Ελένη Γεωργούδη – ics23103

Όλγα Σαρίδου – ics23078

Βαλεντίνος Γούδας – ics23170

Θρασύβουλος Βασδαβάνος – ics23168

Dimensionality Reduction & Clustering

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Applied Informatics UOM 2025-26

TABLE OF CONTENTS

1. Θεωρητικό Υπόβαθρο

Η παρακάτω έρευνα βασίζεται στην ανάλυση δεδομένων εικόνας υψηλής διάστασης, συγκεκριμένα στο σύνολο δεδομένων Fashion MNIST. Επικεντρώνεται στις **προκλήσεις υψηλής διαστασιμότητας** (high dimensionality) στα δεδομένα εικόνας, όπου κάθε εικόνα είναι ένα πολυδιάστατο σημείο.

Για την αντιμετώπιση των προκλήσεων αυτών, χρησιμοποιήθηκαν τεχνικές μείωσης διάστασης χώρου (dimensionality reduction). Ειδικότερα, χρησιμοποιήθηκε η **Ανάλυση Κύριων Συνιστώσεων** (**PCA** - Principal Component Analysis) για γραμμικούς μετασχηματισμούς, ώστε να μειωθούν τα δεδομένα υψηλής διάστασης σε κύριες συνιστώσες που αποτυπώνουν τη διασπορά. Η δεύτερη τεχνική, Στοιβαγμένοι Αυτόματοι Κωδικοποιητές (**SAE - Stacked Autoencoder**), είναι μία τεχνική βασισμένη στα νευρωνικά δίκτυα, η οποία μειώνει τμηματικά τη διάσταση του χώρου. Τέλος, εφαρμόστηκε η **t-SNE** (Distributed Stochastic Neighbor Embedding) η οποία διατηρεί τις τοπικές δομές δεδομένων, διευκολύνοντας την οπτικοποίηση των συστάδων (cluster) των δεδομένων υψηλής διάστασης.

Στην συνέχεια, η έρευνα εξετάζει **αλγορίθμους ομαδοποίησης** (clustering) για τα επεξεργασμένα δεδομένα. Αρχικά, εφαρμόστηκε ο **MiniBatch KMeans** για την αποδοτικότητά του στη διαχείριση μεγάλου όγκου δεδομένων. Επιπλέον, το **DBSCAN** χρησιμοποιήθηκε για τον εντοπισμό συστάδων με διαφορετικά σχήματα και μεγέθη. Η **Aggregate clustering,** μια ιεραρχική τεχνική ομαδοποίησης για να φέρει στην επιφάνεια πληροφορίες σχετικά με τη δομή των δεδομένων.

Η αποδοτικότητα και η αποτελεσματικότητα των παραπάνω μεθόδων ομαδοποίησης αξιολογείται με μετρικές. Αυτές είναι ο δείκτης **Calinski-Harabasz**, για τη διασπορά των συστάδων, ο **δείκτης Silhouette** για τη συνοχή των συστάδων, ο **δείκτης Davies-Bouldin** για την ομοιότητα των συστάδων μεταξύ τους, και τέλος, ο προσαρμοσμένος δείκτης rand (**adjusted Rand index**) για την ακρίβεια που μπορεί να παρέχει σχετικά με κάποια γνωστή αλήθεια αναφοράς (ground truth).

2. Εισαγωγή

Αυτή η έρευνα επικεντρώνεται στο dataset Fashion MNIST, ένα πρότυπο στη μηχανική μάθηση που περιλαμβάνει 70.000 εικόνες σε γκρι κλίμακα (grayscale) σε 10 κατηγορίες μόδας (classes). Κάθε εικόνα έχει διάταξη 28x28 pixels. Το συγκεκριμένο dataset αποτελείται από δεδομένα υψηλής διάστασης με σύνθετα μοτίβα, συνεπάγοντας έτσι προκλήσεις στη μείωση της διάστασης χώρου και στην ομαδοποίηση.

Κύριος στόχος αυτής της έρευνας είναι η αποδοτική μείωση της διάστασης χώρου, και στη συνέχεια η αποτελεσματική ομαδοποίηση (clustering) των δεδομένων εικόνας. Η ομαδοποίηση, ως ένα σημαντικό συστατικό της μη επιβλεπόμενης μάθησης (unsupervised learning) ομαδοποιεί τις εικόνες με τέτοιο τρόπο ώστε αυτές που θα βρίσκονται στην ίδια συστάδα (cluster) να είναι παρόμοιες μεταξύ τους. Για παράδειγμα όλα τα t-shirt να ανηκουν σε μία συστάδα, η οποία δεν θα περιέχει άλλο είδος μπλούζας. Αυτό είναι εξαιρετικά σημαντικό για την κατανόηση των χαρακτηριστικών των δεδομένων και τη διευκόλυνση της ομαδοποίησης των εικόνων.

Εφαρμόστηκαν διάφορες τεχνικές μείωσης διάστασης χώρου (dimensionality reduction). Η Ανάλυση Κύριων Συνιστώσεων (PCA - Principal Component Analysis) μειώνει τον αριθμό των μεταβλητών (διαστάσεων) διατηρώντας το μεγαλύτερο μέρος της πληροφορίας (διακύμανσης) του αρχικού συνόλου δεδομένων. Ο SAE (Stacked Autoencoder) μαθαίνει τα πιο σημαντικά, συμπυκνωμένα χαρακτηριστικά (features) των δεδομένων εισόδου, ώστε αυτά να μπορούν να ανακατασκευαστούν όσο το δυνατόν ακριβέστερα γίνεται. Ο t-SNE, μια τεχνική μη γραμμική, μετατρέπει πολύπλοκα δεδομένα σε 2D ή 3D αναπαραστάσεις, διατηρώντας κυρίως τις τοπικές σχέσεις μεταξύ των σημείων, ώστε να δημιουργούνται ευδιάκριτες συστάδες (cluster) για σκοπούς οπτικοποίησης.

3. Προτεινόμενη Μεθοδολογία

3.1. Προεπισκόπηση και Προεπεξεργασμένα Δεδομένα

3.1.1. Φόρτωση Δεδομένων

Αρχικά, φορτώθηκαν τα δεδομένα με τη χρήση των βιβλιοθηκών numpy και keras, προκειμένου να ανακτηθεί το σύνολο δεδομένων του Fashion MNIST. Το συγκεκριμένο dataset περιλαμβάνει 70.000 ασπρόμαυρες εικόνες, από τις οποίες μετέπειτα θα διαχωριστούν σε 3 κατηγορίες (sets) για την εκπαίδευση του μοντέλου. Κάθε εικόνα είναι 28x28 pixels, τα οποία μπορούν να αποτυπωθούν από 784 μεμονωμένα σημεία.

3.1.2. Κανονικοποίηση Δεδομένων

Η προεπεξεργασία ξεκινάει με την κανονικοποίηση των δεδομένων (normalization) για να κλιμακωθούν (scaling) οι τιμές των pixel κάθε εικόνας μεταξύ 0 και 1. Το συγκεκριμένο βήμα είναι εξαιρετικά σημαντικό ώστε να έχουμε μια ομοιόμορφη κλίμακα εισόδου και ταχύτερη σύγκλιση και σταθερότητα στη εκμάθηση του μοντέλου. Ακόμη, με αυτόν τον τρόπο αποφεύγονται οι μεγάλες αριθμητικές τιμές, οι οποίες θα είχαν δυσανάλογη επίδραση στην εκμάθηση. Συγκεκριμένα, τα pixel διαιρούνται με το 255, αφού οι τιμές των pixel της κάθε εικόνας κυμαίνονται μεταξύ 0-255. Έτσι, το κάθε pixel παίρνει μία τιμή 0 ή 1.

3.1.3. Test, Train και Validation sets

Το σύνολο δεδομένων διαιρέθηκε σε 3 κατηγορίες (κλάσεις): δεδομένα εκπαίδευσης (train set), τα οποία θα εκπαιδεύσουν το μοντέλο, δεδομένα για επικύρωση (validation set), για την ρύθμιση των υπερπαραμέτρων και την ανίχνευση και πρόληψη του Overfitting, και σε δεδομένα για τεστ του μοντέλου μετά την εκπαίδευση (test set). Το test set αποτελείται από 10.000 εικόνες, το train set από το 80% των υπολειπόμενων εικόνων, δηλαδή 48.000 εικόνες, ενώ το validation set από το 20%, 12.000 εικόνες.

3.1.4. Τεχνικές Dimensionality Reduction

Για την αποτελεσματική μείωση διάστασης χώρου, διατηρώντας παράλληλα τις βασικές πληροφορίες του dataset, χρησιμοποιήθηκαν τρεις τεχνικές, καθεμία από τις οποίες προσφέρει διαφορικά οφέλη.

Αρχικά, η Ανάλυση Κύριων Συνιστωσών (PCA) για τη μετατροπή του συνόλου δεδομένων σε ένα σύνολο από γραμμικά μη συσχετισμένων συνιστωσών. Η συγκεκριμένη μέθοδος εστιάζει στις συνιστώσες με τη μεγαλύτερη διασπορά (variance), μειώνοντας τη διαστασιμότητα, διατηρώντας όμως τη μέγιστη δυνατή πληροφορία.

Έπειτα, ο Στοιβαγμένος Αυτόματος Κωδικοποιητής (SAE), μια τεχνική βασισμένη στα νευρωνικά δίκτυα, συμπιέζει τα δεδομένα σε έναν χώρο μικρότερων διαστάσεων, και στη συνέχεια τα ανακατασκευάζει. Συγκεκριμένα, εφαρμόστηκε μια αρχιτεκτονική η οποία συμπιέζει την αρχική διάσταση, δηλαδή κωδικοποιεί τα δεδομένα, από 784 pixels → 128 → 64 → 32. Η τελική διάσταση, η οποία ονομάζεται λανθάνων χώρος (latent space) ισούται με 32 pixels, και αποτελεί τη μέγιστη δυνατή συμπίεση. Αυτή η συμπίεση αναγκάζει το μοντέλο να μάθει και να αποθηκεύσει μόνο τα απαραίτητα και ουσιώδες χαρακτηριστικά. Στη συνέχεια, ανακατασκευάζονται τα δεδομένα, δηλαδή αποκωδικοποιούνται, από 32 → 64 → 128 → 784.

Τέλος, η μέθοδος t-SNE, μια μη γραμμική τεχνική, χρησιμοποιήθηκε για την απεικόνιση των δεδομένων σε έναν χώρο 2 διαστάσεων (2D). Σε αντίθεση με την PCA και SAE, η t-SNE επικεντρώνεται στη διατήρηση των τοπικών σχέσεων και γειτονιών μεταξύ των δεδομένων, διασφαλίζοντας ότι τα παρόμοια δείγματα παραμένουν το ένα κοντά στο άλλο στον μειωμένο χώρο. Με αυτόν τον τρόπο επιτυγχάνεται μια αποτελεσματική οπτικοποίηση των συστάδων που αναδεικνύει τη δομή του dataset, παρόλο που η συγκεκριμένη τεχνική δεν υποστηρίζει την ανακατασκευή των αρχικών δεδομένων από τον χώρο των χαμηλών διαστάσεων.

3.2 Οπτικοποίηση Αλγορίθμων Μείωσης Διαστάσεων

3.2.1 Οπτικοποίηση PCA και SAE

Οι συναρτήσεις που έχουν γραφτεί στα αντίστοιχα μπλοκ κώδικα στα notebooks, έχουν σχεδιαστεί για να οπτικοποιήσουν τα δεδομένα που έχουν αναχθεί σε δύο διαστάσεις χρησιμοποιώντας είτε την **Ανάλυση Κύριων Συνιστωσών (PCA)** (στο `DR_PCA.ipynb`) είτε έναν **Αυτοκωδικοποιητή Στοίβαξης (SAE)** (στο `Deep_DR_SAE.ipynb`).

Το γράφημα παρουσιάζει ένα δισδιάστατο διάγραμμα διασποράς (scatter plot) όπου κάθε σημείο αντιπροσωπεύει ένα δείγμα των δεδομένων.

- Στην περίπτωση της **PCA**, όπως φαίνεται παρακάτω, ο οριζόντιος (άξονας x) και ο κατακόρυφος (άξονας y) άξονας αντιστοιχούν στην πρώτη και τη δεύτερη **Κύρια Συνιστώσα**, αντίστοιχα. Αυτές οι συνιστώσες είναι οι κατευθύνσεις στις οποίες τα δεδομένα παρουσιάζουν τη μεγαλύτερη διακύμανση και αποτελούν γραμμικούς συνδυασμούς των αρχικών χαρακτηριστικών.

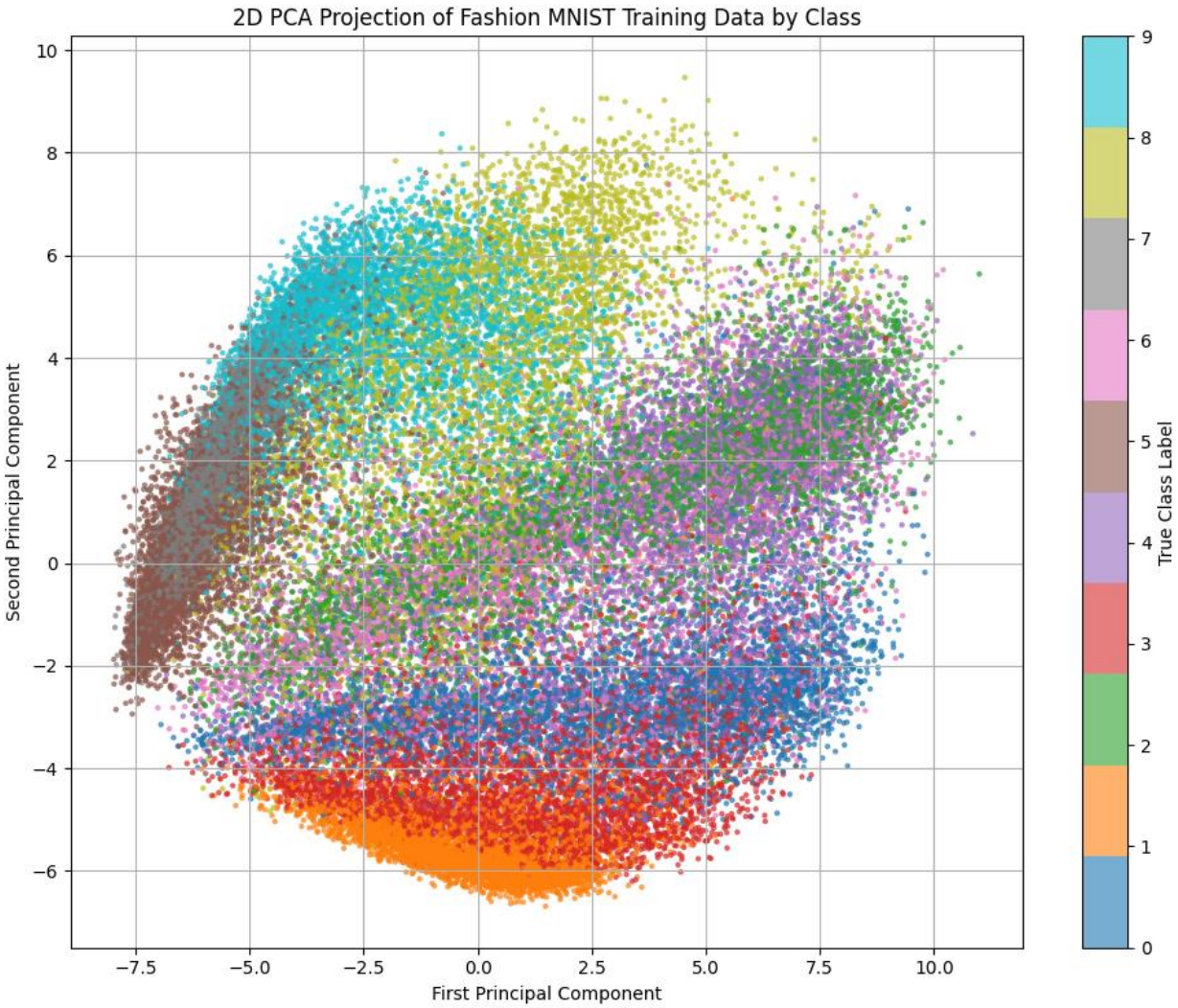


Figure 1: data reduced to two dimensions with PCA

- Για τον **SAE**, όπως φαίνεται στην εικόνα που ακολουθεί, αυτοί οι άξονες αναπαριστούν τις δύο κύριες διαστάσεις στον αναχθέντα χώρο χαρακτηριστικών. Αυτές οι διαστάσεις, που συλλαμβάνονται από το κεντρικό κρυφό επίπεδο του αυτοκωδικοποιητή, αποτυπώνουν σημαντικά μοτίβα ή χαρακτηριστικά από τα δεδομένα υψηλής διάστασης.

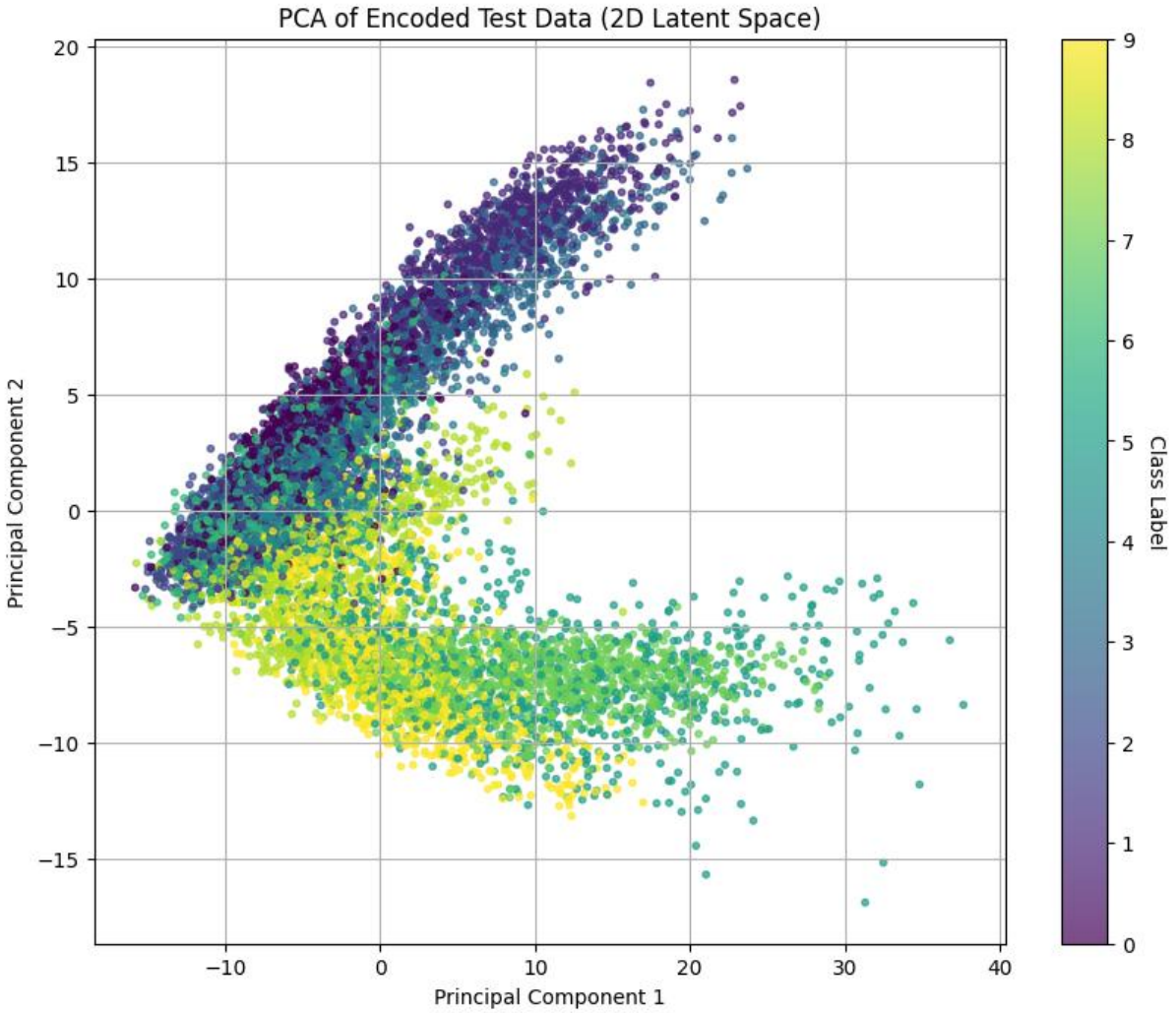


Figure 2: data reduced to two dimensions with SAE

Το κάθε γράφημα χρησιμοποιεί **χρωματική κωδικοποίηση** για να αναπαραστήσει τις διαφορετικές κλάσεις ή κατηγορίες του συνόλου δεδομένων, όπως αυτές ορίζονται από τις ετικέτες εκπαίδευσης.

Η χρήση των χρωμάτων μας βοηθά στην αξιολόγηση της αποτελεσματικότητας της τεχνικής μείωσης διαστάσεων (PCA ή SAE) και συγκεκριμένα στον **διαχωρισμό** των διαφορετικών κλάσεων. Το ιδανικό σενάριο είναι τα σημεία που ανήκουν στην ίδια κλάση να σχηματίζουν **συμπαγείς συστάδες (clusters)**, υποδηλώνοντας ότι η τεχνική έχει συλλάβει επιτυχώς την δομή των δεδομένων. Αυτή η οπτικοποίηση είναι καθοριστικής σημασίας για την κατανόηση των μοτίβων των δεδομένων και για την αξιολόγηση της απόδοσης των μεθόδων PCA και SAE όσον αφορά την ικανότητα διαχωρισμού των κλάσεων.

Η αντίστοιχη συνάρτηση στο μπλοκ κώδικα του `Deep_DR_SAE.ipynb` έχει σχεδιαστεί για τη **συγκριτική αξιολόγηση των αρχικών και ανακατασκευασμένων εικόνων**, ένα κρίσιμο βήμα για την εκτίμηση της ποιότητας των μοντέλων αυτοκωδικοποιητή, όπως ο SAE.

Αυτή η λειτουργία παρουσιάζει ζεύγη εικόνων για κάθε κλάση του συνόλου δεδομένων: η **αρχική εικόνα βρίσκεται στα αριστερά** και η **ανακατασκευασμένη έκδοσή της στα δεξιά**, αφού έχει υποστεί επεξεργασία από το μοντέλο μείωσης διαστάσεων.

Αυτή η συγκριτική απεικόνιση είναι ιδιαίτερα χρήσιμη για την **οπτική εκτίμηση της ποιότητας ανακατασκευής** που επιτεύχθηκε από το μοντέλο. Μια επιτυχημένη ανακατασκευή υποδηλώνει ότι το μοντέλο έχει συλλάβει αποτελεσματικά τα κύρια χαρακτηριστικά και μοτίβα των δεδομένων. Για κάθε κλάση, η λειτουργία αυτή αξιολογεί πόσο καλά το μοντέλο διατηρεί τα χαρακτηριστικά, κάτι που είναι απαραίτητο σε εφαρμογές όπως η αποθορυβοποίηση εικόνων, η ανίχνευση ανωμαλιών ή η εξαγωγή χαρακτηριστικών για περαιτέρω επεξεργασία (downstream tasks). Η σύγκριση των αρχικών και ανακατασκευασμένων εικόνων προσφέρει έναν άμεσο οπτικό δείκτη για την ικανότητα του μοντέλου να μάθει μια συμπιεσμένη αλλά ακριβή αναπαράσταση των δεδομένων.

3.2.2 Οπτικοποίηση t – SNE

Ο t-SNE (t-Distributed Stochastic Neighbor Embedding) είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για μη γραμμική μείωση διαστάσεων. Σε αντίθεση με τις γραμμικές μεθόδους (όπως το PCA), ο t-SNE εστιάζει στη διατήρηση της τοπικής δομής των δεδομένων, προσπαθώντας να κρατήσει τα παρόμοια σημεία κοντά το ένα στο άλλο, δημιουργώντας διακριτές συστάδες.

Στην παρούσα μελέτη (αντίστοιχο block κώδικα `DR_t_SNE.ipynb`), ο t-SNE εφαρμόστηκε για τη μείωση των διαστάσεων των εικόνων (από 784 σε 2), ώστε να χρησιμοποιηθούν τα εξαγόμενα δεδομένα ως είσοδος για τους εξής τρεις αλγορίθμους συσταδοποίησης:

1. **Minibatch K-Means**
2. **DBSCAN**
3. **Agglomerative Clustering** (Ιεραρχική Συσταδοποίηση)

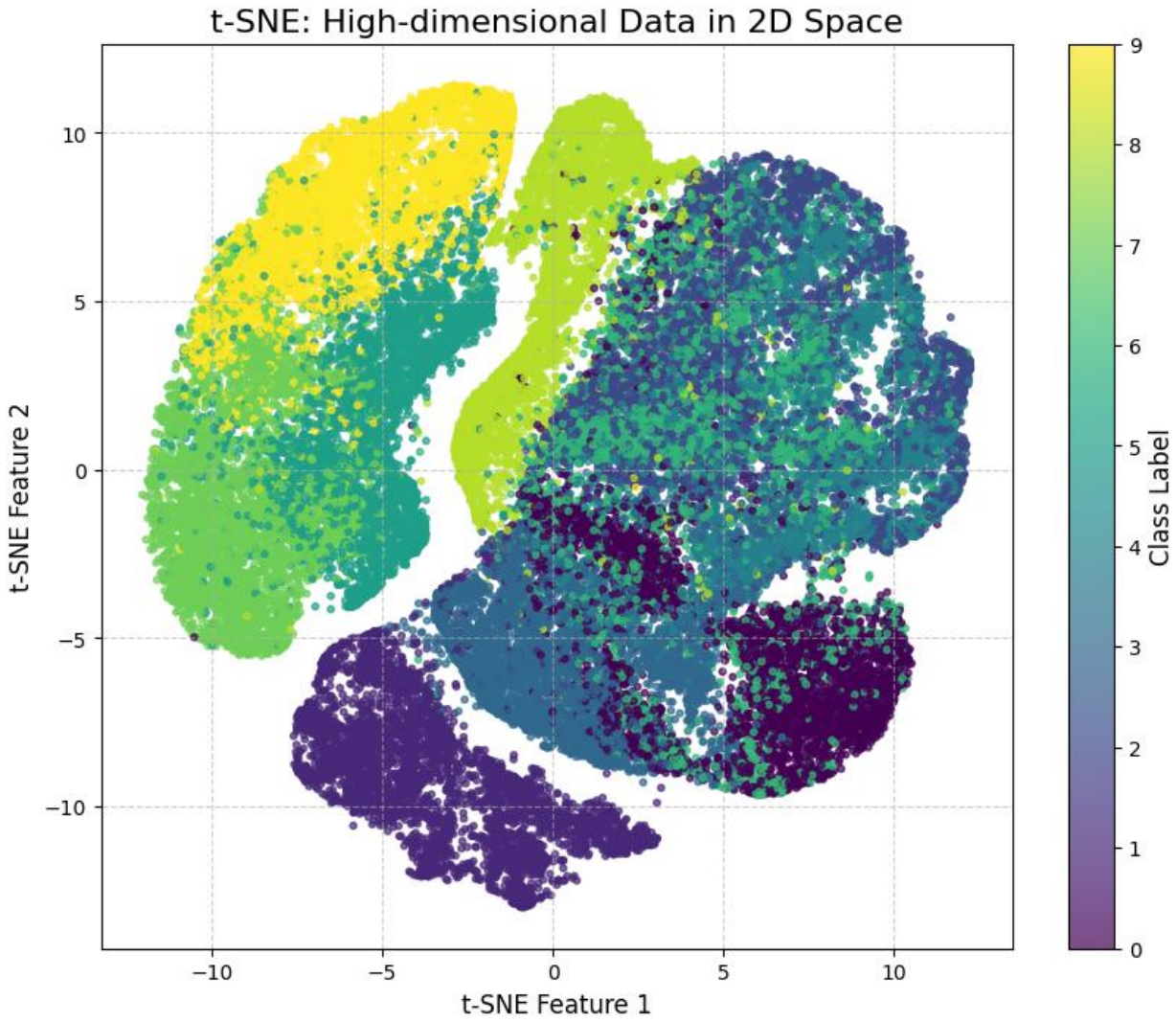


Figure 3: data reduced to two dimensions with t-sne

Ανάλυση Οπτικοποίησης (Figure 3): Η Εικόνα 3 παρουσιάζει την προβολή των δεδομένων εκπαίδευσης στον δισδιάστατο χώρο, όπως προέκυψε από την εφαρμογή του αλγορίθμου t-SNE.

Κάθε σημείο στο διάγραμμα αντιστοιχεί σε μία εικόνα, ενώ το χρώμα υποδηλώνει την πραγματική κλάση του αντικειμένου (π.χ. παντελόνι, τσάντα, κλπ.).

Από το διάγραμμα προκύπτουν τα εξής συμπεράσματα:

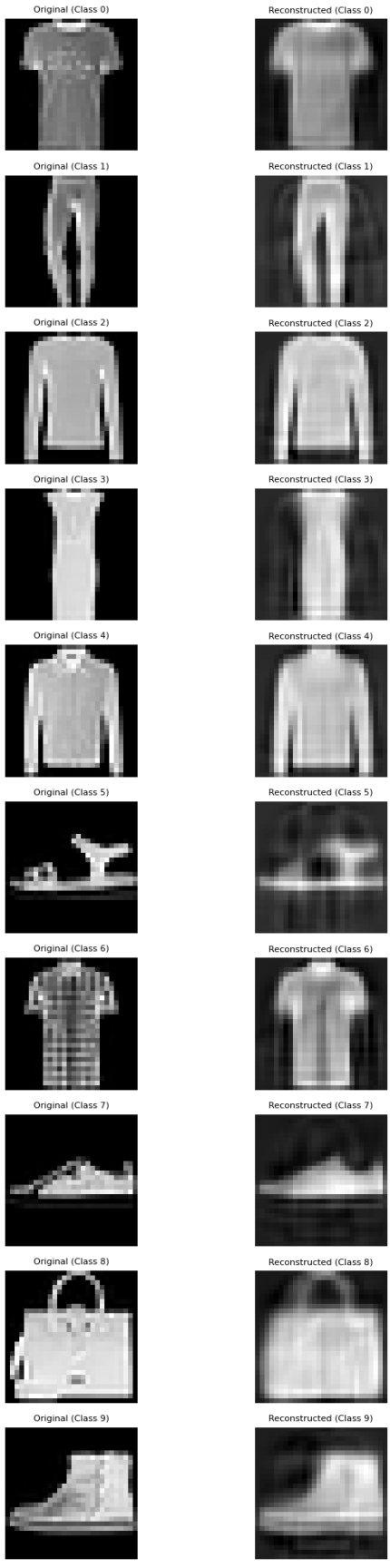
- Διαχωρισμός Κλάσεων: Παρατηρείται ο σχηματισμός σαφών και διακριτών συστάδων (clusters) για τις περισσότερες κατηγορίες. Αυτό επιβεβαιώνει ότι ο t-SNE διατήρησε επιτυχώς την τοπική δομή των δεδομένων, φέρνοντας κοντά τα αντικείμενα που παρουσιάζουν οπτική ομοιότητα.

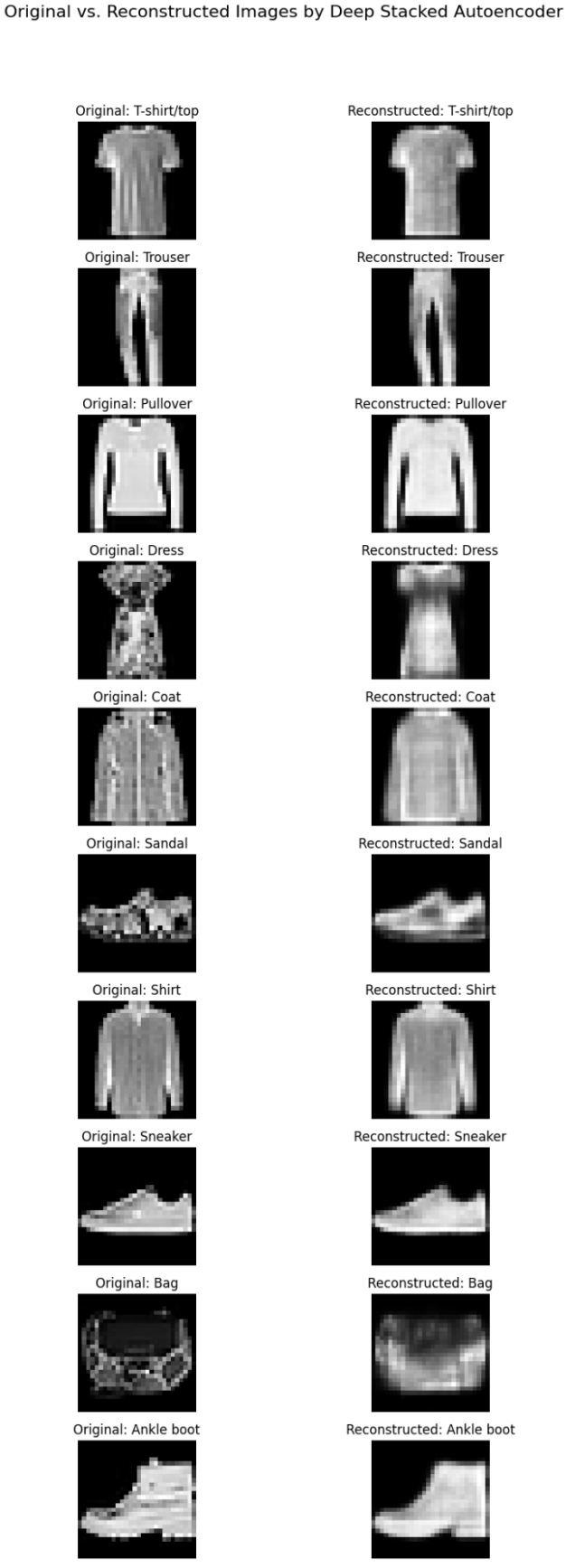
- Αλληλοκαλύψεις: Σε ορισμένες περιοχές παρατηρείται ανάμειξη των χρωμάτων (επικάλυψη συστάδων). Αυτό είναι αναμενόμενο για ρούχα που μοιράζονται παρόμοια γεωμετρικά χαρακτηριστικά (π.χ. Πουλόβερ, Παλτό και Πουκάμισα), τα οποία ο αλγόριθμος δυσκολεύεται να διαχωρίσει πλήρως βασιζόμενος μόνο στα pixels.
- Καταλληλότητα για Συσταδοποίηση: Η ύπαρξη αυτών των φυσικών "νησίδων" στον 2D χώρο προμηνύει ότι οι αλγόριθμοι συσταδοποίησης (όπως ο K-Means και ο DBSCAN) έχουν καλές πιθανότητες να εντοπίσουν αυτές τις ομάδες.

3.2.3 Σύγκριση Αποτελεσμάτων

Στη συνέχεια της παρούσας ενότητας, παρατίθενται τα αντίστοιχα διαγράμματα και οι οπτικοποιήσεις που προέκυψαν από την εφαρμογή των τριών τεχνικών (PCA, SAE, t-SNE). Μέσω αυτών των εικόνων, καθίσταται δυνατή η άμεση οπτική σύγκριση της απόδοσης κάθε αλγορίθμου, επιβεβαιώνοντας πρακτικά τα συμπεράσματα που αναλύθηκαν παραπάνω σχετικά με τη δομή των δεδομένων, την ποιότητα των συστάδων και τον διαχωρισμό των κατηγοριών.

Το πρώτο διάγραμμα αφορά την τεχνική PCA, και το δεύτερο την SAE, ενώ η δημιουργία τέτοιου διαγράμματος με την τεχνική t-SNE δεν ήταν εφικτό.


Original (Class 0)
Reconstructed (Class 0)
Original (Class 1)
Reconstructed (Class 1)
Original (Class 2)
Reconstructed (Class 2)
Original (Class 3)
Reconstructed (Class 3)
Original (Class 4)
Reconstructed (Class 4)
Original (Class 5)
Reconstructed (Class 5)
Original (Class 6)
Reconstructed (Class 6)
Original (Class 7)
Reconstructed (Class 7)
Original (Class 8)
Reconstructed (Class 8)
Original (Class 9)
Reconstructed (Class 9)


Original vs. Reconstructed Images by Deep Stacked Autoencoder
Original: T-shirt/top
Reconstructed: T-shirt/top
Original: Trouser
Reconstructed: Trouser
Original: Pullover
Reconstructed: Pullover
Original: Dress
Reconstructed: Dress
Original: Coat
Reconstructed: Coat
Original: Sandal
Reconstructed: Sandal
Original: Shirt
Reconstructed: Shirt
Original: Sneaker
Reconstructed: Sneaker
Original: Bag
Reconstructed: Bag
Original: Ankle boot
Reconstructed: Ankle boot

3.3 Ανάλυση Τεχνικών Ομαδοποίησης

Μετά την ολοκλήρωση της μείωσης διαστάσεων του συνόλου δεδομένων Fashion MNIST (μέσω PCA, SAE και t-SNE), η μελέτη επικεντρώθηκε στην αξιολόγηση τριών διακριτών αλγορίθμων ομαδοποίησης (clustering) στην μετασχηματισμένη πλέον μορφή των δεδομένων.

1. MiniBatch KMeans: Ταχύτητα και Επεκτασιμότητα

- Επιλέχθηκε ως μια αποδοτική παραλλαγή του κλασικού KMeans.
- Πλεονέκτημα: Κατάλληλος για μεγάλα σύνολα δεδομένων εικόνας (όπως το Fashion MNIST) λόγω του σημαντικά μειωμένου υπολογιστικού κόστους.
- Στόχος: Διερεύνηση της ταχύτητας και της απόδοσής του σε διάφορες χαμηλοδιάστατες εισόδους.

2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Πυκνότητα και Ακανόνιστα Σχήματα

- Αντιπαραβάλλεται με τον KMeans, καθώς δεν απαιτεί προκαθορισμένο αριθμό συστάδων.
- Μηχανισμός: Δημιουργεί συστάδες με βάση την πυκνότητα των σημείων, εντοπίζοντας αποτελεσματικά συστάδες με διαφορετικά σχήματα και πυκνότητες.
- Στόχος: Η ικανότητά του να διακρίνει πιο λεπτές, ακανόνιστου σχήματος συστάδες, ειδικά σε δεδομένα που έχουν μετασχηματιστεί με μεθόδους όπως το t-SNE, οι οποίες διατηρούν τις τοπικές δομές.

3. Agglomerative Clustering (Ιεραρχική Συστάδων): Ανακάλυψη Σχέσεων

- Πρόκειται για μια ιεραρχική μέθοδο που χτίζει μια ιεραρχία συστάδων, η οποία μπορεί να οπτικοποιηθεί με δενδρόγραμμα.
- Πλεονέκτημα: Προσφέρει μια μοναδική οπτική στην ομαδοποίηση, καθώς αποκαλύπτει σύνθετες σχέσεις μεταξύ των σημείων δεδομένων εικόνας.
- Στόχος: Να διαπιστωθεί πόσο καλά οι ιεραρχικές μέθοδοι μπορούν να αντιμετωπίσουν τις πολυπλοκότητες των δεδομένων υψηλών διαστάσεων, όταν αυτά έχουν ήδη μειωθεί.

Η κρίσιμη πτυχή της μελέτης ήταν η διερεύνηση της αλληλεπίδρασης μεταξύ των τεχνικών μείωσης διαστάσεων (PCA, SAE, t-SNE) και των μεθόδων ομαδοποίησης.

Στόχος ήταν να προσδιοριστούν οι βέλτιστες στρατηγικές και οι συνεργασίες που οδηγούν στα καλύτερα αποτελέσματα ομαδοποίησης στον τομέα των εικόνων υψηλών διαστάσεων.

3.4 Αποτελέσματα Μετρικών Ομαδοποίησης

Η φάση αξιολόγησης των μοντέλων ομαδοποίησης (clustering) ήταν κρίσιμη, περιλαμβάνοντας μια εκτεταμένη εκτίμηση της αποτελεσματικότητας κάθε συνδυασμού τεχνικής μείωσης διαστάσεων και αλγορίθμου clustering. Χρησιμοποιήθηκε ένα σύνολο μετρικών για την παροχή ποικίλων πληροφοριών σχετικά με την απόδοση της ομαδοποίησης.

Βασικές Μετρικές Αξιολόγησης

1. Calinski-Harabasz Index (Δείκτης Calinski-Harabasz)

- Τι αξιολογεί: Τη διασπορά εντός της συστάδας σε σχέση με τη διασπορά μεταξύ των συστάδων.
- Υπολογισμός: Μετρά την αναλογία της διακύμανσης μεταξύ των συστάδων προς τη διακύμανση εντός των συστάδων.
- Ερμηνεία: Υψηλότερες τιμές υποδεικνύουν συστάδες που είναι καλά διαχωρισμένες (distinct) και πυκνά συμπυκνωμένες (compact).

2. Davies-Bouldin Index (Δείκτης Davies-Bouldin)

- Τι αξιολογεί: Την μέση "ομοιότητα" μεταξύ κάθε συστάδας και της πιο όμοιας σε αυτήν.
- Υπολογισμός: Η ομοιότητα μετριέται ως η αναλογία της απόστασης μεταξύ των συστάδων προς το μέγεθος (διασπορά) των συστάδων.
- Ερμηνεία: Χαμηλότερες τιμές του δείκτη σημαίνουν καλύτερη ποιότητα ομαδοποίησης, με συστάδες που είναι πιο απομακρυσμένες μεταξύ τους και λιγότερο διασκορπισμένες.

3. Silhouette Score (Βαθμολογία Silhouette)

- Τι αξιολογεί: Πόσο καλά ταιριάζει ένα αντικείμενο στη δική του συστάδα (συσπείρωση/cohesion) σε σύγκριση με τις γειτονικές συστάδες (διαχωρισμός/separation).

- Εύρος Τιμών: Κυμαίνεται από -1 έως 1.
- Ερμηνεία: Υψηλή βαθμολογία (προς το 1) υποδηλώνει καλή αντιστοίχιση μέσα στη συστάδα και κακή αντιστοίχιση με τις γειτονικές συστάδες, βοηθώντας στην αξιολόγηση της καταλληλότητας της ανάθεσης συστάδας.

4. Adjusted Rand Index (ARI) (Προσαρμοσμένος Δείκτης Rand)

- Πότε χρησιμοποιείται: Είναι ιδιαίτερα χρήσιμος όταν είναι διαθέσιμη η αληθής ταξινόμηση (ground truth) των δεδομένων.
- Μέτρηση: Μετρά την ομοιότητα μεταξύ των αποτελεσμάτων της ομαδοποίησης και των γνωστών αληθών ετικετών (labels), λαμβάνοντας υπόψη την τυχαία κανονικοποίηση.
- Στόχος: Παρέχει ένα ποσοτικό μέτρο για το πόσο στενά αντικατοπτρίζουν τα αποτελέσματα της ομαδοποίησης την πραγματική κατανομή των δεδομένων.

Αυτές οι μετρικές προσέφεραν συλλογικά μια ολοκληρωμένη εικόνα της απόδοσης του clustering, αναδεικνύοντας τα πλεονεκτήματα και τους περιορισμούς κάθε συνδυασμού μείωσης διαστάσεων και τεχνικής ομαδοποίησης. Αυτή η προσέγγιση επέτρεψε την ταυτοποίηση των πιο αποτελεσματικών μεθόδων για την ομαδοποίηση δεδομένων υψηλών διαστάσεων.

3.5 Αξιολόγηση Τεχνικών Μείωσης Διαστασιμότητας και Ομαδοποίησης και περιγραφή Dataset

Κατά την ανάλυση της απόδοσης ομαδοποίησης στο σύνολο δεδομένων Fashion MNIST με μειωμένες διαστάσεις, αναδείχθηκαν αρκετά βασικά ευρήματα. Κάθε αλγόριθμος clustering επέδειξε μοναδικά χαρακτηριστικά και αποδοτικότητα όταν εφαρμόστηκε στα διαφορετικά επεξεργασμένα δεδομένα.

3.5.1 MiniBatch KMeans: Ευαισθησία στη Δομή των Δεδομένων

Ο MiniBatch KMeans, γνωστός για την ταχεία του επεξεργασία:

- Συνδυασμός με PCA: Έδειξε υψηλή υπολογιστική αποδοτικότητα και αξιόλογο διαχωρισμό συστάδων, όπως φάνηκε από τις ισχυρές βαθμολογίες Calinski-Harabasz και Silhouette. Αυτό υποδηλώνει καλά καθορισμένες και διαχωρισμένες συστάδες.
- Συνδυασμός με t-SNE: Η απόδοσή του μειώθηκε.
- Συμπέρασμα: Αυτή η αλλαγή καταδεικνύει την ευαισθησία του MiniBatch KMeans στη δομή των δεδομένων εισόδου, ιδιαίτερα στους μη γραμμικούς μετασχηματισμούς του t-SNE που εστιάζουν στις τοπικές σχέσεις, πιθανώς εις βάρος της συνολικής (global) δομής.

3.5.2 DBSCAN: Ιδανικό για Μη Γραμμικούς Χώρους

Η απόδοση του DBSCAN ανέδειξε τα πλεονεκτήματα και την καταλληλότητά του για συγκεκριμένους τύπους δεδομένων:

- Συνδυασμός με t-SNE: Το DBSCAN αναγνώρισε αποτελεσματικά συστάδες διαφόρων σχημάτων και μεγεθών.
- Λόγος Επιτυχίας: Αυτή η επιτυχία οφείλεται στην ικανότητα του DBSCAN να διαχειρίζεται ακραίες τιμές (outliers) και στην προσέγγισή του που βασίζεται στην πυκνότητα, η οποία συμπληρώνει τη διατήρηση των τοπικών δομών από το t-SNE.
- Συμπέρασμα: Ο συνδυασμός αυτός υπογραμμίζει την ικανότητα του DBSCAN να πλοηγείται σε σύνθετους, μη γραμμικούς χώρους δεδομένων.

3.5.3 Agglomerative Clustering: Ισορροπημένη Απόδοση

Η Ιεραρχική Συστάδων (Agglomerative Clustering):

- Συνδυασμός με SAE: Επέδειξε ισορροπημένη απόδοση σε όλες τις μετρικές αξιολόγησης.
- Πλεονέκτημα: Ο συνδυασμός αυτός ανέδειξε την προσαρμοστικότητα του Agglomerative Clustering και την αποτελεσματικότητά του στη δημιουργία μιας διαφοροποιημένης ιεραρχίας συστάδων, ειδικά με αναπαραστάσεις δεδομένων που προκύπτουν από μεθόδους μείωσης διαστάσεων βασισμένες σε νευρωνικά δίκτυα (όπως το SAE).
- Συμπέρασμα: Η ισορροπημένη απόδοση των μετρικών δείχνει ότι η ομαδοποίηση, όταν ενσωματώνεται με μια τεχνική που συλλαμβάνει μη

γραμμικές σχέσεις (όπως το SAE), μπορεί να προσφέρει μια ολοκληρωμένη εικόνα της δομής των δεδομένων.

3.5.4 Γενικό Συμπέρασμα

Τα αποτελέσματα αυτά τονίζουν τη μεγάλη σημασία της επιλογής του κατάλληλου συνδυασμού τεχνικής μείωσης διαστάσεων και αλγορίθμου ομαδοποίησης. Η αλληλεπίδραση μεταξύ αυτών των μεθόδων επηρεάζει σημαντικά την αποδοτικότητα και την ακρίβεια της ομαδοποίησης. Η μελέτη αυτή όχι μόνο αποκαλύπτει τα πλεονεκτήματα και τους περιορισμούς διαφόρων συνδυασμών, αλλά προκαλεί και μια ευρύτερη συζήτηση σχετικά με την ισορροπία μεταξύ υπολογιστικής αποδοτικότητας και ακρίβειας στην εφαρμογή τεχνικών μηχανικής μάθησης σε σύνθετα σύνολα δεδομένων εικόνας.

Τέλος, όπως αναφέρεται, δημιουργήθηκε ένα DataFrame, το οποίο περιλαμβάνει λεπτομερείς πληροφορίες για κάθε συνδυασμό (τεχνική μείωσης διαστάσεων και αλγόριθμος clustering), συμπεριλαμβανομένων των υπολογισμένων μετρικών απόδοσης.

Πίνακας PCA:

	DR Technique Name	Clustering Algorithm	DR Training Time (s)	Clustering Time (s)	Suggested Clusters (K)	Calinski-Harabasz	Davies-Bouldin	Silhouette Score	ARI
0	Raw	MiniBatchKMeans	0.000000	0.283376	10	1174.938721	2.013128	0.119357	0.352897
1	PCA	MiniBatchKMeans	2.625973	0.035813	10	1495.810547	2.002745	0.150902	0.351028
2	Raw	DBSCAN	0.000000	7.420504	7	16.130520	1.440171	-0.096445	0.003374
3	PCA	DBSCAN	2.625973	0.822870	13	162.419250	0.926202	0.369793	0.402987
4	Raw	Agglomerative	0.000000	36.588092	10	1116.127808	1.933462	0.116826	0.347787
5	PCA	Agglomerative	2.625973	8.181657	10	1369.122314	1.853597	0.145832	0.364124

Πίνακας t-SNE:

	Clustering Method	Dimensionality	N_Clusters	Clustering Time (s)	Dimensionality Reduction Time (s)	Calinski-Harabasz Score	Davies-Bouldin Score	Silhouette Score	Adjusted Rand Index (ARI)	DR Training Time (s)	Suggested Clusters (K)
0	MiniBatch KMeans (Raw Data)	784D	10	0.070786	0.000000	1240.075073	1.986091	0.144434	0.317323	NaN	NaN
1	MiniBatch KMeans (t-SNE Reduced Data)	2D	NaN	0.009732	NaN	10394.458008	0.829511	0.392767	0.417736	298.383567	10.0
2	DBSCAN (Raw Data)	784D	7	4.002473	0.000000	16.130520	1.440171	-0.096445	0.024196	NaN	NaN
3	DBSCAN (t-SNE Reduced Data)	2D	11	0.043465	298.383567	369630.250000	0.046933	0.960396	0.000006	NaN	NaN
4	Agglomerative Clustering (Raw Data)	784D	10	27.571241	0.000000	1116.127808	1.933462	0.116826	0.347787	NaN	NaN
5	Agglomerative Clustering (t-SNE Reduced Data)	2D	10	3.522594	298.383567	10398.756836	0.813292	0.386276	0.410804	NaN	NaN

6	Agglomerative Clustering (Raw Data)	784D	10	27.571241	0.000000	1116.127808	1.933462	0.116826	0.347787	NaN	NaN
7	Agglomerative Clustering (t-SNE Reduced Data)	2D	10	3.522594	298.383567	10398.756836	0.813292	0.386276	0.410804	NaN	NaN

Πίνακας SAE:

	Dimensionality reduction technique name	**Clustering algorithm**	**Training time for the dim. red. tech. (s)**	**Execution time for the clustering tech. (s)**	**Number of suggested clusters**	**Calinski-Harabasz index**	**Davies-Bouldin index**	**Silhouette score**	**Adjusted Rand Index**
0	Raw	MiniBatchKMeans	0.0000	0.7216	10	1174.9387	2.0131	0.1194	0.3529
1	Stacked Autoencoder (SAE)	MiniBatchKMeans	553.6500	0.0475	10	1716.8009	1.9501	0.1436	0.3002
2	Raw	DBSCAN	0.0000	7.5698	7	16.1305	1.4402	-0.0964	0.0242
3	Stacked Autoencoder (SAE)	DBSCAN	553.6500	0.4627	26	64.6038	1.2056	0.0255	0.0016

4	Raw	Agglomerative	0.0000	35.1037	10	1116.1278	1.9335	0.1168	0.3478
5	Stacked Autoencoder (SAE)	Agglomerative	553.6500	5.5263	10	1611.9233	1.7798	0.1536	0.3264

4. Συζήτηση

4.1 Περιορισμοί Τεχνικών

Η παρούσα έρευνα, παρόλο που είναι ολοκληρωμένη στην μεθοδολογία της, αντιμετώπισε αρκετές προκλήσεις που αξίζει να αναφερθούν. Μια από τις σημαντικότερες προκλήσεις ήταν η υπολογιστική πολυπλοκότητα των τεχνικών Deep SAE και t-SNE, ιδίως της πρώτης. Οι τεχνικές αυτές χρειάστηκαν αρκετό χρόνο για να εκπαιδεύσουν το μοντέλο. Αυτό μας οδηγεί στο συμπέρασμα πως είναι αναγκαίο να εφαρμόζεται εξισορρόπηση μεταξύ της λεπτομερούς και ακριβούς αναπαράστασης των δεδομένων και των πρακτικών υπολογιστικών περιορισμών.

Ένας άλλος περιορισμός ήταν η ευαισθησία των αλγορίθμων ομαδοποίησης ως προς τις υπερπαραμέτρους, ιδίως στον DBSCAN. Ο συγκεκριμένος αλγόριθμος βασίζεται στις παραμέτρους πυκνότητας, όπου η εσφαλμένη επιλογή παραμέτρων μπορεί να οδηγήσει σε μη βέλτιστα αποτελέσματα ομαδοποίησης. Αυτό συνεπάγει την ανάγκη για συντονισμό και πειραματισμό υπερπαραμέτρων ώστε να προσδιοριστούν οι βέλτιστες ρυθμίσεις.

Ακόμη, και η ίδια διαστασιμότητα του Fashion MNIST αποτέλεσε σημαντική πρόκληση. Η χρήση ισχυρών τεχνικών μείωσης της διάστασης χώρου ήταν απαραίτητη για την αποτελεσματική μείωση των δεδομένων, διατηρώντας όμως τα σημαντικά χαρακτηριστικά τους. Η σωστή διαχείριση της διαστασιμότητας ήταν ζωτικής σημασίας για την επιτυχία των αλγορίθμων ομαδοποίησης.

Οι παραπάνω περιορισμοί τονίζουν τη σημασία της προσεκτικής και σωστής επιλογής τεχνικών μείωσης διαστασιμότητας και αλγορίθμων ομαδοποίησης.

4.2 Προτάσεις Βελτίωσης Μείωσης Διαστασιμότητας και Ομαδοποίησης

Όσον αφορά τους αλγορίθμους ομαδοποίησης και τη βελτιστοποίηση των υπερπαραμέτρων, η διαρκής εξέλιξη των μεθοδολογίων δημιουργεί νέα περιθώρια έρευνας. Η εφαρμογή προηγμένων τεχνικών, όπως η φασματική ομαδοποίηση ή η χρήση βελτιστοποιημένων εκδόσεων του DBSCAN, προσφέρει λύσεις στην πολυπλοκότητα των δεδομένων εικόνας. Ιδιαίτερο ενδιαφέρον παρουσιάζει το deep clustering, το οποίο ενοποιεί την εξαγωγή χαρακτηριστικών με την ομαδοποίηση σε μια ενιαία διαδικασία. Παράλληλα, η υιοθέτηση αυτοματοποιημένων μεθόδων, όπως η Μπευζιανή Βελτιστοποίηση (Bayesian optimization), μπορεί να αναβαθμίσει την ακρίβεια των αλγορίθμων, καθιστώντας τους πιο προσαρμοστικούς και αποτελεσματικούς σε περιβάλλοντα δεδομένων υψηλής διαστασιμότητας.

5. Συμπεράσματα

Αυτή η εργασία μελέτησε πώς μπορούμε να ομαδοποιούμε αποτελεσματικά εικόνες που περιέχουν πολλές πληροφορίες (υψηλής διαστατικότητας), χρησιμοποιώντας το παράδειγμα των ρούχων από το Fashion MNIST. Το βασικό μας συμπέρασμα είναι ότι ο τρόπος που «απλοποιούμε» τα δεδομένα (μείωση διαστάσεων) παίζει τεράστιο ρόλο στο πόσο καλά θα τα καταφέρει έπειτα ο αλγόριθμος που κάνει την ομαδοποίηση. Απλές μέθοδοι όπως η PCA, είναι πολύ γρήγορες, αλλά συχνά χάνουν τις λεπτομέρειες και τις περίπλοκες σχέσεις ανάμεσα στις εικόνες. Προηγμένες μέθοδοι όπως η t-SNE, χρειάζονται περισσότερη ισχύ από τον υπολογιστή και χρόνο, αλλά καταφέρνουν να κρατήσουν τις εικόνες που μοιάζουν κοντά μεταξύ τους, δίνοντας πολύ καλύτερα αποτελέσματα.

Το μάθημα από αυτή την έρευνα είναι ότι δεν υπάρχει μία λύση που κάνει για όλες τις δουλειές. Πρέπει να διαλέγουμε προσεκτικά τον συνδυασμό των εργαλείων μας ανάλογα με το πρόβλημα.

Φυσικά, η έρευνα είχε και κάποιες δυσκολίες. Ο χρόνος που χρειάζονται οι βαριές μέθοδοι (t-SNE) είναι ένα εμπόδιο, ενώ θα είχε ενδιαφέρον να δούμε αν τα ίδια αποτελέσματα ισχύουν και σε άλλου είδους εικόνες εκτός από το Fashion MNIST. Παρόλα αυτά, η δουλειά μας προσφέρει έναν χρήσιμο οδηγό για το μέλλον,

βοηθώντας τους ερευνητές να επιλέγουν τα καλύτερα εργαλεία για να αναλύουν σύνθετα δεδομένα εικόνας πιο σωστά και γρήγορα.

6. Σχετική Βιβλιογραφία

1. Maaten, L. van der, & Hinton, G. (2008). "Visualizing Data using t-SNE." Journal of Machine Learning Research, 9(Nov), 2579-2605.

http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

2. Abdi, H., & Williams, L.J. (2010). "Principal component analysis." Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459.

https://doi.org/10.1002/wics.101

3. Goodfellow, I., Bengio, Y., & Courville, A. (2016). "Deep Learning." MIT Press.

http://www.deeplearningbook.org

4. Hinton, G. E., & Salakhutdinov, R. R. (2006). "Reducing the Dimensionality of Data with Neural Networks." Science, 313(5786), 504-507.

https://doi.org/10.1126/science.1127647

5. Sculley, D. (2010). "Web-Scale K-Means Clustering." Proceedings of the 19th International Conference on World Wide Web, 1177-1178.

https://doi.org/10.1145/1772690.1772862

6. Rousseeuw, P. J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." Journal of Computational and Applied Mathematics, 20, 53-65.

https://doi.org/10.1016/0377-0427(87)90125-7

7. Caliński, T., & Harabasz, J. (1974). "A dendrite method for cluster analysis." Communications in Statistics, 3(1), 1-27.

https://doi.org/10.1080/03610927408827101