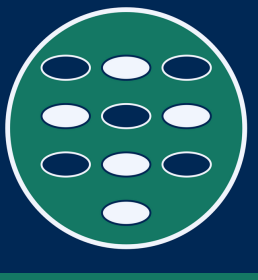
Norges miljø- og biovitenskapelige universitet

Institutt for datavitenskap



Digitalisering på Ås

DAT390

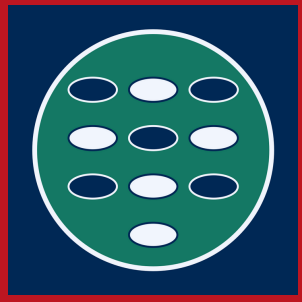
Data science seminar

4 Research ethics and impact

4.3 Analysis of successful papers from Data Science

Noregs miljø- og biovitskaplege universitet

Institutt for datavitskap



Digitalisering på Ås

4 Ethics and impact

4.3 Sample successful papers

DAT390

4th November 2024

Search for the most impactful papers

Norwegian University of Life Sciences

(*or first author, where no corresponding author given)

Google Scholar (by 3.11.2024), with corresponding author* from Data Science.

Publication year 2021

Most cited	**38** citations	A. Jenul *et al.*, "RENT: Repeated elastic net […]," *IEEE Access* **9**: 152333–152346, doi:10.1109/access.2021.3126429, **2021**.
Runner-up	16 citations	A. Jenul *et al.*, doi:10.21105/joss.03323, **2021**.

Publication year 2022

Most cited	**31** citations	R. Helin *et al.*, "On the possible benefits of deep learning for […]," *J. Chemom.* **36**(2): e3374, doi:10.1002/cem.3374, **2022**.
Runner-up	9 citations	M. T. Horsch and B. Schembera, in *Proc. CAOS 2022*, **2022**.

Publication year 2023

Most cited	**13** citations	E. L. Gjelsvik *et al.*, "Current overview and way forward for the […]," *Fuel* **334**: 126696, doi:10.1016/j.fuel.2022.126696, **2023**.
Runner-up	7 citations	M. Horsch *et al.*, doi:10.1109/icapai58366.2023.10193944, **2023**.

A. Jenul *et al.* (2021)

Norwegian University of Life Sciences

RENT - Repeated Elastic Net Technique for Feature Selection

ANNA JENUL[1], (Student Member, IEEE), STEFAN SCHRUNNER[1], (Member, IEEE), KRISTIAN HOVDE LILAND[1], ULF GEIR INDAHL[1], CECILIA MARIE FUTSÆTHER[1], AND OLIVER TOMIC[1]

[1]Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway (e-mail: {anna.jenul,stefan.schrunner,kristian.liland,ulf.indahl,cecilia.futsaether,oliver.tomic}@nmbu.no)

Corresponding author: Anna Jenul (e-mail: anna.jenul@nmbu.no).

This work was partly funded by the Norwegian Cancer Society [grant no. 182672-2016].

ABSTRACT Feature selection is an essential step in data science pipelines to reduce the complexity associated with large datasets. While much research on this topic focuses on optimizing predictive performance, few studies investigate stability in the context of the feature selection process. In this study, we present the Repeated Elastic Net Technique (RENT) for Feature Selection. RENT uses an ensemble of generalized linear models with elastic net regularization, each trained on distinct subsets of the training data. The feature selection is based on three criteria evaluating the weight distributions of features across all elementary models. This fact leads to the selection of features with high stability that improve the robustness of the final model. Furthermore, unlike established feature selectors, RENT provides valuable information for model interpretation concerning the identification of objects in the data that are difficult to predict during training. In our experiments, we benchmark RENT against six established feature selectors on eight multivariate datasets for binary classification and regression. In the experimental comparison, RENT shows a well-balanced trade-off between predictive performance and stability. Finally, we underline the additional interpretational value of RENT with an exploratory post-hoc analysis of a healthcare dataset.

Norwegian University of Life Sciences

A. Jenul *et al.*: Theory-oriented style

Use the introduction to present the formal problem and notation

I. INTRODUCTION

A Predictive task involves a dataset $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_I\} \subseteq \mathbb{R}^N$ and an associated vector of target values $\boldsymbol{y} = \{y_1, \ldots, y_I\} \subseteq \mathbb{T}$, where the target space $\mathbb{T}$ may represent a set of classes (classification task) or a subset of the real numbers (regression task). In this study, our focus lies on generalized linear models (GLMs), which model the target as a linear combination of the inputs with weights $\boldsymbol{\beta} \in \mathbb{R}^N$, followed by a transformation. The N-dimensional input (feature) vectors in the modeling describe object characteristics. Since data acquisition techniques evolve steadily, situations where the number of features (N) exceeds the number of objects (I) often occur.

A feature selector θ_F decomposes the data space into a direct sum of selected features (V_1) and non-selected features (V_2) according to the given feature set $F \subset \{1, \ldots, N\}$,

$$\mathbb{R}^N = V_1 \oplus V_2, \text{ s.t.} V_1 \cong \mathbb{R}^{|F|} \text{ and } V_2 \cong \mathbb{R}^{N-|F|},$$

and projects all objects from $\mathbb{R}^N$ to the subspace V_1, i.e.

$$\theta_F : \mathbb{R}^N \to V_1, \ \theta_F(\boldsymbol{x}) = \text{proj}_{V_1}(\boldsymbol{x}).$$

Analyse the computational complexity of the developed method

Given the first variant, RENT runs an ensemble comprising K independent GLMs, each trained on a number of N features, which delivers a complexity of

$$\mathcal{O}\left(KN^2 \cdot (N + I_{train}^{(K)})\right),$$

where $I_{train}^{(K)} < I_{train}$ denotes the sample size of each subset during RENT training. In addition, hyper-parameter tuning requires training c GLMs, where c is a constant given by the number of level combinations for regularization and cutoff parameters, resulting in

$$\mathcal{O}\left(cN^2 \cdot (N + I_{train})\right).$$

In total, an upper bound to the full runtime complexity of RENT is given by

$$\mathcal{O}\left((K + c) \cdot N^2 \cdot (N + I_{train})\right). \tag{9}$$

R. Helin *et al.* (2022)

Received: 9 March 2021 | Revised: 3 September 2021 | Accepted: 6 September 2021

DOI: 10.1002/cem.3374

SPECIAL ISSUE - RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

On the possible benefits of deep learning for spectral preprocessing

Runar Helin | **Ulf Geir Indahl** | **Oliver Tomic** | **Kristian Hovde Liland**

Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway

Correspondence
Runar Helin, Faculty of Science and Technology, Norwegian University of Life Sciences, Ås 1430, Norway.
Email: runarhel@nmbu.no

Abstract

Preprocessing is a mandatory step in most types of spectroscopy and spectrometry. The choice of preprocessing method depends on the data being analysed, and to get the preprocessing right, domain knowledge or trial and error is required. Given the recent success of deep learning-based methods in numerous applications and their ability to automatically detect patterns in data, we aimed at exploring the possibilities of using such methods for preprocessing. Our study considered a flexible but systematic investigation of spectroscopic preprocessing methods (classical and deep learning-based) combined with pre-

R. Helin *et al.*: Error discussion

TABLE 2 Prediction performance of the model selection phase

Preproc. method	engine	Prediction	RMSEP	AccCV	AccP
Raw		MLP	465.0 ± 3.01	0.910	0.91 ± 0.0016
Raw		CNN	353.0 ± 7.0	0.925	0.926 ± 0.0012
Raw	-	PLS	426.0 ± 0.0	0.913	0.914 ± 0.0
SG-EMSC		MLP	357.0 ± 2.57	0.949	**0.948 ± 0.0007**
SG-EMSC		CNN	323.0 ± 3.96	0.950	**0.948 ± 0.001**
SG-EMSC	-	PLS	415.0 ± 0.0	0.917	0.918 ± 0.0
NN-EMSC		MLP	514.0 ± 5.92	0.929	0.929 ± 0.001
NN-EMSC		CNN	319.0 ± 6.93	0.928	0.922 ± 0.0017
NN-EMSC	MLP	PLS	421.0 ± 0.35	0.915	0.915 ± 0.0001
NN-EMSC	CNN	PLS	406.0 ± 1.57	0.916	0.916 ± 0.0001
NN-NoiseBase		MLP	**316.0 ± 4.34**	0.939	0.938 ± 0.0029
NN-NoiseBase		CNN	336.0 ± 4.2	0.934	0.926 ± 0.0017
NN-NoiseBase	MLP	PLS	413.0 ± 1.1	0.916	0.912 ± 0.0002
NN-NoiseBase	CNN	PLS	425.0 ± 1.57	0.916	0.914 ± 0.0004

"*Note:* […] The RMSEP column is the mean prediction error of 30 repeated trials. The AccCV column is the prediction accuracy for the sevenfold cross-validation. The AccP column is the mean prediction accuracy of 30 repeated trials, including the standard error."

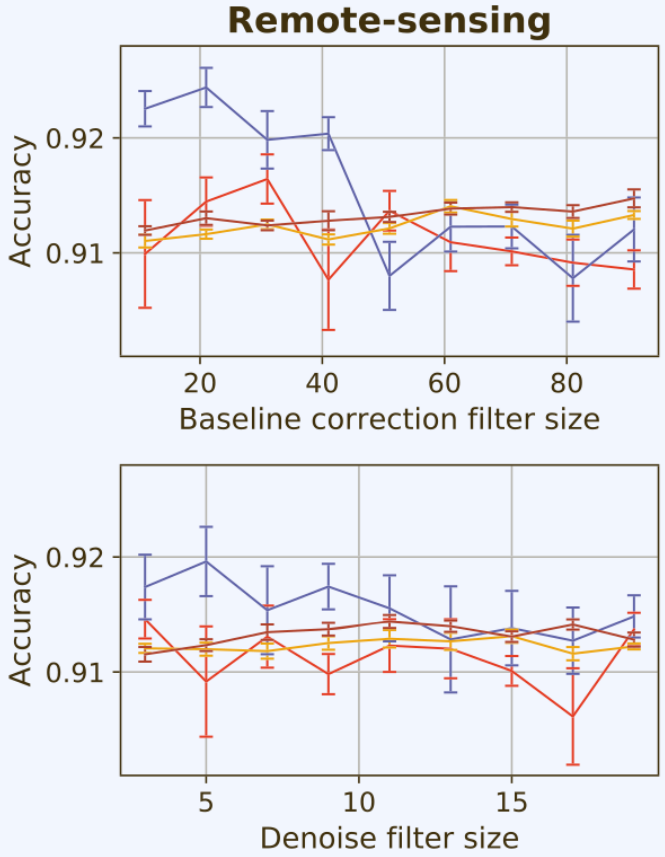


FIGURE 5

Right column: remote-sensing dataset.
Each line shows the mean score with standard error as bars

E. L. Gjelsvik *et al.* (2023)

Norwegian University of Life Sciences

Review article

Current overview and way forward for the use of machine learning in the field of petroleum gas hydrates

Elise Lunde Gjelsvik [a,*], Martin Fossen [b], Kristin Tøndel [a]

[a] *Norwegian University of Life Sciences, Faculty of Science and Technology, Ås, Norway*
[b] *SINTEF AS, Trondheim, Norway*

ARTICLE INFO

Keywords:
Gas hydrates
Machine learning
FT-ICR MS
Chemometrics
Crude oil

ABSTRACT

Gas hydrates represent one of the main flow assurance challenges in the oil and gas industry as they can lead to plugging of pipelines and process equipment. In this paper we present a literature study performed to evaluate the current state of the use of machine learning methods within the field of gas hydrates with specific focus on the oil chemistry. A common analysis technique for crude oils is Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS) which could be a good approach to achieving a better understanding of the chemical composition of hydrates, and the use of machine learning in the field of FT-ICR MS was therefore also examined. Several machine learning methods were identified as promising, their use in the literature was reviewed and a text analysis study was performed to identify the main topics within the publications. The literature search revealed that the publications on the combination of FT-ICR MS, machine learning and gas hydrates is limited to one. Most of the work on gas hydrates is related to thermodynamics, while FT-ICR MS is mostly used for chemical analysis of oils. However, with the combination of FT-ICR MS and machine learning to evaluate samples related to gas hydrates, it could be possible to improve the understanding of the composition of hydrates and thereby identify hydrate active compounds responsible for the differences between oils forming plugging hydrates and oils forming transportable hydrates.

E. L. Gjelsvik *et al.*: Systematic review

Clear formulation of objective and research questions

The objective of this review is to provide an overview of the machine learning methods used within the field of gas hydrates, with specific focus on the oil chemistry. First, we performed a text mining study to show the previous research areas of focus and expose potential gaps within. The aim of text mining is to scrape a web page of text related to a predefined keyword. We accessed all relevant publications from the Scopus Search database [41] and the most common and promising methods in literature are discussed. Additionally, methods

3. Text mining

To achieve an overview of the current status of machine learning methods within the field of petroleum gas hydrates the following questions were defined, of which the answers should give a thorough understanding of the field.

- **Q1**: Within which fields of gas hydrate research are machine learning used?
- **Q2**: What type of machine learning methods are used in the literature?
- **Q3**: What are the challenges in the field of gas hydrates using machine learning?
- **Q4**: How can machine learning improve the field of gas hydrate research?

Well-described methodology for selecting the analysed references

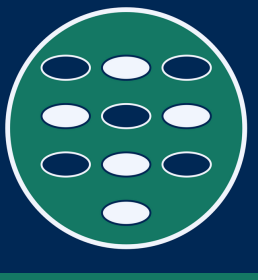
The resulting search phrases were as follows for gas hydrates *'TITLE-ABS-KEY((gas W/1 hydrate*) AND ((machine learning method) OR (method abbreviation)))'* and for FT-ICR MS *'TITLE-ABS-KEY((ft-icr W/1 ms) AND ((machine learning method) OR (method abbreviation)))'*. The 'W/1' ensures that the words are only one term apart and the * allows for different endings of the word, for instance *s* for plural

Subjects	Methods
Gas hydrates FT-ICR MS	Principal Component Analysis (PCA) Partial Least Squares (PLS) Decision Trees (DT) Random Forest Artificial Neural Network (ANN) Support Vector Machine (SVM) Convolutional Neural Network (CNN) Regularisation/LASSO/Elastic Net/Ridge Regression Bayesian Networks (BN) K-Nearest neighbours (KNN)

The results from the two searches, *gas hydrates* and *FT-ICR MS*, with the methods in Table 1, are shown in Fig. 1. From the search of gas hydrates in combination with the methods from Table 1, 184 publications were retrieved and from FT-ICR MS and the methods in Table 1, 104 publications were retrieved. The publications returned by the text mining study are reported in the supporting information.

Norges miljø- og biovitenskapelige universitet

Institutt for datavitenskap



Digitalisering på Ås

DAT390

Data science seminar

4 Research ethics and impact

4.3 Analysis of successful papers from Data Science