



NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
DIVISION OF SIGNALS, CONTROL AND ROBOTICS

Multimodal Remote Sensing Data Classification using Semi-Supervised Variational Autoencoder

DIPLOMA THESIS

of

Pigi Lozou

Supervisor: Petros Maragos
Professor NTUA

Co-Supervisors: Andrea Marinoni Saloua Chlaily
Professor UiT Postdoctoral Researcher UiT

COMPUTER VISION, SPEECH COMMUNICATION AND SIGNAL PROCESSING GROUP
Athens, October 2023



NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
DIVISION OF SIGNALS, CONTROL AND ROBOTICS

Multimodal Remote Sensing Data Classification using Semi-Supervised Variational Autoencoder

DIPLOMA THESIS

of

Pigi Lozou

Supervisor: Petros Maragos
Professor NTUA

Co-Supervisors: Andrea Marinoni Saloua Chlaily
Professor UiT Postdoctoral Researcher UiT

Approved by the examining committee on October 20, 2023:

.....
Petros Maragos
Professor
NTUA

.....
Athanasios Rontogiannis
Associate Professor
NTUA

.....
Alexandros Potamianos
Associate Professor
NTUA

Athens, October 2023

.....

Pigi Lozou

Electrical and Computer Engineer, NTUA

© Pigi Lozou, 2023. All rights reserved.

This work is copyright and may not be reproduced, stored nor distributed in whole or in part for commercial purposes. Permission is hereby granted to reproduce, store and distribute this work for non-profit, educational and research purposes, provided that the source is acknowledged and the present copyright message is retained. Enquiries regarding use for profit should be directed to the author.

The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the National Technical University of Athens.

Abstract

In recent years, the rapid expansion of machine learning has inevitably led to the integration of artificial intelligence into diverse scientific disciplines, where machine learning techniques have played a pivotal role in revolutionizing the processing and analysis of large-scale datasets. This integration has significantly transformed the field of remote sensing.

This thesis contributes to this evolving landscape by presenting a comprehensive investigation into the classification of multimodal remote sensing data using semi-supervised Variational Autoencoder architectures. Variational Autoencoders have emerged as a powerful tool for uncovering the underlying patterns and structures inherent in data, showing significant potential in semi-supervised learning.

The architectural innovation proposed here incorporates a latent feature-level fusion strategy into the Variational Autoencoder framework, enabling the seamless integration of multiple modalities within the realm of remote sensing. Through a series of extensive experiments conducted on dataset representing rural area, we demonstrate the critical impact of encoder selection and latent space dimensionality on classification performance. The semi-supervised Variational Autoencoder models outperformed traditionally used methods such as Support Vector Machines and Random Forests, not only in terms of metrics but also in qualitative performance and uncertainty assessment.

Furthermore, this study provides insights into the strengths and limitations associated with data-level fusion and latent feature-level fusion strategies. As we test the capability of the proposed architectures on progressively larger dataset of urban area, we gain a deeper understanding of the importance of qualitative analysis, which reveals valuable insights about the performance of each fusion strategy.

As we navigate the complex landscape of multimodal data analysis, the framework proposed in this thesis not only offers valuable insights into remote sensing but also opens up exciting possibilities for creative solutions and applications across a spectrum of scientific domains.

- **Keywords:** machine learning, remote sensing, multimodality, variational autoencoder, semi-supervised learning, data fusion

Περίληψη

Τα τελευταία χρόνια, η ταχεία ανάπτυξη της μηχανικής μάθησης οδήγησε στην ενσωμάτωση της τεχνητής νοημοσύνης σε διάφορους επιστημονικούς κλάδους όπου οι τεχνικές μηχανικής μάθησης έφεραν την επανάσταση στην επεξεργασία και ανάλυση συνόλων δεδομένων μεγάλης κλίμακας. Η ενοποίηση αυτή έχει φέρει σημαντικές αλλαγές στον τομέα της τηλεπισκόπησης. Αυτή η διατριβή συμβάλλει σε αυτή την πρόοδο παρουσιάζοντας μια ολοκληρωμένη έρευνα για την ταξινόμηση των πολυτροπικών δεδομένων τηλεπισκόπησης χρησιμοποιώντας ημι-επιβλεπόμενες αρχιτεκτονικές με Variational Autoencoders. Οι Variational Autoencoders έχουν αναδειχθεί ως ένα ισχυρό εργαλείο για την μελέτη των υποκείμενων μοτίβων και δομών που χαρακτηρίζουν τις εικόνες τηλεπισκόπησης, δείχνοντας σημαντικές δυνατότητες στην ημι-εποπτευόμενη μάθηση.

Η αρχιτεκτονική καινοτομία που προτείνουμε, ενσωματώνει τη στρατηγική συγχώνευσης δεδομένων σε επίπεδο latent χαρακτηριστικών στο πλαίσιο του Variational Autoencoder, επιτρέποντας την απρόσκοπτη ενσωμάτωση πολυτροπικών δεδομένων στον τομέα της τηλεπισκόπησης. Μέσω μιας σειράς εκτεταμένων πειραμάτων που πραγματοποιήθηκαν σε σύνολα δεδομένων που απεικονίζουν αγροτικές και αστικές περιοχές, αυτή η μελέτη αναδεικνύει το αντίκτυπο της επιλογής κωδικοποιητή και της διάστασης του latent χώρου στην απόδοση ταξινόμησης. Τα ημι-εποπτευόμενα μοντέλα Variational Autoencoder ξεπέρασαν τις παραδοσιακές μεθόδους, όπως το Support Vector Machine και το Random Forest, όχι μόνο ως προς τις μετρικές αλλά και ως προς την ποιοτική απόδοση και την αβεβαιότητα.

Επιπλέον, παρέχουμε μια ανάλυση των δυνατοτήτων και των περιορισμών που σχετίζονται με τις στρατηγικές σύντηξης πολυτροπικών δεδομένων τόσο σε επίπεδο δεδομένων και σύντηξης σε επίπεδο latent χαρακτηριστικών. Δοκιμάζοντας τις προτεινόμενες αρχιτεκτονικές σε προοδευτικά μεγαλύτερο σύνολο δεδομένων αστικών περιοχών, αποκτούμε μια βαθύτερη κατανόηση της σημασίας της ποιοτικής ανάλυσης, η οποία αποκαλύπτει πολύτιμες πληροφορίες σχετικά με την απόδοση κάθε στρατηγικής σύντηξης.

Καθώς ερευνούμε τον κλάδο της πολυτροπικής ανάλυσης δεδομένων, το πλαίσιο που προτείνεται σε αυτή τη διατριβή όχι μόνο προσφέρει πολύτιμες γνώσεις για την τηλεπισκόπηση αλλά επίσης προσφέρει συναρπαστικές δυνατότητες για δημιουργικές λύσεις και εφαρμογές σε ένα φάσμα επιστημονικών τομέων.

- **Λέξεις Κλειδιά:** μηχανική μάθηση, τηλεπισκόπηση, πολυτροπικά δεδομένα, variational autoencoder, ημι-επιβλεπόμενη μάθηση, σύντηξη δεδομένων

Acknowledgements

This thesis marks the end of my academic journey at the School of Electrical and Computer Engineering, National Technical University of Athens, and the conclusion of my research at the Earth Observation Laboratory of the Arctic University of Norway (UiT). In light of this achievement, I would like to thank my supervising professor Petros Maragos. His insightful lectures and his research work have been a continuous source of inspiration from the very first years of my studies. I want to express a deep thanks to professor Andrea Marinoni, who not only supervised this thesis but also granted me the invaluable opportunity to implement it at the Earth Observation Laboratory of the Arctic University of Norway (UiT). Also, to everyone at the laboratory for their hospitality and assistance, I owe my warmest thanks. Your collaboration and support have been integral to the successful completion of my research. Of course, I would like to express my special gratitude and appreciation to Dr. Saloua Chlaily whose guidance, invaluable advice, stimulating discussions, and generous support greatly enriched my research experience and this thesis. Last but not least, I extend my gratitude to my beloved family and friends for their support and boundless patience throughout my academic journey.

I wish you all the very best!

Pigi Lozou
October 2023

Ευχαριστίες

Με την παρούσα διπλωματική εργασία ολοκληρώνεται ο κύκλος φοίτησής μου στην Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου καθώς και η έρευνα που πραγματοποίησα στο Earth Observation Laboratory του Arctic University of Norway (UiT). Με την αφορμή λοιπόν που μου δίνεται, θα ήθελα να ευχαριστήσω αρχικά τον επιβλέποντα καθηγητή κ. Πέτρο Μαραγκό, ο οποίος, μέσω των εποικοδομητικών του διαλέξεων στο πλαίσιο των προπτυχιακών μαθημάτων αλλά και του γενικότερου ερευνητικού του έργου, με ενέπνευσε από τα πρώτα κιόλας χρόνια των σπουδών μου να ασχοληθώ ενεργά με το συγκεκριμένο αντικείμενο. Φυσικά, τον καθηγητή κ. Andrea Marinoni για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω στο εργαστήριο Earth Observation του Arctic University of Norway (UiT). Επίσης, σε όλους στο εργαστήριο για τη φιλοξενία και τη βοήθειά τους, οφείλω τις θερμότερες ευχαριστίες μου. Η συνεργασία και η υποστήριξή σας ήταν αναπόσπαστο κομμάτι της επιτυχούς ολοκλήρωσης της έρευνάς μου. Ιδιαιτέρες ευχαριστίες στην ερευνήτρια Δρ. Saloua Chlaily για την καθοδήγηση, τις συμβουλές, τις ευχάριστες συζητήσεις και την πολύτιμη βοήθεια που μου προσέφερε καθ' όλη την διάρκεια της συνεργασίας μας. Τέλος, είμαι ευγνώμων στην οικογένειά μου και τους φίλους μου για την αμέριστη στήριξη που μου προσέφεραν και την απεριόριστη υπομονή που έδειξαν καθ' όλη την διάρκεια της ακαδημαϊκής μου σταδιοδρομίας.

Να είστε όλοι καλά!

Πηγή Λόζου
Οκτώβριος 2023

Contents

1	Εκτενής περίληψη στα Ελληνικά	1
1.1	Εισαγωγή	1
1.1.1	Αντικείμενο της διπλωματικής	2
1.2	Θεωρητικό υπόβαθρο	3
1.2.1	Variational Autoencoders	3
1.2.2	Ημι-επιβλεπόμενη μάθηση με VAEs	4
1.3	Μέθοδος	5
1.4	Εφαρμογή σε δεδομένα τηλεπισκόπησης	7
1.4.1	Αγροτική περιοχή	9
1.4.2	Αστική περιοχή	15
1.5	Συζήτηση και συμπεράσματα	18
1.6	Δημοσιεύσεις	18
2	Introduction	19
2.1	Machine learning	20
2.2	Remote sensing	22
2.2.1	Evolution of remote sensing	22
2.2.2	Remote sensors	24
2.2.3	Multimodality	25
2.2.4	Applications of remote sensing	27
2.3	Machine learning for remote sensing	29
2.3.1	Applications of machine learning in remote sensing	29
2.3.2	Challenges of machine learning in remote sensing	30
2.3.3	Machine learning for multimodal remote sensing	31
2.4	Objectives of the thesis	31
2.4.1	Literature review and related work	32
2.4.2	Methodology and contributions	33
3	Theoretical background	34
3.1	Generative and discriminative models	34
3.1.1	Generative models	35
3.1.2	Deep generative models	36

3.2	Fundamentals of VAEs	36
3.2.1	Autoencoders	37
3.2.2	Probabilistic models	38
3.2.3	Latent variables	38
3.2.4	Deep Latent Variable Models	39
3.2.5	Variational inference	40
3.2.6	Variational autoencoder	41
3.3	Semi-supervised classification with VAEs	42
3.3.1	Latent-feature discriminative model	43
3.3.2	Generative semi-supervised model	44
3.3.3	Stacked generative semi-supervised model	46
4	Method	50
4.1	Multimodal data classification with semi-supervised VAE	50
4.1.1	Data-level fusion	50
4.1.2	Latent feature-level fusion - multimodal stacked generative semi-supervised model	51
5	Experimental evaluation	56
5.1	Evaluation methods	56
5.1.1	Accuracy	58
5.1.2	Precision	58
5.1.3	Recall	59
5.1.4	F1-score	59
5.1.5	Kappa coefficient	60
5.1.6	Classification map	60
5.1.7	Uncertainty	60
5.2	Rural area classification	62
5.2.1	Trento dataset	62
5.2.2	Experimental process	62
5.2.3	Results	65
5.2.4	Analysis	68
5.2.5	Comparison with SVM and RF	72
5.3	Urban area classification	74
5.3.1	Houston dataset	74
5.3.2	Experimental process	75

5.3.3	Results	78
5.3.4	Analysis	86
5.4	Discussion	91
6	Conclusion & future work	93
6.1	Conclusion	93
6.2	Future work	94
7	Publications	97
8	References	98

List of Figures

1	Γραφική αναπαράσταση Autoencoder	3
2	Γραφική αναπαράσταση Variational Autoencoder	4
3	Γραφική αναπαράσταση <i>latent-feature discriminative</i> μοντέλου - M1	5
4	Γραφική αναπαράσταση <i>generative semi-supervised</i> μοντέλου - M2	6
5	Γραφική αναπαράσταση <i>stacked generative semi-supervised</i> μοντέλου - M1+M2	6
6	Γραφική αναπαράσταση <i>multimodal stacked generative semi-supervised</i> μον- τέλο - Multi-M1+M2	8
7	Σύνολο δεδομένων Trento: (a) Ψευδής RGB αναπαράσταση με χρήση καναλι- ών HSI, (b) LiDAR DSM	9
8	Σύνολο δεδομένων Houston: (a) Ψευδής RGB αναπαράσταση με χρήση καναλιών HSI, (b) LiDAR DSM, (c) Χρωματική αναπαράσταση multispec- tral LiDAR	10
9	Σύνολο δεδομένων Trento: (a) Ψευδής RGB αναπαράσταση με χρήση καναλι- ών HSI, (b) Ετικέτες, Χάρτες ταξινόμησης για το μοντέλο M1+M2 με: (c) E1 για latent size = 20, (d) E2 για latent size = 15, (e) E3 για latent size = 10, (f) E4 για latent size = 20, (g) E5 για latent size = 15 and patch size = 5x5, (h) και για το μοντέλο Multi-M1+M2 με E2 για latent size = 15 καθώς και για (i) SVM, (j) RF	13
10	Σύνολο δεδομένων Trento: (a) Ψευδής RGB αναπαράσταση με χρήση καναλι- ών HSI, (b) Ετικέτες, Χάρτες αβεβαιότητας για το μοντέλο M1+M2 με (c) E1 για latent size = 20, (d) E2 για latent size = 15, (e) E3 για latent size = 10, (f) E4 για latent size = 20, (g) E5 για latent size = 15 and patch size = 5x5, (h) και για το μοντέλο Multi-M1+M2 με E2 για latent size = 15 and (i) SVM, (j) RF	14
11	Σύνολο δεδομένων Houston: (a) Ψευδής RGB αναπαράσταση με χρήση καναλιών HSI, (b) Ετικέτες, Χάρτες ταξινόμησης για τα μοντέλα: (c) M1+M2 με E2 για latent size = 30, (d) Multi-M1+M2 με E2 για latent size = 30 .	16
12	Σύνολο δεδομένων Houston: (a) Ψευδής RGB αναπαράσταση με χρήση καναλιών HSI, (b) Ετικέτες, Χάρτες αβεβαιότητας για τα μοντέλα: (c) M1+M2 model με E2 για latent size = 30, (d) Multi-M1+M2 με E2 για latent size = 30	17

13	Landsat 1 illustration, (Source)	22
14	Graphical representation of an autoencoder	37
15	Graphical representation of a variational autoencoder	42
16	Graphical representation of <i>latent-feature discriminative</i> model - M1	44
17	Graphical representation of <i>generative semi-supervised</i> model - M2	46
18	Graphical representation of <i>stacked generative semi-supervised</i> model - M1+M2	48
19	Graphical representation of <i>multimodal stacked generative semi-supervised</i> model - Multi-M1+M2 architecture	55
20	Multi-M1+M2 architecture for two modalities	57
21	Trento Dataset: Modalities representation	62
22	Trento dataset: (a) A false-RGB representation using HSI bands, (b) LiDAR DSM, (c) Ground truth labels	63
23	Trento dataset: Test set metrics for the M1+M2 model with encoders E1, E2, E3, E4 and E5 on latent size = 15, (a) Accuracy and F1-score, (b) Precision and recall	66
24	Trento dataset: (a) A false RGB representation using HSI bands, (b) Ground truth labels, Classification maps for M1+M2 model with (c) E1 on latent size = 20, (d) E2 on latent size = 15, (e) E3 on latent size = 10, (f) E4 on latent size = 20, (g) E5 on latent size = 15 and patch size = 5x5, (h) Multi-M1+M2 model with E2 on latent size = 15 and (i) SVM, (j) RF	67
25	Trento dataset: (a) A false RGB representation using HSI bands, (b) Ground truth labels, Homophily-based uncertainty maps for M1+M2 model with (c) E1 on latent size = 20, (d) E2 on latent size = 15, (e) E3 on latent size = 10, (f) E4 on latent size = 20, (g) E5 on latent size = 15 and patch size = 5x5, (h) Multi-M1+M2 model with E2 on latent size = 15 and (i) SVM, (j) RF	69
26	Trento dataset: (a),(b) HSI spectrum and LiDAR channels respectively for vegetation classes, (c),(d) HSI spectrum and LiDAR channels respectively for buildings and roads	71
27	Training elapsed time at Intel Core i7-5500U@2.4GHz, 16GB DDR4 RAM on Ubuntu 20.04.6 LTS, Python 3.8.(<i>Bar heights are represented on a logarithmic scale for enhanced visualization</i>)	74
28	Houston Dataset: Modalities representation	75

29	Houston dataset: (a) A false RGB representation using HSI bands, (b) LiDAR DSM, (c) Color composite of multispectral LiDAR intensity, (d) Ground truth labels	76
30	Houston dataset: (a) A false RGB representation using HSI bands, (b) Ground truth labels, Classification maps for models: (c) M1+M2 with E2 on latent size = 30, (d) Multi-M1+M2 with E2 on latent size = 30	81
31	Houston dataset focused areas: False RGB representation (Left), Ground truth labels (Middle left), Classification maps from M1+M2 model with E2 on latent size = 30 (Middle right) and from Multi-M1+M2 model with E2 on latent size = 30 (Right)	82
32	Houston dataset: (a) A false RGB representation using HSI bands, (b) Ground truth labels, Uncertainty maps for models: (c) M1+M2 with E2 on latent size = 30, (d) Multi-M1+M2 with E2 on latent size = 30	83
33	Houston dataset focused areas: False RGB representation (Left), Ground truth labels (Middle left), Uncertainty maps from M1+M2 model with E2 on latent size = 30 (Middle right) and from Multi-M1+M2 model with E2 on latent size = 30 (Right)	84
34	Houston dataset: Heatmap representation of the normalized Homophily matrix H_h in logarithmic scale	85
35	Houston dataset: HSI spectrum (Left) and LiDAR channels (Right) for selected classes	89
36	Houston dataset: HSI spectrum (Left) and LiDAR channels (Right) for selected classes	90

List of Tables

1	Αρχιτεκτονική των κωδικοποιητών	11
2	Σύνολο δεδομένων Trento: Μετρικές για το μοντέλο M1+M2 με τους κωδικοποιητές E1, E2, E3, και E4 σε διαφορετικά μεγέθη latent χώρου	12
3	Σύνολο δεδομένων Trento: Σύγκριση μετρικών των καλύτερων VAE μοντέλων και των SVM, RF	12
4	Σύνολο δεδομένων Houston: Μετρικές για το μοντέλο M1+M2 με κωδικοποιητή E2 σε διαφορετικά μεγέθη latent χώρου	15
5	Σύνολο δεδομένων Houston: Μετρικές για τα καλύτερα VAE μοντέλα	17
6	Applications of machine learning methods	21
7	Deep learning techniques and their applications	21
8	Examples of active remote sensing sensors and their applications	25
9	Examples of passive remote sensing sensors and their applications	25
10	Applications of remote sensing	28
11	Trento dataset: Color code and name of classes, number of training and test samples	64
12	Encoders' architectures and configuration	64
13	Trento dataset: Test set metrics for the M1+M2 model with encoders E1, E2, E3, and E4 on different latent spaces	65
14	Trento dataset: Comparison of test set metrics for the best VAE models with SVM and RF	66
15	Houston dataset: Color code and name of classes, number of training and test samples	77
16	Houston dataset: Test set metrics for the M1+M2 model with encoder E2 on different latent spaces	78
17	Houston dataset: Test set metrics for the best VAE models	79
18	Houston dataset: Confusion matrix of M1+M2 model with E2	79
19	Houston dataset: Confusion matrix of Multi-M1+M2 model with E2	80

List of Acronyms

VAE	Variational Autoencoder
HSI	Hyperspectral Imaging
LiDAR	Light Detection and Ranging
VAE	Variational Autoencoder
RF	Random Forest
ANNs	Artificial neural networks
CNNs	Generative Adversarial Networks
Radar	Radio Detection and Ranging
Sonar	Sound Navigation and Ranging
SAR	Synthetic Aperture Radar
UAVs	Unmanned Aerial Vehicles
HMM	Hidden Markov Model
GMM	Gaussian Mixture Mode
PDF	Probability Density Functions
DLVM	Deep Latent Variable Models
KL	Kullback-Leibler
ELBO	Evidence Lower Bound
HU	Homophily-based Uncertainty using Energy distance
DSM	Digital Surface Model
DEM	Digital Earth Model
SELU	Scaled Exponential Linear Unit

1 Εκτενής περίληψη στα Ελληνικά

1.1 Εισαγωγή

Τις τελευταίες δεκαετίες, η χρήση των μεθόδων μηχανικής μάθησης έχει φέρει σημαντική επανάσταση στον τομέα της τηλεπισκόπησης, επιτρέποντας την αποτελεσματική επεξεργασία και ανάλυση συνόλων δεδομένων μεγάλης κλίμακας (Zhang and Han, 2020). Με την αυτοματοποίηση εργασιών όπως η ταξινόμηση χρήσης ή κάλυψης Γης και βελτιώνοντας τις προβλέψεις περιβαλλοντικών μεταβλητών, οι τεχνικές μηχανικής μάθησης έχουν εξαλείψει την ανάγκη για εκτεταμένη χειρωνακτική εργασία και ανθρώπινη τεχνογνωσία. Αυτό είχε ως αποτέλεσμα την εξοικονόμηση χρόνου και κόστους για τους επαγγελματίες τηλεπισκόπησης, ενώ παράλληλα επέτρεψε νέες και πιο εξειδικευμένες εφαρμογές σε διάφορα πεδία (Li et al., 2020, Khelifi and Mignotte, 2020, Zhu et al., 2017, Camps-Valls, 2009). Ωστόσο, η χρήση μεθόδων μηχανικής μάθησης σε εφαρμογές τηλεπισκόπησης φέρνει αρκετές προκλήσεις. Η αντιμετώπιση αυτών των προκλήσεων είναι επιτακτική για τη βελτίωση της ακρίβειας και της αποτελεσματικότητας των μεθόδων μηχανικής μάθησης σε εφαρμογές τηλεπισκόπησης.

Μία από τις κύριες προκλήσεις είναι η ποιότητα των δεδομένων τηλεπισκόπησης, η οποία μπορεί να επηρεαστεί από περιβαλλοντικούς παράγοντες, συμπεριλαμβανομένων των ατμοσφαιρικών παρεμβολών, του θορύβου του αισθητήρα και της νεφοκάλυψης. Αυτή η μεταβλητότητα μπορεί να εμποδίσει την ανάπτυξη μοντέλων μηχανικής μάθησης που αποτυπώνουν με ακρίβεια τα εγγενή μοτίβα στα δεδομένα (Yuan et al., 2020). Μάλιστα, η περιορισμένη διαθεσιμότητα των ετικετοποιημένων δεδομένων υψηλής ποιότητας θέτει μια άλλη σημαντική πρόκληση για την αποτελεσματική εκπαίδευση αλγορίθμων μηχανικής μάθησης. Τα δεδομένα τηλεπισκόπησης είναι συχνά πολύπλοκα και δύσκολο να ερμηνευθούν, ειδικά για μη ειδικούς. Το γεγονός αυτό καθιστά την επισήμανση ετικετών χρονοβόρα και δαπανηρή διαδικασία που απαιτεί εξειδικευμένη γνώση. Ως αποτέλεσμα, τα ετικετοποιημένα δεδομένα μπορεί να είναι σπάνια, εμποδίζοντας την απόδοση των αλγορίθμων μηχανικής μάθησης (Yuan et al., 2020, Zhang et al., 2022). Επιπλέον, τα δεδομένα τηλεπισκόπησης μπορεί να είναι τεράστιου μεγέθους, θέτοντας προκλήσεις στην αποθήκευση, την επεξεργασία και την ανάλυση των δεδομένων, ιδιαίτερα για δορυφορικές εικόνες υψηλής ανάλυσης. Επίσης, οι αλγόριθμοι μηχανικής μάθησης ενδέχεται να απαιτούν σημαντικούς υπολογιστικούς πόρους, οι οποίοι μπορεί να είναι περιορισμένοι, ειδικά σε εφαρμογές ενσωματωμένων συστημάτων ή σε εφαρμογές πραγματικού χρόνου (Zhang et al., 2022). Τέλος, τα δεδομένα

τηλεπισκόπησης είναι δυνατό να προέρχονται από πολλαπλούς αισθητήρες. Καθένας από τους αισθητήρες έχει τις απαιτήσεις επεξεργασίας του, γεγονός που καθιστά δύσκολη την ενσωμάτωση δεδομένων από διάφορους αισθητήρες σε ένα μόνο μοντέλο. Τα δεδομένα αυτά ονομάζονται πολυτροπικά (Li et al., 2022).

1.1.1 Αντικείμενο της διπλωματικής

Τα Bayesian νευρωνικά δίκτυα έχουν οδηγήσει σημαντικές προόδους τα τελευταία χρόνια (Wang and Yeung, 2020). Ωστόσο, η εφαρμογή τους στην τηλεπισκόπηση είναι περιορισμένη μέχρι στιγμής (Shirmard et al., 2022). Η Bayesian προσέγγιση των νευρωνικών δικτύων υπόσχεται τη διευκόλυνση της ουσιαστικής ανάλυσης των δεδομένων τηλεπισκόπησης, ιδιαίτερα για την αντιμετώπιση προκλήσεων όπως ο θόρυβος δεδομένων, τα αραιά σύνολα δεδομένων και η απουσία πληροφοριών (Shirmard et al., 2022). Με γνώμονα τις πιο πρόσφατες έρευνες στον τομέα (Wang and Yeung, 2020, Zhang et al., 2022, Yuan et al., 2020, Shirmard et al., 2022), η παρούσα διπλωματική επικεντρώνεται στην ταξινόμηση κάλυψης Γης με αποτελεσματικό χειρισμό των πολυτροπικών και περιορισμένα επισημασμένων δεδομένων. Σκοπεύουμε να διερευνήσουμε τις δυνατότητες της Bayesian θεωρίας σε συνδυασμό με τεχνικές βαθιάς μάθησης για να ξεπεράσουμε αυτήν την πρόκληση.

Συνεισφορές:

Ο στόχος της μελέτης μας είναι να επεκτείνουμε το ημι-επιβλεπόμενο μοντέλο Variational Autoencoder (VAE) που προτείνεται από τους Kingma et al. (2014) για να επιτύχουμε ταξινόμηση πολυτροπικών δεδομένων τηλεπισκόπησης. Σε αντίθεση με το Kingma et al. (2014), το μοντέλο μας λαμβάνει υπόψη τα δεδομένα Light Detection And Ranging (LiDAR) και Hyperspectral Imaging (HSI) για να ταξινομήσει κάθε εικονοστοιχείο της εικόνας. Ερευνούμε επίσης τη συγχώνευση των δύο αισθητήρων σε επίπεδο εικονοστοιχείων και σε επίπεδο latent χαρακτηριστικών. Η προσέγγιση αυτή είναι κατάλληλη για εικόνες τηλεπισκόπησης και μπορεί να βελτιώσει την ακρίβεια ταξινόμησης σε σύγκριση με μεθόδους που βασίζονται σε μεμονωμένους αισθητήρες. Συνοπτικά, η διατριβή κάνει τις ακόλουθες συνεισφορές:

- Προσαρμογή ενός Bayesian μοντέλου βαθιάς μάθησης για να καλύψει τις απαιτήσεις της ταξινόμησης πολυτροπικών δεδομένων τηλεπισκόπησης.
- Αντιμετώπιση της πρόκλησης των περιορισμένα ετικετοποιημένων δεδομένων τηλεπισκόπησης αξιοποιώντας την ημι-επιβλεπόμενη μάθηση.

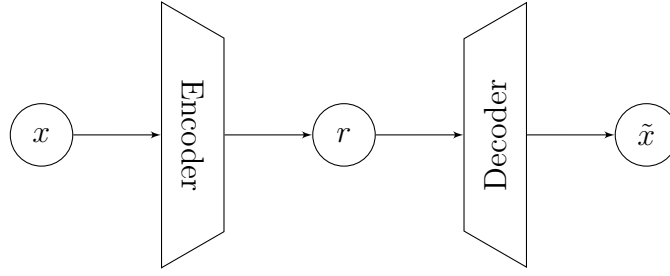


Figure 1: Γραφική αναπαράσταση Autoencoder

- Διερεύνηση των δυνατοτήτων των τεχνικών συγχώνευσης δεδομένων τόσο σε επίπεδο latent χαρακτηριστικών όσο και σε επίπεδο εικονοστοιχείων.

1.2 Θεωρητικό υπόβαθρο

1.2.1 Variational Autoencoders

Οι VAEs ονομάζονται autoencoders (αυτοκωδικοποιητές) επειδή μοιράζονται την ίδια αρχιτεκτονική κωδικοποίησης-αποκωδικοποίησης με τους αυτόκωδικοποιητές (Figure 1, 2). Ωστόσο, σε αντίθεση με τους παραδοσιακούς αυτοκωδικοποιητές, οι VAEs συμπιέζουν δεδομένα εισόδου υψηλών διαστάσεων x σε ένα latent διάνυσμα z αντί για ένα ντετερμινιστικό διάνυσμα r . Οι variational παράμετροι φ βελτιστοποιούνται για να προσεγγίσουν την posterior κατανομή $p_\theta(z|x)$, η οποία με τη σειρά της βοηθά στη βελτιστοποίηση της marginal πιθανότητας $p_\theta(x)$:

$$q_\varphi(z|x) \approx p_\theta(z|x)$$

Τόσο το variational μοντέλο $q_\varphi(z|x)$ όσο και το μοντέλο της κατανομής $p_\theta(x|z)$ μπορούν να παραμετροποιηθούν χρησιμοποιώντας βαθιά νευρωνικά δίκτυα με εκπαιδευμένες παραμέτρους θ και φ αντίστοιχα.

Το κριτήριο βελτιστοποίησης του VAE είναι η μεγιστοποίηση του κατώτατου ορίου πιθανότητας (evidence lower bound - ELBO), και στοχεύει στην καλύτερη προσέγγιση της πραγματικής κατανομής:

$$\mathcal{L}(q_\varphi(z|x)) = \mathbb{E}_{q_\varphi(z|x)} \left[\log \frac{p_\theta(x, z)}{q_\varphi(z|x)} \right] \quad (1)$$

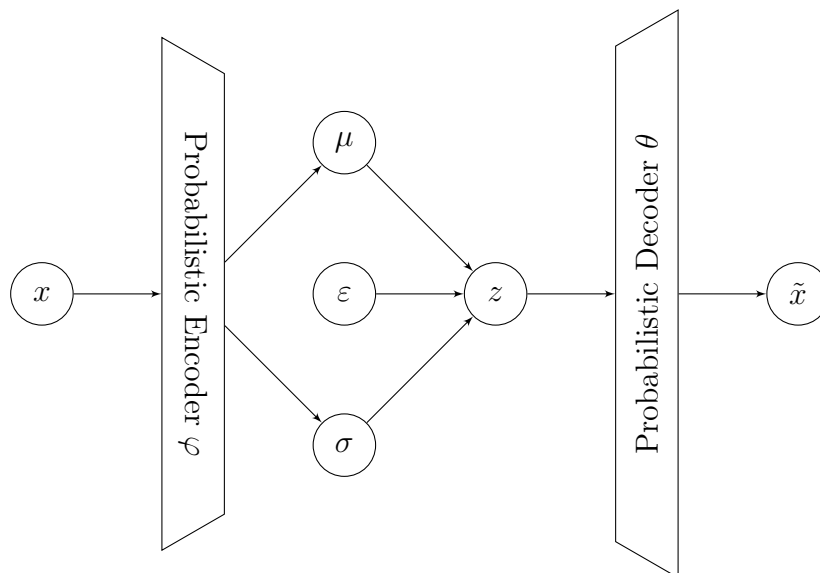


Figure 2: Γραφική αναπαράσταση Variational Autoencoder

1.2.2 Ημι-επιβλεπόμενη μάθηση με VAEs

Τα τελευταία χρόνια, η προσοχή έχει στραφεί στην ημι-επιβλεπόμενη μάθηση λόγω της δυνατότητας της να αντιμετωπίζει προβλήματα ταξινόμησης που περιλαμβάνουν μεγάλο όγκο δεδομένων αλλά και απαιτούν δύσκολες ή δαπανηρές διαδικασίες επισήμανσης. Αξιοποιώντας δεδομένα τόσο με ετικέτες όσο και χωρίς ετικέτες, η ημι-επιβλεπόμενη μάθηση μπορεί να εκτελέσει επιβλεπόμενη ή μη επιβλεπόμενη μάθηση, βελτιώνοντας την διαδικασία μάθησης με δεδομένα χωρίς ετικέτα (Zhu, 2005).

Οι VAEs έχουν πρόσφατα αποκτήσει δημοτικότητα για την αντιμετώπιση ημι-επιβλεπόμενων εργασιών μάθησης. Οι Kingma et al. (2014) προτείνουν τρεις τρόπους χρήσης της latent αναπαράστασης των δεδομένων για τη βελτίωση της απόδοσης ταξινόμησης χρησιμοποιώντας VAE και εμείς θα προσθέσουμε έναν τέταρτο.

1. Το πρώτο μοντέλο (M1), που εισήχθη από τους Kingma et al. (2014), αναφέρεται ως το *latent-feature discriminative* μοντέλο (Figure 3). Το μοντέλο αυτό, περιλαμβάνει εκπαίδευση ενός VAE σε όλα τα παρατηρούμενα δεδομένα x για την εκτέλεση μη επιβλεπόμενης εξαγωγής latent μεταβλητών z . Στη συνέχεια, χρησιμοποιώντας τις ετικέτες y , εκπαιδεύεται ένας ξεχωριστός επιβλεπόμενος ταξινομητής στα δεδομένα με ετικέτα (z, y) .
2. Στην αρχιτεκτονική του δεύτερου μοντέλου (M2) (Kingma et al., 2014), οι πληροφορίες της ετικέτας ενσωματώνονται κατά την εξαγωγή latent χαρακτηριστικών για

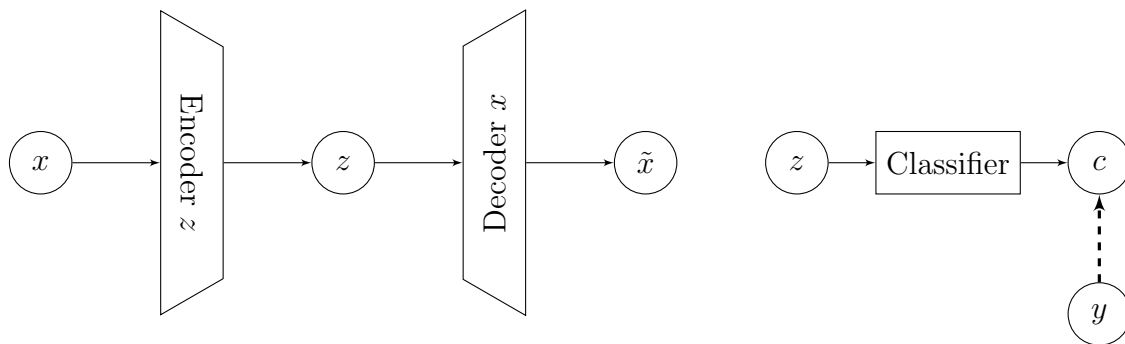


Figure 3: Γραφική αναπαράσταση *latent-feature discriminative* μοντέλου - M1

τη βελτίωση της απόδοσης της ταξινόμησης. Το M2 είναι ένα *generative semi-supervised* μοντέλο (Figure 4) που έχει δύο latent μεταβλητές, z και c . Οι ετικέτες y αντιμετωπίζονται ως είσοδος σε δεδομένα με ετικέτα και ως latent μεταβλητή $c = y$ σε δεδομένα χωρίς ετικέτα.

- Τέλος, το μοντέλο M1+M2 (Kingma et al., 2014), είναι ένα *stacked generative semi-supervised* μοντέλο που συνδυάζει τα πλεονεκτήματα του ημι-επιβλεπόμενου μοντέλου M2 και τη latent αναπαράσταση δεδομένων στο μοντέλο M1 (Figure 5). Για δεδομένα με ετικέτα, οι παρατηρούμενες μεταβλητές x και οι ετικέτες $y = c$ θεωρούνται ως είσοδος και οι latent μεταβλητές είναι z και u . Για δεδομένα χωρίς ετικέτα, μόνο η μεταβλητή x αντιμετωπίζεται ως είσοδος και οι latent μεταβλητές είναι οι z , c και u .

1.3 Μέθοδος

Θέλουμε να αντιμετωπίσουμε ένα πρόβλημα ταξινόμησης το οποίο περιλαμβάνει δεδομένα από πολλαπλές πηγές. Τα δεδομένα αυτά συμβολίζονται ως x_1, x_2, \dots, x_n . Στόχος μας είναι να δημιουργήσουμε ένα σύστημα που όχι μόνο εκμεταλλεύεται τα οφέλη ενός *stacked generative semi-supervised* μοντέλου αλλά και αντιμετωπίζει αποτελεσματικά την πρόκληση του χειρισμού πολυτροπικών δεδομένων.

Η προτεινόμενη προσέγγισή μας, έχει ως στόχο να παρέχει μια καινοτόμο λύση στην ταξινόμηση εικόνων τηλεπισκόπησης. Αποφασίσαμε να εμβαθύνουμε σε προηγμένες αρχιτεκτονικές υιοθετώντας το M1+M2 για ημι-επιβλεπόμενη ταξινόμηση με VAE. Επιδιώκοντας να ταξινομήσουμε πολυτροπικά δεδομένα, επεκτείνουμε αυτήν την αρχιτεκτονική με δύο διακριτούς τρόπους. Πρώτον, μέσω της συγχώνευσης σε επίπεδο δεδομένων και,

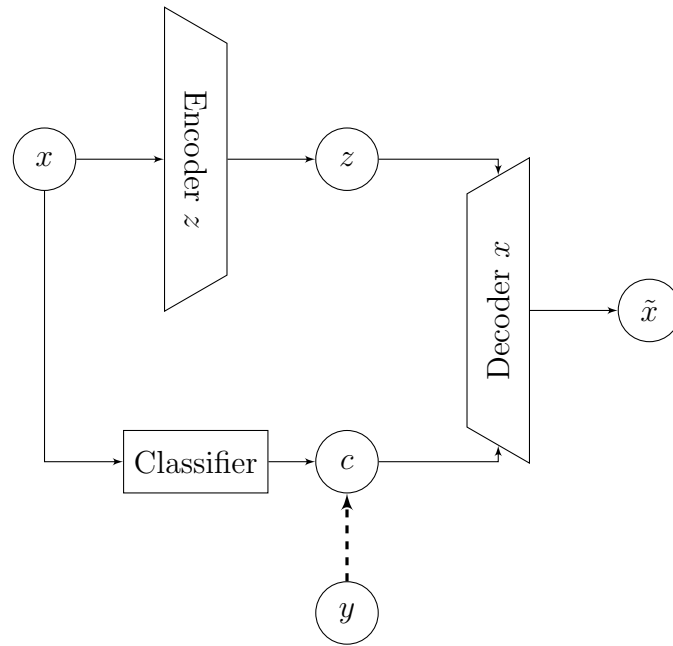


Figure 4: Γραφική αναπαράσταση *generative semi-supervised* μοντέλου - M2

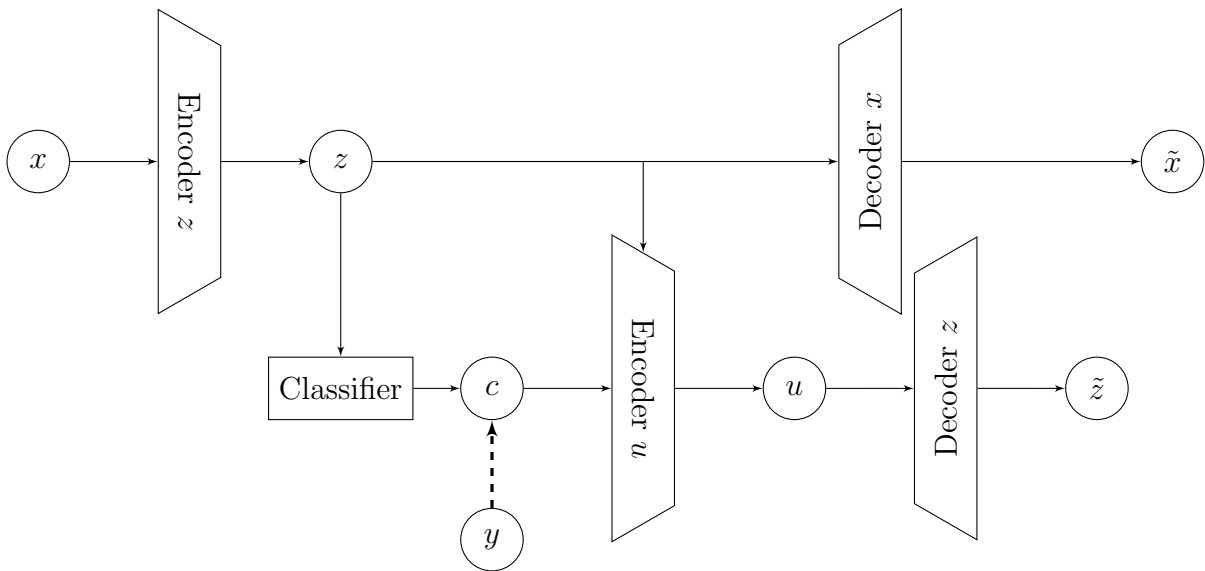


Figure 5: Γραφική αναπαράσταση *stacked generative semi-supervised* μοντέλου - M1+M2

δεύτερον, με την εισαγωγή ενός πρόσθετου επιπέδου latent μεταβλητής για κάθε είδος δεδομένων ώστε πραγματοποιηθεί συγχώνευση σε επίπεδο latent χαρακτηριστικών.

Συγχώνευση σε επίπεδο δεδομένων: Η αρχιτεκτονική M1+M2 έχει σχεδιαστεί για να ενσωματώνει τα δεδομένα σε μια ενιαία είσοδο για το μοντέλο. Αυτά τα δεδομένα, που αντιπροσωπεύονται ως x_1, x_2, \dots, x_n , θα μπορούσαν να είναι διαφόρων τύπων όπως εικόνες, ήχος ή κείμενο. Συνδυάζοντας τα, δημιουργούμε μια ενοποιημένη είσοδο x που τα ενσωματώνει όλα μαζί. Με τον ίδιο τρόπο όπως στην αρχική προσέγγιση που προτάθηκε από τους Kingma et al. (2014), η συνενωμένη είσοδος x χρησιμοποιείται για την εκπαίδευση ενός παραγωγικού μοντέλου και ενός ταξινομητή ταυτόχρονα.

Συγχώνευση σε επίπεδο latent χαρακτηριστικών: Προτείνουμε το *multimodal stacked generative semi-supervised* μοντέλο ή Multi-M1+M2. Το οποίο αντιμετωπίζει κάθε τύπο των παρατηρούμενων δεδομένων x_1, x_2, \dots, x_K ως ξεχωριστή είσοδο (Figure 6). Κάθε μεταβλητή εισόδου κωδικοποιείται σε μια αντίστοιχη latent μεταβλητή z_1, z_2, \dots, z_K . Αυτές οι latent μεταβλητές στη συνέχεια συνενώνονται και τροφοδοτούνται στο υπόλοιπο δίκτυο. Για δεδομένα με ετικέτα, οι παρατηρούμενες μεταβλητές x_1, x_2, \dots, x_K και οι ετικέτες $y = c$ θεωρούνται ως είσοδος και οι latent μεταβλητές είναι z_1, z_2, \dots, z_K και u . Για δεδομένα χωρίς ετικέτα, μόνο οι μεταβλητές x_1, x_2, \dots, x_K αντιμετωπίζονται ως είσοδος και οι latent μεταβλητές είναι z_1, z_2, \dots, z_K, c και u .

1.4 Εφαρμογή σε δεδομένα τηλεπισκόπησης

Στην συνέχεια παρουσιάζουμε τα πειράματά μας, τα οποία περιλαμβάνουν ταξινόμηση κάλυψης Γης αγροτικών και αστικών περιοχών χρησιμοποιώντας ημι-επιβλεπόμενο VAE. Η ταξινόμηση κάλυψης Γης κατηγοριοποιεί την επιφάνεια της Γης με βάση τα χαρακτηριστικά της, συνήθως χρησιμοποιώντας δορυφορικές εικόνες ή αεροφωτογραφίες. Η ανάλυση μας επικεντρώνεται σε δύο αρχιτεκτονικές νευρωνικών δικτύων: M1+M2 και Multi-M1+M2.

Για την αξιολόγηση των μεθόδων αξιοποιούμε δύο πολυτροπικά σύνολα δεδομένων, το Trento Ghamisi et al. (2016) και το Houston (Xu et al., 2019) τα οποία περιλαμβάνουν ταξινόμηση κάλυψης Γης μιας αγροτικής και αστικής περιοχής αντίστοιχα. Τα δύο σύνολα δεδομένων αναλύονται ξεχωριστά. Κάθε σύνολο δεδομένων αποτελείται από δύο είδη δεδομένων: LiDAR και HSI.

Οι εικόνες HSI αποτελούνται από πολλαπλά κανάλια με φασματικές πληροφορίες (Figures 7a,8a). Τα δεδομένα LiDAR παρέχουν πληροφορίες υψομέτρου σε μέτρα πάνω από την επιφάνεια της θάλασσας (Figures 7b,8b). Τα LiDAR του Houston περιλαμβάνουν επίσης multispectral LiDAR. Το multispectral LiDAR αποτελείται από τρία κανάλια τα

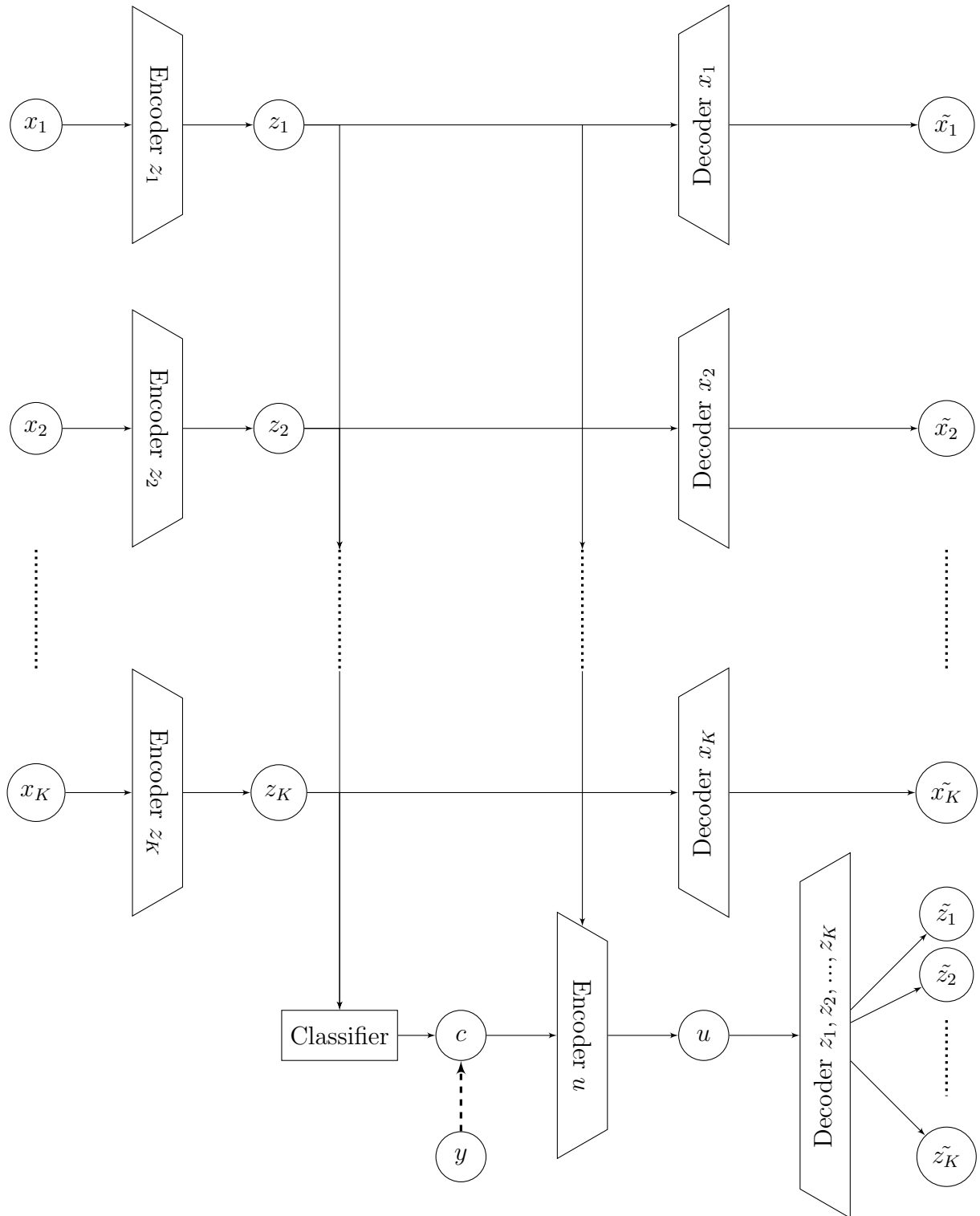


Figure 6: Γραφική αναπαράσταση *multimodal stacked generative semi-supervised* μοντέλο - Multi-M1+M2

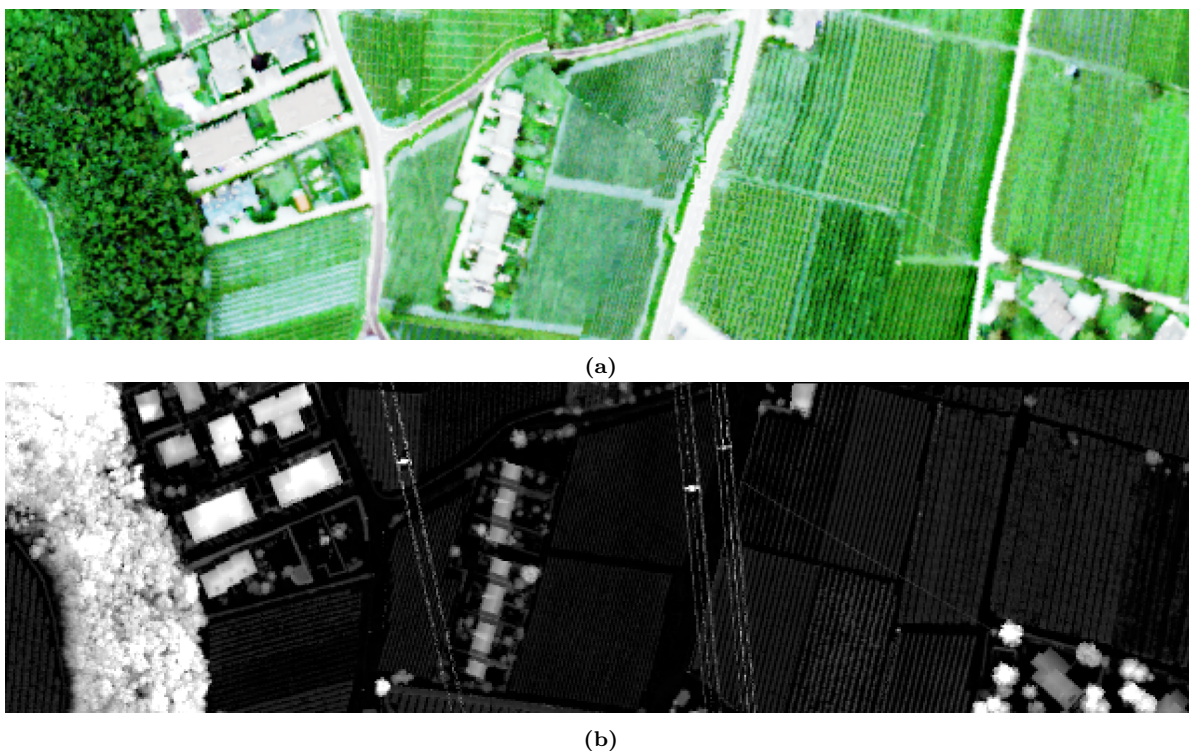


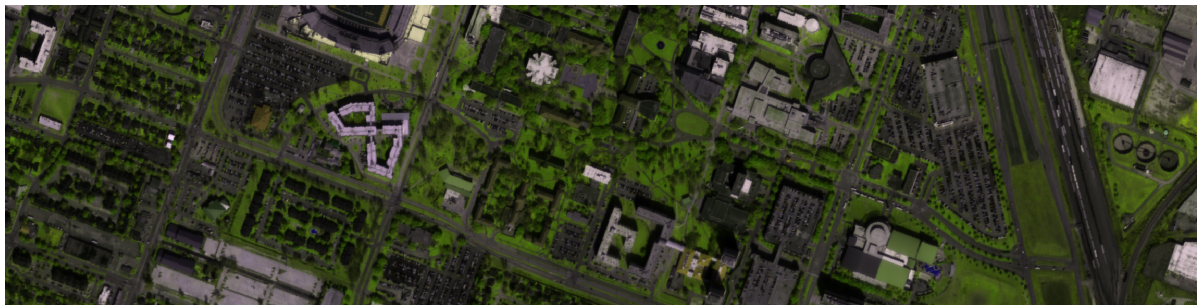
Figure 7: Σύνολο δεδομένων Trento: (a) Ψευδής RGB αναπαράσταση με χρήση καναλιών HSI, (b) LiDAR DSM

οποία παρέχουν πληροφορίες για την φύση της επιφανειας, όπως γυμνή γη ή ανθρωπογενείς δομές (Figure 8c).

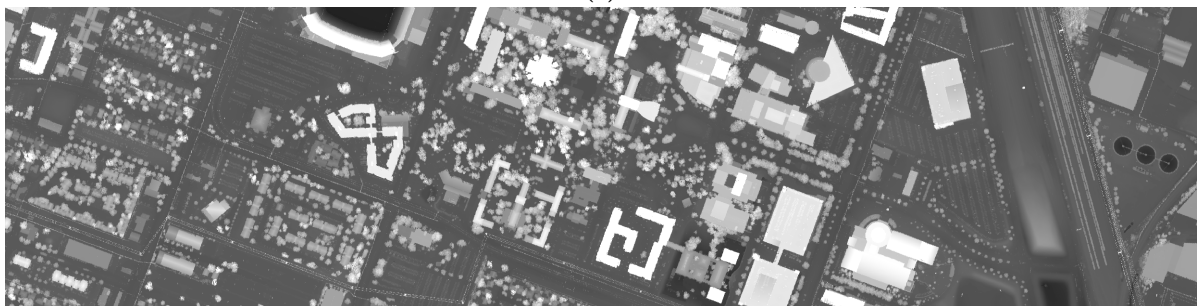
Για να συγκρίνουμε την επίδοση κάθε μοντέλου χρειάζεται να εκτιμήσουμε και να αναλύσουμε τα αποτελέσματα της ταξινόμησης του. Για να κάνουμε διακριτές τις διαφορές πτυχές της απόδοσης των μοντέλων, επιστρατεύουμε τόσο ποσοτικές όσο και ποιοτικές τεχνικές αξιολόγησης. Οι ποσοτικές μετρικές που χρησιμοποιούμε είναι: συνολική ορθότητα (accuracy), ακρίβεια (precision), ανάκληση (recall), βαθμολογία F1 (F1-score) και συντελεστής Kappa (Kappa coefficient). Η ποιοτική αξιολόγηση γίνεται με την χρήση χαρτών ταξινόμησης και χαρτών αβεβαιότητας. Αξιολογούμε το μέτρο της αβεβαιότητας που βασίζεται στην ομοφυλία (homophily based uncertainty - HU) η οποία προσεγγίζει την αβεβαιότητα των προβλέψεων των κλάσεων, σταθμισμένη από την ανομοιοτήτά τους (Chlaily et al., 2023).

1.4.1 Αγροτική περιοχή

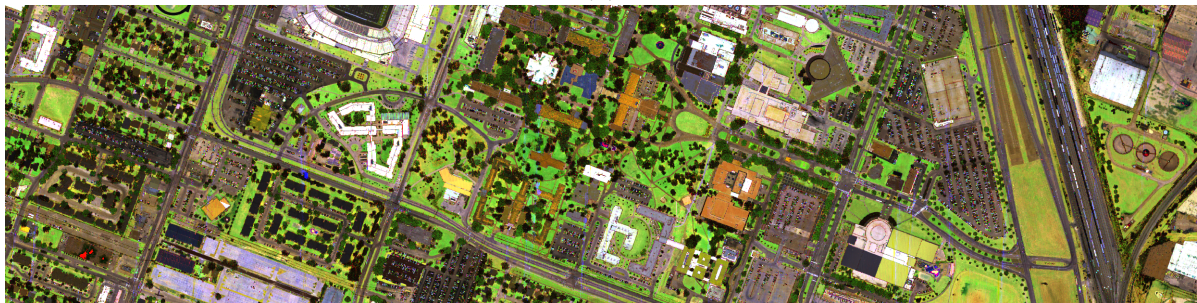
Στην πρώτη σειρά πειραμάτων, χρησιμοποιούμε το μοντέλο M1+M2 ως βάση για το δικό μας μοντέλο ταξινόμησης, προσαρμόζοντάς την αρχιτεκτονική των (Lopez et al., 2020)



(a)



(b)



(c)

Figure 8: Σύνολο δεδομένων Houston: (a) Ψευδής RGB αναπαράσταση με χρήση καναλιών HSI, (b) LiDAR DSM, (c) Χρωματική αναπαράσταση multispectral LiDAR

Table 1: Αρχιτεκτονική των κωδικοποιητών

Encoder	Layers				Pooling		Activation
	#	Type	in	out	Type	Size	
E1	3×	Linear	512	128	-	-	
E2	3×	Linear	1024	256	-	-	
E3	3×	1D Convolutional	32	128	1D Max pooling	2	SELU
E4	3×	1D Convolutional	128	512	1D Max pooling	2	
E5	2×	2D Convolutional	128	256	2D Max pooling	2	

που αρχικά σχεδιάστηκε για ταξινόμηση ψηφίων χρησιμοποιώντας το MNIST. Ωστόσο, εμείς εστιάζουμε στην ταξινόμηση εικονοστοιχείων εικόνων τηλεπισκόπησης, για αυτό τροποποιήσαμε το τμήμα κωδικοποίησης της αρχιτεκτονικής. Για να βελτιώσουμε το μοντέλο μας, πειραματιστήκαμε με διαφορετικούς κωδικοποιητές, μεταβάλλοντας το πλάτος, το βάθος και τους τύπους επιπέδων τους. Δημιουργήσαμε τον Κωδικοποιητή 1 (E1) και τον Κωδικοποιητή 2 (E2) με τρία πλήρως συνδεδεμένα γραμμικά επίπεδα σε διαφορετικά μεγέθη. Ο Κωδικοποιητής 3 (E3) και ο Κωδικοποιητής 4 (E4) χρησιμοποίησαν ένα πλήρως συνδεδεμένο επίπεδο ακολουθούμενο από τρία 1D συνελικτικά επίπεδα σε διαφορετικά μεγέθη. Ο κωδικοποιητής 5 (E5) επεξεργάζεται patches από τη σκηνή χρησιμοποιώντας δύο 2D συνελικτικά επίπεδα. Ο E5 ενσωματώνει χωρικές πληροφορίες, αλλά έχει υψηλότερη πολυπλοκότητα εκπαίδευσης (Table 1).

Προκειμένου να προσδιοριστεί το βέλτιστο μέγεθος latent χώρου, που να αντιπροσωπεύει τα δεδομένα εισόδου, πραγματοποιήσαμε πειράματα με μεγέθη που κυμαίνονται από 10 έως 20. Λαμβάνοντας υπόψη ότι το σύνολο δεδομένων Trento περιλαμβάνει έξι κατηγορίες, στοχεύσαμε σε μια σειρά μεγεθών latent χώρου συγκρίσιμων με τον αριθμό των τάξεων. Αρχικά, εκπαιδεύσαμε την αρχιτεκτονική M1+M2 για τους κωδικοποιητές E1-E4 στα διάφορα μεγέθη latent χώρου. Με βάση την απόδοση αυτών των πειραμάτων, επιλέξαμε το latent μέγεθος με την καλύτερη απόδοση και το χρησιμοποιήσαμε για την εκπαίδευση του E5. Τέλος, επιλέξαμε τον κωδικοποιητή με την καλύτερη απόδοση για την εκπαίδευση της αρχιτεκτονικής Multi-M1+M2. Αφού αναλύσουμε τα ημι-εποπτευόμενα μοντέλα VAE, θέλουμε να τα συγκρίνουμε με την απόδοση των ευρέως χρησιμοποιούμενων μεθόδων Support Vector Machine (SVM) και Random Forest (RF).

Από τα πειράματα που διεξήχθησαν στο σύνολο δεδομένων του Trento, παρατηρούμε ότι η επιλογή του κωδικοποιητή και το μέγεθος του latent χώρου επηρεάζουν σημαντικά την απόδοση της ταξινόμησης. Συγκεκριμένα, η αύξηση του μεγέθους του latent χώρου δεν εξασφαλίζει πάντα βελτιωμένη απόδοση (Table 2). Επιπλέον, η πειραματική μας διαδικασία με τους κωδικοποιητές αποκάλυπτε μία αξιοσημείωτη παρατήρηση: ενώ η

Table 2: Σύνολο δεδομένων Trento: Μετρικές για το μοντέλο M1+M2 με τους κωδικοποιητές E1, E2, E3, και E4 σε διαφορετικά μεγέθη latent χώρου

Encoder	Latent size	Accuracy	Precision	Recall	F1-score	Kappa coeff
E1	10	84.80	73.05	80.51	76.09	78.91
	15	84.86	72.19	80.76	75.63	79.03
	20	84.94	72.38	80.70	75.75	79.13
E2	10	90.44	88.55	92.58	89.94	87.44
	15	92.48	92.17	92.92	92.43	90.01
	20	88.48	87.70	92.28	88.68	85.01
E3	10	90.24	87.92	91.89	89.40	87.13
	15	84.68	71.49	80.50	75.05	78.79
	20	89.52	88.13	91.65	89.36	86.22
E4	10	91.13	89.62	92.54	90.76	88.28
	15	90.70	88.83	91.93	90.01	87.73
	20	91.85	90.10	92.77	91.20	89.21

Table 3: Σύνολο δεδομένων Trento: Σύγκριση μετρικών των καλύτερων VAE μοντέλων και των SVM, RF

Μοντέλο	Accuracy	Precision	Recall	F1-score	Kappa coeff
M1+M2 with E2	92.48	92.17	92.92	92.43	90.01
M1+M2 with E5	96.69	92.55	95.08	93.53	95.59
Multi-M1+M2 with E2	87.49	87.09	91.60	87.85	83.77
SVM	86.34	83.07	88.80	84.83	82.10
RF	88.38	85.38	89.56	86.92	84.62

χρήση χωρικών πληροφοριών στην είσοδο οδηγεί σε βελτιωμένα μετρικά αποτελέσματα, ταυτόχρονα οδηγεί και σε λιγότερο ευδιάκριτα αποτελέσματα, το οποίο είναι ανεπιθύμητο (Figure 9g). Το μοντέλο Multi-M1+M2 που περιλαμβάνει συγχώνευση στο επίπεδο των latent χαρακτηριστικών παρουσιάζει την καλύτερη επίδοση στον διαχωρισμό μεταξύ στενά συναφών κατηγοριών βλάστησης (Figure 9d). Η οπτική ανάλυση των χαρτών ταξινόμησης παρέχει βαθύτερες λεπτομέρειες στη συμπεριφορά του μοντέλου, επισημαίνοντας περιοχές σύγχυσης και εσφαλμένων ταξινομήσεων (Figure 9). Επίσης, οι χάρτες αβεβαιότητας βοηθούν στην κατανόηση της αυτοπεποίθησης του μοντέλου για τις ταξινομήσεις του και στον εντοπισμό περιοχών πιθανής σύγχυσης (Figure 10). Μια πιο λεπτομερής σύγκριση με τα μοντέλα SVM και RF αναδεικνύει την ανώτερη δυνατότητα ταξινόμησης των τεχνικών βαθιάς μάθησης, αλλά και το γεγονός ότι τείνουν να έχουν υποδεέστερη αποδοτικότητα χρόνου (Table 3, Figure 9).

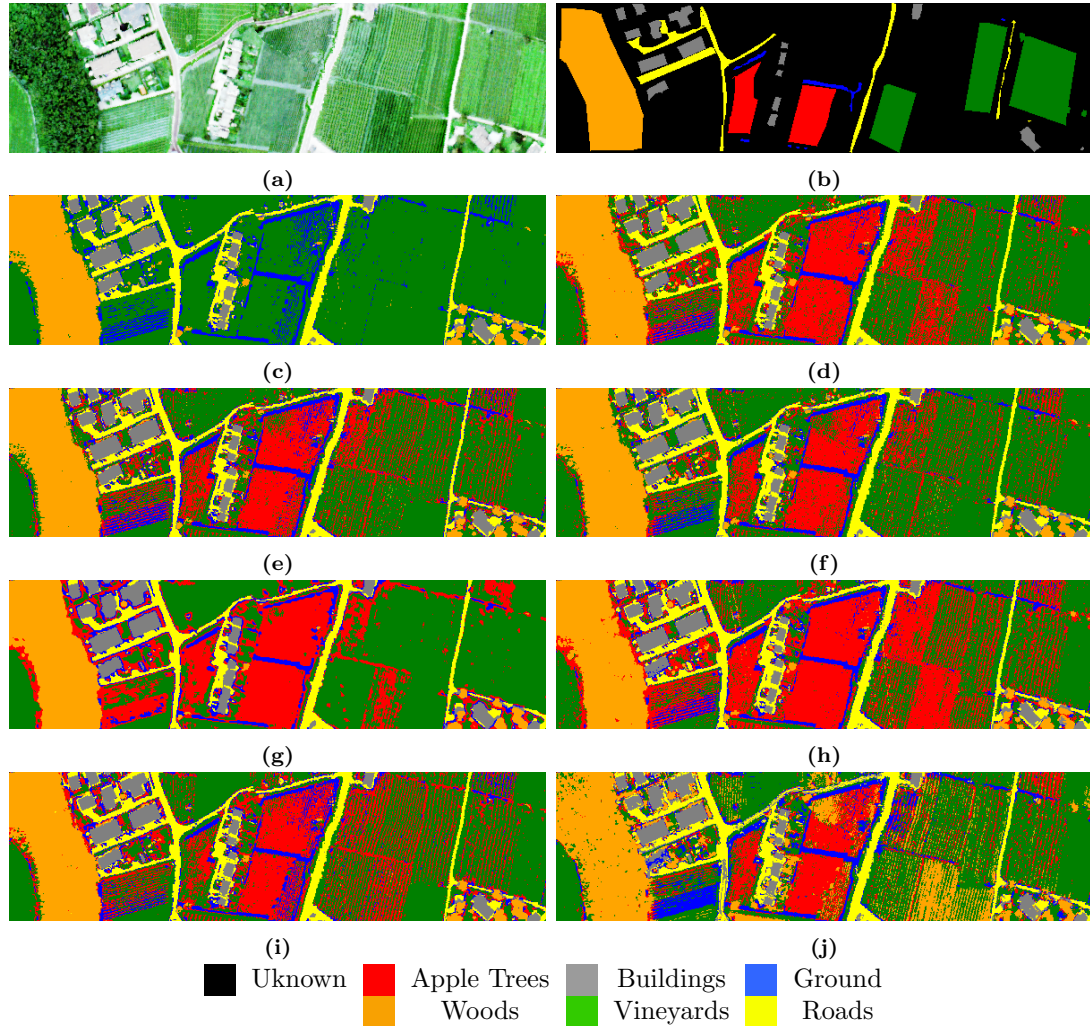


Figure 9: Σύνολο δεδομένων Trento: (a) Ψευδής RGB αναπαράσταση με χρήση καναλιών HSI, (b) Ετικέτες, Χάρτες ταξινόμησης για το μοντέλο M1+M2 με: (c) E1 για latent size = 20, (d) E2 για latent size = 15, (e) E3 για latent size = 10, (f) E4 για latent size = 20, (g) E5 για latent size = 15 and patch size = 5x5, (h) και για το μοντέλο Multi-M1+M2 με E2 για latent size = 15 καθώς και για (i) SVM, (j) RF

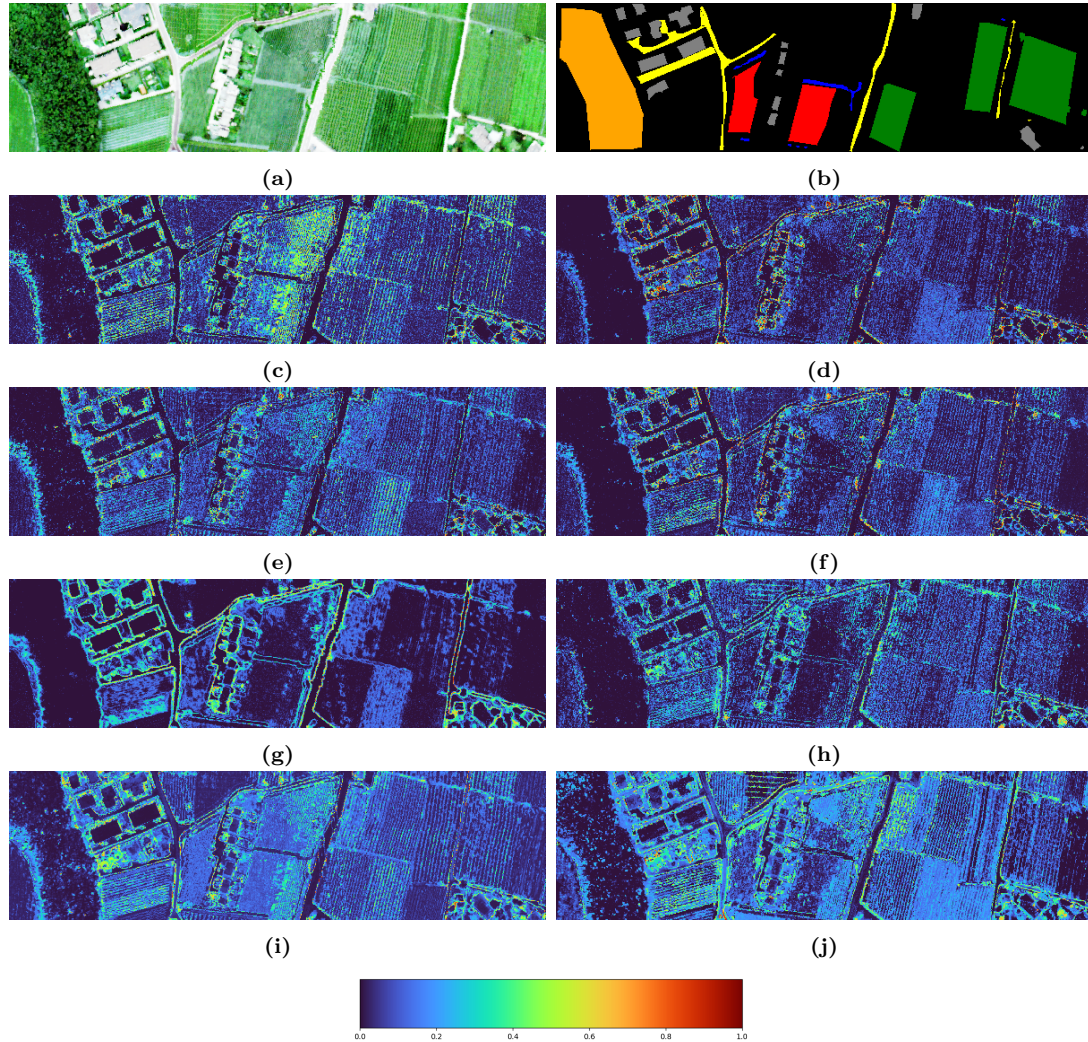


Figure 10: Σύνολο δεδομένων Trento: (a) Ψευδής RGB αναπαράσταση με χρήση καναλιών HSI, (b) Ετικέτες, Χάρτες αβεβαιότητας για το μοντέλο M1+M2 με (c) E1 για latent size = 20, (d) E2 για latent size = 15, (e) E3 για latent size = 10, (f) E4 για latent size = 20, (g) E5 για latent size = 15 and patch size = 5x5, (h) και για το μοντέλο Multi-M1+M2 με E2 για latent size = 15 and (i) SVM, (j) RF

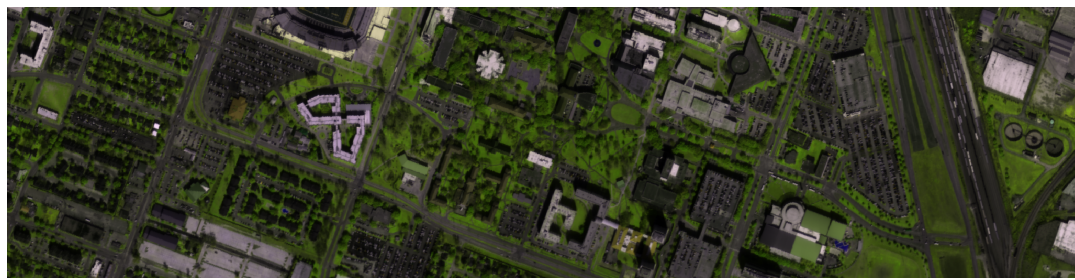
Table 4: Σύνολο δεδομένων Houston: Μετρικές για το μοντέλο M1+M2 με κωδικοποιητή E2 σε διαφορετικά μεγέθη latent χώρου

Latent size	Accuracy	Precision	Recall	F1-score	Kappa coeff
10	64.75	51.42	76.47	58.39	56.06
20	66.03	51.94	76.58	57.84	58.81
30	77.26	64.12	78.96	68.89	71.53
40	64.79	43.72	65.19	49.86	55.62

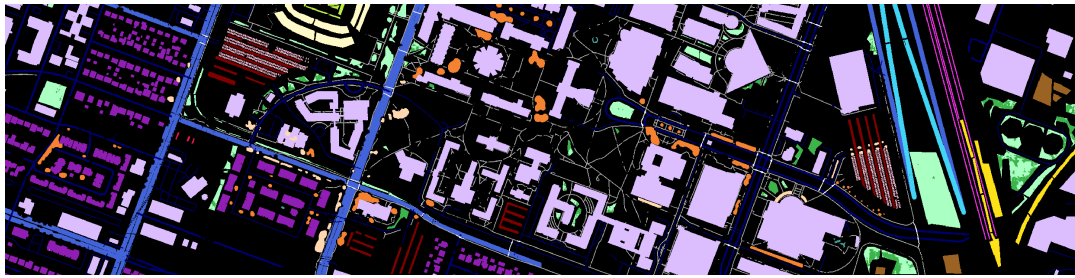
1.4.2 Αστική περιοχή

Στη δεύτερη σειρά πειραμάτων μας, στοχεύουμε να αξιοποιήσουμε τα προηγούμενα αποτελέσματα στο σύνολο δεδομένων αγροτικών περιοχών και να διερευνήσουμε την απόδοση των ημι-επιβλεπόμενων μοντέλων VAE σε ένα πιο σύνθετο σύνολο δεδομένων. Συγκεκριμένα, θέλουμε να αξιολογήσουμε πώς θα αποδώσουν οι διαφορετικοί τρόποι σύντηξης δεδομένων όταν αντιμετωπίζουμε σημαντικά μεγαλύτερο αριθμό κλάσεων και κλάσεις με πιο δυσδιάκριτες διαφορές. Για να το πετύχουμε αυτό, επικεντρωνόμαστε στην εξερεύνηση του κωδικοποιητή E2, ο οποίος έδειξε τα πιο ελπιδοφόρα αποτελέσματα όσον αφορά τις μετρήσεις και την ποσοτική ανάλυση στα προηγούμενα πειράματά μας.

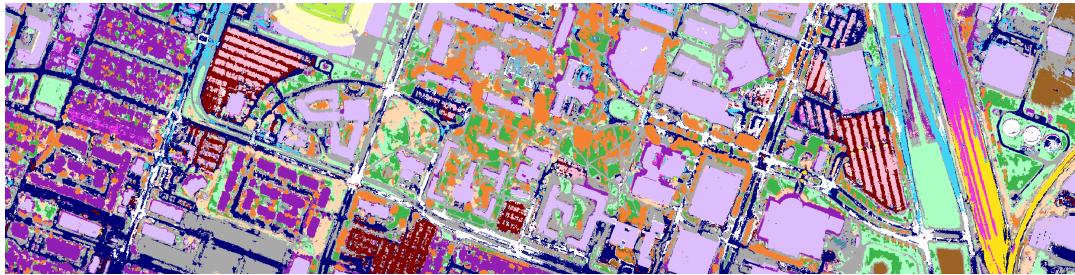
Για το σύνολο δεδομένων Houston, η έρευνά μας σχετικά με την επίδραση του μεγέθους του latent χώρου στην απόδοση της ταξινόμησης αποκαλύπτει ότι ένα βέλτιστο μέγεθος του latent χώρου ισορροπεί την ικανότητα του μοντέλου να αντιλαμβάνεται τις βαθιές πολυπλοκότητες των δεδομένων (Table 4). Το μοντέλο M1+M2 και το Multi-M1+M2 εμφανίζουν ισχυρή απόδοση στον εντοπισμό των περισσότερων αστικών κατηγοριών. Ωστόσο, κατανοούμε ότι, σε ορισμένες περιπτώσεις, η εξισορρόπηση των δεδομένων εισόδου με τη συγχώνευση στο επίπεδο των latent χαρακτηριστικών μπορεί να οδηγήσει σε χειρότερα αποτελέσματα. Η ενσωμάτωση των δεδομένων HSI και LiDAR μέσω της συγχώνευσης στο επίπεδο των latent χαρακτηριστικών παρέχει σημαντικές βελτιώσεις στην διάκριση μεταξύ κατηγοριών με διακριτά φάσματα και LiDAR υπογραφές. Όμως, για κατηγορίες με παρόμοια χαρακτηριστικά HSI ή LiDAR, όπως διάφορα είδη δρόμων, οι μέθοδοι συγχώνευσης είναι λιγότερο αποτελεσματικές (Table 5). Η οπτική ανάλυση των χαρτών ταξινόμησης και των χαρτών αβεβαιότητας βοηθάει στην κατανόηση των δυνατοτήτων των μοντέλων μας, καθώς και στον εντοπισμό περιοχών όπου διάφοροι ταξινομητές αντιμετωπίζουν δυσκολίες (Figures 11, 12).



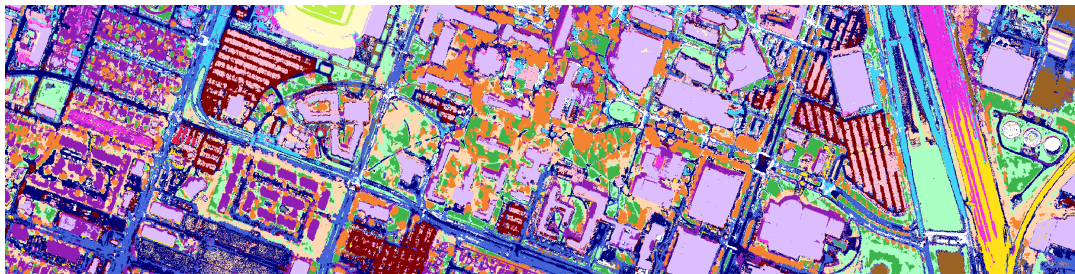
(a)



(b)



(c)



(d)

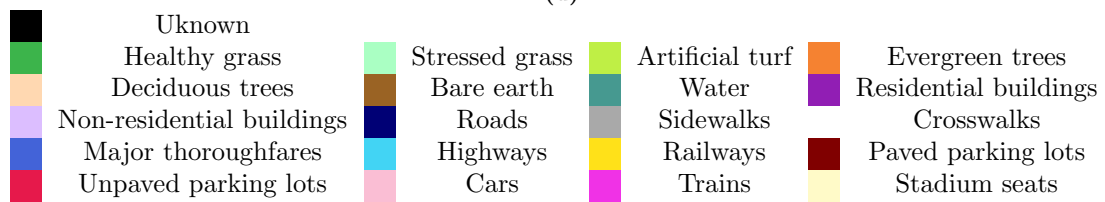
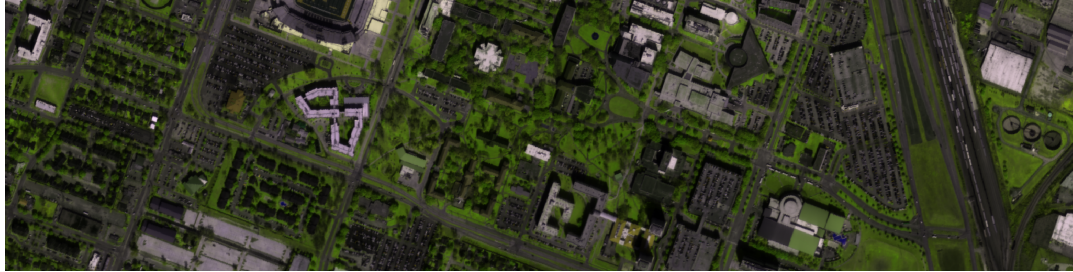


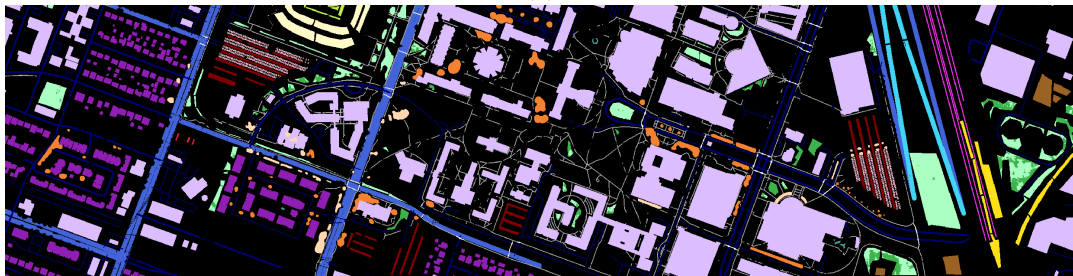
Figure 11: Σύνολο δεδομένων Houston: (a) Ψευδής RGB αναπαράσταση με χρήση καναλιών HSI, (b) Ετικέτες, Χάρτες ταξινόμησης για τα μοντέλα: (c) M1+M2 με E2 για latent size = 30, (d) Multi-M1+M2 με E2 για latent size = 30

Table 5: Σύνολο δεδομένων Houston: Μετρικές για τα καλύτερα VAE μοντέλα

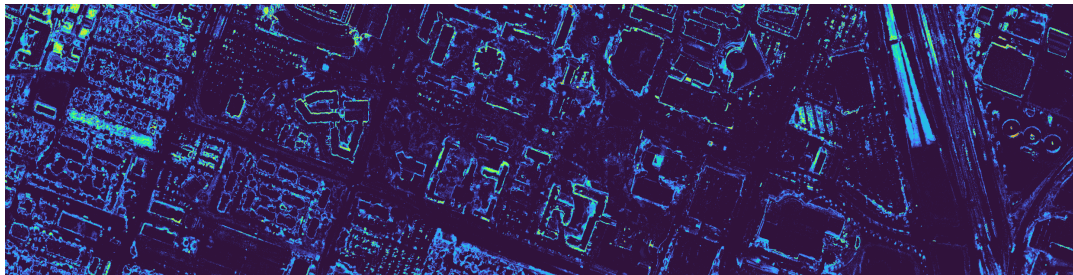
Μοντέλο	Accuracy	Precision	Recall	F1-score	Kappa coeff
M1+M2 with E2	77.26	64.12	78.96	68.89	71.53
Multi-M1+M2 with E2	70.81	57.53	80.23	64.41	64.22



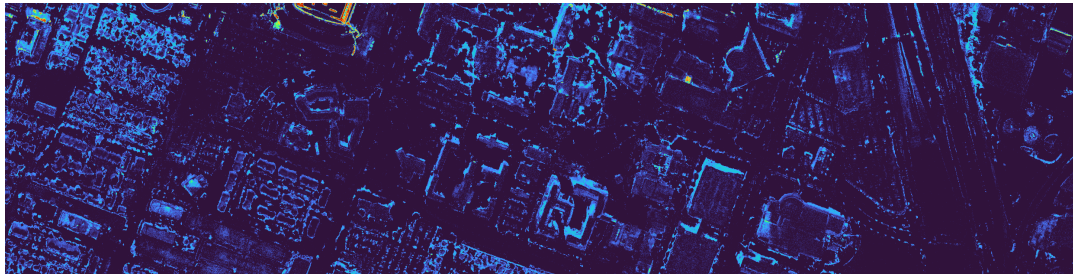
(a)



(b)



(c)



(d)

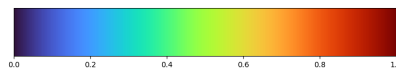


Figure 12: Σύνολο δεδομένων Houston: (a) Ψευδής RGB αναπαράσταση με χρήση καναλιών HSI, (b) Ετικέτες, Χάρτες αβεβαιότητας για τα μοντέλα: (c) M1+M2 model με E2 για latent size = 30, (d) Multi-M1+M2 με E2 για latent size = 30

1.5 Συζήτηση και συμπεράσματα

Γενικά, βλέπουμε ότι οι ταξινομητές έχουν καλύτερη απόδοση στο σύνολο δεδομένων Trento σε σύγκριση με το Houston. Οι διαφορές σε αυτά τα σύνολα δεδομένων περιλαμβάνουν τον τύπο του περιβάλλοντος, την ποικιλία των κλάσεων τους και την φύση των δεδομένων LiDAR.

Εν κατακλείδι, η έρευνά μας έθεσε μια ισχυρή βάση για την ταξινόμηση πολυτροπικών δεδομένων τηλεπισκόπησης χρησιμοποιώντας μια ημι-επιβλεπόμενη προσέγγιση με VAE. Ενσωματώνοντας τόσο ποσοτικές μετρήσεις όσο και ποιοτική ανάλυση, κατανοήσαμε ότι η αύξηση των ποσοτικών μετρήσεων δεν αντιστοιχεί πάντοτε σε υψηλή ποιοτική απόδοση. Τα πειραματικά μας αποτελέσματα απέδειξαν την αποτελεσματικότητα της προτεινόμενης αρχιτεκτονικής του Multi-M1+M2 μοντέλου και ανέδειξαν τα πλεονεκτήματα και τα μειονεκτήματα της προσέγγισής μας στην ταξινόμηση που πολυτροπικών δεδομένων. Καθώς συνεχίζουμε να εξερευνούμε τον περίπλοκο κόσμο της ανάλυσης πολυτροπικών δεδομένων, το προτεινόμενο πλαίσιο ανοίγει τον δρόμο για καινοτόμες λύσεις και νέες εφαρμογές στον τομέα της τηλεπισκόπησης και πέραν αυτού.

1.6 Δημοσιεύσεις

Η έρευνα που πραγματοποίησα στο Earth Observation Laboratory του Arctic University of Norway (UiT) οδήγησε στις ακόλουθες δημοσιεύσεις:

- Chlaily, S., Ratha, D., Lozou, P. and Marinoni A. (2023). On measures of uncertainty in classification, *IEEE Transactions on Signal Processing* **71**: 3710-3725.
- Khachatryan, E., Sandalyuk, N., Lozou, P. (2023). Eddy Detection in the Marginal Ice Zone with Sentinel-1 Data Using YOLOv5, *Remote Sensing* **15**: 2244.

2 Introduction

We live in the Big Data era (Mashey, 1999). Every second, a vast amount of various data are produced, stored, and analyzed. According to IDC prediction by 2025 the global volume of digital data will reach 175 zeta bytes. The manipulation of these volumes creates the need for automation in data analysis which is the role of machine learning (Murphy, 2012). Specifically, machine learning is a term that includes methods for pattern recognition, regression or decision making (Murphy, 2012). In contrast to other methods that use specific predefined instructions, machine learning algorithms make data-driven decisions based on experience extracted from the examples (Bishop and Nasrabadi, 2006).

Machine learning has numerous applications in various fields, including healthcare (Qayyum et al., 2020), marketing (Ma and Sun, 2020), and finance (Henrique et al., 2019), and is particularly useful in tasks such as clustering, regression, and classification. In this thesis, we are interested in exploring the application of machine learning in remote sensing. The integration of machine learning with remote sensing has led to significant advancements and opened up new research opportunities and innovations. For example, according to Scheunders et al. (2018), classification problems accounted for 16% of papers published on remote sensing in ISI Web-Science between 2004 and 2015. These statistics demonstrate the importance of machine learning in remote sensing data analysis, which poses numerous challenges.

The purpose of this introduction is to provide readers with a comprehensive overview of the thesis's focus. It begins by offering a brief overview of machine learning in section 2.1. The discussion then delves into remote sensing, providing details on its various types, as well as its historical context in section 2.2. Additionally, the application of machine learning to remote sensing is explored in section 2.3. The thesis' objectives are then described, followed by a detailed literature review of relevant works in section 2.4. By the end of this introduction, readers will have a clear understanding of the background and context of the research, as well as its significance and contribution to the field of study.

2.1 Machine learning

Depending on the availability and the use of the training data machine learning can be divided into several categories (Mohri et al., 2018). We are going to focus on the three main ones:

- Supervised learning: Labeled training data are used to learn the correspondence between the samples and the labels or target values and make predictions for unseen samples.
- Unsupervised learning: Unlabeled training data are used to learn patterns that can be also identified in unseen samples.
- Semi-Supervised learning: Training data consists of both labeled and unlabeled samples to make predictions in unseen samples.

Machine learning is a rapidly growing field with a wide range of applications. From predicting sales trends (Alon et al., 2001) to identifying fraudulent transactions (Shirgave et al., 2019), machine learning methods can be used to solve a variety of problems. In Table 6, we highlight some commonly used machine learning methods and their applications. These methods can be used to predict patterns in data (Burgess, 1998), classify (Li et al., 2014) or categorize data (Jiang et al., 2012), make decisions based on conditions (Rish et al., 2001), and identify important features in data (Biau and Scornet, 2016). Understanding the applications of machine learning methods can help researchers and practitioners select the appropriate method for their specific problem.

Artificial neural networks (ANNs) are a subset of machine learning systems inspired by the structure and function of the human brain. They consist of layers of interconnected nodes, or "neurons," that process information and make predictions based on input data (Krohn et al., 2019). Deep learning, refers to ANNs with many layers that allow for complex, non-linear relationships between inputs and outputs (Krohn et al., 2019).

Deep learning is a powerful tool that has revolutionized various fields such as computer vision, natural language processing, robotics, and more (Dong et al., 2021, Alzubaidi et al., 2021). In Table 7, we have listed some of the most popular deep learning techniques along with their applications. Each technique is uniquely designed to tackle specific tasks and problems. Understanding their strengths and limitations can help researchers and practitioners choose the right approach for their applications.

Table 6: Applications of machine learning methods

Method	Applications
Linear/Logistic Regression	Model relationships between variables and predicting trends or patterns in data
Decision Trees /Random Forest	Making decisions based on a series of conditions, identifying important features in data
Support Vector Machines	Classifying data into multiple categories, finding the best boundary between classes
K-Nearest Neighbors	Clustering by finding similar data points, making predictions based on similar data
Naive Bayes	Estimating probabilities based on prior knowledge, predicting the likelihood of an event
Artificial Neural Networks	Learning complex patterns in data, making predictions based on large amounts of data

Table 7: Deep learning techniques and their applications

Deep Learning Technique	Applications
Convolutional Neural Networks (CNNs)	Image and video recognition, object detection, segmentation, and classification.
Generative Adversarial Networks (GANs)	Image and video synthesis, super-resolution, style transfer, data augmentation, image inpainting and anomaly detection.
Deep Reinforcement Learning	Decision making in complex environments.
Autoencoders	Data compression, image denoising, feature learning, anomaly detection, image generation and recommendation systems.
Recurrent Neural Networks	Language modeling, natural language processing, machine translation, speech recognition and time series prediction.
Long Short-Term Memory Networks	Sequence modeling, language translation, speech recognition, speech synthesis, text generation and music composition.
Transformers	Language modeling, natural language processing, machine translation and text summarization.
Deep Belief Networks	Unsupervised feature learning, image and audio denoising, dimensionality reduction, anomaly detection and speech processing.

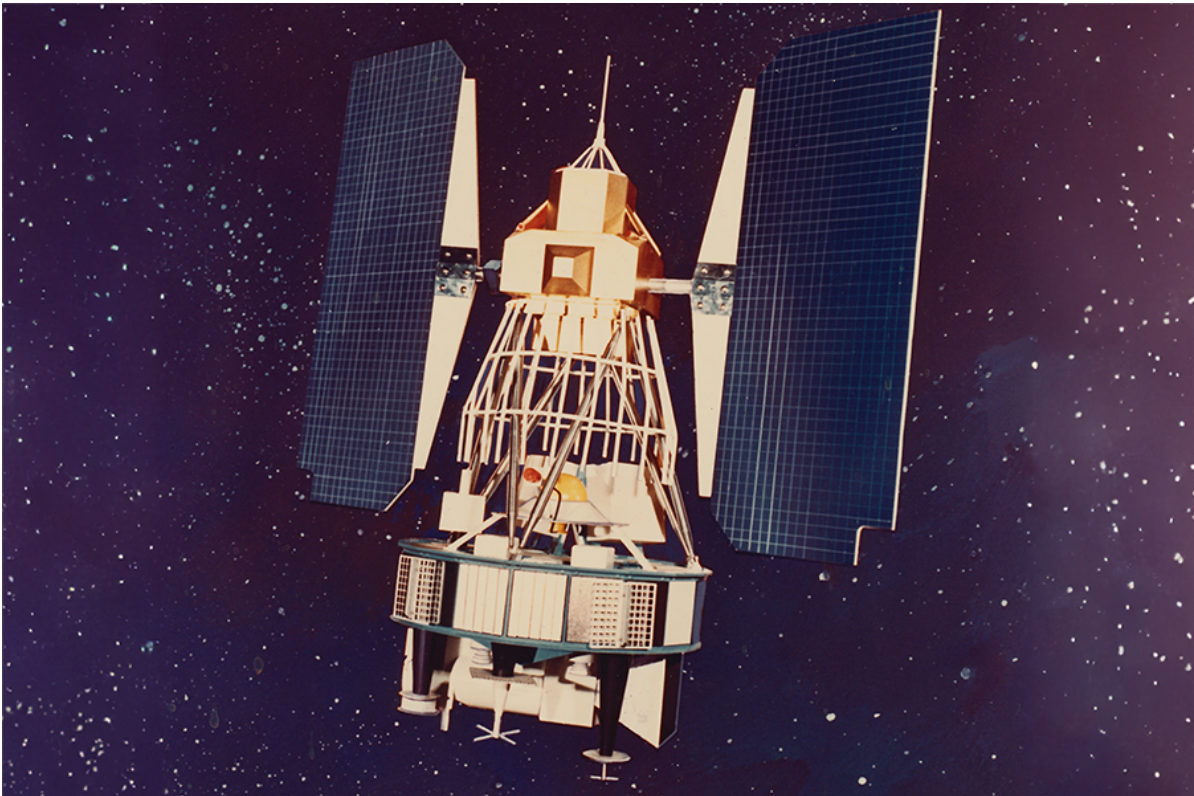


Figure 13: Landsat 1 illustration, ([Source](#))

2.2 Remote sensing

Over the past several decades, space-borne remote sensing has emerged as a primary source of data for a wide range of applications. With the ability to gather information about the Earth's surface without direct physical contact, remote sensing has proven to be a powerful tool. Remote sensing involves the measurement and analysis of electromagnetic radiation reflected or emitted from various sources, such as the Earth's surface, atmosphere, or oceans (Swain and Davis, 1981). As a result, remote sensing has become an essential tool for obtaining critical insights and understanding of various phenomena related to our planet.

2.2.1 Evolution of remote sensing

The history of remote sensing dates back to the early 19th century when photographic technology became available (Emerling, 2013), and people started to use air balloons and kites to observe the Earth's surface (e.g. Gaspard-Felix Tournachon, 1958). Since then, the technology and techniques used in remote sensing have evolved significantly, driven

by the need and interest in understanding our planet. Remote sensing has been applied in various fields such as cartography, agriculture, forestry, meteorology, and military operations.

Campbell and Wynne (2011) give a description of the evolution of remote sensing in the 20th century. During World War I, remote sensing techniques such as air photography were greatly advanced for military surveillance purposes. The use of improved cameras in aircraft during and after the war provided the potential for innovations like mapping areas and photogrammetry (use of photographs to make measurements). With the worldwide economic depression of 1929-1939, remote sensing was first used by the government to examine the impact of economic crises on the environment. During World War II, remote sensing expanded to the use of non-visible spectra for deeper penetration into enemy territories. Research and operations during the war built the theoretical and practical base for using non-visible spectra, specifically infrared and microwave. This technology continued to advance, leading to reliable systems not only for defense and security but also for civilian use.

Remote sensing technology has undergone significant evolution in recent decades. The Landsat 1 satellite (Figure 13), launched in 1972, was the first Earth-orbiting satellite designed for systematic Earth observation, marking the beginning of an era with extended availability of multispectral data (Lauer et al., 1997, Goward and Williams, 1997). Along with routine availability of multispectral data for large regions of the Earth's surface, there was rapid and broad expansion of uses of digital analysis for remote sensing. Landsat has since served as a model for the development of other land observation satellites operated by various organizations worldwide (Loveland and Dwyer, 2012).

Instruments with finer spatial resolution were developed by the early 1980s, and meter and sub-meter resolutions were developed by the early 1990s (Campbell and Wynne, 2011). Such progress, combined with the parallel development of geographic information systems (GIS), which provided the ability to bring remotely sensed data and other geospatial data into a common analytical framework, opened new civil application markets like mapping of urban infrastructure (Jensen and Cowen, 1999), geomorphology (Lo et al., 1997), and flood management (Lin et al., 2016).

Up to that point, the understanding and use of remote sensing was limited between the experts. At the beginning of the 21st century, the increasing capabilities of the internet led to remote sensing products becoming publicly available. This breakthrough in technology has allowed for a more widespread and accessible use of geospatial data

and remote sensing products (e.g. [Google Earth](#), 2005). As a result, the field of remote sensing has experienced substantial growth and innovation, as more researchers and professionals can now access and analyze geospatial data for a broad range of applications ([Gorelick et al., 2017](#)).

2.2.2 Remote sensors

There are two main types of remote sensors: passive and active. Passive remote sensors receive energy from the sunlight that is reflected by the targets or radiation that is naturally emitted from the Earth’s surface ([Campbell and Wynne, 2011](#)), while active sensors emit radiation to the object of interest and analyze the reflected radiation ([Cracknell, 2007](#)).

Active sensors: Active remote sensors use their own energy source to emit radiation towards the target and detect the back-scattered radiation to generate a remote sensing image. The main types of active remote sensing sensors include Light Detection And Ranging (LiDAR), Radio Detection and Ranging (Radar), Sound Navigation and Ranging (Sonar) and Synthetic Aperture Radar (SAR). LiDAR sensors use lasers to emit short pulses of light towards the target and measure the time it takes for the light to reflect back to the sensor, providing high-resolution elevation data for applications such as topographic mapping, forestry, and urban planning ([Dong and Chen, 2017](#)). Radar sensors emit radio waves towards the target and measure the strength and phase of the back-scattered signal to generate an image, which is used for various applications such as short-term weather forecasting and monitoring ([Fukao et al., 2014](#)), agriculture ([Dingle Robertson et al., 2020](#)), and military surveillance ([Fletcher, 1990](#)). Sonar sensors use sound waves to measure the distance and shape of underwater objects, making them ideal for marine exploration, fisheries, and naval operations ([Kolev, 2011](#)). SAR sensors transmit electromagnetic waves which can penetrate clouds, haze, rain and fog and precipitation with very little attenuation. SAR responds to dielectric properties, geometry, roughness of surface providing information about the texture, composition, and other features of the terrain ([Chan and Koo, 2008](#)). Some examples of active sensors and their applications are presented in Table 8.

Passive sensors: Passive remote sensors detect the natural electromagnetic radiation emitted or reflected by the Earth’s surface and atmosphere without any external sources. There are several types of passive remote sensing sensors, including optical, thermal, and polarimeters. Each type of sensor detects a different range of electromagnetic radiation, providing unique information about the region of interest. Optical

Table 8: Examples of active remote sensing sensors and their applications

Sensor type	Radiation emitted	Applications
LiDAR	Light	Topographic mapping, forestry, urban planning
Radar	Radio waves	Weather monitoring, agriculture, military
Sonar	Sound waves	Marine exploration, fisheries, naval operations
SAR	Microwaves	Topographic mapping, disaster monitoring
Altimetry	Microwaves	Ocean studies, sea level monitoring
Sounder	Impulses	Weather forecasts

sensors, such as cameras and scanners, which detect visible and near-infrared light are commonly used for oceanography (Dickey et al., 2006) or to provide information about vegetation properties (Van der Meij et al., 2017), and water resources (Huang et al., 2018). Thermal sensors, which detect infrared radiation emitted by objects have been used in precision agriculture (Khanal et al., 2017) and urban climate tracking (Voogt and Oke, 2003). Polarimeters, which measure the polarization of the incoming light are useful in detecting atmospheric properties such as aerosols (Travis, 1992), biomass tracking, and wetland monitoring (Boerner, 2003). Table 9 provides more examples of the radiation detected and applications for each type of passive remote sensing sensors.

Table 9: Examples of passive remote sensing sensors and their applications

Sensor type	Radiation detected	Applications
Optical	Visible light	Land use/cover mapping, urban planning
Hyperspectral	Visible to infrared light	Mineral mapping, pollution detection
Thermal	Infrared	Surface temperature, wildfire detection
Polarimeters	Polarized light	Aerosol studies, cloud properties,
Passive microwave	Microwave	Atmospheric temperature, precipitation
Radiometers	Infrared, ultraviolet	Surface temperature, atmospheric studies
Spectrometers	Light	Mineral identification, vegetation mapping

2.2.3 Multimodality

As previously discussed, individual sensors are capable of producing diverse and complementary data, resulting in datasets that are referred to as multimodal due to the co-existence of dissimilar modalities. At a human level, the use of multiple sensory information is a natural process through our senses. For instance, when we experience

food, our senses of smell, sight, and taste work together to provide us with a more comprehensive experience.

The utilization of multimodal data holds great potential in various domains. In the realm of communication, the fusion of hand shapes, mouthing patterns, and hand positions has been demonstrated to significantly enhance cued speech recognition (Papadimitriou et al., 2021). In the field of medicine, the integration of diverse data sources such as ultrasound images, laboratory tests, and clinical data has proven to be a robust strategy for the classification and staging of diseases (Ribeiro et al., 2012). For human-robot interactions, the use of multimodal data that combines sensory information from sources like force/torque sensors, visual sensors, and microphones has shown promise in enhancing user-robot interactions (Chalvatzaki et al., 2014, Zlatintsi et al., 2018). Furthermore, in sentiment analysis, the proper integration of audio, text, and visual data can yield notable improvements (Paraskevopoulos et al., 2022). In autonomous driving, datasets encompassing various modalities like RGB images, LiDAR, thermal images, and Radars are becoming increasingly important for improving vehicle automation (Feng et al., 2020). Lastly, in the field of cognitive science, multimodal data collections that incorporate modalities such as eye-tracking, electroencephalography, and webcams hold substantial potential to enhance predictions related to human cognition (Giannakos et al., 2019). In all these contexts, the amalgamation of multiple sensory modalities proves to be a powerful strategy for advancing research and applications.

Similarly, by combining spectral and spatial data, we can obtain a more nuanced understanding of Earth’s processes. Therefore, by using a combination of different sensory data to train a land classifier, we can achieve more accurate classification results. The fusion of observations from different modalities is a promising domain in remote sensing, although it has not been thoroughly investigated (Tsagkatakis et al., 2019). The vast amount of monitoring satellites and the increasing number of sensors have resulted in a considerable data collection for a given scene. The integration of remote sensing sensors provides a more detailed depiction of the observed area, leading to more precise monitoring outcomes. For instance, although the spectral profiles of vegetation on the ground and on the roof may not differ significantly, LiDAR data can provide height information to differentiate them, as demonstrated by (Heiden et al., 2012).

A standardized categorization of fusion problems in remote sensing is frequently delineated (Gómez-Chova et al., 2015, Rasti and Ghamisi, 2020, Mohla et al., 2020, Hong et al., 2021). This taxonomy comprises the following four fusion strategies:

1. Sub-pixel-level fusion, which typically processes data on different spatial scales.

This method involves the use of appropriate transforms to reduce the data-dimensionality from each modality prior to fusing them at the sub-pixel level.

2. Data-level fusion, where the data sets from different modalities are directly fused at the pixel level, necessitating the establishment of a direct pixel correspondence between them.
3. Feature-level fusion, entailing feature extraction for all modalities followed by a fusion at the feature level, which may encompass both the extraction and selection of more appropriate attributes.
4. Decision-level fusion, in which distinct processing paths are adopted for each modality, culminating in fusion at the decision level. This method presumes that all outputs can be combined to improve the accuracy of the outcome.

Depending on the type of modalities being combined, fusion can be divided into homogeneous and heterogeneous categories (Li et al., 2022). Homogeneous fusion, such as spatio-spectral and spatiotemporal fusion, aims to address the spatial-spectral and spatial-temporal resolution trade-offs that occur in optical images due to the imaging process. On the other hand, heterogeneous fusion involves combining data from different imaging mechanisms, such as LiDAR-optical or SAR-optical. Because the imaging mechanisms of these data sources are entirely different, feature-level and decision-level fusion methods are typically employed.

Data fusion involves combining multiple datasets to obtain a comprehensive understanding of a phenomenon. A critical challenge in data fusion is establishing relationships between the different datasets (Gómez-Chova et al., 2015). Deep learning is a technique that can address this challenge by capturing abstract features from remote sensing observations and learning potential associations between different datasets through multi-layer learning. This enables deep learning to represent complex relationships in data fusion. Furthermore, deep learning establishes relationships by extracting abstract features from data samples, which are less sensitive to observation properties such as sensor type and spatial scale. As a result, deep learning models can establish robust relationships between datasets (Yuan et al., 2020).

2.2.4 Applications of remote sensing

Remote sensing technology has been accelerated by advances in sensor technology. These sensors are found on satellites (Kuenzer et al., 2014), Unmanned Aerial Vehicles (UAVs)

Table 10: Applications of remote sensing

Application	Description
Land cover / land use mapping	Analyzing satellite images to accurately map land cover and land use patterns, essential for environmental studies.
Environmental parameter retrieval	Retrieval of environmental parameters, such as temperature, moisture content, and air quality.
Data fusion and down-scaling	Merging data from multiple sources to obtain high spatial, temporal, and spectral resolution of remote sensing data.
Information construction and prediction	Filling in missing information in remote sensing data caused by dead lines, gaps, and cloud cover, to construct information and make predictions.
Object detection and tracking	Providing a better understanding of objects in remote sensing images by detecting and tracking them.
Forecasting	Predicting weather patterns, precipitation, etc., using atmospheric measurements and physical laws.
Change detection	Detecting changes in the landscape using two registered remote sensing images taken at different times.
Soil classification	Categorization of soils according to their unique attributes.

([Adao et al., 2017](#)), and ground observation stations ([Chien et al., 2020](#)), creating space-air-ground Earth observation systems. Remote sensing space technology has been developed to meet the demand for large-region and precise environment and resource applications. Aerial remote sensing technology has been recognized as an effective complement to traditional space-based platforms because of its flexibility, high spatial resolution, and data acquisition on demand. Furthermore, ground observation stations have been established due to the development of new technologies in smartphones and wireless networks, which produce high-frequency on-the-spot observations that enrich the remote sensing data sources. Overall, the space-air-ground observation systems provide massive, multi-source, multimodal, multi-scale, high-dimensional, dynamic-state, and heterogeneous remote sensing big data ([Zhang et al., 2022](#)).

Remote sensing is a powerful tool that has numerous applications in various fields. Useful insights into the applications of remote sensing can be gained from several recent surveys ([Yuan et al., 2020](#), [Ball et al., 2017](#)) that have been conducted. Some of the most prominent applications of remote sensing are summarized in [Table 10](#).

Researchers in the field of signal processing, computer vision and pattern recognition have conducted their researches in application to remote sensing. Among these techniques, traditional feature matching methods, such as the scale-invariant feature transform ([Lowe, 1999](#)), have been adapted to enhance remote sensing tasks like image

registration and change detection (Goncalves et al., 2011, Li, Hu and Ai, 2019). Additionally, the utilization of principal components analysis, a well-established multivariate methodology (Jolliffe, 1990), has proven valuable in assessing the reliability of aerosol parameter retrieval from satellite data (Zubko et al., 2007). Furthermore, the integration of partial differential equation models, used in image processing and segmentation (Sofou and Maragos, 2008), has enabled the enhancement of soil structure analysis and segmentation into homogeneous regions by combining contrast and texture information (Sofou et al., 2005). Moreover, efficient possibilistic clustering algorithms (Xenaki et al., 2015) have found effective application in online unsupervised classification of remote sensing Hyperspectral Imagery (HSI) data (Xenaki et al., 2016). Finally, the implementation of sophisticated spectral unmixing methods has yielded promising results in clustering HSI that represent urban and vegetation areas (Mylona et al., 2017), as well as in the physical interpretation of planetary surface data (Themelis et al., 2012). Signal processing, computer vision and pattern recognition techniques, along with the integration of machine learning, collectively shape the evolving landscape of remote sensing, offering a potent blend of approaches to tackle complex challenges and drive innovation in the field.

2.3 Machine learning for remote sensing

In the last few years, the use of machine learning methods has significantly revolutionized the field of remote sensing by enabling efficient processing and analysis of large-scale datasets (Zhang and Han, 2020). Through the automation of tasks and improving predictions of environmental variables, machine learning techniques have eliminated the need for extensive manual efforts and human expertise. This has resulted in time and cost savings for remote sensing practitioners, while also enabling new and more specific applications in diverse fields (Li et al., 2020, Khelifi and Mignotte, 2020, Zhu et al., 2017, Camps-Valls, 2009). However, using machine learning in remote sensing applications poses several challenges. Addressing these challenges is crucial to improve the accuracy and effectiveness of machine learning in remote sensing applications.

2.3.1 Applications of machine learning in remote sensing

The integration of remote sensing and machine learning has unlocked new possibilities for addressing critical environmental issues such as climate change, natural resource management, and disaster response. For example, in forestry, machine learning algo-

rithms can be used to automatically identify tree species (Yu et al., 2017) and monitor the spatiotemporal changes in forest cover (Tariq et al., 2023) or crop yield (Cheng et al., 2022). Additionally, in water resource management, remote sensing data can be utilized to monitor water quality and detect pollution (Wang et al., 2017) while machine learning algorithms can also predict wildfire spread (Huot et al., 2022). Deep learning algorithms have also shown very promising results for sea-ice and permafrost tracking (Khaleghian et al., 2021) as well as oceanic eddy detection in polar regions (Khachatrian et al., 2023).

The integration of remote sensing and machine learning has also facilitated progress in urban planning, where deep learning algorithms can process high-resolution imagery to analyze urban growth and building footprints (Li, Liu, Wang, Li, Jia and Gui, 2019), road network mapping (Senchuri et al., 2021) and estimate livelihood (Ratledge et al., 2022). Furthermore, in the energy sector, machine learning techniques can be used to analyze satellite imagery to estimate the capacity of solar energy farms (Ravishankar et al., 2022), identify potential locations for renewable energy installations (Majidi Nezhad et al., 2021), such as rooftop photovoltaic (Zhong et al., 2021). The combination of remote sensing and machine learning has led to a wide range of applications across various fields and opened up new avenues for research, development, and innovation.

2.3.2 Challenges of machine learning in remote sensing

However, there are several challenges associated with using machine learning in remote sensing applications. One of the primary challenges is the quality of the remote sensing data, which can be influenced by environmental factors, including atmospheric interference, sensor noise, and cloud cover. Such variability can hinder the development of machine learning models that accurately capture the inherent patterns in the data (Yuan et al., 2020). Additionally, the limited availability of high-quality labeled data poses another significant challenge for effectively training machine learning algorithms. Remote sensing data is often complex and difficult to interpret, especially for non-experts, which makes labeling it a time-consuming and expensive process requiring specialized knowledge and expertise. As a result, labeled data may be scarce, impeding the effectiveness of machine learning algorithms (Yuan et al., 2020, Zhang et al., 2022). Furthermore, remote sensing data can be massive in size, posing challenges in storing, processing, and analyzing the data, particularly for high-resolution satellite imagery. Machine learning algorithms can also demand substantial computational resources, which may be limited, especially in on-board or real-time applications (Zhang et al., 2022). Finally, in light of the challenges associated with machine learning in remote sensing, the concept of mul-

timodality becomes even more significant. Remote sensing data may involve multiple sensors, each with its processing requirements, making it challenging to integrate data from various sensors into a single model (Li et al., 2022).

Overall, the challenges of using machine learning in remote sensing applications are significant, but they can be overcome with careful data preparation, algorithm design, and interpretation.

2.3.3 Machine learning for multimodal remote sensing

The development of artificial intelligence has led to the emergence of deep learning as a promising approach for modeling complex relationships of real-world observations and the desired output. Deep learning achieves this by autonomously performing feature extraction and fusion. Depending on the types of observed data being combined and the corresponding objectives, deep learning-based fusion of multiple modalities in remote sensing can be accomplished through a unified framework (Li et al., 2022).

Recent studies have demonstrated the effectiveness of deep learning-based multimodal data fusion in a variety of remote sensing applications. For instance, Sharma et al. (2020) proposed a multimodal version of the YOLO architecture for vehicle detection using RGB and infrared data. In another study, Wu et al. (2021) used CNNs for multimodal data fusion and classification of urban areas. Additionally, several studies have proposed an efficient and effective framework for fusing HSI and LiDAR data using CNNs (Hang et al., 2020, Zhang et al., 2021, Xie et al., 2022). These findings highlight the potential of deep learning-based multimodal data fusion for enhancing the accuracy and reliability of remote sensing applications.

2.4 Objectives of the thesis

Bayesian neural networks have made significant advancements in recent years (Wang and Yeung, 2020). Nevertheless, their application in remote sensing has been limited thus far (Shirmard et al., 2022). These approaches hold promise for facilitating the meaningful analysis of remote sensing data, particularly in addressing challenges like data noise, sparse datasets, and missing information (Shirmard et al., 2022). Driven by the latest surveys in the field (Wang and Yeung, 2020, Zhang et al., 2022, Yuan et al., 2020, Shirmard et al., 2022), the focus of this thesis is to improve the land cover classification by effectively handling the multimodal and limited labeled remote sensing data. To this end, we intend to investigate the potential of Bayesian theory coupled

with deep learning techniques to overcome this challenge.

2.4.1 Literature review and related work

Kingma and Welling (2013) introduced Variational Autoencoder (VAE) as a way to implement variational Bayes algorithm with deep neural networks. VAEs have shown great promise in capturing the underlying patterns and structures of remote sensing images, such as HSI and SAR images. VAEs have been utilized in various applications, such as desertification detection, soil classification, and multispectral image classification. Zerrouki et al. (2021) used VAEs to extract features for desertification detection in multi-temporal satellite images, while Harefa and Zhou (2021) applied VAEs for dimensionality reduction in soil classification with laser-induced breakdown spectroscopy. Valero et al. (2021) proposed a VAE architecture for feature extraction and classification of multispectral Sentinel-2 images. In addition, Ma et al. (2022) integrated a VAE model with a GAN to better align cross-modal features. Shen et al. (2020) used VAE to extract spatial and semantic features of the remote sensing image for achieving the image captioning task.

Classification is a crucial task in remote sensing, and VAEs have also been employed for this purpose. Wang et al. (2020) and Li et al. (2021) proposed conditional VAEs with an adversarial training process for HSI classification.

Meanwhile, semi-supervised techniques have gained interest due to the difficulty of data availability. Kingma et al. (2014) introduced a semi-supervised learning model with stacked VAEs and applied it for classification in MNIST handwritten digit database, using this model Connors and Vatsavai (2017) constructed an auxiliary semi-supervised VAE that takes temporal dependencies into account for the domain of change detection. Cenggoro et al. (2017) used variational semi-supervised learning to solve the imbalance problem in land use/land cover classification using Landsat 7 satellite images with six bands. Moreover, Thoreau et al. (2022) introduced a semi-supervised model that combines conventional neural network layers with physics-based layers for the semantic segmentation of remote sensing HSI. The study by Arun et al. (2022) utilized a VAE to map the spatial and spectral data of UAVs and HSI to a common latent space. This allowed for the creation of a latent graph generator-based classifier that could use both labeled and unlabeled samples for prediction purposes.

2.4.2 Methodology and contributions

The aim of our study is to extend the semi-supervised model proposed by Kingma et al. (2014) to achieve pixel-wise classification of multimodal remote sensing data. Unlike Kingma et al. (2014), our model takes into account LiDAR and HSI data to provide a classification for each point in the image. We also investigate the fusion of the two modalities on either the pixel level or feature level. This approach is more suitable for remote sensing imagery and can improve classification accuracy in comparison to single modality-based methods. In brief, the thesis makes the following contributions:

- Adapting a Bayesian deep learning model to cater to the requirements of remote sensing multimodal data classification.
- Addressing the challenge of limited labeled data in remote sensing by leveraging semi-supervised learning.
- Investigating the potential of fusion techniques at both latent feature and data levels.

3 Theoretical background

This chapter serves as an introduction to the fundamental concepts used in this thesis. We begin by providing an overview of the basic generative modeling concepts in section 3.1. We then delve into the specifics of VAEs in section 3.2. Finally, we discuss how VAEs are utilized for semi-supervised classification in section 3.3. Through this comprehensive explanation of key concepts, readers will gain a deeper understanding of the methods and approaches used in the thesis.

3.1 Generative and discriminative models

Discriminative and generative models are two fundamental types of machine learning models that differ in their approach to modeling the underlying probability distribution of the observed variables x based on the corresponding labels y . The goal of a discriminative model is to find the decision boundary that separates the different classes of data in the input space. For example, in the binary case the decision boundary can be represented by a function $f(x)$ such that if $f(x) > 0$, the input x is classified as belonging to one class, and if $f(x) < 0$, it is classified as belonging to the other class (Jebara, 2012). From a probabilistic point of view, the discriminative model learns the posterior probability distribution of the output variable given the input variable, denoted as $p(y|x)$ (Ng and Jordan, 2001). Contemporary machine learning classifiers primarily consist of various models, such as logistic regression, support vector machine (SVM), supervised feed-forward deep neural networks, nearest neighbors, conditional random fields, and others (Harshvardhan et al., 2020). The common objective of these models is to perform discriminative classification.

In contrast, a generative model learns the joint probability distribution of both the inputs x and the labels y , denoted as $p(x, y)$ (Jebara, 2012). Once the model has learned this distribution, it can be used to generate new samples of data that are similar to the training data. Generative models are currently employed as effective feature extraction tools in addition to their conventional use in pattern recognition followed by generation, regression, clustering, classification, recommendations, topic modeling, text generation, and other applications (Harshvardhan et al., 2020). To summarize:

- **Discriminative:** Learning the boundary between different classes of data.
- **Generative:** Modeling the underlying probability distribution of the data.

Discriminative models have some advantages over generative models. For instance, discriminative models are generally faster at predicting new data points, while generative models often require iterative solutions. Additionally, discriminative models usually have better predictive performance since they are trained to predict the class label instead of the joint distribution of input vectors and targets (Ulusoy and Bishop, 2005). Furthermore, training generative models, especially deep generative models, is a lengthier process since it involves learning a higher number of correlations to create a probability distribution that resembles the original data (Harshvardhan et al., 2020) as well as making strong assumptions about the data structure (Dimakis, 2022). This is in contrast to discriminative models that simply label instances to their most probable classes.

On the other hand, generative models offer several advantages compared to discriminative models. First, they can effectively handle missing or partially labeled data, and can leverage a large quantity of unlabeled data to supplement small amounts of expensive labeled data. Additionally, generative models can handle compositional data, without needing to see all possible combinations during training. This is not the case for standard discriminative models (Ulusoy and Bishop, 2005). Generative models also allow for predicting missing data parts because they provide an understanding of the data manifold through manifold learning. Furthermore, generative models are versatile, allowing for multimodal outputs and a single input to perform various tasks without separate training. Finally, generative models have been shown to enhance data quality in studies where they generate super-resolution images, among other applications (Turhan and Bilge, 2018).

3.1.1 Generative models

One of the earliest and highly influential generative models is the Hidden Markov Model (HMM), which was introduced in the 1960s (Rabiner, 1989). This model has found extensive application in signal processing, particularly in speech recognition. The HMM is a generative model that models the joint distribution of hidden states and observations as a Markov process. It is mainly applied in speech recognition (Rabiner, 1989), but it has also been used in other domains such as financial and statistical applications (Bhar and Hamori, 2004), biological sequence modeling (Krogh et al., 1994), and more.

Another important generative model is Gaussian Mixture Model (GMM). GMM is a type of parametric probability density function that uses a combination of Gaussian component densities, where the model's parameters are derived through the Expectation-Maximization algorithm or Maximum A Posteriori estimation from a prior model (Reynolds

et al., 2009). In biometric systems, GMMs are typically used as a way to represent the probability distribution of continuous measurements or features. GMMs have found use in a wide range of applications, including image and video clustering (Yang and Ahuja, 1998), speech recognition (Burget et al., 2010), language identification (Torres-Carrasquillo et al., 2002), anomaly detection (Zong et al., 2018), and others.

3.1.2 Deep generative models

Recently, there have been remarkable advances in the field of generative models. One of the most significant developments is the use of deep neural networks to learn generative models (Rezende et al., 2014). According to (Ruthotto and Haber, 2021) deep generative models refer to neural networks with numerous hidden layers and are trained to approximate complex, high-dimensional probability distributions. Essentially, the objective in training these models is to learn a probability distribution that is either unknown or difficult to calculate due to its complexity, using only a limited number of samples.

Deep neural networks have emerged as a powerful tool for learning generative models, which can generate new data samples that are similar to those in the original dataset. Two popular types of generative models are VAEs (Kingma and Welling, 2013) and GANs (Goodfellow et al., 2020). GANs have been particularly successful in generating high-quality images (Karras et al., 2020) and have also shown promise in audio and text generation (Zhang et al., 2017). Additionally, GANs can be used for data editing (Zhu et al., 2020) and data augmentation (Frid-Adar et al., 2018, Waheed et al., 2020). On the other hand, VAEs are capable of learning complex patterns in data and can extract informative features from it (Nishizaki, 2017, Kuznetsov et al., 2020). Moreover, they have been successfully applied to identify anomalies in data and detect potential frauds (Pol et al., 2019, An and Cho, 2015, Park et al., 2018). This is because VAEs learn to represent the data in a lower dimensional latent space and can identify samples that deviate significantly from the learned distribution.

3.2 Fundamentals of VAEs

VAEs are among the most widely used deep generative frameworks. This model does not rely on strong assumptions and can be trained quickly via back-propagation (Kingma and Welling, 2013). Although VAEs introduce some approximation errors, the errors are typically small, especially when using high-capacity models. These characteristics have contributed to the rapid popularity and growth of VAEs in the research community

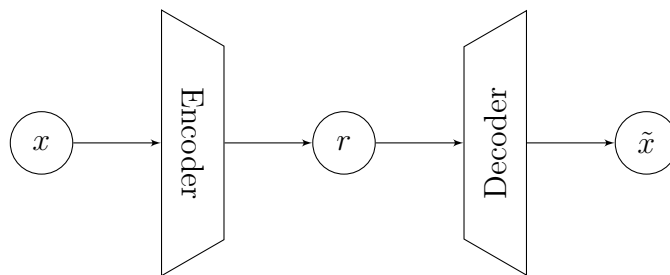


Figure 14: Graphical representation of an autoencoder

(Ruthotto and Haber, 2021). To understand in-depth the idea of VAEs 3.2.6, we need to introduce some basic concepts like Autoencoders 3.2.1, Probabilistic Models 3.2.2, Latent Variables 3.2.3 and Variational Inference 3.2.5.

3.2.1 Autoencoders

Autoencoders are a type of unsupervised artificial neural network architecture . They are used to learn low-dimensional representations of unlabeled data, and consist of two primary components: an encoder and a decoder. Observed variables x used as input. The encoder maps the input data x to a compressed representation r , while the decoder maps the compressed representation back to the original input data \tilde{x} (e.g., Baldi, 2012) (see Figure 14). The quality of the encoding is evaluated by how well the input can be regenerated by the decoder. It’s worth highlighting that an autoencoder is not considered a generative model, because it only reconstructs the given input.

Autoencoders have gained widespread attention in recent years due to their ability to perform a variety of tasks, such as dimensionality reduction, feature extraction, and denoising. Dimensionality reduction is a key application of autoencoders, where the aim is to map high-dimensional input data to a lower-dimensional space that preserves as much information as possible. This is particularly useful when dealing with high-dimensional data such as images or text, where reducing the dimensionality can lead to more efficient storage and computation (Ryu et al., 2020). Autoencoders are also used for denoising, where the aim is to remove noise from the input data. This is achieved by training the autoencoder to map noisy input data to the corresponding clean input data (Vincent et al., 2008, Alain and Bengio, 2014). Feature extraction is another important application of autoencoders. In this context, the encoder is trained to learn a compressed representation of the input data that captures the most important features or patterns in the data (Zabalza et al., 2016). These learned features can then be used for downstream tasks such as classification, clustering, or regression.

3.2.2 Probabilistic models

Probabilistic models can help understand natural and artificial phenomena, predict future events, and make automated decisions. These models are mathematical descriptions of the phenomena and incorporate uncertainty in the form of probability distributions (Murphy, 2022). Probabilistic models can include both continuous and discrete variables, and the most complete models specify the relationships between variables in the form of a joint probability distribution (Murphy, 2022).

So, probabilistic models aim to model a probability distribution for a set of observed variables x . The true distribution of the data, $p^*(x)$, is unknown, so a model $p_\theta(x)$ with parameters θ is chosen to approximate it. The goal of learning is to find the value of θ that makes $p_\theta(x)$ as close as possible to $p^*(x)$ for any observed x . To achieve this, $p_\theta(x)$ must be flexible enough to adapt to the data, but also take into account prior knowledge about the distribution of the data (Kingma et al., 2019).

Frequently, for tasks like classification or regression, the goal is a conditional model $p_\theta(y|x)$ that estimates the underlying conditional distribution $p^*(y|x)$, which represents the distribution of a variable y given an observed variable x . To achieve this, a model $p_\theta(y|x)$ with parameters θ is selected to approximate the unknown distribution $p^*(y|x)$ (Kingma et al., 2019).

Neural networks can be used as a type of function approximator that are flexible and computationally scalable. Neural networks are particularly useful for probabilistic models, such as modeling Probability Density Functions (PDFs). This is because they allow for stochastic gradient-based optimization, which makes it possible to scale to large models and large datasets. Deep learning has been shown to be effective for many classification and regression problems. For instance, in image classification, deep neural networks are used to parameterize a categorical distribution:

$$p_\theta(y|x) = \text{Categorical}(y; \text{NeuralNet}(x))$$

over a class label y , conditioned on an image x , $\text{NeuralNet}(\cdot)$ represents a deep neural network. (Kingma et al., 2019).

3.2.3 Latent variables

Latent variables are variables that cannot be directly observed but carry information about the observable world. For instance, a person's intelligence cannot be measured

directly, but it can be inferred through IQ tests, making intelligence a latent variable with respect to IQ scores (Borsboom, 2008). However, latent variables do not need to correspond to real-world phenomena. In machine learning, latent variables are used to elegantly model a variety of applications. For example, in an image recognition model, latent variables can represent textural and shape information that is not always easily discernible by humans. The use of latent variables allows for more efficient and accurate models in many fields, including psychology and health research (Cai, 2012)

To formalize this, let us assume that x is an observed variable and z the corresponding latent variable that can be sampled from a PDF $p_\theta(z)$. The distribution $p_\theta(z)$ is often called the prior distribution over z . In case of unconditional modeling marginal distribution (or marginal likelihood or the model evidence) $p_\theta(x)$ over x is given by:

$$p_\theta(x) = \int p_\theta(x, z) dz \quad (2)$$

where $p_\theta(x, z)$ denotes the joint distribution over both x and z .

3.2.4 Deep Latent Variable Models

A probabilistic model that aims to model the joint distribution $p_\theta(x, z)$ with parameters θ is a latent variable model. Deep Latent Variable Models (DLVM) refers to a type of latent variable models $p_\theta(x, z)$ that uses neural networks to parameterize distributions with trainable parameters θ . A major advantage of DLVMs is that even if the individual factors in the model are simple, like conditional Gaussian, the marginal distribution $p_\theta(x)$ can still be very complex and contain almost any kind of dependency (Kingma et al., 2019). Therefore, DLVMs are useful for approximating complex underlying distributions. The most common DLVM is one that implies conditionally dependent variables x and z . So, the joint distribution $p_\theta(x, z)$ is specified as factorization with the following structure:

$$p_\theta(x, z) = p_\theta(x|z)p_\theta(z) \quad (3)$$

The primary challenge in DLVMs is that the integral of the marginal probability $p_\theta(x)$ as described in (2) is intractable (Blei et al., 2017). As a result, it's not possible to differentiate it with respect to the parameters and optimize it, as is possible with fully observed models. The intractability of the marginal probability is related to the intractability of the posterior distribution $p_\theta(z|x)$ through Bayes rule and factorization

in (3):

$$p_{\theta}(z|x) = \frac{p_{\theta}(x, z)}{p_{\theta}(x)} = \frac{p_{\theta}(x|z)p_{\theta}(z)}{p_{\theta}(x)} \quad (4)$$

While the joint distribution $p_{\theta}(x, z)$ is efficient to compute, both the marginal likelihood $p_{\theta}(x)$ and the posterior $p_{\theta}(z|x)$ are intractable in DLVMs. However, variational inference can be used to estimate them.

3.2.5 Variational inference

Variational inference is a method for approximating the conditional density of latent variables given observed variables or posterior distribution (Bishop and Nasrabadi, 2006). The objective of variational inference is to estimate an approximate posterior distribution, $q_{\varphi}(z|x)$, that is computationally tractable. Among the most common approaches for evaluating the similarity between the posterior and approximate posterior distributions (Bhattacharyya, 1946) is the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951), which has been widely used in machine learning. The KL divergence between the posterior and approximate posterior distributions can be written as:

$$D_{KL}(q_{\varphi}(z|x)||p_{\theta}(z|x)) = \mathbb{E}_{q_{\varphi}(z|x)}[\log q_{\varphi}(z|x) - \log p_{\theta}(z|x)] \quad (5)$$

$$= \mathbb{E}_{q_{\varphi}(z|x)}[\log q_{\varphi}(z|x) - \log p_{\theta}(x, z)] + \log p_{\theta}(x) \quad (6)$$

$$= \mathcal{L}(q_{\varphi}(z|x)) + \log p_{\theta}(x) \quad (7)$$

Since the KL divergence is non-negative, we can deduce that $\log p_{\theta}(x) \geq \mathcal{L}(q_{\varphi}(z|x))$. This inequality motivates the introduction of the Evidence Lower Bound (ELBO):

$$\mathcal{L}(q_{\varphi}(z|x)) = \mathbb{E}_{q_{\varphi}(z|x)}[\log \frac{p_{\theta}(x, z)}{q_{\varphi}(z|x)}] \quad (8)$$

To obtain a good approximation of the posterior distribution, we need to minimize the KL divergence. However, the KL divergence (7) still involves the intractable term $p_{\theta}(x)$, so it cannot be optimized directly. Instead, we can maximize the ELBO (8), which is equivalent to minimizing the KL divergence (Blei et al., 2017, Bishop and Nasrabadi,

2006). Thus, we aim to maximize the ELBO as follows:

$$q^*(z|x) = \arg \min_{q_\varphi(z|x) \in Q} (D_{KL}(q_\varphi(z|x) || p_\theta(z|x))) \quad (9)$$

$$= \arg \max_{q_\varphi(z|x) \in Q} (\mathcal{L}(q_\varphi(z|x))) \quad (10)$$

Here, Q is a family of simple distributions, such as Gaussian and Bernoulli distributions.

3.2.6 Variational autoencoder

VAEs were introduced by Kingma and Welling (2013) as a method for performing efficient variational inference using artificial neural networks. VAEs are named autoencoders because they share the same encoding-decoding architecture as autoencoders (see Figures 14, 15). However, unlike traditional autoencoders, VAEs compress high-dimensional input data x to a latent vector z rather than a deterministic vector r (Section 3.2.1). The variational parameters φ are optimized to approximate the posterior distribution $p_\theta(z|x)$, which in turn helps optimize the marginal likelihood $p_\theta(x)$:

$$q_\varphi(z|x) \approx p_\theta(z|x)$$

Both variational model $q_\varphi(z|x)$ and the conditional distribution $p_\theta(x|z)$ are parameterized using deep neural networks with trainable parameters φ and θ respectively. The optimization objective of the VAE is maximizing ELBO, expressed in (8), which aims to better approximate the true distribution.

To optimize the ELBO using standard gradient-based techniques, Kingma and Welling (2013) introduced the Auto-Encoding Variational Bayes algorithm to efficiently compute the gradient of the ELBO. They assume that the distributions $p_\theta(z)$ and $p_\theta(x|z)$ are differentiable almost everywhere with respect to both θ and z . For a chosen approximate posterior $q_\varphi(z|x)$, they use the *reparameterization trick*, where the random variable $z = q_\varphi(z|x)$ is re-parameterized with a differentiable transformation $g_\varphi(\varepsilon, x)$ of a noise variable ε , such that $z = g_\varphi(\varepsilon, x)$. This trick allows gradients to be calculated with respect to mini-batches of data (see Figure 15).

In order to simplify the calculations, Kingma and Welling (2013) assume the variational approximate posterior $q_\varphi(z|x)$ to be a multivariate Gaussian with diagonal covariance matrix. As for the prior $p_\theta(z)$ they assume for simplicity a multivariate Gaussian

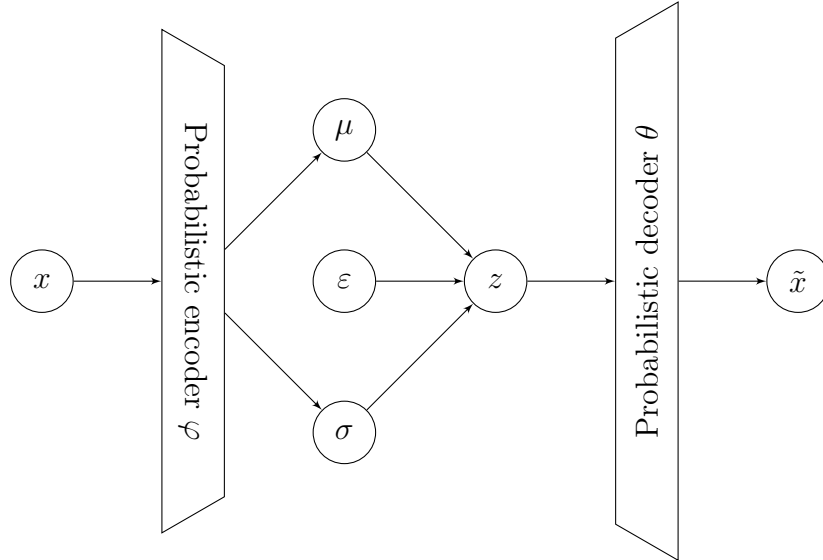


Figure 15: Graphical representation of a variational autoencoder

with diagonal covariance matrix $\mathcal{N}(z; 0, I)$ although, a full covariance matrix is possible:

$$q_{\varphi}(z|x) = \mathcal{N}(z; \mu, \sigma^2 I), \quad (11)$$

$$p_{\theta}(z) = \mathcal{N}(z; 0, I), \quad (12)$$

$$z = \mu + \sigma * \varepsilon, \quad (13)$$

$$\varepsilon \sim \mathcal{N}(0, I) \quad (14)$$

3.3 Semi-supervised classification with VAEs

In recent years, semi-supervised learning has drawn significant attention due to its potential to address classification tasks that involve large amounts of data but require difficult, expensive, or expert labeling procedures. By leveraging both labeled and unlabeled data, semi-supervised learning can perform supervised or unsupervised learning tasks, with a particular focus on enhancing supervised learning tasks with unlabeled data (Zhu, 2005).

Semi-supervised learning has a wide range of practical applications, including tumor classification on methylation data (Tran et al., 2022), automatic speech recognition (Zhang et al., 2020), and semantic image segmentation (Papandreou et al., 2015). In

these cases, unlabeled data are abundant and easy to collect, but labeling the entire dataset is very expensive and time-consuming. Semi-supervised learning can help overcome these challenges by leveraging the unlabeled data to improve the accuracy of the classification tasks. For example, in tumor classification on methylation data, semi-supervised learning can improve the accuracy of tumor diagnosis by leveraging unlabeled data to identify important features that distinguish between tumor and normal samples.

VAEs and generative models have recently gained popularity for tackling semi-supervised learning tasks. To this end, Kingma et al. (2014) proposed three frameworks that utilize the latent representation of the data to enhance classification performance using VAEs.

3.3.1 Latent-feature discriminative model

The first framework (M1), introduced by Kingma et al. (2014), referred to as the *latent-feature discriminative* model (Figure 16), involves training a VAE on all the observed data x to perform unsupervised extraction of latent variables z . Then using the labels y a separate classifier is trained on the embedding of the labeled data (z, y) . The VAE is trained using the assumptions (11-14) along with:

$$p_{\theta}(x|z) = f(x; z, \theta) \quad (15)$$

where f is a suitable function, such as a Bernoulli or Gaussian.

The objective function of the model of this model, $\mathcal{J}_{M1}(x)$ is the negative of ELBO of the marginal distribution $p_{\theta}(x)$ on the approximate posterior $q_{\varphi}(z|x)$. For a single data point, it is expressed as:

$$\mathcal{L}(q_{\varphi}(z|x)) = -\mathcal{J}_{M1}(x) \quad (16)$$

By minimizing $\mathcal{J}_{M1}(x)$, i.e, maximizing ELBO, the VAE learns a compact representation of the data that captures the most relevant features for classification. The embedding of the labeled data (z, y) is then used to train a separate classifier, such as SVM, to predict the class labels of the unlabeled data. This framework has shown promising results on various classification tasks, especially in scenarios where labeled data is scarce.

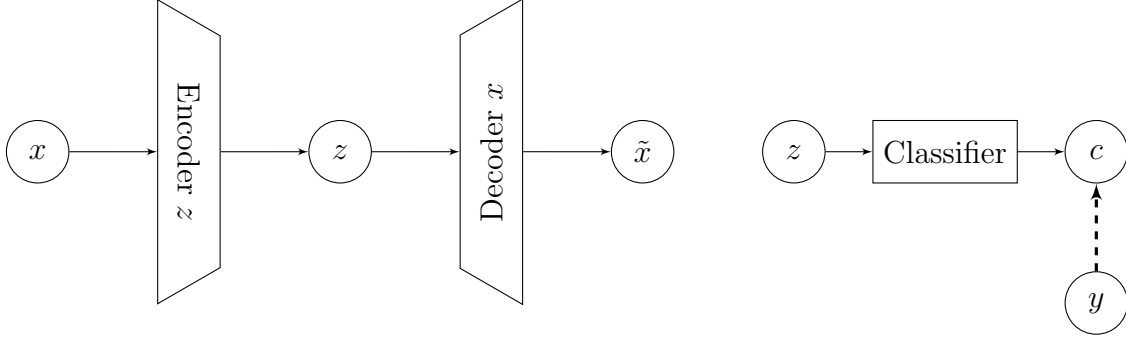


Figure 16: Graphical representation of *latent-feature discriminative* model - M1

3.3.2 Generative semi-supervised model

In the second model's architecture (M2), introduced by Kingma et al. (2014), the label information is incorporated during feature extraction to improve performance on the classification task. M2 is a *generative semi-supervised* model (Figure 17) that has two latent variables, z and c . The labels y are treated as input on labeled data and as a latent variable $c = y$ on unlabeled data. The joint distribution is factorized with the following structure:

$$p_{\theta}(x, z, c) = p_{\theta}(x|z, c)p(c)p(z) \quad (17)$$

The generative process for the data is defined by:

$$p(c) = \text{Cat}(c|\pi) \quad (18)$$

$$p(z) = \mathcal{N}(z|0, I) \quad (19)$$

$$p_{\theta}(x|z, c) = f(x; z, c, \theta) \quad (20)$$

Here, $\text{Cat}(c|\pi)$ is a multinomial categorical distribution with probabilities π , and as before, f is a simple and suitable function, such as a Bernoulli or Gaussian. The variational model for each of the latent variables z and c has a factorized form:

$$q_{\varphi}(z, c|x) = q_{\varphi}(z|x)q_{\varphi}(c|x) \quad (21)$$

The factor distributions are specified as Gaussian and Categorical:

$$q_\varphi(z|x, c) = \mathcal{N}(z|\mu_\varphi(x, c), \sigma_\varphi^2(x)I) \quad (22)$$

$$q_\varphi(c|x) = \text{Cat}(c|\pi_\varphi(x)) \quad (23)$$

To calculate the objective function of the model, two cases are considered: labeled data and unlabeled data.

- For labeled data, input variables are x and $c = y$, and the latent variable is z . The objective function $\mathcal{J}_{M2}(x)$ of the model is the negative of ELBO of the marginal distribution $p_\theta(x, c)$ on the approximate posterior $q_\varphi(z|x, c)$. Using the factorized form from (17), the objective function for a single data point is expressed as:

$$\log(p_\theta(x, c)) \geq E_{q_\varphi(z|x, c)}[\log(\frac{p_\theta(x|z, c)p(c)p(z)}{q_\varphi(z|x, c)})] \quad (24)$$

$$= -\mathcal{L}_{M2}(x, c) \quad (25)$$

- For unlabeled data, only variable x is treated as input, and z and c are treated as unknown latent variables. By considering again the marginal distribution of the input(s), the objective function $\mathcal{U}_{M2}(x)$ is the negative ELBO of the marginal distribution $p_\theta(x)$ on the approximate posterior $q_\varphi(z, c|x)$. Using the factorized forms from (17) and (21) the objective function for a single data point is expressed as:

$$\log(p_\theta(x)) \geq \mathbb{E}_{q_\varphi(z, c|x)}[\log(\frac{p_\theta(x|z, c)p(c)p(z)}{q_\varphi(z|x)q_\varphi(c|x)})] \quad (26)$$

$$= -\mathcal{U}_{M2}(x) \quad (27)$$

Hence, the objective function of the *generative semi-supervised* model is computed by the bound on the marginal likelihood, encompassing the complete dataset. This is represented as follows:

$$\mathcal{J}_{M2} = \sum_{x \in \text{labeled}} \mathcal{L}_{M2}(x, c) + \sum_{x \in \text{unlabeled}} \mathcal{U}_{M2}(x) \quad (28)$$

By minimizing $\mathcal{J}_{M2}(x)$ thus maximizing ELBO $\mathcal{L}_{M2}(x, c)$ and $\mathcal{U}_{M2}(x)$ the VAE

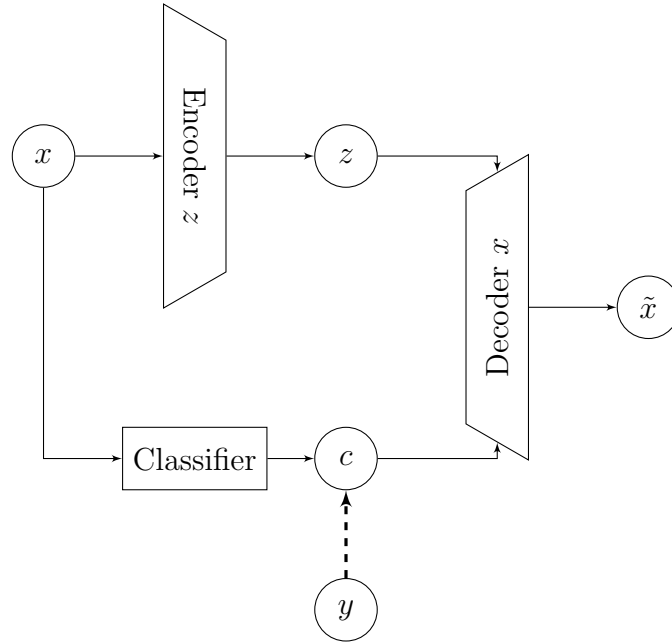


Figure 17: Graphical representation of *generative semi-supervised* model - M2

learns a latent representation of the data that captures the most relevant features for classification while also solves the classification task.

3.3.3 Stacked generative semi-supervised model

Lastly, the M1+M2 model, introduced by [Kingma et al. \(2014\)](#), is a *stacked generative semi-supervised* model that combines the advantages of semi-supervised generative model M2 and the latent representation of data in model M1 (see [Figure 18](#)). The input variables x are expressed from two levels of latent variables z, c, u whose joint distribution $p_{\theta}(x, z, c, u)$ is factorized into:

$$p_{\theta}(x, z, c, u) = p_{\theta}(x|z)p_{\theta}(z|c, u)p(c)p(u) \quad (29)$$

The generative process for the data is defined as follows:

$$p(c) = \text{Cat}(c|\pi) \quad (30)$$

$$p(u) = \mathcal{N}(u|0, I) \quad (31)$$

$$p_\theta(z|c, u) = f(z; c, u, \theta) \quad (32)$$

$$p_\theta(x|z) = f(x; z, \theta) \quad (33)$$

The variational model for each of the latent variables (z , c , and u) has a factorized form:

$$q_\varphi(z, c, u|x) = q_\varphi(z|x)q_\varphi(c|z)q_\varphi(u|z, c) \quad (34)$$

following the corresponding assumptions with the M2. The factor distributions are specified as Gaussian and Categorical:

$$q_\varphi(u|z, c) = \mathcal{N}(u|\mu_\varphi(z, c), \sigma_\varphi^2(z)I) \quad (35)$$

$$q_\varphi(c|z) = \text{Cat}(c|\pi_\varphi(z)) \quad (36)$$

$$q_\varphi(z|x) = \mathcal{N}(z|\mu_\varphi(x), \sigma_\varphi^2(x)I) \quad (37)$$

To calculate the objective function of the semi-supervised model (M1+M2), we need to consider both labeled and unlabeled data, just like in model M2.

- For labeled data, observed variables x and the labels $y = c$ are considered as input, and latent variables are z and u . The variational model is $q_\varphi(z, u|x, c) = q_\varphi(z|x)q_\varphi(u|z, c)$. The objective function $\mathcal{L}_{M1+M2}(x, c)$ is the negative of the ELBO of the marginal distribution $p_\theta(x, c)$ on the approximate posterior $q_\varphi(z, u|x, c)$. Using the factorized form from (29), the objective function for a single data point is expressed as:

$$\log p_\theta(x, c) \geq E_{q_\varphi(z, u|x, c)} \left[\log \left(\frac{p_\theta(x|z)p_\theta(z|c, u)p_\theta(c)p_\theta(u)}{q_\varphi(z|x)q_\varphi(u|z, c)} \right) \right] \quad (38)$$

$$= -\mathcal{L}_{M1+M2}(x, c) \quad (39)$$

- For unlabeled data, only variable x is treated as input, and latent variables are z ,

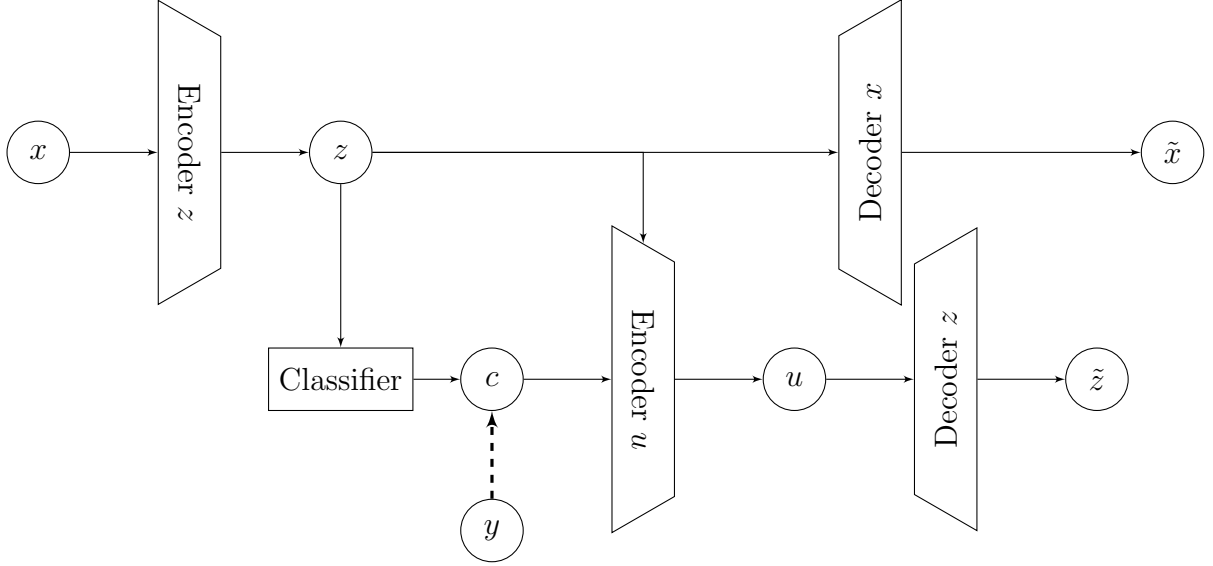


Figure 18: Graphical representation of *stacked generative semi-supervised* model - M1+M2

c , and u . The objective function $\mathcal{U}_{M1+M2}(x)$ is the negative ELBO of the marginal distribution $p_\theta(x)$ on the approximate posterior $q_\phi(z, c, u|x)$. Using the factorized forms from (29) and (34), for a single data point, it is given by:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z, c, u|x)} \left[\log \left(\frac{p_\theta(x|z)p_\theta(z|c, u)p_\theta(c)p_\theta(u)}{q_\phi(z|x)q_\phi(c|z)q_\phi(u|z, c)} \right) \right] \quad (40)$$

$$= -\mathcal{U}_{M1+M2}(x). \quad (41)$$

Hence, the objective function of the semi-supervised model for the *stacked generative semi-supervised* model is calculated as the marginal likelihood for the entire dataset, given by the following equation:

$$\mathcal{J}_{M1+M2} = \sum_{x \in \text{labeled}} \mathcal{L}_{M1+M2}(x, c) + \sum_{x \in \text{unlabeled}} \mathcal{U}_{M1+M2}(x) \quad (42)$$

By minimizing $\mathcal{J}_{M1+M2}(x)$ thus maximizing ELBO $\mathcal{L}_{M1+M2}(x, c)$ and $\mathcal{U}_{M1+M2}(x)$, the VAE learns two levels of latent representation of the data that captures the most relevant features for classification while also solves the classification task.

As discussed in the section (2.4.1), these models' architectures have served as the foundation for numerous classification scenarios. However, a modification for multimodal data has yet to be introduced. This is precisely what we aim to accomplish in the next

chapter.

4 Method

As previously addressed in section 2.4.1, the semi-supervised architectures explained in section 3.3 have laid the groundwork for various classification scenarios. However, an adaptation tailored specifically to multimodal data has not yet been presented. Our primary objective in this chapter is precisely to fill this gap by introducing a framework capable of accommodating multimodal datasets. Multimodality refers to data encompassing two or more distinct modalities. Consequently, our initial stride involves the introduction of models suitable for handling two modalities, with the subsequent potential for facile expansion to accommodate additional modalities.

4.1 Multimodal data classification with semi-supervised VAE

Problem definition: We are presented with a classification problem that involves data in multiple modalities, denoted as x_1, x_2, \dots, x_n . Our objective is to create a model that by combining the advantages of deep generative models and effectively managing multiple modalities, provides an innovative solution to the classification problem using limited labeled data.

We have decided to delve into the realm of advanced architectures by adopting M1+M2 for semi-supervised classification with VAE. In our pursuit of classifying multimodal data, we have extended this architecture in two distinct manners. Firstly, through data-level fusion, and secondly, by introducing an additional layer of latent variables for each modality to enable latent feature-level fusion.

4.1.1 Data-level fusion

M1+M2's architecture is designed to integrate multiple modalities into a single input for the model. These modalities, represented as x_1, x_2, \dots, x_n , could be different types of data such as images, audio, or text. By concatenating them, we create a unified input x that incorporates all the modalities. In the same manner as in the original approach described in section 3.3.3, the concatenated input x is used to train a generative model and a classifier simultaneously.

However, in some cases, the modalities may have variations in shape or the amount of information they provide. For instance, let's say we have two modalities: images and textual descriptions. The images may have different dimensions or resolutions, while

the textual descriptions may vary in length or level of detail. When we concatenate these modalities, we implicitly assume that the importance of each modality is directly proportional to its shape or size. This assumption may not always be true.

In reality, the importance of each modality might not solely depend on its shape or size. Other factors, such as the inherent relevance or quality of the information provided, could also play a significant role. To address this, we aim to find a way to input each modality separately and ensure that they have a balanced influence on the overall model’s output. By exploring alternative fusion methods, we can potentially achieve a more nuanced integration of the modalities.

4.1.2 Latent feature-level fusion - multimodal stacked generative semi-supervised model

We propose the Multi-M1+M2, a *multimodal stacked generative semi-supervised* model that treats each modality of the observed data x_1, x_2, \dots, x_K as a separate input (Figure 19). Each input variable is encoded into a corresponding latent variable z_1, z_2, \dots, z_K . These latent variables are then concatenated and fed into the rest of the network, following a similar methodology as described in section 3.3.3.

Consider our dataset $\{x_1^{(j)}, x_2^{(j)}, \dots, x_K^{(j)}, c^{(j)}\}_{j=1}^N$ consisting of N i.i.d. samples of K different modalities x_1, x_2, \dots, x_K and their corresponding class labels $c^{(j)} \in \{0, 1, 2, \dots, C\}$. The data generation process involves unobserved latent variables z_1, z_2, \dots, z_K and u . are obtained from prior distributions $p_{\theta_1}^*(z_1), p_{\theta_2}^*(z_2), \dots, p_{\theta_K}^*(z_K)$. Then, the values $x_1^{(j)}, x_2^{(j)}, \dots, x_K^{(j)}$ are generated from conditional distributions $p_{\theta_1^*}(x_1|z_1), p_{\theta_2^*}(x_2|z_2), \dots, p_{\theta_K^*}(x_K|z_K)$ respectively. The prior and conditional distributions are parameterized by families of distributions: $p_{\theta_i}(z_i), p_{\theta_i}(x_i|z_i)$ for $i = 1, 2, \dots, K$. This results in a deep generative model with one stochastic variable for each modality, whose joint distribution is factorized with the following structure:

$$p_{\theta}(x_1, x_2, \dots, x_K, z_1, z_2, \dots, z_K, c, u) = p_{\theta}(z_1, z_2, \dots, z_K|c, u)p(c)p(u) \prod_{i=1}^K p_{\theta_i}(x_i|z_i) \quad (43)$$

The generative process for the data is defined by:

$$p(c) = \text{Cat}(c|\pi) \quad (44)$$

$$p(u) = \mathcal{N}(u|0, I) \quad (45)$$

$$p_\theta(z_1, z_2, \dots, z_K|c, u) = \mathcal{N}(z_1, z_2, \dots, z_K|\mu_\theta(c, u), \sigma_\theta^2(c, u)) \quad (46)$$

$$p_{\theta_i}(x_i|z_i) = \text{Cat}(x_i|\pi_{\theta_i}(z_i)), i = 1, 2, \dots, K \quad (47)$$

To address the task of multi-class classification, we utilize the prior distribution $p(c) = \text{Cat}(c|\pi)$, which represents a multinomial categorical distribution with probabilities $\pi = 1/C$ assigned to each of the C classes. The conditional distributions $p(x_i|z_i) = \text{Cat}(x_i|\pi_{\theta_i}(z_i))$ in our model are Bernoulli distributions. The probabilities for these distributions are denoted as $\pi_{\theta_i}(z_i)$. Following the same approach as with M2 and M1+M2, we make the assumption that the variational model can be factorized as follows:

$$q_\varphi(z_1, z_2, \dots, z_K, c, u|x_1, x_2, \dots, x_K) = q_\varphi(c|z_1, z_2, \dots, z_K)q_\varphi(u|z_1, z_2, \dots, z_K, c) \prod_{i=1}^K q_{\varphi_i}(z_i|x_i) \quad (48)$$

In this factorization, we specify the distributions as Gaussian and Categorical:

$$q_\varphi(u|c, z_1, z_2, \dots, z_K) = \mathcal{N}(u|\mu_\varphi(c, z_1, z_2, \dots, z_K), \sigma_\varphi^2(z_1, z_2, \dots, z_K)I) \quad (49)$$

$$q_\varphi(c|z_1, z_2, \dots, z_K) = \text{Cat}(c|\pi_\varphi(z_1, z_2, \dots, z_K)) \quad (50)$$

$$q_{\varphi_i}(z_i|x_i) = \mathcal{N}(z_i|\mu_{\varphi_i}(x_i), \sigma_{\varphi_i}^2(x_i)I), i = 1, 2, \dots, K \quad (51)$$

To calculate the objective function of the model, two cases are considered: labeled data and unlabeled data.

- For labeled data, the input variables are $x_1, x_2, \dots, x_K, c = y$, and the latent variables are z_1, z_2, \dots, z_K, u . The variational model can be expressed as:

$$q_\varphi(z_1, z_2, \dots, z_K, u|c, x_1, x_2, \dots, x_K) = q_\varphi(u|z_1, z_2, \dots, z_K, c) \prod_{i=1}^K q_{\varphi_i}(x_i|z_i) \quad (52)$$

The objective function $\mathcal{L}(x)$ of the model is the negative of ELBO of the marginal distribution $p_\theta(x_1, x_2, \dots, x_K, c)$ on the approximate posterior $q_\varphi(z_1, z_2, \dots, z_K, u | x_1, x_2, \dots, x_K, c)$. Using the factorized form from (43), the objective function for a single data point is expressed as:

$$\begin{aligned} \log p_\theta(x_1, x_2, \dots, x_K, c) &\geq \\ \mathbb{E}_{q_\varphi(z_1, z_2, \dots, z_K, u | x_1, x_2, \dots, x_K, c)} &\left[\log \frac{p_\theta(z_1, z_2, \dots, z_K, |c, u) p_\theta(c) p_\theta(u) \prod_{i=1}^K p_{\theta_i}(x_i | z_i)}{q_\varphi(u | z_1, z_2, \dots, z_K, c) \prod_{i=1}^K q_{\varphi_i}(x_i | z_i)} \right] \\ &= -\mathcal{L}(x_1, x_2, \dots, x_K, c) \end{aligned} \quad (53)$$

- For unlabeled data, only the variables x_1, x_2, \dots, x_K are treated as input, while $z_1, z_2, \dots, z_K, c, u$ are treated as unknown latent variables. Similar to the labeled data case, the objective function $\mathcal{U}(x_1, x_2, \dots, x_K)$ is the negative ELBO of the marginal distribution $p_\theta(x_1, x_2, \dots, x_K)$ on the approximate posterior $q_\varphi(z_1, z_2, \dots, z_K, c, u | x_1, x_2, \dots, x_K)$. Using the factorized forms from (43) and (48) the objective function for a single data point is expressed as:

$$\begin{aligned} \log p_\theta(x_1, x_2, \dots, x_K) &\geq \\ \mathbb{E}_{q_\varphi(z_1, z_2, \dots, z_K, c, u | x_1, x_2, \dots, x_K)} &\left[\log \frac{p_\theta(z_1, z_2, \dots, z_K, |c, u) p_\theta(c) p_\theta(u) \prod_{i=1}^K p_{\theta_i}(x_i | z_i)}{q_{\varphi_i}(c | z_1, z_2, \dots, z_K,) q_\varphi(u | z_1, z_2, \dots, z_K, c) \prod_{i=1}^K q_{\varphi_i}(x_i | z_i)} \right] \\ &= -\mathcal{U}(x_1, x_2, \dots, x_K) \end{aligned} \quad (54)$$

Therefore, the objective of the semi-supervised model for the stacked multimodal model, calculated as the marginal likelihood for the entire dataset, is as follows:

$$\mathcal{J} = \sum_{labeled} \mathcal{L}(x_1, x_2, \dots, x_K, c) + \sum_{unlabeled} \mathcal{U}(x_1, x_2, \dots, x_K) \quad (55)$$

In this problem formulation, the term $q_\varphi(c | z_1, z_2, \dots, z_K) = \text{Cat}(c | \pi_\varphi(z_1, z_2, \dots, z_K))$ represents the classifier. Like in any classification problem, the output is a categorical distribution that provides a score for each class.

By minimizing \mathcal{J} , the VAE aims to maximize the ELBO \mathcal{L} for labeled data and \mathcal{U} for unlabeled data. This process allows the VAE to learn latent representation for each

modality in the data. These latent representations capture the essential features needed for classification while simultaneously addressing the classification task.

In summary, we have proposed a technique for conducting data-level fusion with the semi-supervised VAE model M1+M2. Additionally, we have introduced an extension of this model, named Multi-M1+M2 that enables latent feature-level fusion.

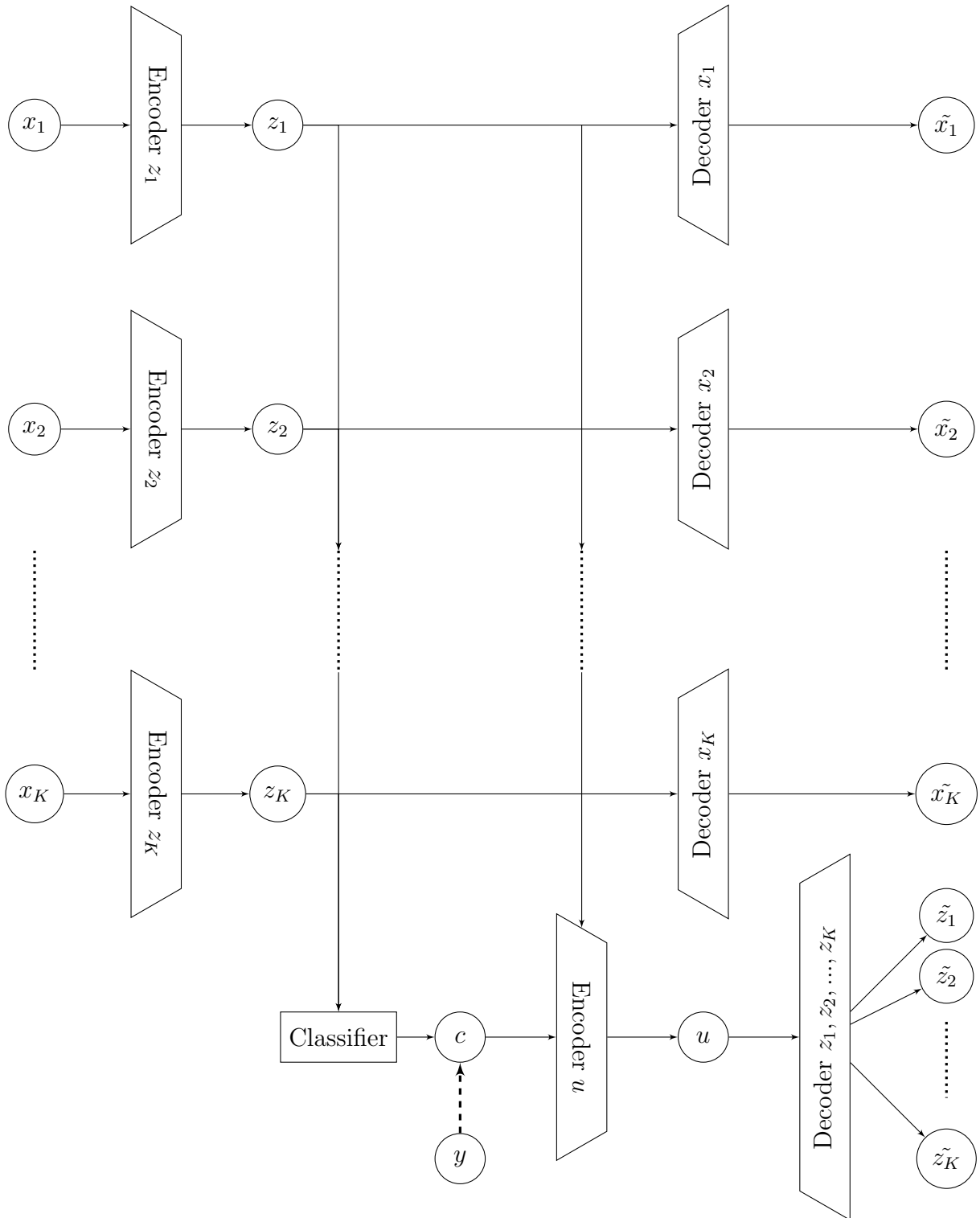


Figure 19: Graphical representation of *multimodal stacked generative semi-supervised model* - Multi-M1+M2 architecture

5 Experimental evaluation

This section presents the experiments, results, and analysis of multimodal data classification using semi-supervised VAE in two land cover pixel-wise classification scenarios. Land cover classification involves categorizing the Earth’s surface into different classes based on its physical or biological characteristics. It involves analyzing satellite imagery or aerial photographs to determine the distribution of land cover types in a particular area or region. Pixel-wise classification is implemented which classifies individual pixels in an image into different categories based on their characteristics.

The neural networks employed for analysis are the ones described in section 4: (1) M1+M2 architecture, which treats both modalities as a single input, and (2) Multi-M1+M2 architecture, designed to process two separate modalities independently as depicted in Figure 20. Both of these architectures are implemented using PyTorch (Paszke et al., 2019). We compared M1+M2 and Multi-M1+M2 classification results with the well-known Support Vector Machines (SVM) and Random Forest (RF) classification methods. SVM and RF are implemented using scikit-learn (Pedregosa et al., 2011).

The primary objective of this experimental section is to assess the architectures’ quality and robustness. We achieve this by utilizing two multimodal datasets, namely Trento and Houston which represent an urban and rural area respectively. The two datasets are analyzed separately. Each dataset consists of two modalities: LiDAR and HSI and more details regarding each dataset will be provided in subsections 5.2.1 and 5.3.1 respectively.

Section 5.1 will present the evaluation methods applied for comparing the developed architectures. Furthermore, section 5.2 and section 5.3 elaborate on the experimental procedures and analysis carried out on the Trento and Houston datasets respectively. Lastly, section 5.4 offers an overall comparison on the architectures performance in both of the datasets.

5.1 Evaluation methods

To compare each method’s performance we need to assess and analyze its classification results. To achieve this, we employed both quantitative and qualitative evaluation methods to capture different aspects. The quantitative evaluation utilized metrics such as accuracy, precision, recall, F1-score, and Kappa coefficient. Meanwhile, the qualitative

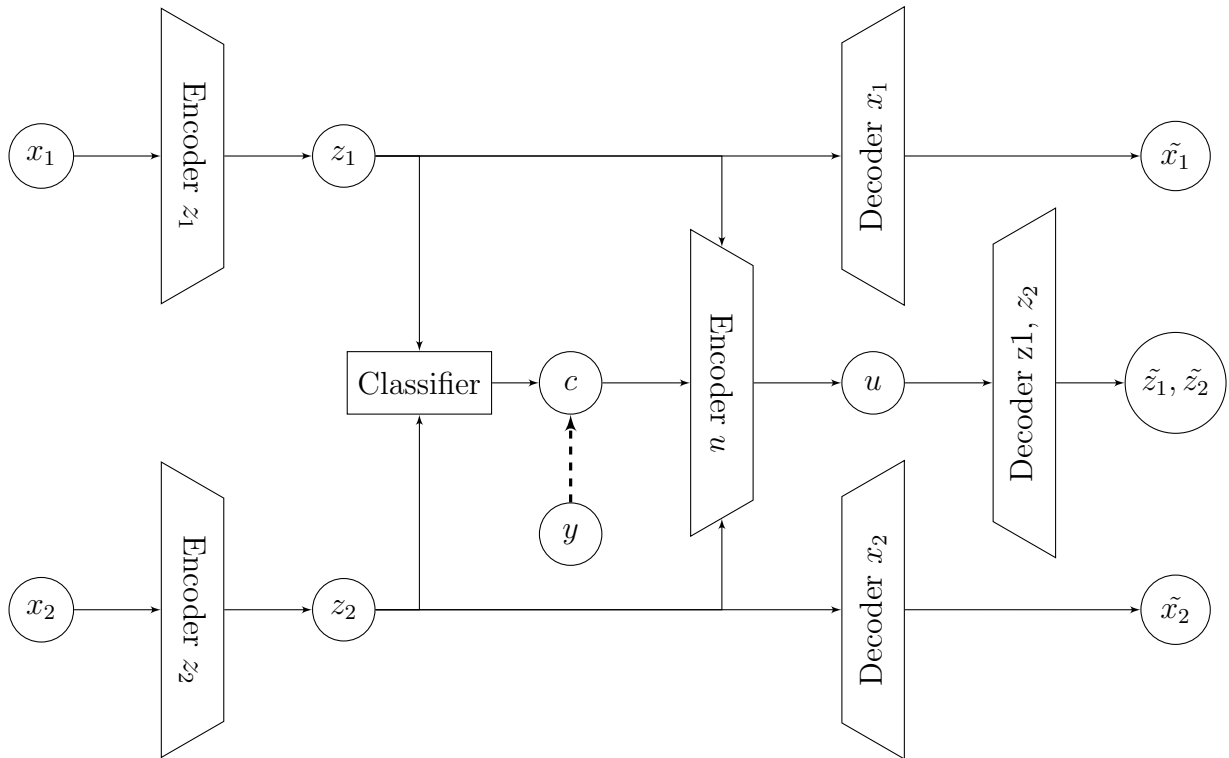


Figure 20: Multi-M1+M2 architecture for two modalities

assessment involved an examination of the classification maps and uncertainty maps. These offer valuable insights into each architectures' outcomes. The following section outlines the evaluation methods under consideration.

5.1.1 Accuracy

Accuracy, in the context of classification, is a measure of how well a classification model or algorithm correctly identifies or predicts the target classes or categories. It is typically calculated as the ratio of the number of correct predictions to the total number of predictions made by the model. The formula to calculate accuracy (Hossin and Sulaiman, 2015) is:

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total number of Predictions}} \quad (56)$$

Accuracy is a commonly used evaluation metric, especially when the classes are balanced, meaning they have similar proportions in the dataset (Hand, 2012). However, accuracy alone may not provide a complete picture of the model's performance, particularly in scenarios where the class distribution is imbalanced or the cost of misclassification varies across classes.

In imbalanced datasets, where one class dominates the sample, accuracy can be misleading. For instance, if 95 of instances belong to Class A and 5 belong to Class B, a naive model that predicts all instances as Class A would have a high accuracy of 95%, but it would fail to correctly predict any instances of Class B. In such cases, additional evaluation metrics like precision, recall, F1-score, or Kappa coefficient may be more informative (Hand, 2012).

5.1.2 Precision

Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive. Precision is calculated using the following formula (Hossin and Sulaiman, 2015):

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (57)$$

True Positives are the number of instances correctly predicted as positive, while False Positives are the number of instances incorrectly predicted as positive when they are actually negative. In the case of multiple classes we calculate as precision the average of the precision for each class, this method is called macro averaging (Hossin and Sulaiman,

2015).

Precision focuses on the quality of positive predictions and disregards instances that are incorrectly predicted as positive. Precision is particularly useful in situations where the cost of false positives is high.

5.1.3 Recall

Recall, also known as sensitivity or true positive rate, is a performance metric used in classification and information retrieval tasks to evaluate the ability of a model to correctly identify positive instances. It measures the proportion of true positive instances that are correctly predicted as positive out of all instances that are actually positive. Recall is calculated using the following formula (Hossin and Sulaiman, 2015):

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (58)$$

True Positives represent the instances that are correctly predicted as positive, while False Negatives represent the instances that are incorrectly predicted as negative when they are actually positive. For multiple classes we calculate recall with macro averaging for all the classes (Hossin and Sulaiman, 2015).

Recall focuses on the ability of the model to identify positive instances and avoid false negatives. Recall is particularly important in scenarios where missing positive instances carries significant consequences.

5.1.4 F1-score

The F1-score is a commonly used performance metric in classification tasks that combines precision and recall into a single value. It provides a balanced assessment of the model's accuracy in predicting both positive and negative instances. The F1-score is calculated using the following formula (Hossin and Sulaiman, 2015):

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (59)$$

We choose to evaluate on F1-score since it is particularly useful when the class distribution is imbalanced or when both false positives and false negatives have similar costs or implications. It gives equal importance to precision and recall, taking into account both the positive and negative class predictions.

5.1.5 Kappa coefficient

The Kappa coefficient (Cohen, 1960), also known as Cohen’s Kappa, is a statistical measure that assesses the level of agreement between two raters or evaluators when classifying categorical items. The Kappa coefficient takes into account the agreement between the observed agreement and the expected agreement that occurs due to chance. It corrects for the possibility of agreement occurring randomly and provides a more reliable measure of agreement than simply calculating the raw agreement rate. The Kappa coefficient ranges from -1 to 1, where: A value of 1 indicates perfect agreement between the raters or classifiers. A value of 0 indicates agreement that is no better than chance. A negative value indicates agreement that is worse than chance. The formula to calculate the Kappa coefficient is as follows (Sim and Wright, 2005):

$$Kappa = \frac{P_o - P_e}{1 - P_e} \quad (60)$$

Where P_o represents the observed agreement, which is the proportion of agreement between the raters. P_e represents the expected agreement, which is the proportion of agreement expected by chance. To calculate the expected agreement P_e , the Kappa coefficient considers the marginal probabilities of each rater’s classifications and assumes independence between the raters. It is calculated as the product of the proportions of items assigned to each category by each rater.

5.1.6 Classification map

A classification map is a visual representation of the categorization of classes within a specific geographic area. It provides a spatial depiction of the distribution and extent of various classes or categories of interest. These maps enable decision-makers, researchers, and planners to analyze and understand the spatial patterns and relationships between different classes, identify areas of interest, monitor changes over time, and make informed decisions based on the information conveyed by the map.

5.1.7 Uncertainty

An uncertainty map is a spatial representation that depicts the level of uncertainty or confidence associated with the classification or prediction of different features or classes within a specific area. Uncertainty serves as a valuable tool for decision-making by providing information about the reliability or uncertainty of the classification results

(Chlaily et al., 2023). It can offer insights into areas where the classification results may be less reliable, indicating the need for further investigation or data collection (Amodei et al., 2016). This information is crucial for making informed decisions and utilizing the classification results cautiously in subsequent applications or decision-making processes. By incorporating uncertainty maps into the analysis and interpretation of classification results, users can gain a better understanding of the limitations and potential errors in the classification process.

To assess and quantify uncertainty in classification results, various approaches and techniques can be employed. These include measures such as variance, Gini-Simpson index, or Shannon entropy (Shadman Roodposhti et al., 2019). In our analysis, we utilize the Homophily-based Uncertainty using Energy distance (HU) approach that Chlaily et al. (2023) proposed:

$$HU(y_*) \triangleq \frac{p_*^T H \odot H p_*}{p_{\max}^T H \odot H p_{\max}} = \frac{\sum_{i=1}^C \sum_{j=1}^C p_{*i} p_{*j} d(q_i, q_j)^2}{p_{\max}^T H \odot H p_{\max}} \quad (61)$$

Where p_* is a probability vector associated with a classifier's outcome y_* . Symbols \cdot^T and \odot denote the transpose operator and Hadamard product, respectively. Also, H denotes the distance/similarity measure between the probability distributions q_i and q_j corresponding to classes i and j , respectively. And where $p_{\max} = \operatorname{argmax}(p_*^T H \odot H p_*)$.

The concept of HU considers the likeness between classes and can be interpreted as a weighted summation. In this scheme, greater weights are allocated to classes that are farther apart. This is achieved through the utilization of a matrix denoted as H . This matrix quantifies the Energy distance between the probability distributions of classes.

$$H = (d(q_i, q_j))_{1 \leq i, j \leq C} = \left(\sqrt{\int \|Q_i(x) - Q_j(x)\|^2 dx} \right)_{(i,j) \in \{1, \dots, C\}^2} \quad (62)$$

Where C is the number of classes and Q_i and Q_j are the cumulative distribution functions for classes i and j , respectively. Consequently, HU quantifies the extent to which a classifier deviates from one that confuses distant classes.

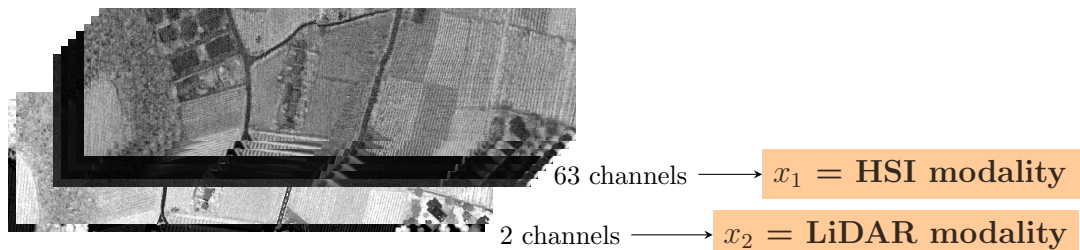


Figure 21: Trento Dataset: Modalities representation

5.2 Rural area classification

5.2.1 Trento dataset

The Trento dataset represents a rural area located in the city of Trento, Italy (Ghamisi et al., 2016). This dataset comprises two modalities: HSI data and LiDAR Digital Surface Model (DSM) data (Figure 21). The HSI consists of 63 bands spanning the spectrum of $402.89 - 989.09nm$, with a spectral resolution of $9.2nm$. Figure 22a illustrates a false-RGB representation of the entire scene using HSI bands. The LiDAR DSM data provides elevation information in meters above sea level, with a spatial resolution of $1m$ (Figure 22b). The dataset size is 600×166 pixels.

The scene has been partially labeled with six classes, as outlined in Table 11 and illustrated in Figure 22c. These classes offer a general indication of the presence of woods, apple trees, vineyards, ground areas, and more detailed identification of human-made structures. Table 11 gives also information about the number of training and testing samples for different classes. Additionally, we have utilized 1000 randomly selected unlabeled data points for unsupervised learning purposes. We use 20% of the training test for validation during the training process.

5.2.2 Experimental process

In our initial series of experiments, we employed the M1+M2 model as a basis for our classification model. The code implementation for these experiments was based on the repository provided by Lopez et al. (2020), which originally implemented the M1+M2 model for handwritten digit classification using the MNIST dataset. However, our focus was on point-wise remote sensing image classification, so we made modifications to the encoding part of the original architecture.

To explore different possibilities and improve our model, we experimented with various encoders. These encoders were varied in terms of their width, depth, and layer

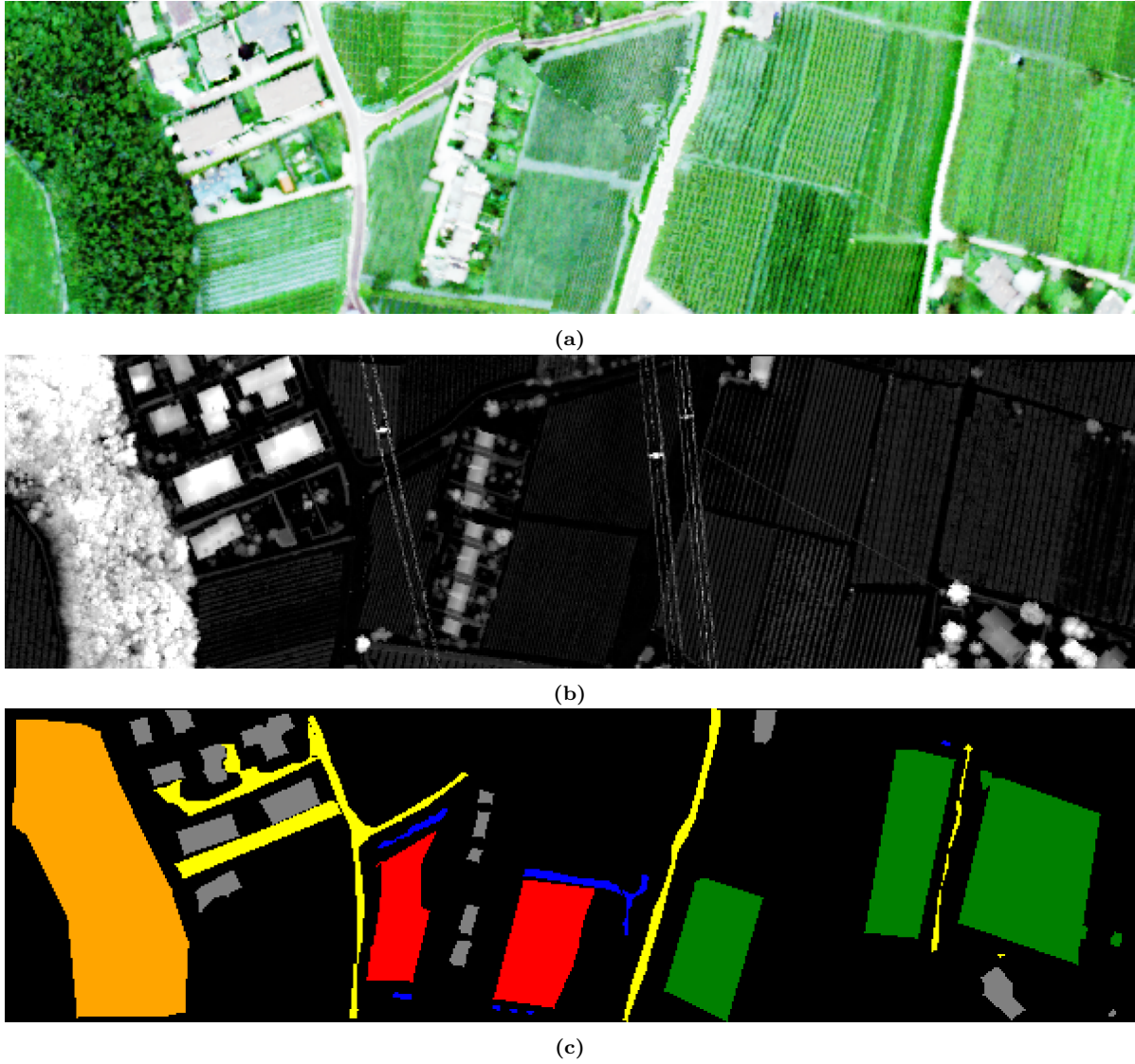


Figure 22: Trento dataset: (a) A false-RGB representation using HSI bands, (b) LiDAR DSM, (c) Ground truth labels

Table 11: Trento dataset: Color code and name of classes, number of training and test samples


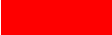





No	Class		Number of Samples		
	Color	Name	Labelled	Training	Testing
c_0		Unknown	69386	1000	0
c_1		Apple Trees	4034	129	3905
c_2		Buildings	2903	125	2778
c_3		Ground	479	105	374
c_4		Woods	9123	154	8969
c_5		Vineyards	10501	184	10317
c_6		Roads	3174	122	3252
Total			99600	1819	29395

Table 12: Encoders' architectures and configuration

Encoder	#	Layers		Pooling		Activation	
		Type	in	out	Type		Size
E1	3×	Linear	512	128	-	-	
E2	3×	Linear	1024	256	-	-	
E3	3×	1D Convolutional	32	128	1D Max pooling	2	SELU
E4	3×	1D Convolutional	128	512	1D Max pooling	2	
E5	2×	2D Convolutional	128	256	2D Max pooling	2	

types. Table 12 contains details about each encoder's architecture. Specifically, we implemented Encoder 1 (E1) and Encoder 2 (E2), which consisted of three fully-connected linear layers with increasing feature sizes. In contrast, Encoder 3 (E3) and Encoder 4 (E4) utilized a fully connected layer to input the data into three 1D convolutional layers of increasing size. Additionally, Encoder 5 (E5) was designed to process patches of size $n \times n$ from the scene and incorporated spatial information through two 2D convolutional layers. It is important to note that E5 had a higher training complexity compared to the other encoders due to its patching nature. Every encoder utilized the Scaled Exponential Linear Unit (SELU) as its activation function. We prefer SELU over other activation functions since it addresses the vanishing and exploding gradient problems that can occur in deep networks, promoting more stable and efficient training Goceri (2019). We also incorporated a dropout rate of 0.1 for effective regularization purposes.

In order to determine the optimal latent space size, which represents the input data, we conducted experiments with latent spaces ranging from 10 to 20. Considering that the Trento dataset comprises six classes, we aimed for a range of latent space sizes comparable to the number of classes. Initially, we trained the M1+M2 architecture

Table 13: Trento dataset: Test set metrics for the M1+M2 model with encoders E1, E2, E3, and E4 on different latent spaces

Encoder	Latent size	Accuracy	Precision	Recall	F1-score	Kappa coeff
E1	10	84.80	<i>73.05</i>	80.51	<i>76.09</i>	78.91
	15	84.86	72.19	<i>80.76</i>	75.63	79.03
	20	<i>84.94</i>	72.38	80.70	75.75	<i>79.13</i>
E2	10	90.44	88.55	92.58	89.94	87.44
	15	92.48	92.17	92.92	92.43	90.01
	20	88.48	87.70	92.28	88.68	85.01
E3	10	<i>90.24</i>	87.92	<i>91.89</i>	<i>89.40</i>	<i>87.13</i>
	15	84.68	71.49	80.50	75.05	78.79
	20	89.52	<i>88.13</i>	91.65	89.36	86.22
E4	10	91.13	89.62	92.54	90.76	88.28
	15	90.70	88.83	91.93	90.01	87.73
	20	<i>91.85</i>	<i>90.10</i>	<i>92.77</i>	<i>91.20</i>	<i>89.21</i>

for encoders E1-E4 across the various latent space sizes. Based on the performance of these experiments, we selected the best-performing latent size and utilized it to train E5. Finally, we selected the best-performing encoder based on the evaluation methods described in section 5.1 to train the Multi-M1+M2 architecture.

Throughout all our experiments, we employed the Adam optimizer with an initial learning rate of 1e-4 and trained the models for 500 epochs.

5.2.3 Results

Table 13 presents the classification accuracy, precision, recall, F1-score, and Kappa coefficient for each latent space and different encoders. The highest score for each encoder is marked in italics, while the best performing combination of encoder and latent size is marked in bold. By selecting an appropriate latent size and autoencoder, we can strike a balance between performance and computational efficiency.

Regarding the impact of latent size on performance, it is observed that increasing the size of the latent space for E1 positively affects its performance, although this effect is not as pronounced for the other models. Specifically, for E2, a latent size of 15 yields the optimal results, while for E3, a latent size of 10, and for E4, a latent size of 20.

E1, which features the smallest linear layers, exhibits the lower performance. Comparatively, convolutional encoders E3 and E4 produce results that are on par with the larger linear encoder E2. Notably, E2 consistently outperforms the others across all metrics. Moreover, the convolutional nature of E3 and E4 comes with increased com-

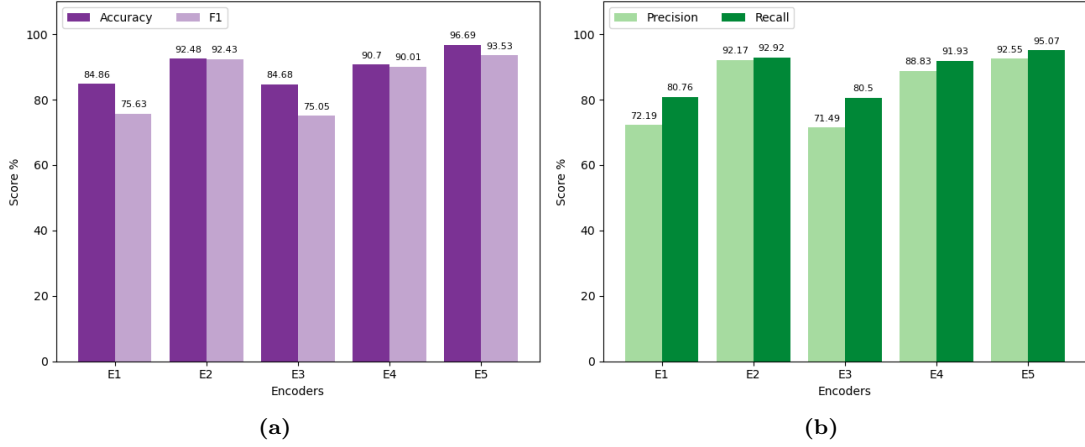


Figure 23: Trento dataset: Test set metrics for the M1+M2 model with encoders E1, E2, E3, E4 and E5 on latent size = 15, (a) Accuracy and F1-score, (b) Precision and recall

Table 14: Trento dataset: Comparison of test set metrics for the best VAE models with SVM and RF

Model	Accuracy	Precision	Recall	F1-score	Kappa coeff
M1+M2 with E2	92.48	92.17	92.92	92.43	90.01
M1+M2 with E5	96.69	92.55	95.08	93.53	95.59
Multi-M1+M2 with E2	87.49	87.09	91.60	87.85	83.77
SVM	86.34	83.07	88.80	84.83	82.10
RF	88.38	85.38	89.56	86.92	84.62

computational complexity. It is interesting to emphasize that greater complexity does not necessarily translate to improved performance in this context.

Based on the findings from the experiments in encoders E1 through E4, a latent size of 15 is selected for the training of the more computationally expensive E5, which utilizes patches of size 5x5.

Figure 23 depicts bar plots illustrating the classification accuracy, precision, recall, and F1-score across various encoders within the M1+M2 architecture’s latent space set at 15. Focusing on Figure 23a, we observe that the F1-score consistently registers lower values than accuracy. This difference arises due to accuracy’s failure to account for data imbalances. Notably, certain encoders like E2 and E4 exhibit narrower gaps between these metrics, indicating their effectiveness in addressing data imbalance issues. In Figure 23b, another pattern becomes evident: recall consistently outperforms precision. This pattern signifies our architecture’s effectiveness in identifying a majority of correct outcomes, though it may struggle to deliver a higher ratio of relevant-to-irrelevant results.

Table 14 presents the performance metrics for different models, including M1+M2 with the best performing encoder using patching (E5) and without patching (E2), as well

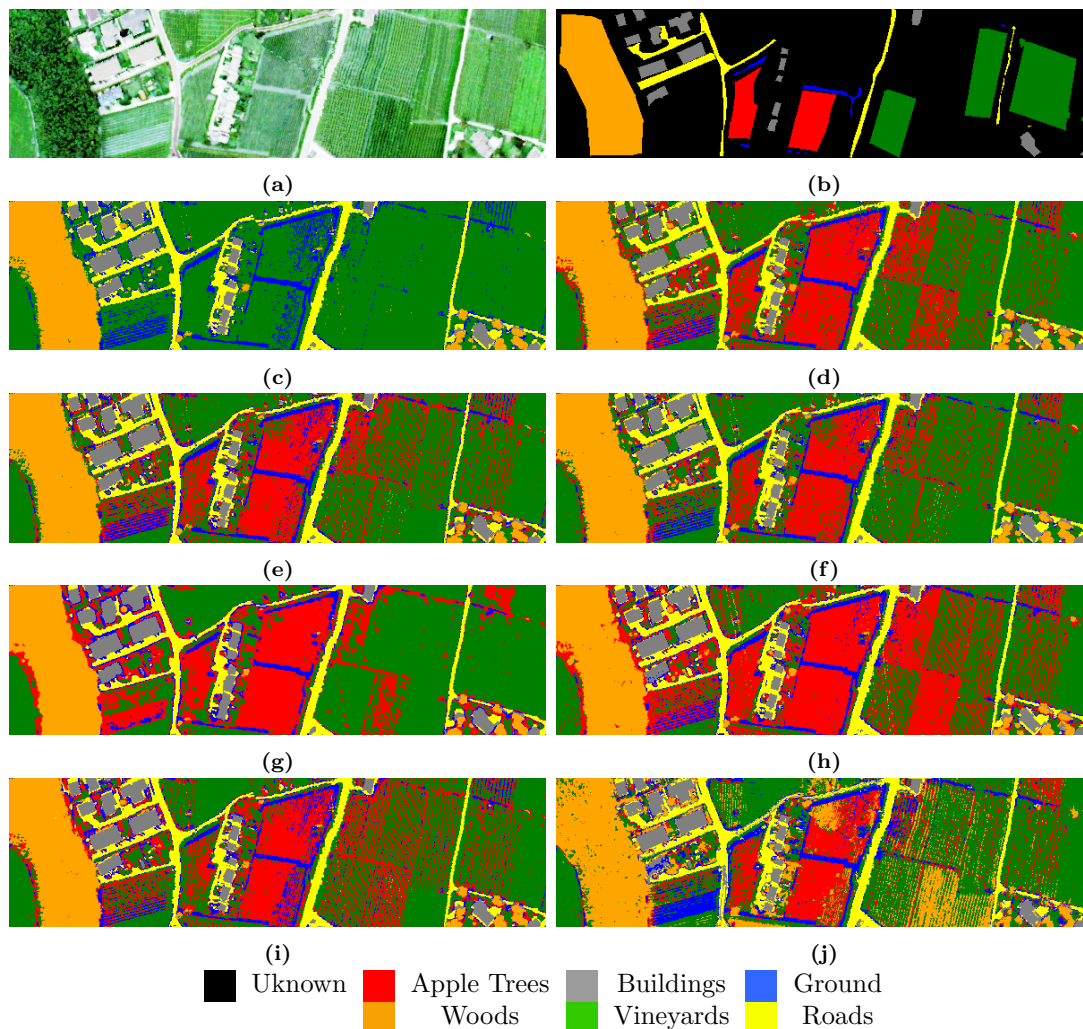


Figure 24: Trento dataset: (a) A false RGB representation using HSI bands, (b) Ground truth labels, Classification maps for M1+M2 model with (c) E1 on latent size = 20, (d) E2 on latent size = 15, (e) E3 on latent size = 10, (f) E4 on latent size = 20, (g) E5 on latent size = 15 and patch size = 5x5, (h) Multi-M1+M2 model with E2 on latent size = 15 and (i) SVM, (j) RF

as the Multi-M1+M2 model, SVM, and RF. The purpose of this table is to compare the performance of these models based on various metrics. By examining these metrics, we can gain insights into the effectiveness of different models and make informed decisions regarding our selection.

In addition to the quantitative metrics presented in the previous section, it is also important to consider the visual representation of the classification maps to gain a deeper understanding of the model's performance. Figure 24 provides an overview of the classification maps for the entire dataset, including both labeled and unlabeled data points. The latent sizes with the highest metrics for E1, E2, E3, and E4 were selected

for display in this figure.

While E5 demonstrates the best metrics on the test set, a closer examination of Figures 24c-24g reveals that the classification maps generated by this encoder exhibit a higher degree of blurriness when applied to the entire image. This blurriness may indicate a loss of detail and precision in the classification results. To mitigate this issue, we have chosen to incorporate E2 into the Multi-M1+M2 model. By referring to Figure 24h, Figure 24i, and Figure 24j, we can observe the classification maps for the results obtained using the Multi-M1+M2 model, SVM, and RF, respectively.

Figure 25 presents the uncertainty maps for the selected classifications using the Homophily uncertainty approach. The Homophily matrix in Equation 63 is defined using the energy distance.

In the uncertainty maps, lower uncertainty values are represented by blue colors, indicating areas where the classifier exhibits higher confidence in its classifications. As the uncertainty values increase, the colors transition from blue to red, indicating areas where the classifier is more uncertain due to the presence of closely related classes or a small number of classes. As the uncertainty values approach red, it signifies higher confusion between multiple or more distant classes.

$$H_{\text{Trento}} = \begin{matrix} & \begin{matrix} c_1 & c_2 & c_3 & c_4 & c_5 & c_6 \end{matrix} \\ \begin{matrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \end{matrix} & \begin{bmatrix} 0 & 0.89 & 0.58 & 0.35 & 0.36 & 0.88 \\ 0.89 & 0 & 0.56 & 0.85 & 1 & 0.33 \\ 0.58 & 0.56 & 0 & 0.6 & 0.73 & 0.65 \\ 0.35 & 0.85 & 0.6 & 0 & 0.51 & 0.91 \\ 0.36 & 1 & 0.73 & 0.51 & 0 & 0.95 \\ 0.88 & 0.33 & 0.65 & 0.91 & 0.95 & 0 \end{bmatrix} \end{matrix} \quad (63)$$

5.2.4 Analysis

Comparing the classification maps (Figure 24), VAE models demonstrate high accuracy for the classes of Buildings c_2 and Roads c_6 . Additionally, these classes exhibit low uncertainty, showing a high level of confidence in the classification results (Figure 25) which indicates that there is no confusion or confusion between a very small amount or close classes. It is worth noting that Buildings c_2 and Roads c_6 show low energy distances between them and very high energy distances between the rest of the classes (Equation 63). While these classes are closely related due to their human-made nature

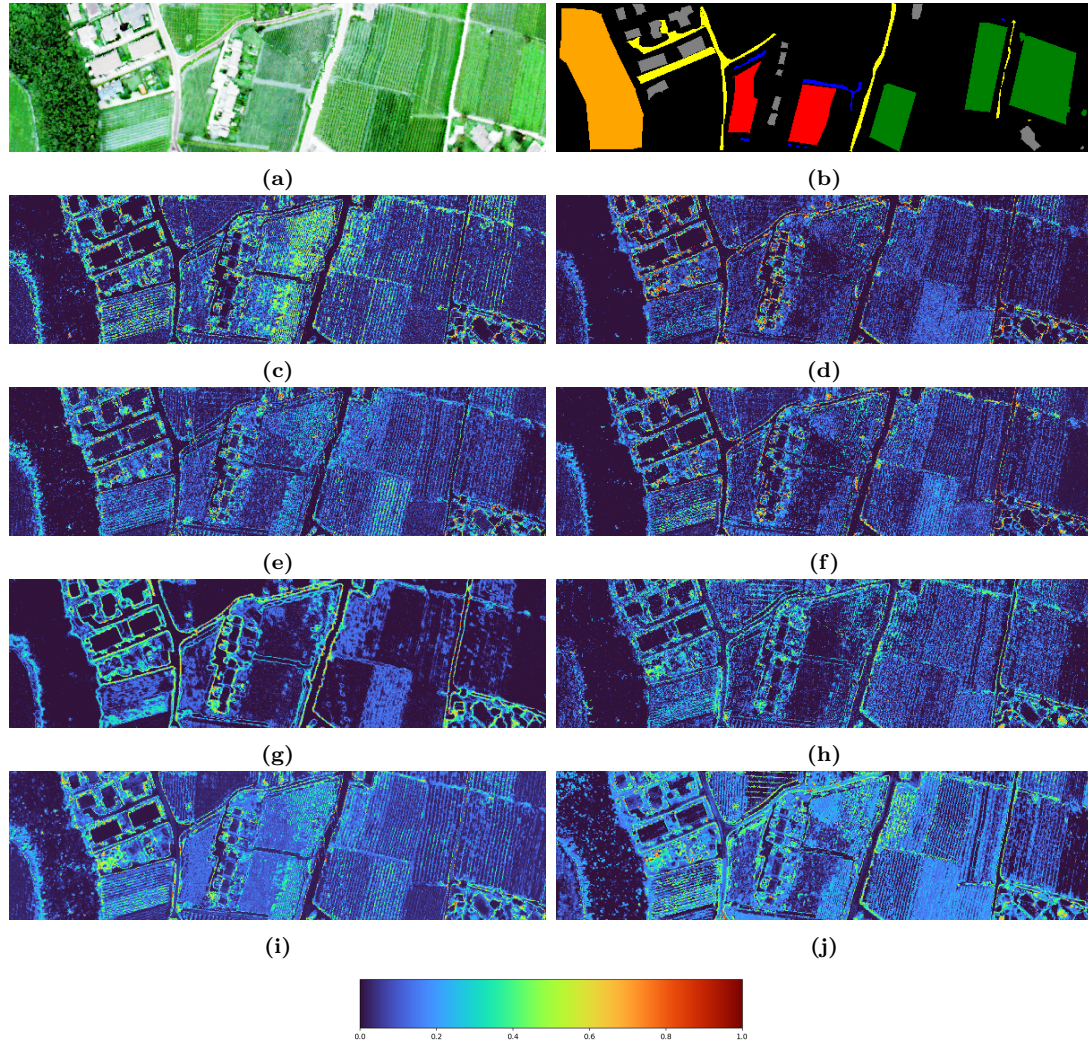


Figure 25: Trento dataset: (a) A false RGB representation using HSI bands, (b) Ground truth labels, Homophily-based uncertainty maps for M1+M2 model with (c) E1 on latent size = 20, (d) E2 on latent size = 15, (e) E3 on latent size = 10, (f) E4 on latent size = 20, (g) E5 on latent size = 15 and patch size = 5x5, (h) Multi-M1+M2 model with E2 on latent size = 15 and (i) SVM, (j) RF

and use of similar materials, they exhibit significant differences in terms of their elevation which results in accurate classification.

Models M1+M2 with E2, E3, E4, and particularly with E1 (Figure 24c-24f), as well as the Multi-M1+M2 model, successfully detect Ground c_3 areas. The Multi-M1+M2 model exhibits a particularly sharp Ground c_3 detection. It is worth noting that, apart from the ground spaces between the fields, there are detections between the Apple Trees c_1 . It's key to recognize that these detections may be considered incorrect based on the way labels are created. However, in reality, there is indeed ground between the vegetation. These areas exhibit higher uncertainty, indicating confusion between distant classes.

The use of patching of E5 results in a degradation of the definition of the Buildings c_2 , Roads c_6 , and Ground c_3 classes, as it smooths out their boundaries (Figure 24g). This effect is also evident in the corresponding uncertainty map (Figure 25g), where increased uncertainty is observed at the edges of Buildings c_2 , Roads c_6 , and Ground c_3 .

The class of Woods c_4 is successfully detected by all the M1+M2, Multi-M1+M2 models (Figure 24c-24h), even in areas outside the main forested region with low uncertainty, showing again a high level of confidence in the classification results (Figure 25). The vegetation classes of Apple Trees c_1 , Vineyards c_5 and Woods c_4 show low energy distances between them. Upon closer inspection, a few wood pixels are classified as Vineyards c_5 or Apple Trees c_1 . These predictions exhibit a slightly higher level of uncertainty (Figure 25c-24h), indicating the potential confusion among those classes. The M1+M2 model with patching encoder E5 provides a smooth classification of the large Woods c_4 area.

The classes of Apple Trees c_1 and Vineyards c_5 exhibit the most variation. The M1+M2 model with E1 fails to detect any Apple Trees c_1 , classifying all of them as Vineyards c_5 (Figure 24c). Evidently, this model lacks the capacity to learn the differences between these two vegetation classes. By increasing the size of the layers, the model with E2 achieves much better results in Apple Trees c_1 classification (Figure 24d). With convolutional encoders E3 and E4 (Figure 24e-24f), there is still a mixture between Vineyards c_5 and Apple Trees c_1 . The spatial information provided by E5 appears to improve the mixing between these two classes (Figure 24g), while also reducing the levels of uncertainty (Figure 25g), but at the cost of producing a bulkier result. The Multi-M1+M2 model demonstrates the best separation between these two classes (Figure 24h), indicating that latent feature-level fusion has helped improve the

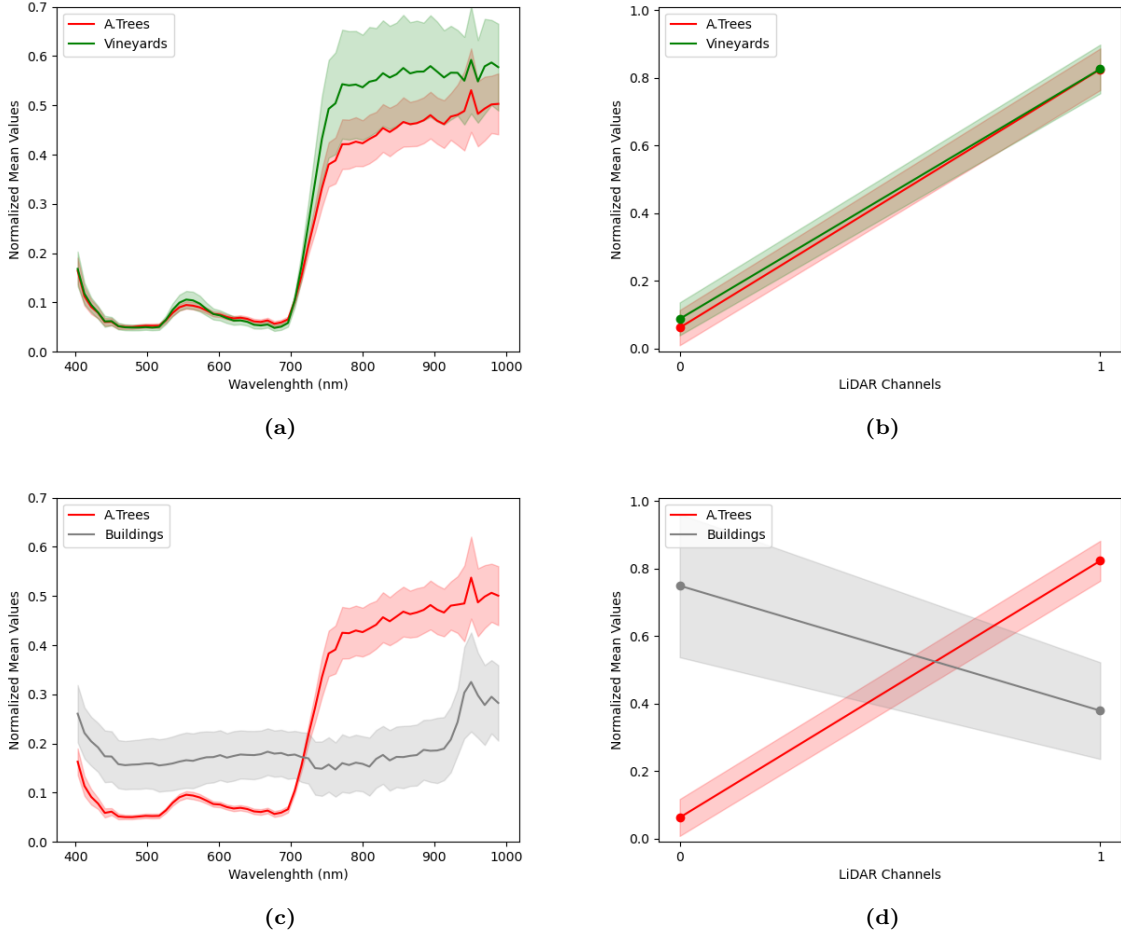


Figure 26: Trento dataset: (a),(b) HSI spectrum and LiDAR channels respectively for vegetation classes, (c),(d) HSI spectrum and LiDAR channels respectively for buildings and roads

distinguishability between the two vegetation classes. An interpretation of our outcome could stand from the fact that HSI dimensionality significantly surpasses that of LiDAR. Hence, a feature-based approach proves to be more effective as it adeptly balances the contributions from each modality.

Figure 26 shows the mean values of the channels for HSI and LiDAR data for a selection of Trento dataset’s classes.

When comparing the HSI and LiDAR of Buildings c_2 and Apple Trees c_1 , minimal overlap is observed (Figure 26c, 26d). What’s interesting is that none of the classifiers get confused between these two classes. This clear distinction between the HSI and LiDAR characteristics of these classes contributes to their accurate separation.

In contrast, the classes that exhibit a higher degree of mixing during classification are Vineyards c_5 and Apple Trees c_1 , which are both vegetation classes. A comparison

of the mean values obtained from HSI and LiDAR for these two classes (see Figure 26a and Figure 26b) reveals a substantial degree of similarity. Specifically, the first half of the mean HSI signatures and the second channel of LiDAR exhibit significant overlap. This overlapping phenomenon presents a considerable challenge in effectively distinguishing between these classes.

Nevertheless, it is noteworthy that none of the models are able to achieve a complete and distinct separation between these classes. This observation implies that solely relying on HSI and LiDAR data may not be sufficient for discriminating between classes like Apple Trees c_1 and Vineyards c_5 . Additionally, the possibility exists that this overlapping is a result of incorrect labeling.

Lastly, it is worth significant to point out that the M1+M2 model with 2D patched encoder exhibits the highest metrics (Table 14). However, the classification results of the other classifiers are preferable since they maintain important spatial details. This observation underlines the limitation of relying solely on quantitative metrics as an indicator of model performance.

These findings highlight the effect of VAEs and deep semi-supervised learning in the classification process, as it has strong learning capabilities that can aid in distinguishing between classes. By leveraging the unique characteristics of each data source using latent feature-level fusion, it is possible to enhance the performance and reliability of the classification outcomes. Moreover, it's worth emphasizing that standard metrics appeared to be able to provide only an initial overview of performance. Relying solely on metrics calculated on a very small labeled dataset is not advisable. Instead, a more comprehensive evaluation can be achieved by delving deeper into the qualitative results.

5.2.5 Comparison with SVM and RF

After analyzing the semi-supervised VAE models we want to compare them with the performance of widely used methods of SVM and RF. Similarly with M1+M2 and Multi-M1+M2, SVM and RF classifier has high accuracy for classes of Buildings c_2 and Roads c_6 (Figure 24) while exhibiting low uncertainty, indicating a high level of confidence in the classification results (Figure 25). SVM and RF model has an amplified effect of detecting Ground c_3 between the Apple Trees c_1 (Figure 24i, 24j). As mentioned above there is indeed Ground c_3 between the vegetation but RF appears to be even more aggressive in Ground c_3 detection, as it detects the entire Apple Trees c_1 area as Ground c_3 (Figure 24j). These areas exhibit higher uncertainty than M1+M2 and Multi-M1+M2 models, indicating the confusion between a bigger number or more distant

classes.

Furthermore, the RF classifier appears to have a greater confusion than M1+M2, Multi-M1+M2 and SVM when it comes to Woods c_4 , as it detects some Vineyards c_5 as Woods c_4 and identifies numerous Woods c_4 areas in regions that should contain Vineyards c_5 and Apple Trees c_1 . The challenging classes of Apple Trees c_1 and Vineyards c_5 seem to be more weakly separated in SVM than the Multi-M1+M2 model. While, the RF classifier further aggravates the mixing of classes, detecting Apple Trees c_1 , Vineyards c_5 , and Woods c_4 in areas that consist solely of Vineyards c_5 or Apple Trees c_1 (Figure 24j), with these areas exhibiting higher uncertainty (Figure 25j).

There are different aspects to the reason why SVM and RF show inferior performance than the deep learning approach. One factor that can affect the performance of SVM compared to neural networks is the complexity of the problem. SVMs are known for their ability to handle high-dimensional feature spaces and perform well in cases where the number of features is larger than the number of samples. However, in cases where the problem is highly complex and the relationship between the features and the target variable is nonlinear, neural networks with their ability to learn complex patterns and relationships may outperform SVM (Coltekin and Rama, 2018). Furthermore, RF may be more sensitive to noisy or irrelevant features in the dataset compared to neural networks. RF constructs decision trees based on random subsets of features, and if the dataset contains noisy or irrelevant features, these may be included in the decision-making process. In contrast, neural networks can learn to ignore irrelevant features through the process of training and regularization, leading to potentially better performance in the presence of noisy data (Chen and Ishwaran, 2012).

Shallow learning techniques such as SVM and RF exhibit an advantage in terms of efficiency when it comes to training duration. This stands in sharp contrast to deep learning methods, which require more extensive computational resources and time for training. Although, when the computational resources are available neural networks can be trained efficiently using parallel processing and can handle large amounts of data more effectively (Huang et al., 2012).

In Figure 27 we can see the elapsed time for training the deep learning models we developed along with SVM and RF as implemented from scikit-learn (Pedregosa et al., 2011). We conducted our experiments on a computer system equipped with the following specifications: an Intel Core i7-5500U CPU running @2.4GHz, 16GB of DDR4 RAM. The experiments were executed on a Ubuntu operating system (Version 20.04.6 LTS),

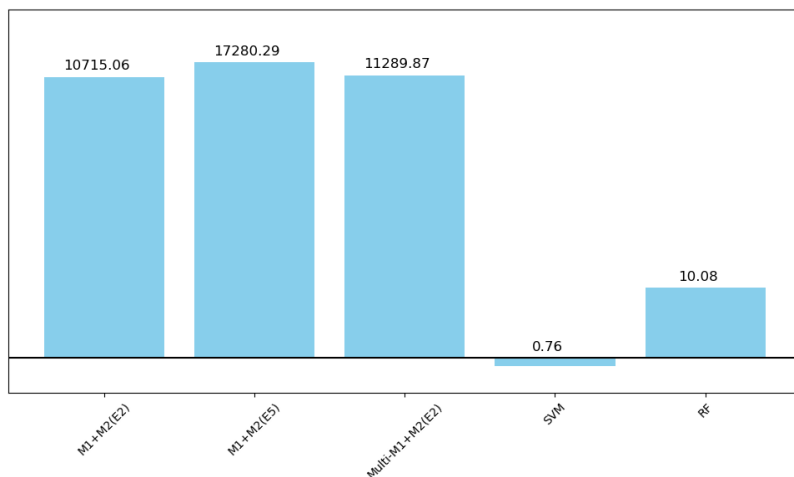


Figure 27: Training elapsed time at Intel Core i7-5500U@2.4GHz, 16GB DDR4 RAM on Ubuntu 20.04.6 LTS, Python 3.8. (Bar heights are represented on a logarithmic scale for enhanced visualization)

and we employed Python 3.8 for our computations. Elapsed time measurements were taken using Python’s ‘time’ module. The absolute time values may vary on different systems. Training M1+M2 with the spatial information from encoder E5 takes nearly five hours, while models M1+M2 and Multi M1+M2 with encoder E2 require two and a half hours, in contrast to SVM and RF methods, which only take a few seconds.

Overall, in comparison to widely adopted shallow learning techniques, VAE models demonstrated superior results while entailing longer training times. The next step is to further evaluate the capabilities of the selected M1+M2 and Multi-M1+M2 and assess their potential for real-world applications. Moving forward we will employ a dataset with a larger number of classes and more challenging class differentiation.

5.3 Urban area classification

5.3.1 Houston dataset

The Houston dataset 2018 was collected to capture urban areas around the University of Houston, USA campus and its surroundings. This dataset was originally distributed for the 2018 Geoscience and Remote Sensing Society Data Fusion Contest (Xu et al., 2019). The dataset includes two modalities: HSI and LiDAR (Figure 28). The HSI consists of 48 bands ranging from 380 to 1050 nm with a spectral resolution of 0.5m. Figure 29a presents a false-RGB representation of the total scene using HSI bands. The image size is 4172 x 1202 pixels. The LiDAR data consists of seven channels. The four

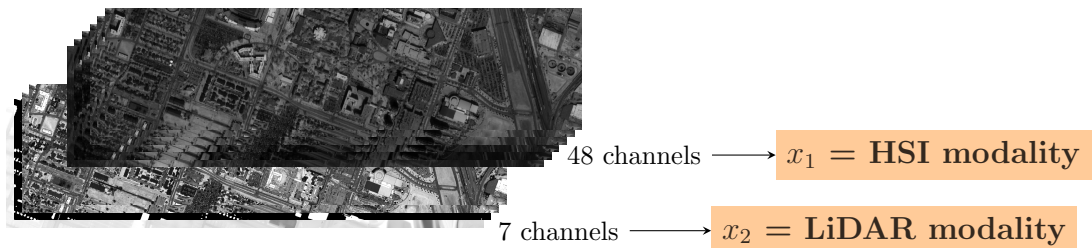


Figure 28: Houston Dataset: Modalities representation

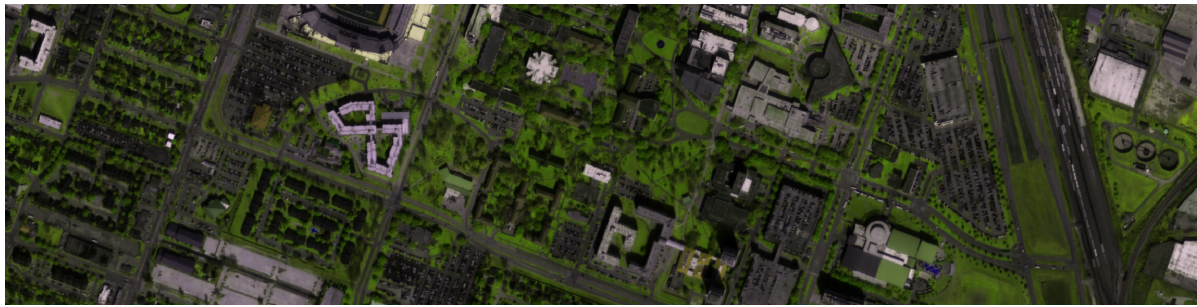
elevation channels represent the DSM and the Digital Earth Model (DEM) (Figure 29b). The other three channels provide intensity rasters at three different laser wavelengths that include information about the surface nature, such as bare earth or human-made structures (Figure 29c). The size of the LiDAR dataset is 8344 x 2404 pixels, and it has been down-sampled by a factor of two to match the size of the HSI, resulting in a final spatial resolution of 1m.

The Houston dataset has been partially labeled with twenty classes of the urban scene, as outlined in Table 15 and illustrated in Figure 29d.. These classes provide a detailed separation between similar classes, such as residential and non-residential buildings, different parts of the road, various types of vegetation, and trees. Additionally, some relatively rare objects like cars, trains, and stadium seats were labeled to test the limits of both the sensor devices and the classification algorithms.

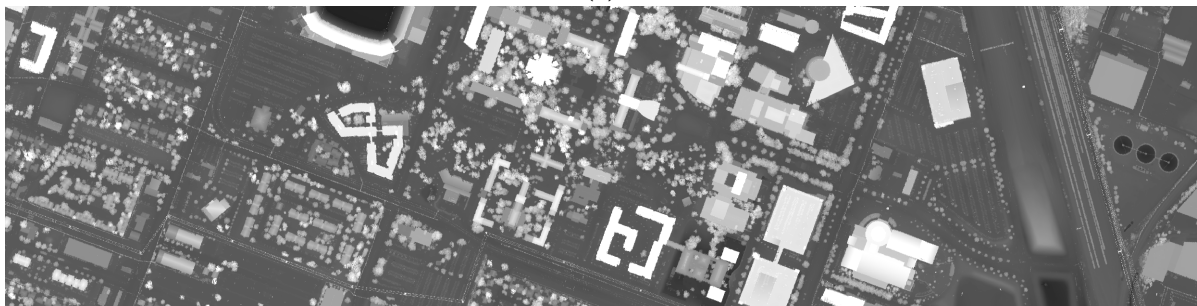
For training and validation purposes, a random split was performed with 2000 labeled datapoints per class with the exception of classes Water c_7 and Unpaved parking lots c_{17} , which lacked a sufficient number of labeled samples. Consequently, we opted for 1000 samples for Water c_7 and 500 samples for Unpaved parking lots c_{17} . In addition, 5000 unlabeled datapoints were used for unsupervised training (Table 15). The main goal of this split was to ensure an equal representation of all the classes during training. The remaining labeled data were used for testing purposes.

5.3.2 Experimental process

In our second series of experiments, we aim to build upon our previous results on the rural area dataset and investigate the performance of semi-supervised VAE models on a more complex dataset. Specifically, we want to assess how the different fusion methods will perform when deal with a significantly larger number of classes and more challenging distinctions between classes. To achieve this, we will focus on exploring the encoder E2, which exhibited the most promising results in terms of metrics and quantitative analysis



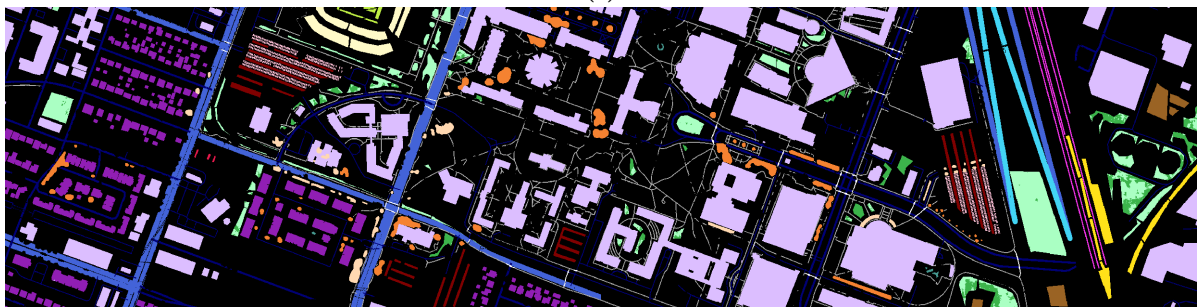
(a)



(b)



(c)



(d)

Figure 29: Houston dataset: (a) A false RGB representation using HSI bands, (b) LiDAR DSM, (c) Color composite of multispectral LiDAR intensity, (d) Ground truth labels

Table 15: Houston dataset: Color code and name of classes, number of training and test samples










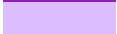










No	Color	Class	Number of Samples		
		Name	Labeled	Training	Testing
c_0		Unknown	3712226	5000	0
c_1		Healthy grass	39196	2000	37196
c_2		Stressed grass	130008	2000	128008
c_3		Artificial turf	2736	2000	736
c_4		Evergreen trees	54322	2000	52322
c_5		Deciduous trees	20172	2000	28172
c_6		Bare earth	18064	2000	16064
c_7		Water	1064	1000	64
c_8		Residential buildings	158995	2000	156995
c_9		Non-residential buildings	894669	2000	892669
c_{10}		Roads	183283	2000	181283
c_{11}		Sidewalks	136035	2000	134035
c_{12}		Crosswalks	6059	2000	2059
c_{13}		Major thoroughfares	185438	2000	183438
c_{14}		Highways	39438	2000	37438
c_{15}		Railways	27748	2000	25748
c_{16}		Paved parking lots	45932	2000	43932
c_{17}		Unpaved parking lots	587	500	87
c_{18}		Cars	26289	2000	24289
c_{19}		Trains	21479	2000	19479
c_{20}		Stadium seats	27296	2000	1976410
Total			5731136	42500	5688636

Table 16: Houston dataset: Test set metrics for the M1+M2 model with encoder E2 on different latent spaces

Latent size	Accuracy	Precision	Recall	F1-score	Kappa coeff
10	64.75	51.42	76.47	58.39	56.06
20	66.03	51.94	76.58	57.84	58.81
30	77.26	64.12	78.96	68.89	71.53
40	64.79	43.72	65.19	49.86	55.62

in our previous experiments.

To ensure an optimal latent space for the Houston dataset, which contains a considerably larger number of classes compared to the Trento dataset, we increased the size of the latent space to test values of 10, 20, 30, and 40. We trained the M1+M2 model with E2 for all four latent sizes and selected the latent space that yields the best metrics. Once we had determined the optimal latent size, we trained the Multi-M1+M2 model using this chosen latent size.

Throughout all of our experiments, we utilized the Adam optimizer with an initial learning rate of $1e-4$. The models were trained for a total of 500 epochs to ensure sufficient convergence and capture the underlying patterns in the data.

By conducting these experiments, we aimed to gain insights into the performance of our models in a more complex dataset with a larger number of classes and more challenging class differentiation. This allowed us to further evaluate the capabilities of our models and assess their potential for real-world applications.

5.3.3 Results

Table 16 provides a comprehensive analysis of the classification performance of model M1+M2 with E2, showcasing evaluation metrics such as classification accuracy, precision, recall, F1-score, and the Kappa coefficient for each corresponding latent space size. The highest score for each metric is highlighted in bold, allowing us to identify the most effective configurations.

Our investigation into the impact of latent space size on performance reveals an intriguing trend. Initially, as the latent space size increases from 10 to 30, we observe a progressive improvement in model performance, indicating that a larger latent space allows for more meaningful representations and enhanced learning capabilities. However, beyond a latent size of 30, the performance starts to deteriorate, suggesting that an excessively large latent space may lead to overfitting or loss of important information.

Table 17: Houston dataset: Test set metrics for the best VAE models

Model	Accuracy	Precision	Recall	F1-score	Kappa coeff
M1+M2 with E2	77.26	64.12	78.96	68.89	71.53
Multi-M1+M2 with E2	70.81	57.53	80.23	64.41	64.22

Table 18: Houston dataset: Confusion matrix of M1+M2 model with E2

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}	c_{14}	c_{15}	c_{16}	c_{17}	c_{18}	c_{19}	c_{20}
c_1	.99	0	0	.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c_2	.04	.93	0	0	0	0	0	0	0	0	.02	0	0	0	0	0	0	0	0	0
c_3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c_4	.01	0	0	.96	0	0	0	.01	0	0	.02	0	0	0	0	0	0	0	0	0
c_5	0	.02	0	0	.91	0	0	.03	.01	.01	.02	0	0	0	0	0	0	0	0	0
c_6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c_7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
c_8	0	0	0	.01	.01	0	0	.94	.02	0	.01	0	0	0	0	0	0	0	0	0
c_9	0	0	0	0	.01	0	0	.04	.85	.01	.05	.01	0	.01	0	0	0	0	.01	0
c_{10}	0	0	0	0	0	0	0	.02	0	.61	.13	.18	0	.03	.01	.01	0	0	0	0
c_{11}	.02	.03	0	.02	.01	0	0	.03	.01	.09	.72	.05	0	.01	0	0	0	0	0	0
c_{12}	0	0	0	0	0	0	0	0	0	.03	.04	.93	0	0	0	0	0	0	0	0
c_{13}	0	.01	0	0	0	0	0	0	0	.38	.22	.26	0	.1	0	.01	0	0	0	0
c_{14}	0	0	0	0	0	0	0	0	0	.02	.04	.06	0	.87	0	0	0	0	0	0
c_{15}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.99	0	0	0	0	0
c_{16}	0	0	0	0	0	0	0	0	0	.02	0	.01	0	0	0	.95	0	.02	0	0
c_{17}	0	0	0	0	0	0	0	0	0	.86	.14	0	0	0	0	0	0	0	0	0
c_{18}	0	0	0	0	.01	0	0	0	.01	0	.01	.01	0	0	0	.04	0	.93	0	0
c_{19}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.99	0
c_{20}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Based on the compelling findings from these experiments, we conclude that a latent space size of 30 emerges as the optimal choice for our dataset. This particular configuration strikes a balance, enabling the model to effectively capture the underlying complexities of the data without succumbing to the pitfalls of excessive dimensionality. Thus, the chosen latent size of 30 is expected to facilitate robust and accurate learning on the given dataset.

Table 17 displays a comparison of performance metrics for the models including M1+M2 and the Multi-M1+M2 model utilizing the most effective encoder E2. The primary objective of this table is to assess and contrast the effectiveness of these models based on a range of performance metrics. By analyzing these metrics, we can draw valuable insights into how these models scale and perform in different scenarios. Additionally, it offers us valuable guidance on selecting the most appropriate model for specific tasks, considering their respective scalability and overall effectiveness.

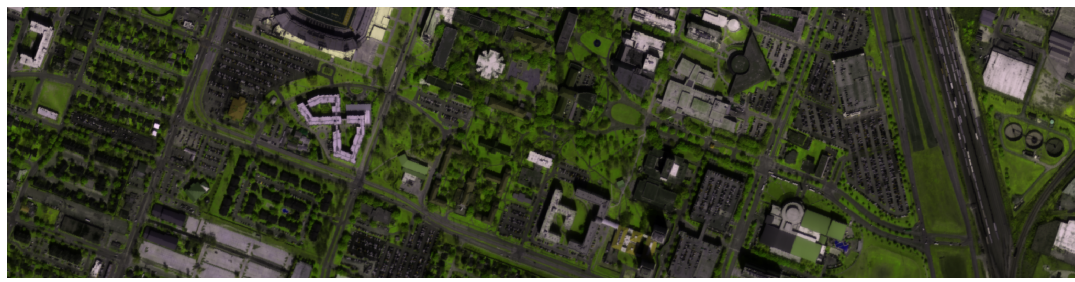
Table 19: Houston dataset: Confusion matrix of Multi-M1+M2 model with E2

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}	c_{14}	c_{15}	c_{16}	c_{17}	c_{18}	c_{19}	c_{20}
c_1	.99	0	0	.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c_2	.06	.9	0	0	.01	0	0	0	0	.01	0	0	.01	0	0	0	0	0	0	0
c_3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c_4	.01	0	0	.97	.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c_5	0	.01	0	0	.94	0	0	.01	.01	.01	0	0	0	0	0	0	0	.01	0	0
c_6	0	0	0	0	0	.99	0	.01	0	0	0	0	0	0	0	0	0	0	0	0
c_7	0	0	0	0	0	0	.99	0	0	0	0	0	0	0	0	0	0	0	0	0
c_8	0	0	0	.02	.06	0	0	.81	.06	.02	0	0	0	0	0	.01	0	0	.01	0
c_9	0	0	0	0	.01	0	0	.07	.75	.04	0	.01	.01	0	0	.02	0	.06	.02	.01
c_{10}	0	.01	0	.01	.02	.01	0	.05	0	.49	0	.1	.21	.04	.01	.03	0	.03	0	0
c_{11}	.02	.1	0	.03	.11	.04	0	.1	0	.23	0	.09	.21	.03	0	.01	0	.03	0	0
c_{12}	0	.01	0	0	0	.01	0	0	0	.1	0	.73	.1	.02	0	.02	0	.01	0	0
c_{13}	0	.03	0	0	.01	.01	0	.01	0	.15	0	.09	.57	.08	0	.01	0	.03	0	0
c_{14}	0	0	0	0	0	0	0	0	0	.02	0	.01	.05	.9	0	0	0	0	.01	0
c_{15}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.99	0	0	0	0	0
c_{16}	0	0	0	0	0	0	0	0	0	.03	0	0	.01	0	0	.94	0	.03	0	0
c_{17}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
c_{18}	0	0	0	0	.01	0	0	0	0	.01	0	.01	0	0	0	.07	0	.9	0	0
c_{19}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.01	.98	0
c_{20}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

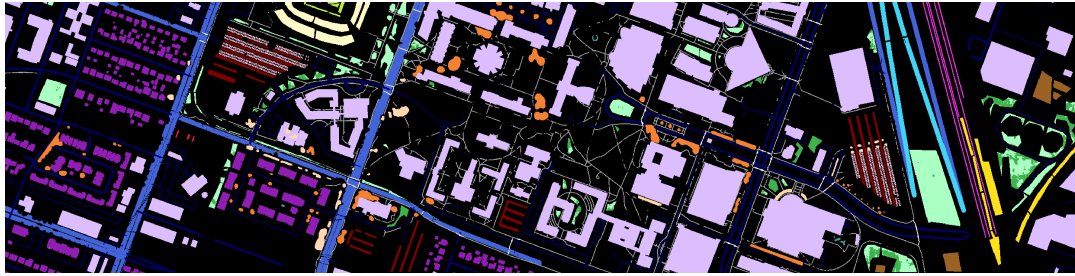
To give more details on the performance for each class we present corresponding confusion matrices in Tables 18 and 19. Each matrix’s rows correspond to the actual conditions or ground truth, while its columns depict the predicted conditions. These matrices have been normalized based on the true conditions (rows) to ensure accurate representation.

Figure 30 offers an overview of the classification maps generated for the entire dataset, encompassing both labeled and unlabeled data points. These maps showcase the results obtained from utilizing the M1+M2 model as well as the Multi-M1+M2 model with E2. These visualizations facilitate the identification of potential patterns, clusters, or misclassifications, which might not be evident solely through quantitative metrics.

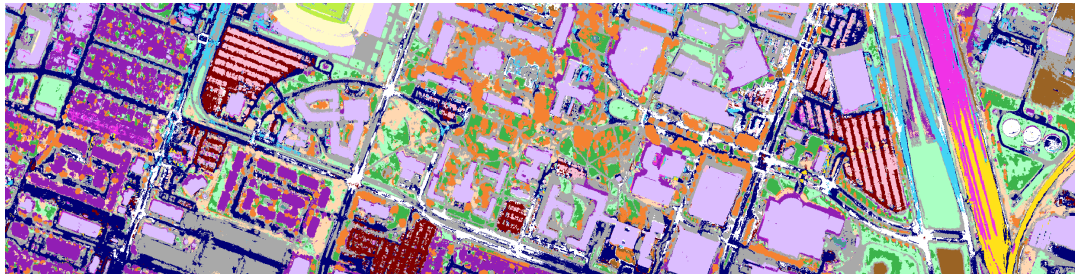
Figure 32, showcases the uncertainty maps for the selected classifications. These maps were generated using the innovative HU approach. The figure offers a visual representation of the uncertainty inherent in the selected classifications. The HU approach provides valuable insights into the level of confidence and the quality of classification as explained earlier (Section 5.1.7). To calculate the Homophily matrix, we used the energy distance as our main measurement. In Figure 34, a heatmap that shows the normalized energy distance H_h on a logarithmic scale is presented. The colors on the heatmap range from blue, which indicates an energy distance closer to 1.0×10^{-17} , to



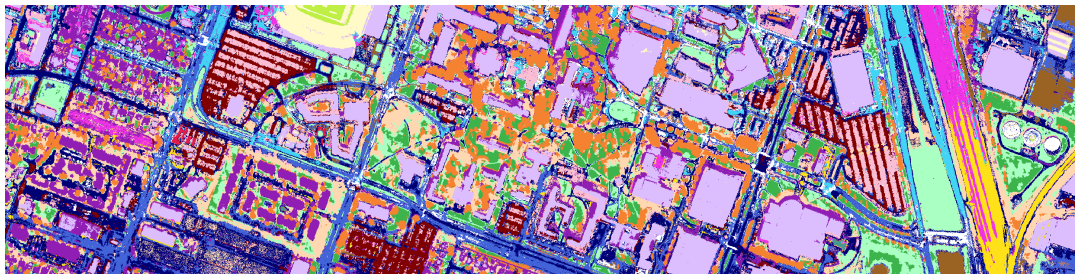
(a)



(b)



(c)



(d)

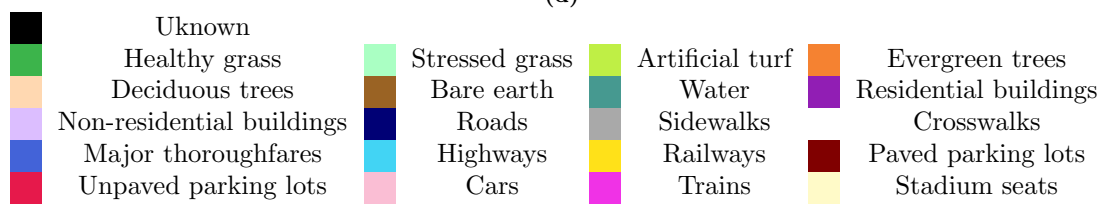


Figure 30: Houston dataset: (a) A false RGB representation using HSI bands, (b) Ground truth labels, Classification maps for models: (c) M1+M2 with E2 on latent size = 30, (d) Multi-M1+M2 with E2 on latent size = 30

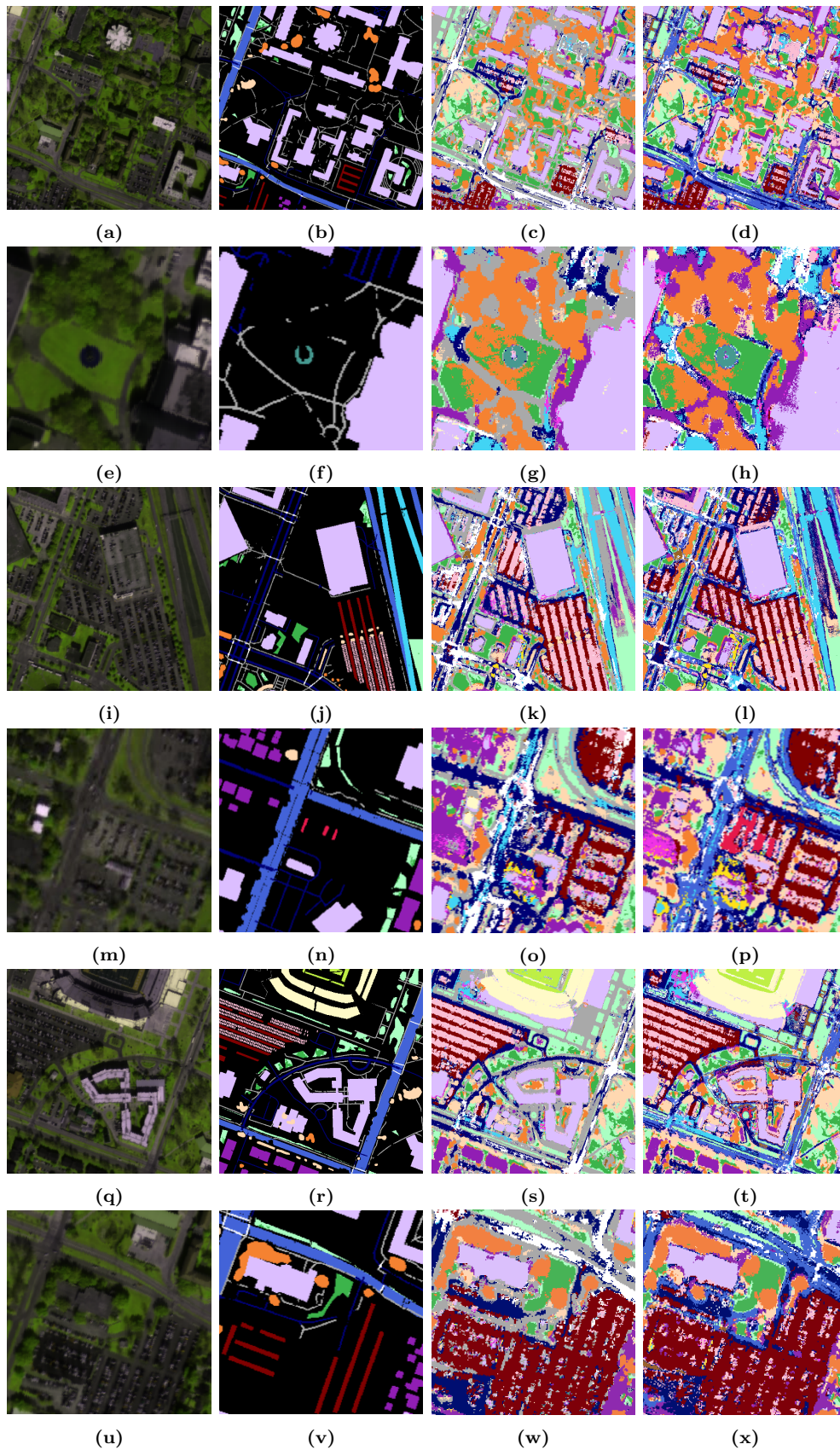


Figure 31: Houston dataset focused areas: False RGB representation (Left), Ground truth labels (Middle left), Classification maps from M1+M2 model with E2 on latent size = 30 (Middle right) and from Multi-M1+M2 model with E2 on latent size = 30 (Right)

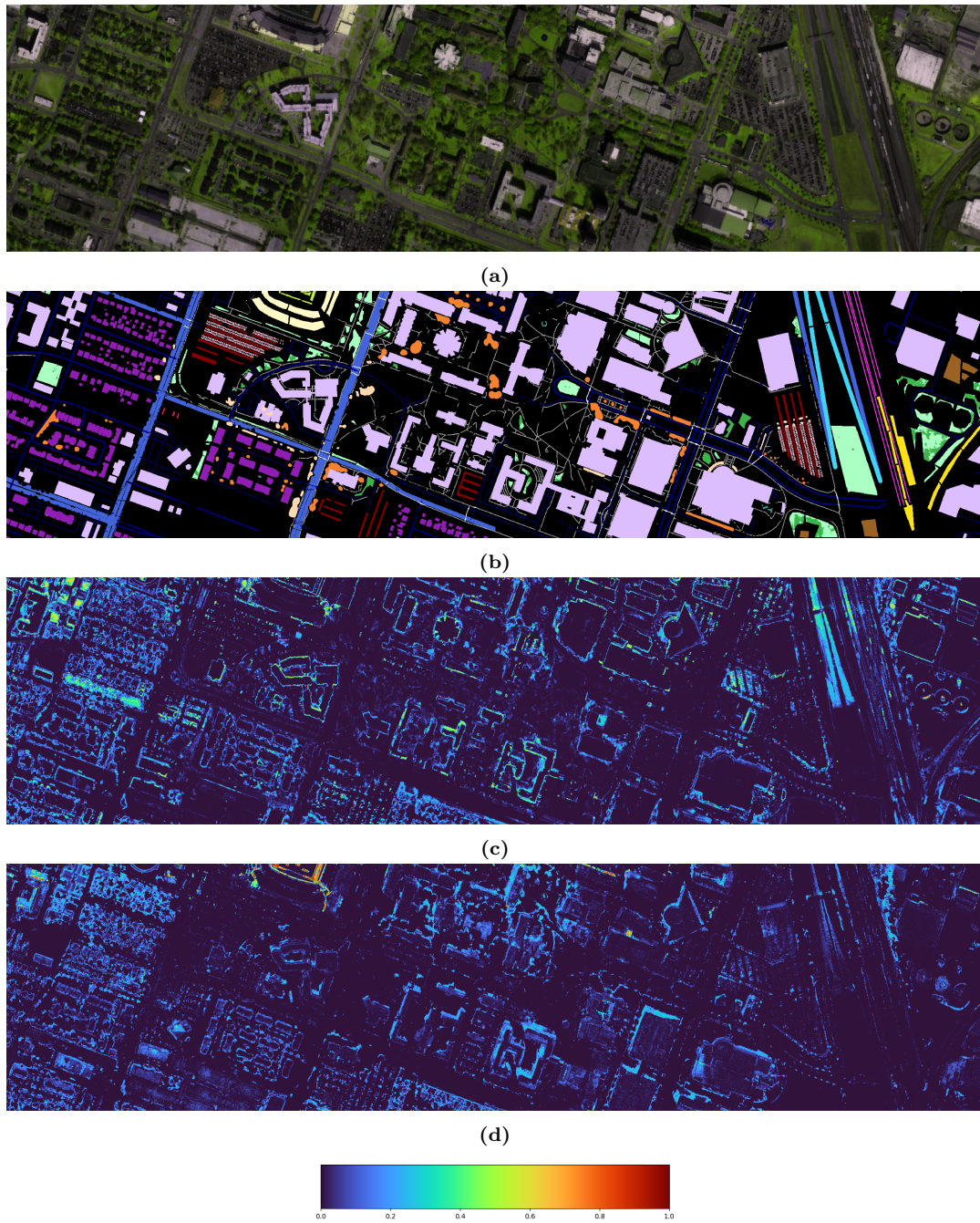


Figure 32: Houston dataset: (a) A false RGB representation using HSI bands, (b) Ground truth labels, Uncertainty maps for models: (c) M1+M2 with E2 on latent size = 30, (d) Multi-M1+M2 with E2 on latent size = 30

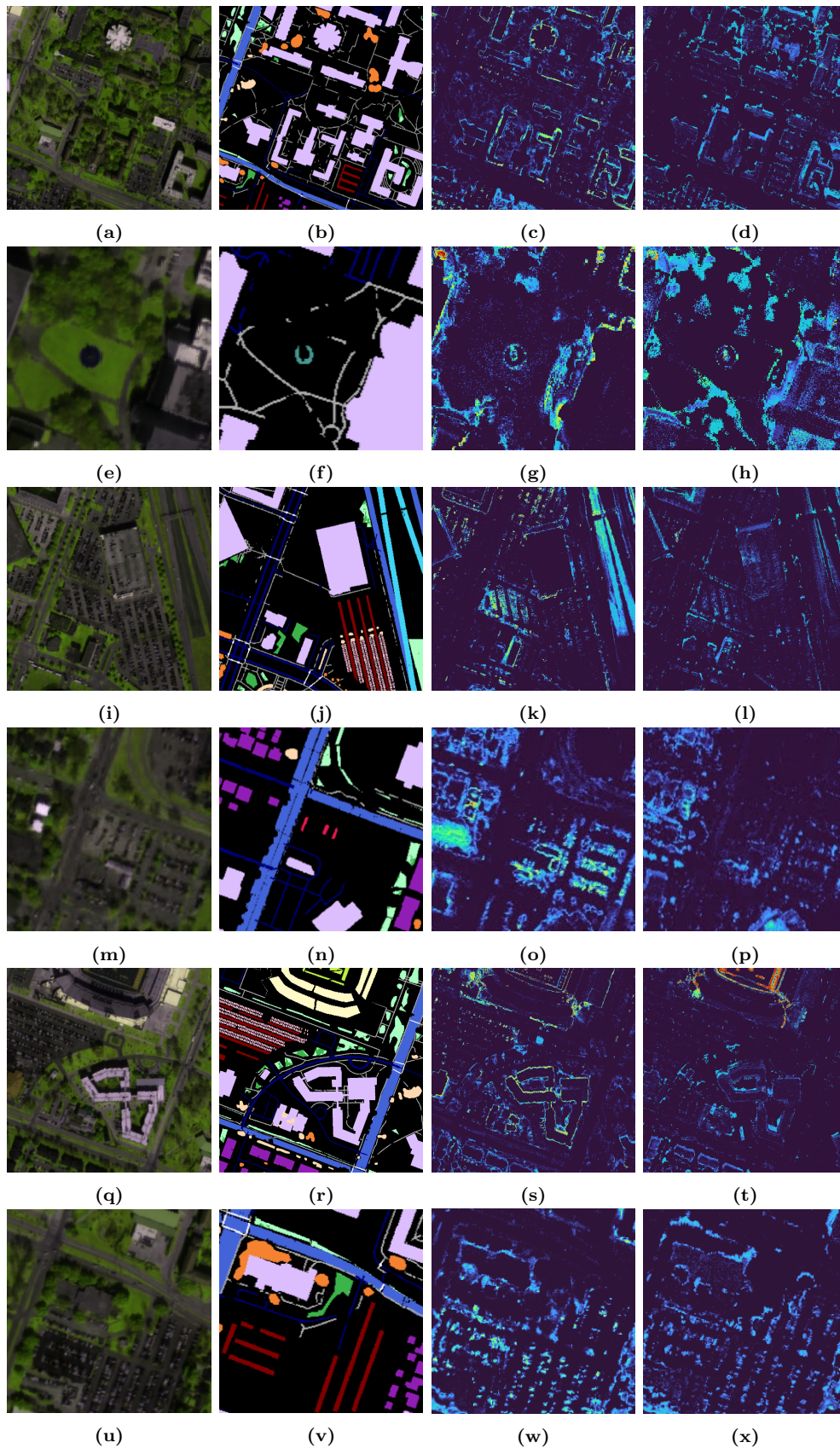


Figure 33: Houston dataset focused areas: False RGB representation (Left), Ground truth labels (Middle left), Uncertainty maps from M1+M2 model with E2 on latent size = 30 (Middle right) and from Multi-M1+M2 model with E2 on latent size = 30 (Right)

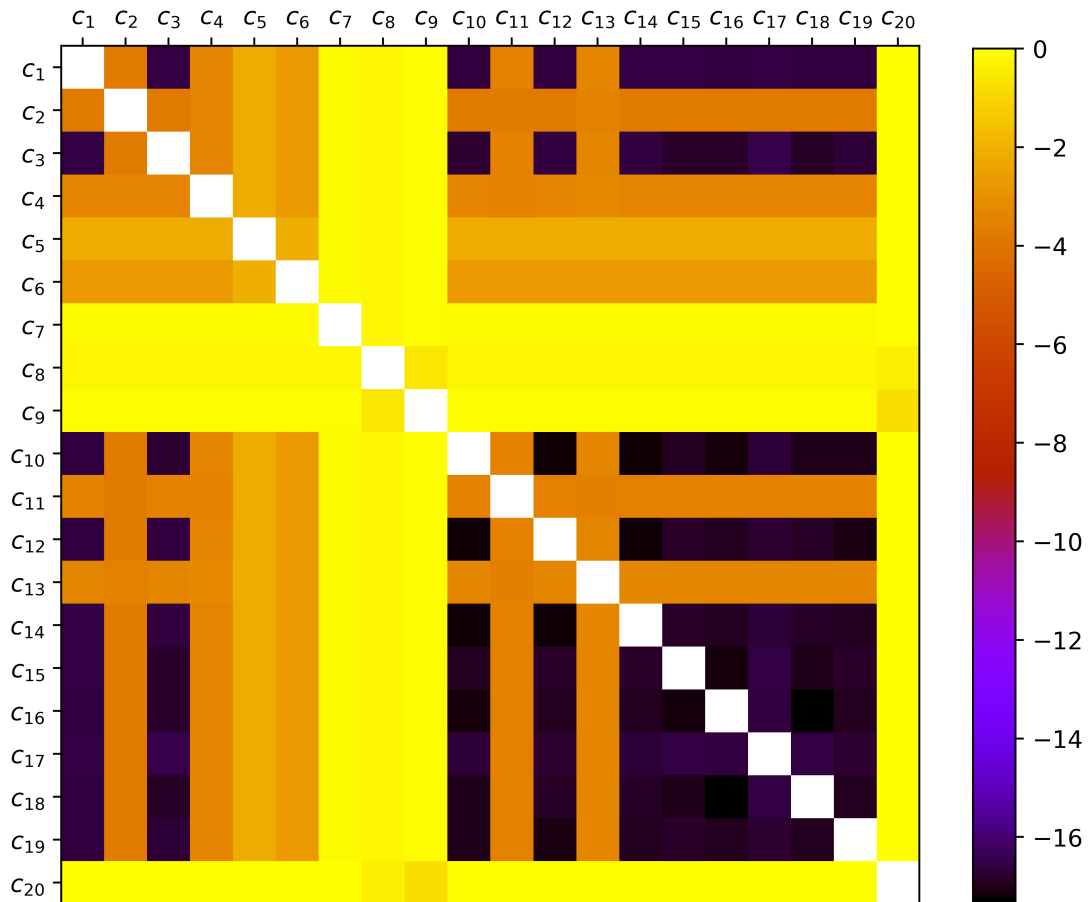


Figure 34: Houston dataset: Heatmap representation of the normalized Homophily matrix H_h in logarithmic scale

yellow, which signifies an energy distance closer to 0. Since the logarithm of zero is not defined, zero values have white color. We chose this representation because it helps us focus on the relative values of class distances, which are more important for our analysis than their absolute values.

Dealing with the complex Houston dataset, characterized by an extensive number of classes, underscores the necessity of conducting a detailed per-class analysis. We want to facilitate a more comprehensive comparison of M1+M2 and Multi-M1+M2 associated with each individual class. Thus, except for the confusion matrices we also focus on some areas of higher interest. Figure 31 showcases these areas' classification maps and Figure 33 the corresponding uncertainties.

5.3.4 Analysis

We base our analysis on the impact of latent feature-level fusion and data-level fusion on the utilization of modalities. Figures 35 and 36 provide a visual representation of mean channel values for both HSI and LiDAR data, specifically focusing on selected classes within the Houston dataset, which will serve as the basis for our subsequent analysis.

The classification of various vegetation classes, including Healthy grass c_1 , Stressed grass c_2 , Artificial turf c_3 , Evergreen trees c_4 , and Deciduous trees c_5 , exhibits a remarkable level of accuracy (Tables 18, 19). Notably, the M1+M2 and Multi-M1+M2 models achieve accuracy rates exceeding 91% on the test set. The Multi-M1+M2 model demonstrates slightly higher accuracy for Evergreen trees c_4 and Deciduous trees c_5 , albeit slightly lower for Stressed grass c_2 . When we narrow our focus to the central Houston area, characterized by abundant vegetation (Figure 31a), it becomes evident that the M1+M2 model occasionally misclassifies certain regions containing Evergreen trees c_4 and Deciduous trees c_5 as Sidewalks c_{11} (Figure 31c). Conversely, the Multi-M1+M2 model exhibits improved differentiation between Evergreen trees c_4 and Deciduous trees c_5 (Figure 31d). These classes display overlapping HSI signatures in the latter portion of their spectral profiles (Figure 35a). By adjusting the relative importance of each modality through latent feature-level fusion, we enhance the separation of these classes, simplifying their discrimination. Both classifiers demonstrate minimal uncertainty for these cases (Figures 33c,33d), signifying limited confusion.

The Unpaved parking lots c_{17} class presents an intriguing case, as the M1+M2 model misclassifies it as Roads c_{10} and Sidewalks c_{11} with low uncertainty, while the Multi-M1+M2 method accurately identifies it successfully with minimal uncertainty (Figures 31o,31p and Figures 33o,33p). Neither the HSI data nor the LiDAR data for

these classes do not overlap significantly (Figures 36c,36d). This indicates why latent feature-level fusion enhanced the ability of the classifier to distinguish the characteristics of these classes.

The Cars c_{18} and Paved parking lots c_{16} classes exhibit fairly accurate detection by both classifiers (Tables 18,19). However the levels of uncertainty for the two classifiers vary. Multi-M1+M2 has significantly lower predictions indicating more confident classification (Figures 33p, 33o) Also, a small percentage of these classes is occasionally confused between them, resulting in pixelated detections. This confusion may arise from the proximity of these classes in labeling, as well as the diminutive size of cars within the image, making them susceptible to human labeling errors. Focusing further in certain unlabeled regions (Figures 31w,31x) we can see that Multi-M1+M2 successfully detects the Cars c_{18} in contrast with M1+M2 which detects Sidewalks c_{11} . These areas are not included in the labels so the efficacy of Multi-M1+M2 is not reflected in its accuracy (Table 19).

Lastly, both classifiers correctly detect the limited Stadium seats c_{20} regions. The M1+M2 model occasionally identifies some Sidewalks c_{11} inside the stadium (Figures 31s,31t). Intuitively, areas with Stadium seats c_{20} are more likely to be present than Sidewalks c_{11} within a stadium. Multi-M1+M2 manages to detect Stadium seats c_{20} in the stadium but with higher uncertainty (Figure 33t), indicating potential confusion with distant classes in this area. It's important to mention that also Stadium seats c_{20} has a substantial energy distance from most of the other classes (Equation 34) which means that confusion with most of classes will result in higher uncertainty.

Both classifiers excel in effectively detecting Bare earth c_6 (Tables 18,19), a common strength shared across both models, even extending to areas beyond labeled regions (Figure 30). This is an example of a class that either selection of encoding fusion can work effectively.

In the context of identifying the limited Water c_7 regions, all tested classifiers perform adeptly (Tables 18,19). However, it is noteworthy that the M1+M2 model exhibits substantially higher uncertainty in proximity to Water c_7 areas (Figures 33g, 33h). This discrepancy can be attributed to M1+M2 potentially misclassifying a greater number of classes as Water c_7 , plus due to a significant energy distance between the Water c_7 class and others (Equation 34).

The asphalt classes, encompassing Roads c_{10} , Sidewalks c_{11} , Crosswalks c_{12} , Major thoroughfares c_{13} , and Highways c_{14} , present a challenge due to varying sizes, city locations, and colors. Both architectures show high confusion between these classes.

M1+M2 assigns big part of Roads c_{10} as Crosswalks c_{12} while Multi-M1+M2 mixes Roads c_{10} with Major thoroughfares c_{13} and both classifiers detect a part of Major thoroughfares c_{13} as Highways c_{14} (Figures 31k,31l, 31o, 31p). Also, Multi-M1+M2 misses all of the Sidewalks c_{11} and M1+M2 all of the Major thoroughfares c_{13} . By inspecting the HSI and LiDAR of these classes (Figures 35c, 35d) we see that this is happening because all modalities highly overlap making this kind of data very challenging for classification. While both classifiers exhibit relatively low uncertainties (Figures 33k,33l, 33o, 33p), even for misclassifications. Which shows that the complexities of these classes even when comes to uncertainty remain apparent since classifiers are falsely confident. On the bright side we can note that these classes are confused mainly between them and not with other more irrelevant classes which also defies the low levels of uncertainty.

All classifiers excel in identifying labeled Railways c_{15} , with both M1+M2 and Multi-M1+M2 achieving detection accuracy exceeding 97% (Tables 18,19). But, a closer examination of classification maps (Figures 31k,31l) reveals some illogical Trains c_{19} placements in the urban center, particularly pronounced with the Multi-M1+M2 model. These defections are, in fact, areas with parked cars, as evident from the false RGB image of Houston. This possibly occurred due to the fact that cars and trains are man made metallic objects thus they share some similar characteristics in the spectral and LiDAR signatures (Figures 36a,36b) so this might occur to mixing them in some areas.

Both classifiers demonstrate satisfactory detection of buildings. Notably, the M1+M2 model outperforms in terms of accuracy the Multi-M1+M2 model in both building classes (Tables 18,19). Both models occasionally misclassify some Residential buildings c_8 as Non-residential buildings c_9 , and vice versa (Table 19). This phenomenon can be attributed to the amount of overlap in the LiDAR modality for these two classes (Figure 35f). Data-level fusion places more emphasis on the considerably larger HSI modality, giving the LiDAR channels a smaller role in the classification decision. As a result, confusion between these classes increases with latent feature-level fusion. Notably, both Residential buildings c_8 and Non-residential buildings c_9 classes exhibit a significant energy distance from other classes (Equation 34). Therefore, it is reassuring to observe that Multi-M1+M2 has lower levels of uncertainty compared to M1+M2 in the Residential buildings c_8 and Non-residential buildings c_9 area (Figures 33c 33d ,33s, 33t). This indicates that Multi-M1+M2 confuses between smaller amount or closer classes than M1+M2.

In terms of metrics, M1+M2 outperforms Multi-M1+M2 except for recall (Table

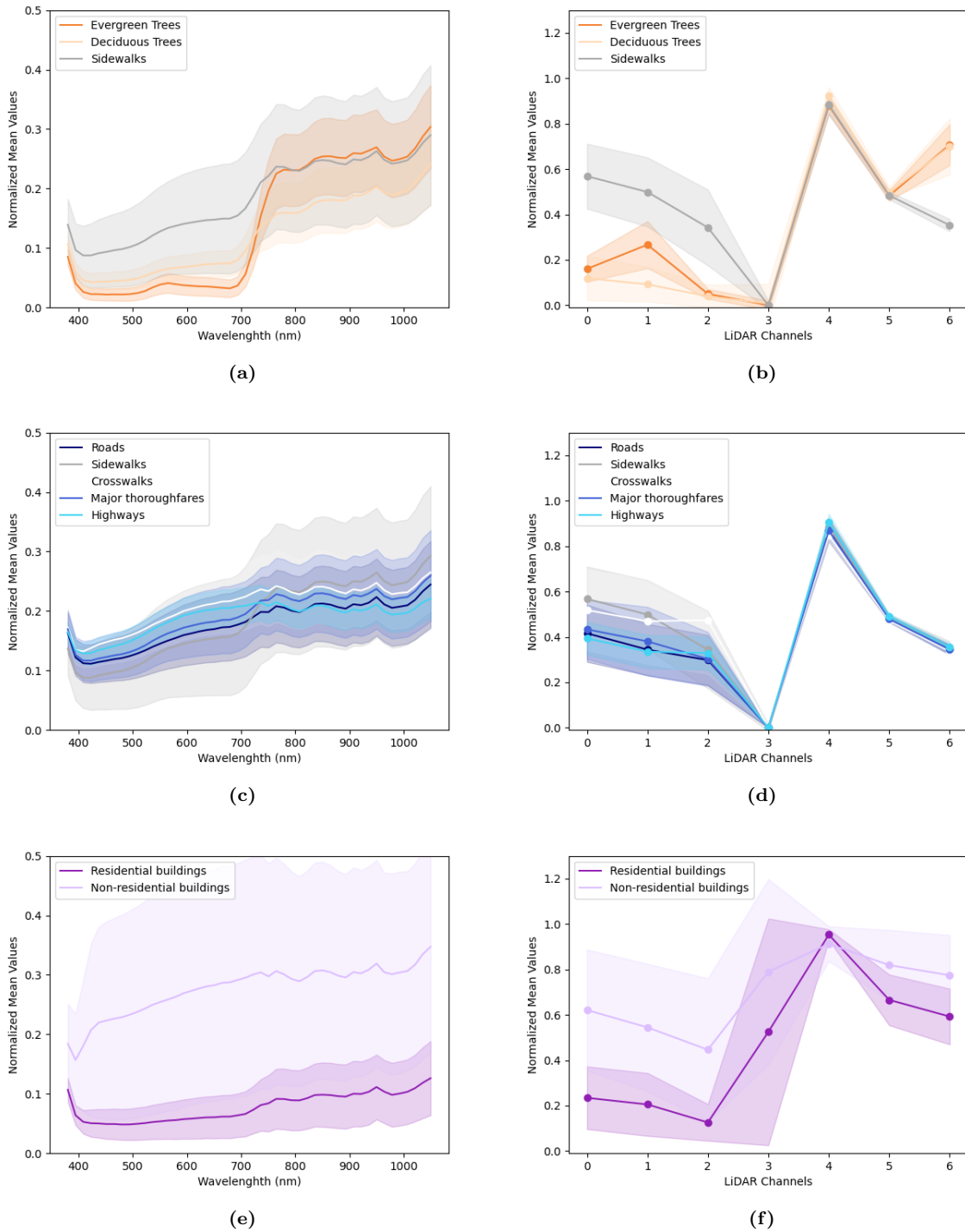


Figure 35: Houston dataset: HSI spectrum (Left) and LiDAR channels (Right) for selected classes

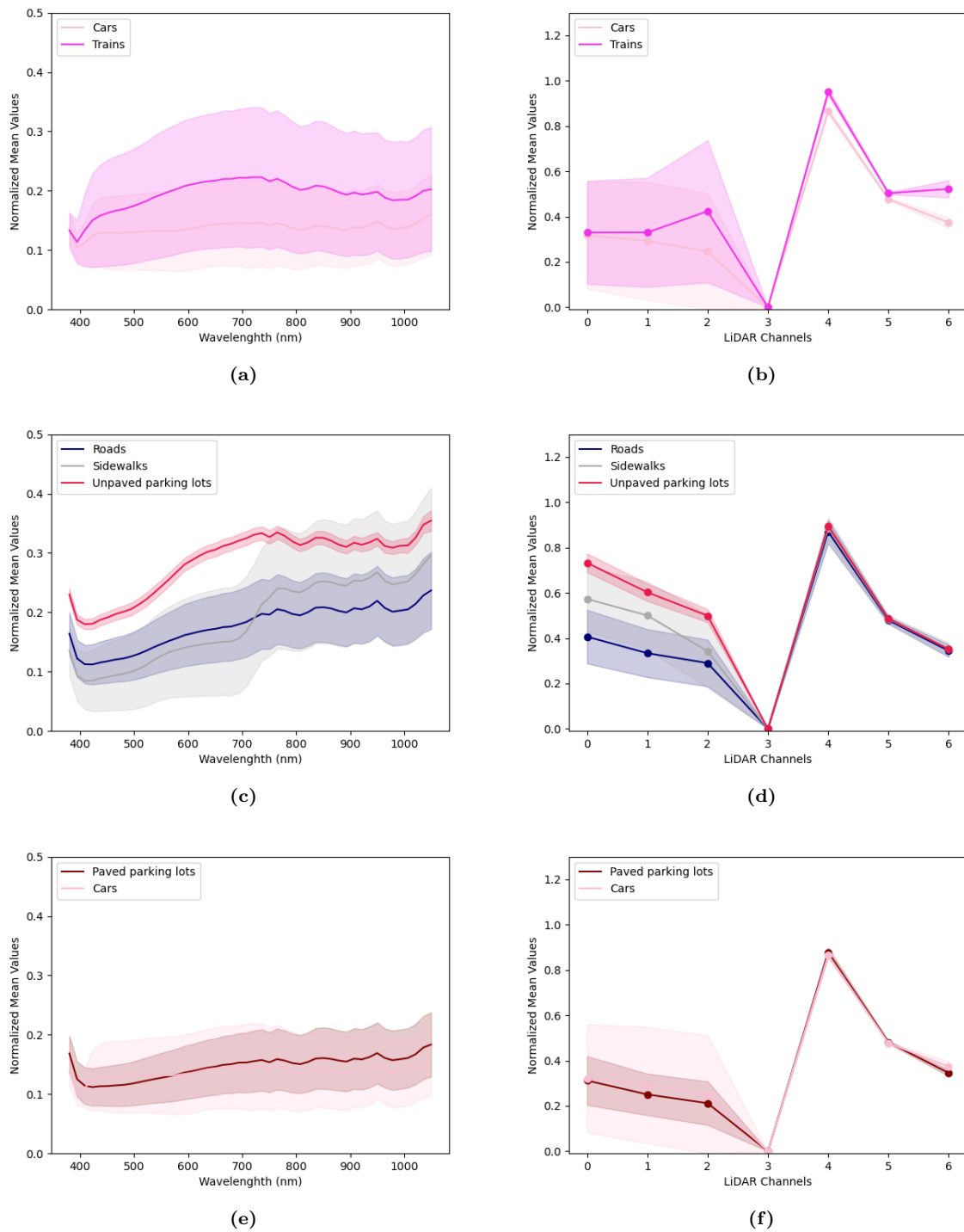


Figure 36: Houston dataset: HSI spectrum (Left) and LiDAR channels (Right) for selected classes

17). Lower accuracy and precision may be attributed to the fact that Multi-M1+M2 faced more challenges in distinguishing between the various asphalt classes in the labeled dataset compared to M1+M2. Additionally, M1+M2 struggled to detect both the Unpaved parking lots c_{17} and Major thoroughfares c_{13} classes (Table 18), whereas Multi-M1+M2 only had issues with detecting Sidewalks c_{11} (Table 19). Especially when dealing with a complex and extensive dataset, it's not advisable to solely rely on metrics for drawing conclusions. Instead, a more comprehensive and detailed analysis of the entire dataset's area is necessary.

Our findings underscored the effectiveness and limitations of both fusion approaches in certain contexts. The Multi-M1+M2 model proved to be a valuable tool for enhancing class separation, particularly among vegetation classes. This approach significantly reduced confusion and minimized uncertainty, even among these closely related classes. Multi-M1+M2 showcases enhanced capabilities in distinguishing complex classes like parking lots, thanks to its latent feature-level fusion approach, which leverages the strengths of both HSI and LiDAR data. In some cases, particularly with the detection of stadium seating and buildings, Multi-M1+M2 exhibits higher uncertainty, potentially due to confusion with high energy distance classes. However, in general it seems more confident about its detections than M1+M2. The M1+M2 model demonstrated its strength in consistently separating between buildings. However, this fusion method exhibited higher uncertainty in proximity to building and water areas, potentially stemming from confusion between more high energy distance classes. Asphalt classes remained a challenge for both models, revealing the complexities of these classes, yet highlighting the classifiers' confidence even when making erroneous judgments. Lastly, parking lots and cars were accurately distinguished by both models, but occasionally exhibited confusion between the two, largely due to their close proximity and small car sizes.

The choice between these models depends on the specific classification task and the degree of class complexity and the quality of the data, with each model offering unique advantages and trade-offs.

5.4 Discussion

Our experiments and analysis provide valuable insights into the strengths and limitations of VAEs when combining features and data for pixel-wise classification in rural and urban areas. This offers guidance for improving future multimodal classification techniques and

applications.

The classifiers performed better, both in terms of numbers and the overall quality of results, when classifying the Trento dataset compared to the Houston dataset. The Trento dataset represents rural environment mainly characterized by large vegetation areas. In contrast, the Houston dataset presents a more diverse landscape, featuring an extensive mix of earth, vegetation and human-made structures. Additionally, the urban nature of the Houston dataset results in finer details within its classes. Furthermore, the LiDAR data in the Trento dataset consists solely of DSM channels, while the Houston dataset includes both DEM and DSM data and multispectral LiDAR. It is probable that these additional LiDAR channels do not significantly contribute to data separation, especially when considering the Multi-M1+M2 scenario, where they have a notable impact on feature generation.

Despite achieving higher metrics for Trento, we decided not to use spatial information for pixel classification in Houston. This decision was made to maintain sharper results. We anticipate that patching would have an even more detrimental effect in Houston, given that some classes are as narrow as a single pixel.

6 Conclusion & future work

6.1 Conclusion

In this thesis, we have addressed the gap in the existing literature by introducing a novel framework for multimodal data classification using a semi-supervised VAE. We began by reviewing the landscape of semi-supervised VAE architectures and their applications in classification scenarios, highlighting the absence of a tailored adaptation for multimodal data. Our primary objective was to develop a model capable of accommodating diverse modalities within the remote sensing domain.

Through the formulation of our problem and the development of the Multi-M1+M2 architecture, we have demonstrated the potential of our approach in handling multimodal data. By extending the M1+M2 framework to incorporate two levels of fusion – data-level and latent feature-level – we have achieved a comprehensive diverse integration of modalities. The data-level fusion method involved concatenating multiple modalities into a single input, while the latent feature-level fusion method introduced separate latent variables for each modality. These strategies not only allowed us to harness the advantages of a generative model but also effectively manage some of the challenges posed by multimodal data.

We conducted an extensive analysis of rural and urban area classification using various encoders on two distinct datasets: Trento and Houston. Our results and analysis shed light on the strengths and limitations of different models, latent space sizes, and fusion strategies.

From the experiments conducted on the Trento dataset, we observed that the choice of encoder and latent size significantly affects classification performance. While increasing the latent size does not always guarantee improved performance. Furthermore, our experimentation with encoders also revealed a noteworthy observation: while leveraging spatial information in our inputs yielded enhanced metrics, it resulted in less distinct classifications, an undesirable outcome. The Multi-M1+M2 model incorporating latent feature-level fusion demonstrated improved separation between closely related vegetation classes. The visual analysis of classification maps provided deeper insights into model behavior, highlighting areas of confusion and misclassifications. Additionally, the uncertainty maps facilitated understanding the model’s confidence in its classifications and identifying regions of potential confusion. A more in-depth comparison between

SVM and RF models highlights the superior classification potential of deep learning techniques but that they tend to typically exhibit suboptimal time efficiency.

For the Houston dataset, our investigation into the impact of latent size on classification performance revealed an optimal latent size effectively balanced the model's ability to capture underlying data complexities. The M1+M2 and Multi-M1+M2 model exhibited strong performance in detecting most urban classes. Although, we understood, that in some cases, the balancing of modalities with latent feature-level fusion can lead to worse results. The integration of HSI and LiDAR data through latent feature-level fusion provided notable improvements in distinguishing between classes with distinct spectral and LiDAR signatures. However, for classes with highly analogous HSI or LiDAR properties, such as different types of roads, the fusion methods were less effective. The visual analysis of classification and uncertainty maps aided in comprehending the models' strengths, as well as highlighting areas where different classifiers struggled.

In conclusion, our research has laid a strong foundation for the classification of remote sensing multimodal data using a semi-supervised VAE approach. By integrating quantitative metrics and qualitative analysis, we gained a comprehensive understanding that elevated quantitative measures do not invariably align with superior qualitative performance. Our experimental results proved the effectiveness of the proposed Multi-M1+M2 model's architecture and highlighted the strengths and weaknesses of our approach in the robustness of classification tasks involving multimodal data. As we continue to delve into the intricate landscape of multimodal data analysis, the proposed framework opens the door to innovative solutions and impact applications within the field of remote sensing and beyond.

6.2 Future work

While our research has made a stride in the realm of multimodal data classification using VAE, there are several avenues for further exploration and refinement. The following areas present opportunities for future work:

- **Balanced loss function** : In the context of multimodal data we can use a balanced loss function to achieve a desirable distinction in different data modalities during learning. Ensuring equitable weightage to each source of information.
- **Attention is all you need**: The attention mechanism is a technique that allows the model to focus on different parts of the input data when making predictions.

The self-attention mechanism, in particular, is an improvement of the attention mechanism that is better at capturing internal correlations within the data or features (Vaswani et al., 2017). By incorporating an attention mechanism into the VAE, the model might be able to assign different weights to different modalities based on their relevance to the classification task.

- Training performance: One of the limitations of Bayesian deep learning models is computational complexity. One approach to mitigate these issues is variational dropout, which is a generalization of Gaussian dropout where the dropout rates are learned (Kingma et al., 2015). Dimensionality reduction techniques can also help reduce the computational complexity of Bayesian deep learning methods (Jospin et al., 2022).
- Improvement on the reconstruction and generation performance of VAEs: An approach to further enhance the semi-supervised VAE is to find ways to have a better reconstruction and generation performance. Some approaches are to combine the strengths of both GANs and VAEs (Zemouri, 2020) and the use of dilated convolutions in the decoder part (Yang et al., 2017), leading to improved performance and can potentially benefit our semi-supervised VAE models.
- Additional modalities and scalability: Our experiments were conducted in datasets with two modalities. Extending it to accommodate more modalities could enhance its versatility and applicability to a broader range of tasks. Investigating the scalability of the proposed architecture to handle an arbitrary number of modalities is an intriguing direction for future research. Although in the sector of remote sensing it is challenging to find a dataset that incorporates more than two aligned modalities.
- Enhancing data exploration: Applying the proposed architecture to more real-world applications within the remote sensing domain, such as sea-ice classification or anomaly detection, would demonstrate the practical utility of our model and its potential for addressing complex challenges in remote sensing data analysis. Broadening the scope of this study not only could involve gathering and evaluating the suggested architectures across a wider array of remote sensing data types, such as SAR or Polarimetry, but also could include temporal data, such as time series of satellite images, in order enable tracking changes and patterns in areas over time.

- Other domains: Our versatile approach can be adapted and applied to various multimodal data types, such as medical imaging, natural language processing, and more, to tackle complex classification challenges across different fields.

Incorporating these directions into future research could enhance the performance of VAE which will lead to more accurate and robust classification models contributing to better urban planning, resource management, and environmental monitoring.

7 Publications

The research I conducted in Earth Observation Laboratory of the Arctic University of Norway (UiT) has led to the following publications:

- Chlaily, S., Ratha, D., Lozou, P. and Marinoni A. (2023). On measures of uncertainty in classification, *IEEE Transactions on Signal Processing* **71**:3710-3725.
- Khachatryan, E., Sandalyuk, N., Lozou, P. (2023). Eddy Detection in the Marginal Ice Zone with Sentinel-1 Data Using YOLOv5, *Remote Sensing* **15**:2244.

8 References

References

- Adao, T., Hruvska, J., Padua, L., Bessa, J., Peres, E., Morais, R. and Sousa, J. J. (2017). Hyperspectral imaging: A review on uav-based sensors, data processing and applications for agriculture and forestry, *Remote Sensing* **9**: 1110.
- Alain, G. and Bengio, Y. (2014). What regularized auto-encoders learn from the data-generating distribution, *The Journal of Machine Learning Research* **15**: 3563–3593.
- Alon, I., Qi, M. and Sadowski, R. J. (2001). Forecasting aggregate retail sales:: a comparison of artificial neural networks and traditional methods, *Journal of Retailing and Consumer Services* **8**: 147–156.
- Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M. and Farhan, L. (2021). Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions, *Journal of Big Data* **8**: 1–74.
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mané, D. (2016). Concrete problems in ai safety, *arXiv preprint arXiv:1606.06565* .
- An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability, *Special Lecture on IE* **2**: 1–18.
- Arun, P., Sadeh, R., Avneri, A., Tubul, Y., Camino, C., Buddhiraju, K. M., Porwal, A., Lati, R. N., Zarco-Tejada, P. J., Peleg, Z. et al. (2022). Multimodal earth observation data fusion: Graph-based approach in shared latent space, *Information Fusion* **78**: 20–39.
- Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures, *International Conference on Machine Learning (ICML) Workshop on Unsupervised and Transfer Learning*.

- Ball, J. E., Anderson, D. T. and Chan, C. S. (2017). Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community, *Journal of Applied Remote Sensing* **11**: 042609–042609.
- Bhar, R. and Hamori, S. (2004). *Hidden Markov models: Applications to financial economics*, Springer Science & Business Media.
- Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations, *Sankhya: The Indian Journal of Statistics (1933-1960)* **7**: 401–406.
- Biau, G. and Scornet, E. (2016). A random forest guided tour, *Test* **25**: 197–227.
- Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, Springer.
- Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: A review for statisticians, *Journal of the American Statistical Association* **112**: 859–877.
- Boerner, W. (2003). Recent advances in extra-wide-band polarimetry, interferometry and polarimetric interferometry in synthetic aperture remote sensing and its applications, *IEE Proceedings-Radar, Sonar and Navigation* **150**: 113–124.
- Borsboom, D. (2008). Latent variable theory, *Measurement: Interdisciplinary Research and Perspectives* **6**: 25–53.
- Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery* **2**: 121–167.
- Burget, L., Schwarz, P., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A., Glembek, O., Goel, N., Karafiat, M., Povey, D. et al. (2010). Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Cai, L. (2012). Latent variable modeling, *Shanghai archives of psychiatry* **24**: 118.
- Campbell, J. B. and Wynne, R. H. (2011). *Introduction to remote sensing*, Guilford Press.
- Camps-Valls, G. (2009). Machine learning in remote sensing data processing, *IEEE International Workshop on Machine Learning for Signal Processing*.

- Cenggoro, T. W., Isa, S. M., Kusuma, G. P. and Pardamean, B. (2017). Classification of imbalanced land-use/land-cover data using variational semi-supervised learning, *International Conference on Innovative and Creative Information Technology (ICITech)*.
- Chalvatzaki, G. G., Pavlakos, G., Maninis, K., Papageorgiou, X. S., Pitsikalis, V., Tzafestas, C. S. and Maragos, P. (2014). Towards an intelligent robotic walker for assisted living using multimodal sensorial data, *International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*.
- Chan, Y. K. and Koo, V. (2008). An introduction to synthetic aperture radar (sar), *Progress In Electromagnetics Research* **2**: 27–60.
- Chen, X. and Ishwaran, H. (2012). Random forests for genomic data analysis, *Genomics* **99**: 323–329.
- Cheng, M., Jiao, X., Shi, L., Penuelas, J., Kumar, L., Nie, C., Wu, T., Liu, K., Wu, W. and Jin, X. (2022). High-resolution crop yield and water productivity dataset generated using random forest and remote sensing, *Scientific Data* **9**: 641.
- Chien, S. A., Davies, A. G., Doubleday, J., Tran, D. Q., McLaren, D., Chi, W. and Maillard, A. (2020). Automated volcano monitoring using multiple space and ground sensors, *Journal of Aerospace Information Systems* **17**: 214–228.
- Chlaily, S., Ratha, D., Lozou, P. and Marinoni, A. (2023). On measures of uncertainty in classification, *IEEE Transactions on Signal Processing* **71**: 3710–3725.
- Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**: 37–46.
- Coltekin, C. and Rama, T. (2018). Tubingen-oslo at semeval-2018 task 2: Svms perform better than rnns in emoji prediction, *International Workshop on Semantic Evaluation*.
- Connors, C. and Vatsavai, R. R. (2017). Semi-supervised deep generative models for change detection in very high resolution imagery, *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*.
- Cracknell, A. P. (2007). *Introduction to remote sensing*, CRC Press.

- Dickey, T., Lewis, M. and Chang, G. (2006). Optical oceanography: recent advances and future directions using global remote sensing and in situ observations, *Reviews of Geophysics* **44**.
- Dimakis, A. G. (2022). *Deep Generative Models and Inverse Problems*, Cambridge University Press.
- Dingle Robertson, L., Davidson, A., McNairn, H., Hosseini, M., Mitchell, S., De Abellera, D., Veron, S. and Cosh, M. H. (2020). Synthetic aperture radar (sar) image processing for operational space-based agriculture mapping, *International Journal of Remote Sensing* **41**: 7112–7144.
- Dong, P. and Chen, Q. (2017). *LiDAR remote sensing and applications*, CRC Press.
- Dong, S., Wang, P. and Abbas, K. (2021). A survey on deep learning and its applications, *Computer Science Review* **40**: 100379.
- Emerling, J. (2013). *Photography: History and theory*, Routledge.
- Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W. and Dietmayer, K. (2020). Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges, *IEEE Transactions on Intelligent Transportation Systems* **22**: 1341–1360.
- Fletcher, R. J. (1990). Military radar defence lines of northern north america: an historical geography, *Polar Record* **26**: 265–276.
- Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J. and Greenspan, H. (2018). Synthetic data augmentation using gan for improved liver lesion classification, *IEEE International Symposium on Biomedical Imaging (ISBI)*.
- Fukao, S., Hamazu, K. and Doviak, R. J. (2014). *Radar for meteorological and atmospheric observations*, Springer.
- Ghamisi, P., Höfle, B. and Zhu, X. X. (2016). Hyperspectral and lidar data fusion using extinction profiles and deep convolutional neural network, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **10**: 3011–3024.
- Giannakos, M. N., Sharma, K., Pappas, I. O., Kostakos, V. and Velloso, E. (2019). Multimodal data as a means to understand the learning experience, *International Journal of Information Management* **48**: 108–119.

- Goceri, E. (2019). Analysis of deep networks with residual blocks and different activation functions: classification of skin diseases, *International Conference on Image Processing Theory, Tools and Applications (IPTA)*.
- Gómez-Chova, L., Tuia, D., Moser, G. and Camps-Valls, G. (2015). Multimodal classification of remote sensing images: A review and future directions, *Proceedings of the IEEE* **103**: 1560–1584.
- Goncalves, H., Corte-Real, L. and Goncalves, J. A. (2011). Automatic image registration through image segmentation and sift, *IEEE Transactions on Geoscience and Remote Sensing* **49**: 2589–2600.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2020). Generative adversarial networks, *Communications of the ACM* **63**: 139–144.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. and Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone, *Remote Sensing of Environment* **202**: 18–27.
- Goward, S. N. and Williams, D. L. (1997). Landsat and earth systems science: development of terrestrial monitoring, *Photogrammetric Engineering and Remote Sensing* **63**: 887–900.
- Hand, D. (2012). Assessing the performance of classification methods, *International Statistical Review* **80**: 400–414.
- Hang, R., Li, Z., Ghamisi, P., Hong, D., Xia, G. and Liu, Q. (2020). Classification of hyperspectral and lidar data using coupled cnns, *IEEE Transactions on Geoscience and Remote Sensing* **58**: 4939–4950.
- Harefa, E. and Zhou, W. (2021). Performing sequential forward selection and variational autoencoder techniques in soil classification based on laser-induced breakdown spectroscopy, *Analytical Methods* **13**: 4926–4933.
- Harshvardhan, G., Gourisaria, M. K., Pandey, M. and Rautaray, S. S. (2020). A comprehensive survey and analysis of generative models in machine learning, *Computer Science Review* **38**: 100285.

- Heiden, U., Heldens, W., Roessner, S., Segl, K., Esch, T. and Mueller, A. (2012). Urban structure type characterization using hyperspectral remote sensing and height information, *Landscape and urban Planning* **105**: 361–375.
- Henrique, B. M., Sobreiro, V. A. and Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction, *Expert Systems with Applications* **124**: 226–251.
- Hong, D., Hu, J., Yao, J., Chanussot, J. and Zhu, X. X. (2021). Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model, *ISPRS Journal of Photogrammetry and Remote Sensing* **178**: 68–80.
- Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations, *International Journal of Data Mining & Knowledge Management Process* **5**: 1.
- Huang, C., Chen, Y., Zhang, S. and Wu, J. (2018). Detecting, extracting, and monitoring surface water from space using optical sensors: A review, *Reviews of Geophysics* **56**: 333–360.
- Huang, G.-B., Zhou, H., Ding, X. and Zhang, R. (2012). Extreme learning machine for regression and multiclass classification, *IEEE Transactions on Systems Man and Cybernetics* **42**: 513–529.
- Huot, F., Hu, R. L., Goyal, N., Sankar, T., Ihme, M. and Chen, Y.-F. (2022). Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data, *IEEE Transactions on Geoscience and Remote Sensing* **60**: 1–13.
- Jebara, T. (2012). *Machine learning: discriminative and generative*, Springer Science & Business Media.
- Jensen, J. R. and Cowen, D. C. (1999). Remote sensing of urban/suburban infrastructure and socio-economic attributes, *Photogrammetric Engineering and Remote Sensing* **65**: 611–622.
- Jiang, S., Pang, G., Wu, M. and Kuang, L. (2012). An improved k-nearest-neighbor algorithm for text categorization, *Expert Systems with Applications* **39**: 1503–1509.

- Jolliffe, I. T. (1990). Principal component analysis: a beginner's guide - i. introduction and application, *Weather* **45**: 375–382.
- Jospin, L. V., Laga, H., Boussaid, F., Buntine, W. and Bennamoun, M. (2022). Hands-on bayesian neural networks—a tutorial for deep learning users, *IEEE Computational Intelligence Magazine* **17**: 29–48.
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. and Aila, T. (2020). Analyzing and improving the image quality of stylegan, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Khachatryan, E., Sandalyuk, N. and Lozou, P. (2023). Eddy detection in the marginal ice zone with sentinel-1 data using yolov5, *Remote Sensing* **15**: 2244.
- Khaleghian, S., Kramer, T., Everett, A., Kiarbech, A., Hughes, N., Eltoft, T. and Marinoni, A. (2021). Synthetic aperture radar data analysis by deep learning for automatic sea ice classification, *European Conference on Synthetic Aperture Radar (EUSAR)*, pp. 1–6.
- Khanal, S., Fulton, J. and Shearer, S. (2017). An overview of current and potential applications of thermal remote sensing in precision agriculture, *Computers and Electronics in Agriculture* **139**: 22–32.
- Khelifi, L. and Mignotte, M. (2020). Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis, *IEEE Access* **8**: 126385–126400.
- Kingma, D. P., Mohamed, S., Jimenez Rezende, D. and Welling, M. (2014). Semi-supervised learning with deep generative models, *Advances in Neural Information Processing Systems* **27**: 3581–3589.
- Kingma, D. P., Salimans, T. and Welling, M. (2015). Variational dropout and the local reparameterization trick, *Conference on Neural Information Processing Systems (NIPS)*.
- Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes, *arXiv preprint arXiv:1312.6114* .
- Kingma, D. P., Welling, M. et al. (2019). An introduction to variational autoencoders, *Foundations and Trends in Machine Learning* **12**: 307–392.

- Kolev, N. (2011). *Sonar systems*, Books on Demand.
- Krogh, A., Brown, M., Mian, I. S., Sjölander, K. and Haussler, D. (1994). Hidden markov models in computational biology: Applications to protein modeling, *Journal of Molecular Biology* **235**: 1501–1531.
- Krohn, J., Beyleveld, G. and Bassens, A. (2019). *Deep Learning Illustrated*, Addison-Wesley Professional.
- Kuenzer, C., Ottinger, M., Wegmann, M., Guo, H., Wang, C., Zhang, J., Dech, S. and Wikelski, M. (2014). Earth observation satellite sensors for biodiversity monitoring: potentials and bottlenecks, *International Journal of Remote Sensing* **35**: 6599–6647.
- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *The Annals of Mathematical Statistics* **22**: 79–86.
- Kuznetsov, V., Moskalenko, V. and Zolotykh, N. Y. (2020). Electrocardiogram generation and feature extraction using a variational autoencoder, *arXiv preprint arXiv:2002.00254* .
- Lauer, D. T., Morain, S. A. and Salomonson, V. V. (1997). The landsat program: Its origins, evolution, and impacts, *Photogrammetric Engineering and Remote Sensing* **63**: 831–838.
- Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B. and Chanussot, J. (2022). Deep learning in multimodal remote sensing data fusion: A comprehensive review, *International Journal of Applied Earth Observation and Geoinformation* **112**: 102926.
- Li, J., Hu, Q. and Ai, M. (2019). Rift: Multi-modal image matching based on radiation-variation insensitive feature transform, *IEEE Transactions on Image Processing* **29**: 3296–3310.
- Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D. D. and Chen, M. (2014). Medical image classification with convolutional neural network, *International Conference on Control Automation Robotics & Vision (ICARCV)*.
- Li, W., Liu, H., Wang, Y., Li, Z., Jia, Y. and Gui, G. (2019). Deep learning-based classification methods for remote sensing images in urban built-up areas, *IEEE Access* **7**: 36274–36284.

- Li, Y., Zhang, Y. and Zhu, Z. (2020). Error-tolerant deep learning for remote sensing image scene classification, *IEEE Transactions on Cybernetics* **51**: 1756–1768.
- Li, Z., Zhu, X., Xin, Z., Guo, F., Cui, X. and Wang, L. (2021). Variational generative adversarial network with crossed spatial and spectral interactions for hyperspectral image classification, *Remote Sensing* **13**: 3131.
- Lin, L., Di, L., Yu, E. G., Kang, L., Shrestha, R., Rahman, M. S., Tang, J., Deng, M., Sun, Z., Zhang, C. et al. (2016). A review of remote sensing in flood assessment, *International Conference on Agro-Geoinformatics*.
- Lo, C. P., Quattrochi, D. A. and Luvall, J. C. (1997). Application of high-resolution thermal infrared remote sensing and gis to assess the urban heat island effect, *International journal of Remote sensing* **18**: 287–304.
- Lopez, R., Boyeau, P., Yosef, N., Jordan, M. and Regier, J. (2020). Decision-making with auto-encoding variational bayes, *Advances in Neural Information Processing Systems (NIPS)*.
- Loveland, T. R. and Dwyer, J. L. (2012). Landsat: Building a strong future, *Remote Sensing of Environment* **122**: 22–29.
- Lowe, D. G. (1999). Object recognition from local scale-invariant features, *IEEE International Conference on Computer Vision (ICCV)*.
- Ma, L. and Sun, B. (2020). Machine learning and ai in marketing—connecting computing power to human insights, *International Journal of Research in Marketing* **37**: 481–504.
- Ma, S., Liu, C., Li, Z. and Yang, W. (2022). Integrating adversarial generative network with variational autoencoders towards cross-modal alignment for zero-shot remote sensing image scene classification, *Remote Sensing* **14**: 4533.
- Majidi Nezhad, M., Nastasi, B., Groppi, D., Lamagna, M., Piras, G. and Astiaso Garcia, D. (2021). Green energy sources assessment using sentinel-1 satellite remote sensing, *Frontiers in Energy Research* **9**: 649305.
- Mashey, J. R. (1999). Big data and the next wave of {InfraStress} problems, solutions, opportunities, *USENIX Annual Technical Conference (USENIX)*.

- Mohla, S., Pande, S., Banerjee, B. and Chaudhuri, S. (2020). Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification, *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.
- Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2018). *Foundations of Machine Learning*, MIT Press.
- Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*, MIT Press.
- Murphy, K. P. (2022). *Probabilistic machine learning: an introduction*, MIT Press.
- Mylona, E. A., Sykioti, O. A., Koutroumbas, K. D. and Rontogiannis, A. A. (2017). Spectral unmixing-based clustering of high-spatial resolution hyperspectral imagery, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **10**: 3711–3721.
- Ng, A. and Jordan, M. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, *Conference on Neural Information Processing Systems (NIPS)*.
- Nishizaki, H. (2017). Data augmentation and feature extraction using variational autoencoder for acoustic modeling, *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.
- Papadimitriou, K., Parelli, M., Sapountzaki, G., Pavlakos, G., Maragos, P. and Potamianos, G. (2021). Multimodal fusion and sequence learning for cued speech recognition from videos, *International Conference on Human-Computer Interaction (HCI)*.
- Papandreou, G., Chen, L.-C., Murphy, K. P. and Yuille, A. L. (2015). Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation, *IEEE International Conference on Computer Vision (ICCV)*.
- Paraskevopoulos, G., Georgiou, E. and Potamianos, A. (2022). Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

- Park, D., Hoshi, Y. and Kemp, C. C. (2018). A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder, *IEEE Robotics and Automation Letters* **3**: 1544–1551.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library, *Conference on Neural Information Processing Systems (NIPS)*.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-learn: Machine learning in python, *Journal of Machine Learning Research* **12**: 2825–2830.
- Pol, A. A., Berger, V., Germain, C., Cerminara, G. and Pierini, M. (2019). Anomaly detection with conditional variational autoencoders, *IEEE International Conference on Machine Learning and Applications (ICMLA)*.
- Qayyum, A., Qadir, J., Bilal, M. and Al-Fuqaha, A. (2020). Secure and robust machine learning for healthcare: A survey, *IEEE Reviews in Biomedical Engineering* **14**: 156–180.
- Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition, *Proceedings of the IEEE* **77**: 257–286.
- Rasti, B. and Ghamisi, P. (2020). Remote sensing image classification using subspace sensor fusion, *Information Fusion* **64**: 121–130.
- Ratledge, N., Cadamuro, G., de la Cuesta, B., Stigler, M. and Burke, M. (2022). Using machine learning to assess the livelihood impact of electricity access, *Nature* **611**: 491–495.
- Ravishankar, R., AlMahmoud, E., Habib, A. and de Weck, O. L. (2022). Capacity estimation of solar farms using deep learning on high-resolution satellite imagery, *Remote Sensing* **15**: 210.
- Reynolds, D. A. et al. (2009). Gaussian mixture models, *Encyclopedia of Biometrics* **741**: 659–663.

- Rezende, D. J., Mohamed, S. and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models, *International Conference on Machine Learning (ICML)*.
- Ribeiro, R. T., Marinho, R. T. and Sanches, J. M. (2012). Classification and staging of chronic liver disease from multimodal data, *IEEE Transactions on Biomedical Engineering* **60**: 1336–1344.
- Rish, I. et al. (2001). An empirical study of the naive bayes classifier, *International Joint Conferences on Artificial Intelligence Workshop on Empirical Methods in Artificial Intelligence*.
- Ruthotto, L. and Haber, E. (2021). An introduction to deep generative modeling, *GAMM-Mitteilungen* **44**: e202100008.
- Ryu, S., Choi, H., Lee, H. and Kim, H. (2020). Convolutional autoencoder based feature extraction and clustering for customer load analysis, *IEEE Transactions on Power Systems* **35**: 1048–1060.
- Scheunders, P., Tuia, D. and Moser, G. (2018). *Data processing and analysis methodology*, Elsevier.
- Senchuri, R., Kuras, A. and Burud, I. (2021). Machine learning methods for road edge detection on fused airborne hyperspectral and lidar data, *IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*.
- Shadman Roodposhti, M., Aryal, J., Lucieer, A. and Bryan, B. A. (2019). Uncertainty assessment of hyperspectral image classification: Deep learning vs. random forest, *Entropy* **21**: 78.
- Sharma, M., Dhanaraj, M., Karnam, S., Chachlakis, D., Ptucha, R., Markopoulos, P. and Saber, E. (2020). Yolors: Object detection in multimodal remote sensing imagery, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **14**: 1497–1508.
- Shen, X., Liu, B., Zhou, Y., Zhao, J. and Liu, M. (2020). Remote sensing image captioning via variational autoencoder and reinforcement learning, *Knowledge-Based Systems* **203**: 105920.

- Shirgave, S., Awati, C., More, R. and Patil, S. (2019). A review on credit card fraud detection using machine learning, *International Journal of Scientific & Technology Research* **8**: 1217–1220.
- Shirmard, H., Farahbakhsh, E., Müller, R. D. and Chandra, R. (2022). A review of machine learning in processing remote sensing data for mineral exploration, *Remote Sensing of Environment* **268**: 112750.
- Sim, J. and Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements, *Physical Therapy* **85**: 257–268.
- Sofou, A., Evangelopoulos, G. and Maragos, P. (2005). Soil image segmentation and texture analysis: a computer vision approach, *IEEE Geoscience and Remote Sensing Letters* **2**: 394–398.
- Sofou, A. and Maragos, P. (2008). Generalized flooding and multicue pde-based image segmentation, *IEEE Transactions on Image Processing* **17**: 364–376.
- Swain, P. H. and Davis, S. M. (1981). Remote sensing: The quantitative approach, *IEEE Transactions on Pattern Analysis & Machine Intelligence* **3**: 713–714.
- Tariq, A., Jiango, Y., Li, Q., Gao, J., Lu, L., Soufan, W., Almutairi, K. F. and Habibur Rahman, M. (2023). Modelling, mapping and monitoring of forest cover changes, using support vector machine, kernel logistic regression and naive bayes tree models with optical remote sensing data, *Heliyon* **9**: e13212.
- Themelis, K. E., Schmidt, F., Sykioti, O., Rontogiannis, A. A., Koutroumbas, K. D. and Daglis, I. A. (2012). On the unmixing of mex/omega hyperspectral data, *Planetary and Space Science* **68**: 34–41.
- Thoreau, R., Risser, L., Achard, V., Berthelot, B. and Briottet, X. (2022). p3vae: a physics-integrated generative model. application to the semantic segmentation of optical remote sensing images, *arXiv preprint arXiv:2210.10418* .
- Torres-Carrasquillo, P. A., Reynolds, D. A. and Deller, J. R. (2002). Language identification using gaussian mixture model tokenization, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

- Tran, Q. T., Alom, M. Z. and Orr, B. A. (2022). Comprehensive study of semi-supervised learning for dna methylation-based supervised classification of central nervous system tumors, *BMC Bioinformatics* **23**: 1–17.
- Travis, L. D. (1992). Remote sensing of aerosols with the earth-observing scanning polarimeter, *Polarization and Remote Sensing*.
- Tsagkatakis, G., Aidini, A., Fotiadou, K., Giannopoulos, M., Pentari, A. and Tsakalides, P. (2019). Survey of deep-learning approaches for remote sensing observation enhancement, *Sensors* **19**: 3929.
- Turhan, C. G. and Bilge, H. S. (2018). Recent trends in deep generative models: a review, *International Conference on Computer Science and Engineering (UBMK)*.
- Ulusoy, I. and Bishop, C. M. (2005). Generative versus discriminative methods for object recognition, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Valero, S., Agulló, F. and Inglada, J. (2021). Unsupervised learning of low dimensional satellite image representations via variational autoencoders, *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*.
- Van der Meij, B., Kooistra, L., Suomalainen, J., Barel, J. M. and De Deyn, G. B. (2017). Remote sensing of plant trait responses to field-based plant–soil feedback using uav-based optical sensors, *Biogeosciences* **14**: 733–749.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017). Attention is all you need, *Conference on Neural Information Processing Systems (NIPS)*.
- Vincent, P., Larochelle, H., Bengio, Y. and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders, *International Conference on Machine Learning (ICML)*.
- Voogt, J. A. and Oke, T. R. (2003). Thermal remote sensing of urban climates, *Remote Sensing of Environment* **86**: 370–384.
- Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F. and Pinheiro, P. R. (2020). Covidgan: data augmentation using auxiliary classifier gan for improved covid-19 detection, *IEEE Access* **8**: 91916–91923.

- Wang, H. and Yeung, D.-Y. (2020). A survey on bayesian deep learning, *ACM Computing Surveys (CSUR)* **53**: 1–37.
- Wang, X., Tan, K., Du, Q., Chen, Y. and Du, P. (2020). Cva2e: A conditional variational autoencoder with an adversarial training process for hyperspectral imagery classification, *IEEE Transactions on Geoscience and Remote Sensing* **58**: 5676–5692.
- Wang, X., Zhang, F. and Ding, J. (2017). Evaluation of water quality based on a machine learning algorithm and water quality index for the ebinur lake watershed, china, *Scientific Reports* **7**: 12858.
- Wu, X., Hong, D. and Chanussot, J. (2021). Convolutional neural networks for multimodal remote sensing data classification, *IEEE Transactions on Geoscience and Remote Sensing* **60**: 1–10.
- Xenaki, S. D., Koutroumbas, K. D. and Rontogiannis, A. A. (2015). A novel adaptive possibilistic clustering algorithm, *IEEE Transactions on Fuzzy Systems* **24**: 791–810.
- Xenaki, S. D., Koutroumbas, K. D. and Rontogiannis, A. A. (2016). Hyperspectral image clustering using a novel efficient online possibilistic algorithm, *European Signal Processing Conference (EUSIPCO)*.
- Xie, L., Feng, X., Zhang, C., Dong, Y., Huang, J. and Liu, K. (2022). Identification of urban functional areas based on the multimodal deep learning fusion of high-resolution remote sensing images and social perception data, *Buildings* **12**: 556.
- Xu, Y., Du, B., Zhang, L., Cerra, D., Pato, M., Carmona, E., Prasad, S., Yokoya, N., Hänsch, R. and Le Saux, B. (2019). Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 ieee grss data fusion contest, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **12**: 1709–1724.
- Yang, M.-H. and Ahuja, N. (1998). Gaussian mixture model for human skin color and its applications in image and video databases, *Storage and Retrieval for Image and Video Databases*.
- Yang, Z., Hu, Z., Salakhutdinov, R. and Berg-Kirkpatrick, T. (2017). Improved variational autoencoders for text modeling using dilated convolutions, *PMLR International Conference on Machine Learning*.

- Yu, X., Hyypä, J., Litkey, P., Kaartinen, H., Vastaranta, M. and Holopainen, M. (2017). Single-sensor solution to tree species classification using multispectral airborne laser scanning, *Remote Sensing* **9**: 108.
- Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J. et al. (2020). Deep learning in environmental remote sensing: Achievements and challenges, *Remote Sensing of Environment* **241**: 111716.
- Zabalza, J., Ren, J., Zheng, J., Zhao, H., Qing, C., Yang, Z., Du, P. and Marshall, S. (2016). Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging, *Neurocomputing* **185**: 1–10.
- Zemouri, R. (2020). Semi-supervised adversarial variational autoencoder, *Machine Learning and Knowledge Extraction* **2**: 361–378.
- Zerrouki, Y., Harrou, F., Zerrouki, N., Dairi, A. and Sun, Y. (2021). Desertification detection using an improved variational autoencoder-based approach through etm-landsat satellite data, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **14**: 202–213.
- Zhang, M., Li, W., Tao, R., Li, H. and Du, Q. (2021). Information fusion for classification of hyperspectral and lidar data using ip-cnn, *IEEE Transactions on Geoscience and Remote Sensing* **60**: 1–12.
- Zhang, X. and Han, L. (2020). How well do deep learning-based methods for land cover classification and object detection perform on high resolution remote sensing imagery?, *Remote Sensing* **12**: 417.
- Zhang, X., Zhou, Y. and Luo, J. (2022). Deep learning for processing and analysis of remote sensing big data: A technical review, *Big Earth Data* **6**: 527–560.
- Zhang, Y., Gan, Z., Fan, K., Chen, Z., Heno, R., Shen, D. and Carin, L. (2017). Adversarial feature matching for text generation, *PMLR International Conference on Machine Learning*.
- Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C.-C., Pang, R., Le, Q. V. and Wu, Y. (2020). Pushing the limits of semi-supervised learning for automatic speech recognition, *arXiv preprint arXiv:2010.10504* .

- Zhong, T., Zhang, Z., Chen, M., Zhang, K., Zhou, Z., Zhu, R., Wang, Y., Lü, G. and Yan, J. (2021). A city-scale estimation of rooftop solar photovoltaic potential based on deep learning, *Applied Energy* **298**: 117–132.
- Zhu, J., Shen, Y., Zhao, D. and Zhou, B. (2020). In-domain gan inversion for real image editing, *European Conference Computer Vision (ECCV)*.
- Zhu, X. (2005). Semi-supervised learning literature survey, *World* **10**: 10.
- Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F. and Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources, *IEEE Geoscience and Remote Sensing Magazine* **5**: 8–36.
- Zlatintsi, A., Rodomagoulakis, I., Koutras, P., Dometios, A., Pitsikalis, V., Tzafestas, C. S. and Maragos, P. (2018). Multimodal signal processing and learning aspects of human-robot interaction for an assistive bathing robot, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D. and Chen, H. (2018). Deep autoencoding gaussian mixture model for unsupervised anomaly detection, *International Conference on Learning Representations*.
- Zubko, V., Kaufman, Y. J., Burg, R. I. and Martins, J. V. (2007). Principal component analysis of remote sensing of aerosols over oceans, *IEEE Transactions on Geoscience and Remote Sensing* **45**: 730–745.