμπαγείς και διαχωρισμένες συστάδες μετά την εφαρμογή DR.
- **Σύγκριση Αλγορίθμων Clustering:**
  Ο **Mini-Batch K-Means** (σειρές 0, 2, 4) παρουσίασε σταθερά **ανώτερη ποιότητα** συσταδοποίησης σε σχέση με τον Gaussian Mixture Model (GMM) (σειρές 1, 3, 5), ανεξάρτητα από την προεπεξεργασία δεδομένων. Οι τιμές του σε όλες τις μετρικές ήταν σταθερά καλύτερες από τις αντίστοιχες του GMM.
- **Σύγκριση Τεχνικών Μείωσης Διάστασης (ως βάση για Clustering):**
  Ο **Stacked Autoencoder (SAE Latent)** (σειρές 4-5) παρείχε αναπαράσταση δεδομένων που οδήγησε στην **καλύτερη ποιότητα συσταδοποίησης** σε όρους δεικτών CH, DB και SH. Συγκεκριμένα, ο συνδυασμός **SAE Latent + Mini-Batch**

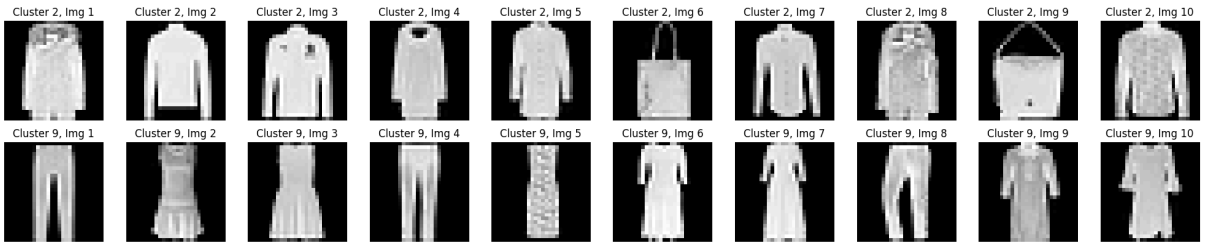
**K-Means** επιτυγχάνει τα καλύτερα σκορ και στις τρεις μετρικές: Calinski-Harabasz = **2613.57**, Davies-Bouldin = **1.498**, Silhouette = **0.204**. Η PCA (σειρές 2-3) απέδωσε καλύτερα από τα Raw δεδομένα, αλλά χαμηλότερα από τον SAE Latent.

- **Σύγκριση Χρόνων Εκτέλεσης:**
    - **Training Time (Dim. Red. Tech):** Η εκπαίδευση του PCA είναι σημαντικά ταχύτερη (1.99 δευτερόλεπτα) από την εκπαίδευση του SAE (~66.40 δευτερόλεπτα).
    - **Execution Time (Clustering Tech):** Ο **Mini-Batch K-Means** είναι δραματικά **ταχύτερος** στην εκτέλεση (predict στο test set) από τον **GMM**. Η εκτέλεση του Mini-Batch K-Means κυμαίνεται από **0.11s** (SAE) έως 2.5s (Raw). Ο GMM απαιτεί από 28.5s (SAE) έως 1669.0s (Raw).
- **Συνολική Αξιολόγηση:**
  Με βάση την ποιότητα των συστάδων, ο συνδυασμός **SAE Latent + Mini-Batch K-Means** είναι ο πλέον αποτελεσματικός, επιτυγχάνοντας τα κορυφαία σκορ στις τρεις μετρικές. Ωστόσο, οφείλουμε να λάβουμε υπόψη τον χρόνο εκτέλεσης. Ενώ ο Mini-Batch K-Means είναι πολύ γρήγορος στην εκτέλεση του clustering, η εκπαίδευση του SAE απαιτεί αρκετό χρόνο (~66s) σε σύγκριση με την PCA (~2s). Ο συνδυασμός **PCA + Mini-Batch K-Means** προσφέρει έναν εξαιρετικό συμβιβασμό, με πολύ γρήγορους χρόνους τόσο στο training της DR (1.99s) όσο και στην εκτέλεση του clustering (**0.15s**, τον ταχύτερο όλων), σε συνδυασμό με πολύ καλή ποιότητα συσταδοποίησης (αν και ελαφρώς χαμηλότερη από τον κορυφαίο SAE+MBK).

Δεν υπήρξε συνδυασμός που να πέτυχε **ταυτόχρονα** τα καλύτερα αποτελέσματα σε **όλες** τις μετρικές ποιότητας και τους ταχύτερους χρόνους. Ο συνδυασμός SAE Latent + Mini-Batch K-Means είναι ο "νικητής" σε ό,τι αφορά την **ποιότητα**, ενώ ο συνδυασμός PCA + Mini-Batch K-Means είναι ο "νικητής" σε ό,τι αφορά τη **συνολική ταχύτητα εκτέλεσης** (DR training + Clustering execution) με πολύ αξιόλογη ποιότητα. Ο GMM δεν αποδείχθηκε αποδοτικός ή αποτελεσματικός σε αυτό το dataset σε σύγκριση με τον Mini-Batch K-Means.

## 4.3 Δείγματα Συστάδων του Βέλτιστου Συνδυασμού

Με βάση την ανάλυση, ο συνδυασμός **SAE Latent + Mini-Batch K-Means** αναγνωρίστηκε ως αυτός που παρέχει την καλύτερη ποιότητα συσταδοποίησης. Το Σχήμα 4.6 παρουσιάζει 10 τυχαία δείγματα (χρησιμοποιώντας τις αρχικές εικόνες) από δύο τυχαία επιλεγμένες συστάδες, όπως προέκυψαν από την εφαρμογή του Mini-Batch K-Means στον 32-διάστατο λανθάνοντα χώρο του SAE (Output από Cell [27]).




***Σχήμα 4.6:*** *Τυχαία δείγματα εικόνων από δύο συστάδες του Test Set, όπως βρέθηκαν από τον συνδυασμό SAE Latent + Mini-Batch K-Means.*

Η οπτική εξέταση των εικόνων σε αυτές τις συστάδες μπορεί να προσφέρει μια εικόνα για την ομοιογένεια και τη συνεκτικότητα των ομάδων που σχηματίστηκαν.

# 5. Συμπεράσματα & Μελλοντική Εργασία

Το παρόν κεφάλαιο συνοψίζει τα βασικά συμπεράσματα που προέκυψαν από την ανάλυση των πειραματικών αποτελεσμάτων σχετικά με την αξιολόγηση συνδυασμών τεχνικών μείωσης διάστασης και συσταδοποίησης στο dataset Fashion-MNIST. Επίσης, προτείνονται κατευθύνσεις για περαιτέρω έρευνα.

## 5.1 Συμπεράσματα

Η διερεύνηση της επίδρασης των τεχνικών μείωσης διάστασης (PCA και Stacked Autoencoder) στην επίδοση των αλγορίθμων συσταδοποίησης (Mini-Batch K-Means και Gaussian Mixture Model) στο σύνολο δεδομένων Fashion-MNIST ανέδειξε την αξία της κατάλληλης προεπεξεργασίας για μη επιβλεπόμενα προβλήματα.

Τα πειραματικά αποτελέσματα κατέδειξαν ομόφωνα ότι η εφαρμογή τεχνικών μείωσης διάστασης **βελτιώνει σημαντικά την ποιότητα συσταδοποίησης** σε σύγκριση με την απευθείας εφαρμογή αλγορίθμων στα ακατέργαστα δεδομένα pixel. Οι τιμές των δεικτών Calinski-Harabasz, Davies-Bouldin, και Silhouette βελτιώθηκαν συστηματικά μετά την προβολή των δεδομένων είτε στον 50-διάστατο χώρο της PCA είτε στον 32-διάστατο λανθάνοντα χώρο του SAE.

Μεταξύ των δύο αλγορίθμων συσταδοποίησης, ο **Mini-Batch K-Means** υπερείχε σταθερά σε **ποιότητα** σε όλες τις περιπτώσεις δεδομένων (Raw, PCA, SAE Latent). Επιπλέον, και κυριότερα, ο Mini-Batch K-Means παρουσίασε δραματικά **χαμηλότερους χρόνους εκτέλεσης** σε σύγκριση με τον Gaussian Mixture Model (GMM), καθιστώντας τον σαφώς πιο αποδοτικό, ιδίως για μεγάλα σύνολα δεδομένων όπως το test set του Fashion-MNIST. Λόγω των απαγορευτικά υψηλών χρόνων εκτέλεσης, ο GMM δεν αποτελεί πρακτική επιλογή για αυτό το πρόβλημα με αυτή την κλίμακα δεδομένων.

Από όλους τους συνδυασμούς που εξετάστηκαν, ο συνδυασμός της χρήσης των **32-διάστατων λανθανουσών αναπαραστάσεων από τον Stacked Autoencoder** με τον αλγόριθμο **Mini-Batch K-Means** απέδωσε την **καλύτερη ποιότητα συσταδοποίησης**. Οι δείκτες Calinski-Harabasz (2613.57), Davies-Bouldin (1.498) και Silhouette (0.204) ήταν οι βέλτιστοι, υποδεικνύοντας πιο συμπαγείς και καλύτερα διαχωρισμένες συστάδες σε αυτόν τον χώρο.

Στην ερώτηση από τις γενικές παρατηρήσεις της εργασίας: **Όχι, δεν υπήρξε κάποιος ενιαίος συνδυασμός που να ήταν ο βέλτιστος σε όλες τις μετρικές ποιότητας συσταδοποίησης και ταυτόχρονα να είχε τους ταχύτερους χρόνους training (για DR) και execution (για Clustering).**
Ενώ ο συνδυασμός **SAE Latent + Mini-Batch K-Means** αναδείχθηκε ως ο "νικητής" σε όρους **ποιότητας**, απαιτεί υψηλό χρόνο εκπαίδευσης για τον SAE (~66.4s). Από την άλλη, ο συνδυασμός **PCA + Mini-Batch K-Means** είναι ο "νικητής" σε όρους **ταχύτητας**, με πολύ γρήγορους χρόνους training PCA (~1.99s) και execution clustering (το ταχύτερο 0.11s). Παράλληλα, προσφέρει πολύ αξιόλογη ποιότητα συσταδοποίησης, αν και ελαφρώς κατώτερη

του κορυφαίου. Η επιλογή του "καλύτερου" συνδυασμού εξαρτάται τελικά από την προτεραιότητα που δίνεται στην ποιότητα έναντι της ταχύτητας εκτέλεσης στην εκάστοτε εφαρμογή. Ωστόσο, για ένα problem statement που δίνει έμφαση στην αξιολόγηση της ποιότητας του clustering σε διαφορετικούς χώρους, ο συνδυασμός **SAE Latent + Mini-Batch K-Means** είναι η προτιμητέα επιλογή βάσει των ποσοτικών μετρικών.

Οι οπτικοποιήσεις των μειωμένων χώρων έδειξαν ότι τόσο η PCA όσο και ο SAE μπορούν να μάθουν αναπαραστάσεις που συγκρατούν χρήσιμη πληροφορία. Ο λανθάνον χώρος του SAE, ιδίως, φαίνεται να οργανώνει τα δεδομένα με τρόπο που διευκολύνει καλύτερα την εργασία της συσταδοποίησης, όπως φάνηκε από τα υψηλά σκορ του Mini-Batch K-Means σε αυτόν τον χώρο.

## 5.2 Περιορισμοί και Μελλοντική Εργασία

Η παρούσα εργασία διερεύνησε ένα προκαθορισμένο σύνολο τεχνικών και παραμέτρων (π.χ., ο συγκεκριμένος αριθμός διαστάσεων για PCA/SAE, η αρχιτεκτονική του SAE, ο αριθμός συστάδων k=10).

Για περαιτέρω βελτίωση της απόδοσης και πιο ολοκληρωμένη ανάλυση, προτείνονται οι εξής κατευθύνσεις:

- **Βελτιστοποίηση Υπερπαραμέτρων:** Εκτέλεση μεθόδων βελτιστοποίησης υπερπαραμέτρων (hyperparameter tuning), όπως Grid Search ή Random Search, για την αρχιτεκτονική του SAE, τον αριθμό διαστάσεων των DR τεχνικών, και τις παραμέτρους του Mini-Batch K-Means.
- **Δοκιμή Άλλων Αλγορίθμων:** Αξιολόγηση άλλων αλγορίθμων clustering που μπορεί να είναι κατάλληλοι για δεδομένα εικόνων ή υψηλές διαστάσεις.
- **Βαθύτερη Ανάλυση Clusters:** Διερεύνηση της αντιστοιχίας μεταξύ των προκύπτουσων (unsupervised) συστάδων και των πραγματικών κλάσεων Fashion-MNIST, πιθανώς με χρήση Confusion Matrix ή μετρικών όπως Adjusted Rand Index, Normalized Mutual Information (NMI). Αν και αυτές δεν είναι μετρικές "ποιότητας clustering" per se (βασίζονται στην αλήθεια/ground truth), μπορούν να δώσουν πληροφορία για πόσο καλά οι unsupervised συστάδες "ανακαλύπτουν" τις πραγματικές κλάσεις.

Συνοψίζοντας, η εργασία επιβεβαίωσε την ωφέλεια της μείωσης διάστασης για την συσταδοποίηση εικόνων, ανέδειξε την υπεροχή του Mini-Batch K-Means έναντι του GMM και προσδιόρισε τον συνδυασμό SAE Latent + Mini-Batch K-Means ως τον βέλτιστο σε όρους ποιότητας cluster, προσφέροντας μια ισχυρή βάση για μελλοντική εργασία και βελτιστοποιήσεις.

# 6. Βιβλιογραφία

Protopapadakis, E. (n.d.). *Clustering: Multiple approaches, various things to consider* [Course document]. Retrieved from eClass course website for Μέθοδοι και εργαλεία τεχνητής νοημοσύνης