Learning in the Compressed Data Domain: Application to Milk Quality Prediction

Dixon Vimalajeewa*, Chamil Kulatunga, Donagh P. Berry

Telecommunications Software and Systems Group, Arclabs Research and Innovation Centre, Waterford Institute of Technology, Carriganore, Waterford, Ireland

Teagasc, Animal & Grassland Research and Innovation Centre, Moorepark, Fermoy, Co. Cork, Ireland

(dvimalajeewa, ckulatunga) @tssg.org , Donagh.Berry@teagasc.ie

Abstract

Smart dairy farming has become one of the most exciting and challenging area in cloud-based data analytics. Transfer of raw data from all farms to a central cloud is currently not feasible as applications are generating more data while internet connectivity is lacking in rural farms. As a solution, Fog computing has become a key factor to process data near the farm and derive farm insights by exchanging data between on-farm applications and transferring some data to the cloud. In this context, learning in the compressed data domain, where decompression is not necessary, is highly desirable as it minimizes the energy used for communication/computation, reduces required memory/storage, and improves application latency. Mid-infrared spectroscopy (MIRS) is used globally to predict several milk quality parameters as well as deriving many animal-level phenotypes. Therefore, compressed learning on MIRS data is beneficial both in terms of data processing in the Fog, as well as storing large data sets in the cloud. In this paper, we used principal component analysis and wavelet transform as two techniques for compressed learning to convert MIRS data into a compressed data domain. The study derives near lossless compression parameters for both techniques to transform MIRS data without impacting the prediction accuracy for a selection of milk quality traits.

Keywords: Compressed learning, MIRS, principal component analysis, wavelet transformation, partial least squares regression, fog computing.

1. Introduction

Even though smart farming is advancing with the recent developments of Internet of Things (IoT), cloud-based computing, and deep learning, it has become one of the most challenging industrial sectors in big data analytics due to the limitations of ICT infrastructures [47]. However, according to the statistics from the Food and Agriculture Organization of the United Nations (FAO), smart farming will be a key contributor to sustainable intensification in agriculture to feed the 9.2 billion human population by 2050 [1]. There is also a growing interest in pasture-based smart dairy farming in the countries like New Zealand and Ireland, which tend to be in less direct competition with human edible protein and energy sources. Therefore, more harmonized research is needed to optimally utilize ICT infrastructures in precision dairy farming to minimize consumed storage space, communication and computations to facilitate contemporary analytics providing near real-time insights on-farm [37]. This is where the notion of effective data compression approaches are important.

Most sensor-based technologies and IoT platforms are designed today to collate and store vast quantities of raw data readings from different sources in geographically distributed farms. Many computational facilities for data analytic applications such as MyAgCentral[1] are now seeking computational resources in cluster-based servers in large centralized data centres. At same the time, the Agricultural Information Management Standards of FAO (AIMS) has already started developing standards and maintaining interoperable trans-national databases for open agricultural data. Therefore farm data will be aggregated as big datasets and there is a requirement to store these data for long-term analytical purposes. This is beneficial since aggregation of data, which extracts a large number of descriptive features in temporally and spatially diverse domains, contributes to an improved learning accuracy. Therefore, compression of such data without a loss of accuracy is vital in terms of the ***storage*** requirement.

Dissemination of data in its raw format (i.e., in the measurement domain) into large cloud-based data centers is generally not feasible for most farms due to high energy consumption, time criticality of the applications, and

* Corresponding author *Email address:* `dvimalajeewa@tssg.org` (Dixon Vimalajeewa)

[1] http://www.sageinsights.com

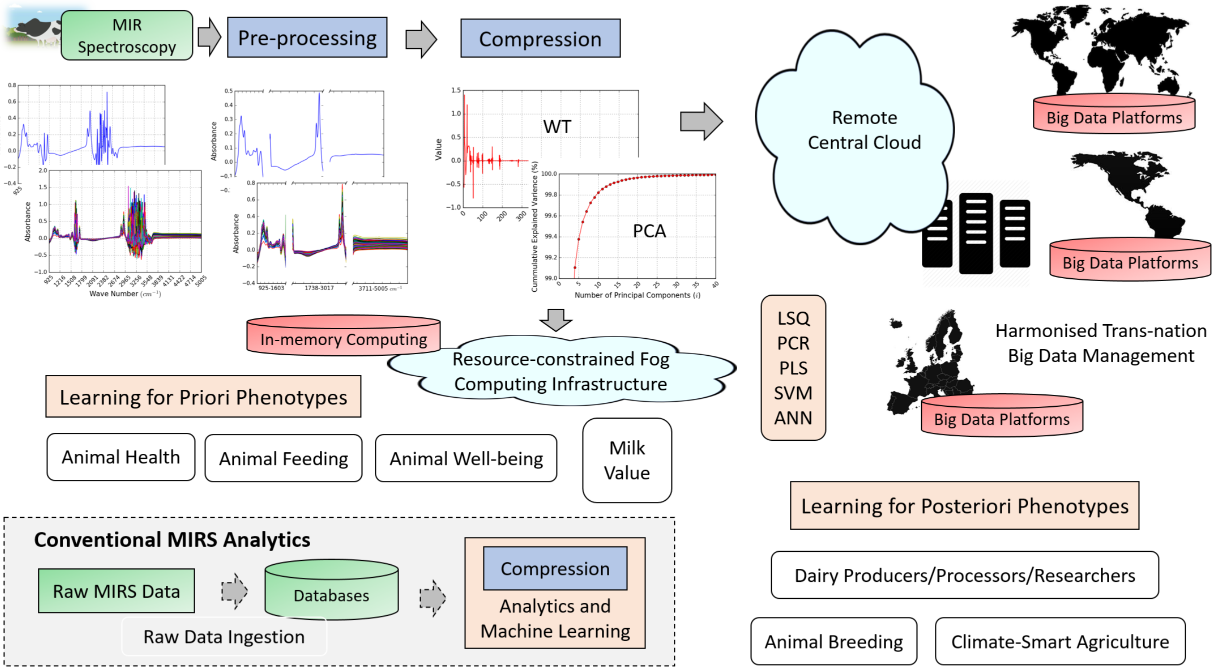

MIR Spectroscopy
Pre-processing
Compression
WT
PCA
Remote Central Cloud
Big Data Platforms
Big Data Platforms
In-memory Computing
Resource-constrained Fog Computing Infrastructure
Learning for Priori Phenotypes
LSQ
PCR
PLS
SVM
ANN
Harmonised Trans-nation Big Data Management
Big Data Platforms
Animal Health
Animal Feeding
Animal Well-being
Milk Value
Learning for Posteriori Phenotypes
Conventional MIRS Analytics
Raw MIRS Data
Databases
Raw Data Ingestion
Compression
Analytics and Machine Learning
Dairy Producers/Processors/Researchers
Animal Breeding
Climate-Smart Agriculture
Absorbance
Wave Number (cm⁻¹)
Value
Cumulative Explained Variance (%)
Number of Principal Components (i)

Fig. 1. PCA and WT for MIRS have been applied to avoid overfitting and de-noising in the conventional spectrometry analysis. The two techniques can be used for data compression in future distributed analytics platforms with compressed learning.

the poor/costly rural internet connectivity. For example, if a disease detection system is centralized, it may slow down the farmers' response because of the necessity to transfer vast quantity of data readings to a remote cloud and wait for the outcome to return. Therefore, compression of data is also important in terms of the ***communication*** efficiency.

However, the key challenge today is whether the centralized storage and computational technologies (with communication networks) contributing to smart dairy farming will still not be sufficient to deliver the future demand without an advanced data analytic infrastructure closer to remote farm management systems. Therefore, a scalable computational infrastructure under constrained resources (proximate to the farm) is essential. In such a constrained infrastructure, compression is a key performance factor also for ***computational*** efficiency in addition to storage and communication.

Emergence of Fog Computing: With the increase in the amount of data generated from connected sensors, there is a demand to move processing capabilities closer to the data sources, which is in contrast to centralizing raw data in a large data centre. This phenomenon of distributing computations towards the data was first termed as data gravity by *Dave McCrory* in 2010 and is now being realized with new technologies such as Fog (i.e., edge) computing [3] and cloudlets [21]. Fog computing can enable datasets to be processed at the extreme edge of the internet. This computational infrastructure may collectively be formed by low computational proximate devices located near or within the farm. Therefore Fog computing will be a key enabler for many farm analytics to run using scalable in-memory data processing platforms like Spark[2], Flink[2], Storm[2] and H20[3] with in-memory databases like Ignite[2] and SAP HANA. Therefore raw-data compression near the data source is a desirable requirement for near future.

As a result, machine learning models, which have targeted highly-provisioned cloud infrastructures, must be re-designed for these resource-constrained infrastructures to minimize storage, communication and computational requirements. New distributed machine learning paradigms like compressed learning [4] and attribute-distributed learning [50] have significant potential to develop effective learning models [49] rather than centralizing all raw datasets from the farms. The main motivation of the present paper is to validate a compressed learning approach [4] for milk quality analysis based on Mid-infrared spectroscopy (MIRS) technology, which can effectively overcome those three challenges in Fog computing. With compressed learning, any machine learning algorithm can be used in a low-dimensional (i.e. latent) space without decompressing data, while optimizing the resource requirement as well as learning efficiency and accuracy of outcomes. Even though the compressed learning approach has been widely used in many fields for learning from complex data sources, such as high-resolution image and video processing and text analysis [25, 30], its applicability is new to the MIRS based milk quality analysis.

MIRS is the most economical technology used for assessing milk quality. Therefore, MIRS spectra in predictive models are frequently used to develop farm decision-support tools for efficient milk data processing. For instance, the OptiMIR[4] project has used MIRS of milk recordings in an innovative way to observe different characteristics

[2] http://apache.org [3] http://www.h2o.ai [4] https://www.keep.eu/keep/project-ext/21114/OptiMIR

of cows such as energy balance and early detection of diseases. Also, the routinely obtained MIRS of milk can be used for deriving novel models to quantify milk composition on both an animal basis and on bulk tank samples as well as derive milk related herd-level phenotypes [28]. In addition, variation in MIRS of milk can be used as an indicator in predicting animal characteristics such as the physiological state of an animal and its feed efficiency. The collaborative use of MIRS milk data from different farms can also improve the accuracy of the predictions. Therefore processing vast quantities of milk samples with Fog computing is highly desirable for MIRS milk quality analysis in the future smart dairy farming.

Conventional MIRS analysis [42, 43] has been conducted based on co-located data processing by a single computational facility. As shown in Fig. 1, the raw data are directly collated into a repository, mostly by non-experts of data science, and later analyzed by the domain-specific data science experts. This significantly increases the computation and power resource requirement on the cloud using raw MIRS data. In modern distributed processing infrastructures, data pre-processing such as de-noising and dimensionality reduction can be carried out closer to data sources. It would, in turn, improve three forms of resource efficiency of the system and reduce the input cost compared to the conventional approach. Water absorbance data collected using MIRS technology, for instance, hampers the accuracy of milk quality prediction. Removal of these data using distributed computing, prior to sending to the cloud would potentially improve the model accuracy as well as reduce the amount of data in the cloud. Therefore, interpretation of biological data on the edge, using domain-specific knowledge of MIRS, would optimize resource utilization both in big data analytics as well as Fog computing [3].

Compression techniques such as Principal Component Analysis (PCA) [19], Wavelet Transform (WT) [44], manifold and deep learning methods [14] have been widely used for learning from compressed data. The present study investigates the linear learnability of MIRS milk quality data for a selection of milk quality traits in a compressed data domain. We examine in detail PCA and WT as two compression techniques, which have been widely used for MIRS data analysis [42]. The two compression techniques will provide a near lossless compression for many of the currently investigated milk quality parameters. The study concludes with the generalized/harmonized compression parameters required for the two compression techniques to perform compressed learning, i.e. the number of principal components (l) in PCA and the number of wavelet coefficients (r) in WT. We also discuss the additional factors to be required for de-compression, if needed. The impact on the MIRS prediction accuracy at different compression levels was investigated using Partial Least Squares (PLS) linear regression modelling, which has frequently been used in milk MIRS-based predictions [43]. We discuss the importance of sample size in PCA and WT-based compressions and the benefits of supervised compression in compressed learning. Furthermore, we compare our PCA and WT-based compressed learning approaches with state-of-the-art neural network based deep learning techniques such as auto-encoder, GoogleNet and ResNet. While the root mean square error obtained for our approach is comparable for certain features, it is typically higher compared to these techniques. However, the use of deep learning techniques requires a large amount of resources that makes their deployment unsuitable for our resource constrained environments.

Section 1 introduces the paper by discussing the importance of learning in the compressed data domain for MIRS-based analytics from the perspective of Fog computing and big data analytics. Section 2 presents the related works in compressed domain machine learning approaches and applications. Section 3 describes the MIRS techniques used in predicting milk quality traits. Compressibility analysis of MIRS data using PCA and WT is given in Section 4. Section 5 presents the performance statistics of applying the PLS on compressed MIRS data. Section 6 discusses generated results, applicability and a comparison with a state-of-the-art techniques while Section 7 concludes the paper.

2. Related Works

The concept of learning in the compressed data domain has been used in a vast range of applications such as hyperspectral image analysis in neuro-science [7] and geo-sciences [41], feature selection in video processing [25, 14], machine learning applications in mobile computing [35], distributed data fusions in sensor networks [32, 30], as well as classification of complex and big data structures (e.g. text and images) [31, 11]. Generally, the primary purpose of using machine learning in the compressed data domain (in the rest of the paper we refer to simply as compressed learning) was based on a few main reasons: 1) efficient access to large data volumes in big data computations, 2) energy-efficient communications between constrained devices, and 3) computations in resource limited Fog computing environments. In general, the main categories of compressed learning techniques comprise of the PCA, WT, and deep neural networks as compression methods and the learning methods such as regression and classification. The related works presented here have shown that compressed learning has effectively minimized communications, memory, and data storage, while also reducing the learning complexity and hence the processing time of the applications.

A universal framework for compressed learning, in association with compressed sensing, has been presented periodically in the literature [7]. Dimensionality reduction and data compression have been applied on measurement data based on different basis functions mainly using Fourier and Wavelet. To avoid complete reconstruction of time domain signal of electroencephalograms based on random projections, [36] provided a comprehensive analysis of a methodology and mathematical framework for compressed learning with data sparsity. Additionally, [26] explained a compressed signal processing approach to adequately preserve the similarity metric of pattern recognition in electroencephalograms. The generality of random projections on Nyquist-domain data enables significant reductions in computation.

In order to accurately reconstruct a signal from the Nyquist-domain, the highest frequency of a signal should be less than half of the sampling rate [38]. However, Donoho in 2004 proposed a compressed sensing approach, stating that with the knowledge of signal's sparsity, a signal can be recovered even with fewer samples [6]. This compressed sensing approach combines signal acquisition and compression into one step (i.e. compression at the time of sampling) instead of performing in two steps (traditional sampling) [38]. Hence, compressed sensing reduces potentially the computational, storage, and communication in higher dimensional data processing compared to the traditional data sampling and compression. Therefore, compressed sensing has gained much attention in the recent past for compressed learning with higher dimensional data such as photography, holography and facial recognition [33].

The requirement of dimensionality reduction of big data for subsequent use in machine learning were discussed and classical PCA has failed as a strategy when the number of observations is very large. This has resulted in issues of memory and storage limitations for single processor computers. As an alternative, [49] proposed a new PCA approach based on scanning data by rows. The study [8] outlined compressed linear algebra (CLA) for in-memory operations such as matrix-vector multiplication in compressed data domain. Also [8] documented the drawbacks of heavy weight compression algorithms due to computational complexity in decompressing and lightweight algorithms because of poor compression ratios while making a clear case for the operation of linear algebra directly using matrices with compressed data.

Learning from feature extraction has been extensively used in image and video analysis. A novel technique for constructing high resolution images from low resolution images and recognition of such images using a singular value decomposition (SVD) based PCA approach have been investigated in [14]. Moreover, a SVD-based approach to extract potential global features from facial images given in [15] used special properties of singular values of an image to devise a compact, global feature for image-representation. Also the authors of [15] theoretically proved that leading singular values can be used as rotation-shift-scale-invariant global features of an image. Texture image retrieval and classification based on SVD was investigated in [17], while [11] proposed a texture-based image classification approach based on cross-covariance matrix of image textures. The authors of [11] claim to have reduced the processing time of image classifications by using the compressed domain cross-covariance vectors of the original image data. Sometimes PCA and WT have been jointly used in compressed learning. For instance, PCA has been used to accelerate WT and eye location verification based on the features extracted from facial images using WT-salient maps in [16]. Moreover, a WT-based salient feature extraction approach has been presented in [13]. Another approach to mimic the human visual system's salient detection in images using wavelet-based salient patch detection is given in [18].

High resolution space-borne optical images were analyzed when proposing an efficient ship detection approach using compressed learning with a Deep Neural Network (DNN) algorithm [41]. Only the relevant information was extracted using WT from space-borne optical images to observe ship positions with less detection time. Similarly, [32] proposed an energy-efficient ship navigation method based on compressed domain learning. A large amount of sensor-based ship navigation data was compressed on-board using PCA. Also [32] applied regression analysis in an on-shore located data centre to derive optimal navigation paths based on the compressed data. A lossless dictionary-based compressed learning approach for unsupervised feature learning for text data was discussed in [31] as well as also the applicability of k-grams based compressed data for many tasks in text processing.

Compressed learning has been used extensively in feature learning applications in sequential video frames. The method presented in [25] can separate (as background and foreground of a sequence of video frames) a large set of raw data using a small amount of information based on prior knowledge. The authors of [25] named the protocol Compressive Online Robust Principal Component Analysis (CORPCA) and stated that it can be used to extract only significant features from high dimensional data of time-variant processes by taking a single instance at a time (i.e., a frame). In CORPCA, compression is performed recursively using the compressive information that is extracted from its previous stage. In [30], PCA was applied in compressed domain for re-enforcement learning. This approach reduced the table sizes of state space and action space thereby minimizing the memory spaces and learning times. The study [30] also showed that PCA reduced communications in multi-agent distributed learning environment.

Compressed learning and models are becoming more popular in big data and mobile computing platforms. DNN models are commonly used in mobile applications. However, such applications are too large to fit into constrained mobile computing resources. Therefore, compressed versions of DNN models were introduced in [35], which have the same properties as their corresponding original models and provide an energy-efficient platform to run those models. In addition, distributed computing frameworks such as Apache Spark have been combined with a compressed data representation framework developed by the *Saccinct* project [5]. This framework enables to

[5] http://succinct.cs.berkeley.edu

query data stores in a compressed data domain, so that Spark users benefit in searching point queries directly on a compressed representation of the input data. Deep learning techniques have achieved higher classification accuracy than the traditional compression techniques. An application of auto-encoder technique for hyper-spherical image classification has been provided in [48].

Different spectrometric analyses have benefited from compression techniques to perform further learning. The study [27] discussed a general overview of MIRS applications as a phenotyping tool for deriving milk quality traits, while [10] explained the use of MIRS with WT and PCA in a quantification of the extra-virgin olive oil adulteration process. MIR (and NIR) spectroscopic techniques were used in [23] to determine the dry matter content of tea; WT and PLS chemometric techniques were performed to determine the tea dry matter content.

In general, compressed learning has been used extensively in a vast range of spectrometry applications. However, spectrometry analysis towards a generalized (harmonized) compressed learning process for MIRS-based milk quality monitoring has not been thoroughly investigated. Even though the two compression algorithms, PCA and WT, have been used [42], application of data compression where data are generated (near sensing) has not been thoroughly studied for use in future analytics platforms.

3. MIRS FOR PREDICTING MILK QUALITY TRAITS

Fourier Transform (FT)-MIRS is the prominent MIRS approach currently used in routine milk testing. Globally, milk samples from individual dairy cows and bulk samples are routinely taken to assess milk quality which can be subsequently used by dairy producers and processors in making management decisions but also by breeders to identify genetically elite candidate parents of the next generation. The milk quality information originates from predictions from the transmittance of light in the mid-infrared region (i.e. $2500 - 25000nm; 900 - 5000cm^{-1}$) of the electromagnetic spectrum. The outcome of the MIRS analysis is a spectrum for each sample and this transmittance value is available for each wavelength irrespective of its information content [28]. Milk protein, fat, lactose, urea, minerals, acetone, ketone bodies, casein are some of the most reported milk quality parameters predicted from MIRS [28], which are used in deriving many priori and posteriori phenotypes by the stakeholders.

Some compression algorithms like Lempel-Ziv-Welch (LZW) deliver the objective of data compression (i.e., data are compressed without losing information), but de-compression is still necessary to convert the data into its original (measurement) domain because statistical learning cannot be applied to compressed data. This type of compression method help in optimizing issues such as storage and communication difficulties. However, de-compression brings back the original data dimensionality with irrelevant and redundant MIRS data. Thus, the complexity of learning from the original data remains unchanged. Therefore, such lossless compression algorithms increase only the computational cost of de-compression without making any contribution to the learning process. In conventional cloud systems, decompression happens in high-end servers where energy and computational power are generally not a constraint. However, in Fog computing, decompression may be performed at a resource constrained Fog node [3]. Therefore, compared to the general compression-decompression approaches, the compressed learning concept in MIRS using PCA and WT offers an effective methodology in a resource constrained infrastructure.

The quality and the dimensionality of MIRS data are crucial factors for machine learning. High dimensionality and multi-collinearity (i.e. correlated data) also limits the use of multiple linear regression. As an early approach to compressed learning, [5] proposed a better lossless data compression technique using only a few and potentially

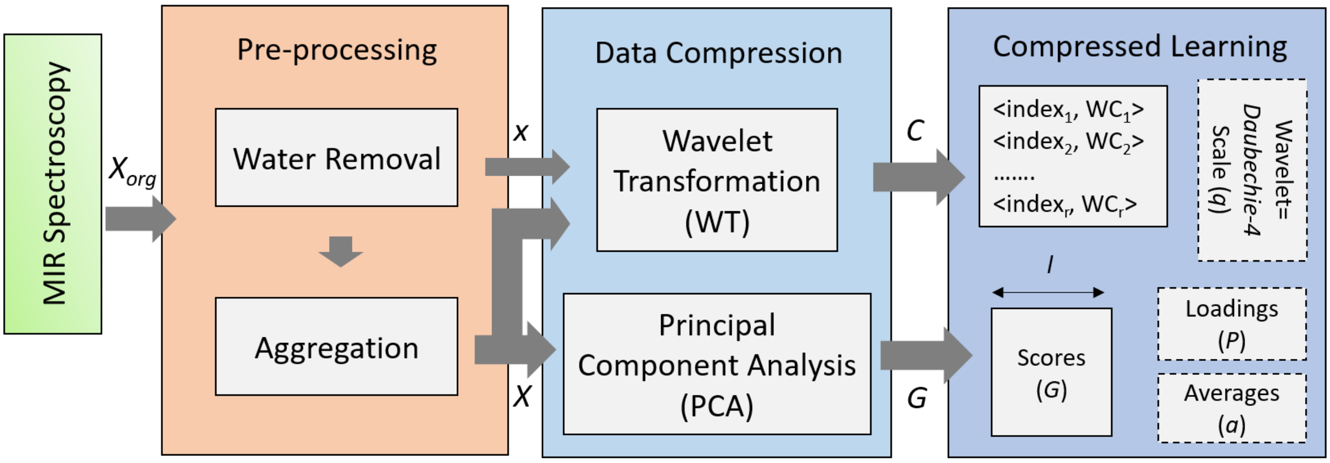


Fig. 2. Pipeline of the Compressed Learning framework: Data pre-processing/aggregation is performed at a very early stage. Compression of MIRS data is carried out using PCA or WT irrespective of the intended milk quality trait (unsupervised). Data in the compressed domain (i.e. scores in PCA or indexed WCs in WT) will be used by different machine learning applications.

Table 1
Mathematical notations used in the paper to represent MIRS dataset and PCA, WT, PLS algorithms.

Notation	Description
X_{org}	Original MIR spectra with values in absorbance
X	Water removed MIR spectra
n	Number of samples in the gold standard
m	Number of wavenumbers in X after removing water
Y	Target variables of milk quality components in %
k	Number of selected milk quality components
(x, y)	A sample (a row) of X and Y in the gold standard
l	Number of PCs selected using PCA
G	Scores matrix of PCA in compressed domain
P	PCA loadings matrix for data recovery
a	Averages of selected l columns
r	Number of WCs selected using WT compression
C	WT after thresholding in compressed domain
q	Level of scale in WT
X'	Reconstructed MIR spectra from the compressed domain
u	Number of Latent Variables in PLS

disjoint sets of highly significant WT coefficients in an orthonormal basis. In addition to using WT, in the present study, we attempted to compare compression performances with PCA as a widely used technique in MIRS. There are many well-defined pre-treatment techniques (e.g. scaling, scatter correction, etc.) in MIRS analytics to undertake quality control of data which we need to apply before the compression process. For example, the spectrum contains dissolved water absorbance ($O = H$ bonds) in the $1500 - 1800cm^{-1}$ and $2900 - 3800cm^{-1}$ ranges, and these regions are not useful in the prediction of milk quality traits.

In compressed learning, the original data can be recovered with the recovery algorithm of PCA and WT only if it is needed. However, by performing all analytical processes in the compressed domain, we can eliminate the additional cost of de-compression, which is in contrast with some other user-interacted data compression applications like in multimedia. Since WT uses a known basis ($Daubechie - 4$ at scale q in Fig. 2), no additional information is needed to decompress. However, in PCA, the loading matrix and the column averages (a and P in Fig. 2) of the original data matrix are required for the decompression as we explain later in the paper.

The compression level of a MIR spectrum (X) depends on the target response variable (Y) of the learning algorithm. For example, an analytical engine for animal health status can be run in one computational sub-system while another analytical algorithm for milk quality may be run in another sub-system. In the present study we investigate a generalization approach of compression of MIRS data only using X (Fig. 2), which is the important research question in compressed learning for MIRS milk quality monitoring. However, we briefly discuss the possibility of further compression based on a known Y within the discussion section.

The data used in the paper originated from the Teagasc research dairy farm at Moorepark, Ireland where MIR spectra were collected and the composition of milk was determined using FOSS MilkScan prediction equations. The input data matrix contained the spectra of 712 different milk samples in the wavenumber region $925 - 5005cm^{-1}$ with a resolution of $3.853cm^{-1}$; wavenumbers were rounded to the nearest integer. As a result, the given spectrum contained 1060 transmittance data points. Therefore, the original MIRS spectra used (called gold standard) to develop linear prediction models was a (712×1060) size matrix and denoted by X_{org}. We converted them to absorbance values by taking log_{10} of the reciprocal of the given transmittance values. Absorbance indicates the amount of absorption of electromagnetic radiation when the MIR light penetrates through the milk sample. Higher absorbance values indicate that the MIR light penetrates less at certain wavenumbers according to the molecular bonds. In addition, percentages of the selected milk nutrient components; lactose, fat, protein and urea, corresponding to each sample were stored in a column matrix ($Y_{n \times k}$, where $n = 712$ and $k = 5$). PLS model calibration and validation were applied on to these gold standard data (Y) to derive our generalized compression parameters.

4. COMPRESSIBILITY OF MIRS MILK QUALITY DATA

This section will discuss the compressibility (unsupervised or general) of MIRS dataset using PCA and WT. First, we will investigate the data redundancy of the available MIR spectra and hence the compressibility of such data, which can be improved using PCA and WT without noticeable information loss. Second, we will discuss the selection of main input (or compression) parameters required for the PCA compression (number of principal components (l)) and WT compression methods (number of wavelet coefficients (r)). Since there is no prior information regarding the learning purposes (e.g. regression, classification) which the compressed data will be used for, the compression should be performed by preserving the original properties of the MIRS data as much as possible. Therefore, the selection of compression parameters is important and should be performed carefully. In order to select reliable values for l and r, their impact on the quality of compression was examined. The variance explained by

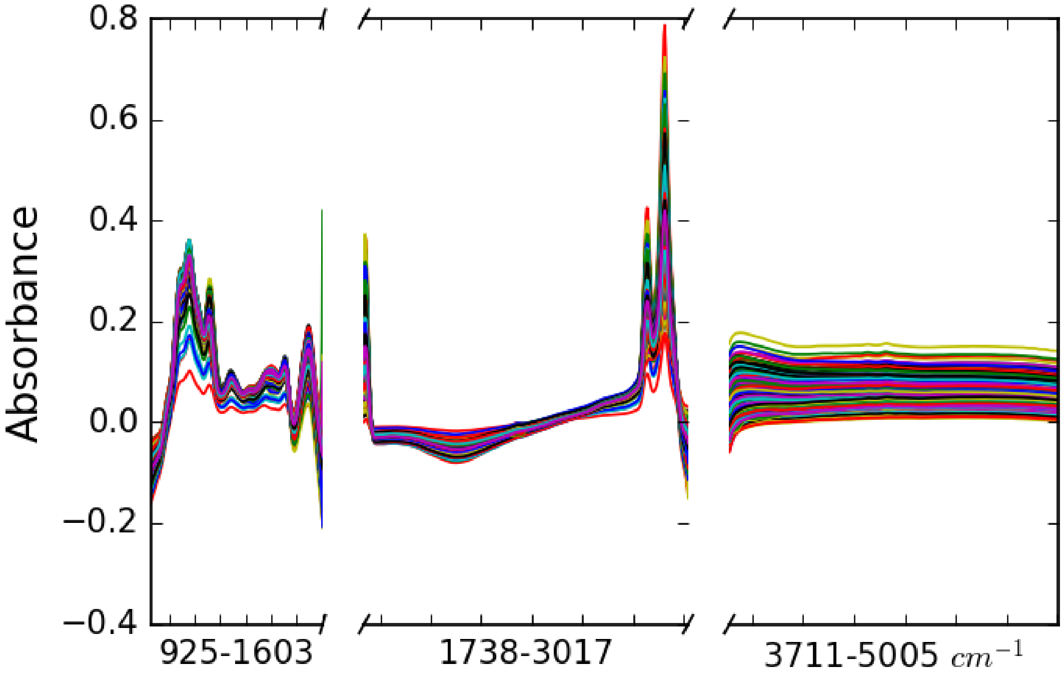


Fig. 3. Water-free MIR spectra (X) of 712 milk samples in the wave region $925 - 5005 cm^{-1}$. Pre-processing has reduced the feature-space dimensionality from 1060 wavenumbers to 847 due to removal of water absorption.

the principal components and the reconstruction error were used to quantify the quality of compression. We have used compression ratio as the final evaluation metric of our approach as it indicates computation, communication and storage performance of the analytics infrastructure. The notations used to represent different matrices, vectors, and values in MIRS dataset, PCA, WT and PLS algorithms are given in Table 1.

4.1. Pre-Processing of the MIRS Data

In spectrometry analysis, dissolved water adds unnecessary variability to the MIR spectra and could affect the resulting prediction accuracy. Most possibly, this effect is a random fluctuation or a systematic shift of the spectra. For instance, milk spectrum indicates two random sharp fluctuation regions, which occur in the wavenumber regions $1500 - 1800 cm^{-1}$ and $2900 - 3800 cm^{-1}$ per visual observation. Those regions are the water absorbance regions according to the pure water spectrum at $25°C$. In distributed analytics, we precisely identify those two regions based on PLS model calibration on our gold standard data and suggest these regions should be removed in the pre-processing stage before the compression. Therefore, based on our systematic identification of the two water regions, the corresponding wavenumbers can be removed from all raw MIR spectra in the measurement data domain.

In order to identify the water regions, we selected visually observable bare minimum water regions as $1464 - 1849 cm^{-1}$ and $2890 - 3814 cm^{-1}$ and removed these from X_{org}. Then we progressively recaptured one wavenumber at a time from the discarded regions to our predictors. In each step, the impact of the addition was quantified based on cross-validated root mean squared error ($RMSE_{CV}$) of the PLS predictive algorithm (explained in Section 5). The predictive error indicated a noticeable increase as the water absorbance regions began to be included in the prediction. Our finalised wave regions removed were $1607 - 1734 cm^{-1}$ and $3021 - 3707 cm^{-1}$. By removing the wavenumbers which were in the water absorbance regions, the dimensionality of water free spectrum (X) became 712×847 (i.e., $m = 847$), which reduced the amount of unwanted MIRS data by 20.1% and Fig. 3 represents the water absorbance regions removed spectra. The pre-processing stage could precisely remove the wavenumbers from the original spectra to obtain X, which is then fed into the compression stage.

In order to reliably develop our prediction model, we applied pre-treatment processes to the gold standard data. We mean-centred and scaled the values of X so that the mean and the standard deviation (SD) of each wavelength was 0 and 1, respectively. Scaling was not a compulsory approach in MIRS data since all the features were in the units of absorptions. However, this standardization could avoid confusion when using widely available machine learning libraries. We verified the normality of each response variable using *Shapiro* similarity check as a pre-requirement for applying PLS regression [34]. Outliers in Y, which were identified as when the difference between the value and its mean is more than three times the SD of a target variable, were removed from the data. Gold standard data (i.e., Y) were not available for all the samples and were therefore not considered if missing. After applying these pre-treatments, the final number of samples used for fat prediction was 701 and for lactose, protein and urea was 704.

4.2. *Compression with Principal Component Analysis*

The existence of strong correlations among the feature vectors makes MIRS predictions unamenable to simple analytic techniques like multiple linear regression due to matrix singularity (n=712<m=847) and multi-collinearity (correlations among feature variables), which could contribute to over-fitting. To overcome these issues, mostly PCA has been used for dimensionality reduction in MIRS, while WT has been used for de-noising [39]. However, both techniques can also be used for de-noising as well as for dimensionality reduction. In addition, PCA can particularly be useful as a data visualization tool and for feature extraction while WT can be used as an accelerating tool for efficient feature extraction by PCA. Therefore, the order of applying the techniques in a resource-constrained distributed computational infrastructure is important, but this has not been a concern for users of MIR spectroscopic analytics in the past.

Application of PCA in most of the higher dimensional data studies was variance based [45]. Feature vectors, which explain a significant portion of the variance in the original data (motive to capture only significant information as possible), are extracted based on the correlations among the different predictive variables (columns of X). Once a certain number of PCs are selected, this forms a low dimensional subspace of data such that every selected component is orthogonal to the other with minimum loss of information. Because of neglecting components, which contribute little to explaining the variability in the data, this concept has been used for dimensionality reduction of multi-dimensional data in a vast range of applications [19].

From a mathematical point of view, suppose $n < m$ in the feature matrix $X_{n\times m}$, where n and m are integers. PCA computes a new set of transformed variables called principal components (PCs) as linear combinations of the original variables. The first PC ($PC1$) accounts for the largest possible variance in X while the second PC ($PC2$), which explains the second largest variance in X, is computed to be orthogonal to $PC1$. The third PC ($PC3$) is derived to be orthogonal to both $PC1$ and $PC2$. The remaining PCs are computed in the same way and the transformed values of these PCs are called scores. The total number of PCs that can be generated from X is the minimum of n and m. In our MIRS data, $PC1, PC2$ and $PC3$ respectively explained $6.9\%, 5.6\%$ and 4.8% of the variability in X_{org}, and $68.5\%, 23.0\%$ and 4.9% of the variability in X. The Singular Value Decomposition (SVD) technique was used in the present study to compute PCs [19].

The new feature space of X ($G_{n\times l}$ - compressed domain data) is formed by selecting the columns from G which correspond to the first l largest singular values in D. The value of l is decided upon based on a threshold of the cumulative variance of PCs. The coefficients of the linear combinations are contained in $P_{l\times l}$, which we need to transform $G_{n\times l}$ back to the original domain X' (i.e. when the column average vector a of length l is provided).

We used the R package *pls 2.6-0* [29] which was developed to calculate the PCs from our MIRS dataset. PCA of X_{org} and X gives 712 PCs, which is the minimum of $n = 712$ for both 1060 or $m = 847$. The proportion of variance accounted by the different number of PCs were studied and also the loss of information from recovering the original data from those PCs were quantified by using the reconstruction error for both X_{org} and X. Fig. 4 (left) shows the cumulative percentage of variance from 99% onwards explained by PCs of both X_{org} and X. Fig. 4 (right) also

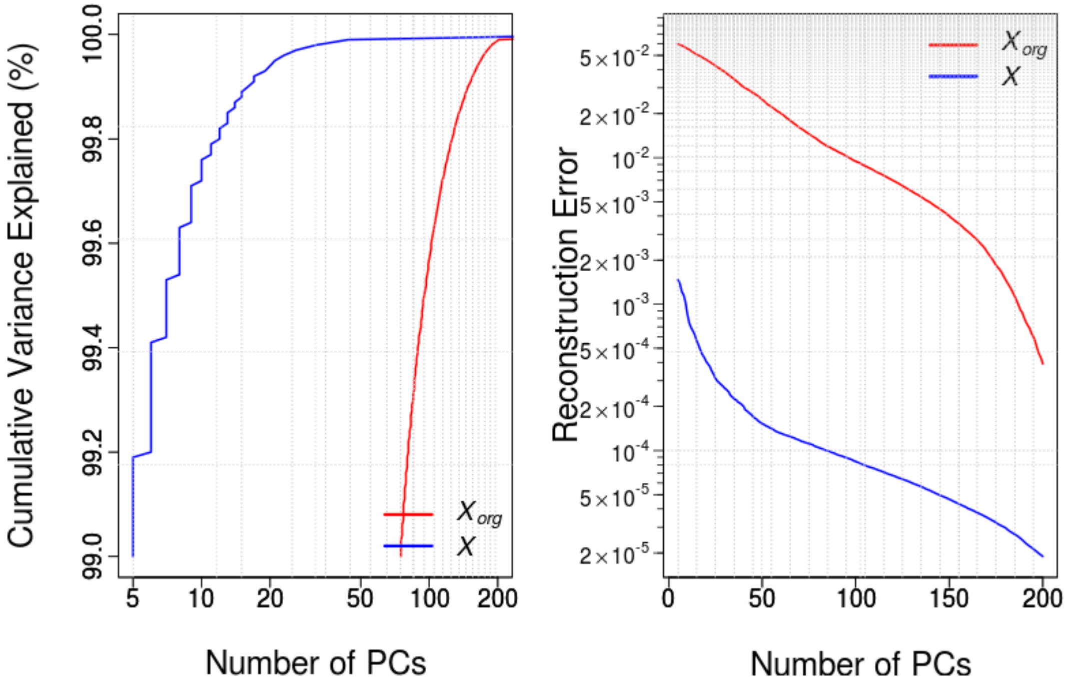


Fig. 4. Cumulative variance explained and the reconstruction error at different number of PCs of the original (X_{org}) and water removed (X) spectra.

shows the reconstruction error ($\sum_{i=1}^{n}(\sqrt{\sum_{j=1}^{m}(\delta x_{i,j}^2)/m})/n$) at different PCs, where $\delta x_{i,j}$ is the difference between the original value (X) and the reconstructed value (X') of a data point for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m$. For example, the number of PCs needed to explain 99.9% of total variance were 145 and 20 for the original and water removed spectra, respectively. Then the dimensions of the compressed domain data of X can be reduced to 712×15 providing a compression ratio (defined as $\frac{c}{m} \times 100$, where $c = l, r$, for the rest of the paper) of 98.2% having a reconstruction error of 3.05×10^{-4}. The cumulative variance explained by the PCs of X was above 99.99% and its increment was less than 10^{-3} after the first 100 PCs. Therefore, the optimal value for l was selected as 100 with minimum loss of information (0.01%) in our analysis. The reconstruction error corresponding to the first 100 PCs of X was 6.28×10^{-5} which guaranteed that the amount of information loss was small (the compression ratio was 85.96%).

Fig. 4 shows that PCA can significantly reduce the dimensionality of the MIRS data at different accuracy levels. The results also show that the water-related wavelengths contribute a significant amount of variability in the dataset, which should be removed based on our concluded wavenumber regions prior to compression. Hence, a significant amount of communication and computation energy can be saved for the benefit of future Fog and big data analytics. PCA-compressed data also minimizes over-fitting where the compressed domain data ($G_{n\times l}$) can directly be used in subsequent linear regression models.

However, our presented PCA compressed data may not have removed high frequency noise, while Wavelet compression in the next section can remove such noise in MIRS data. Since PCA is an unsupervised learning approach, it only accounted for collinearity among feature variables. However, in most of the real-world datasets, including our MIRS data, collinearity between response and feature variables also exist. In such situations, supervised dimension reduction techniques can be used and the optimal number of PCs required to generate a stable prediction model might be further reduced as shown in Section 6.

4.3. Compression with Wavelet Transformation

WT can be applied to a single or any finite group of spectra and analysed on any scale with orthogonal basis functions [44]. Every basis function consists of two types of functions : 1) wavelet function (mother wavelet), which is a high-pass filter capturing sharp behaviours (called details), and 2) scaling function (father wavelet), which is a collection of scaling functions capturing more general behaviours (called approximations) and act as a low-pass filter. In general, the data passes through these two filters and then generates approximate and detailed signals at a certain scale (q). The outcome of the high-pass filter is taken as Wavelet Coefficients (WCs), representing high frequency components. When the scale is higher, WCs are increased while the Scaling Coefficients (SC) are reduced. The number of filtering steps might deteriorate the transformed signal and may affect reconstruction (de-compression) after a certain scale [23], which we need to select for our MIRS data compression.

When selecting a basis function, the important properties to be considered are orthogonality of basis, preservation of data sparsity, independence between wavelet coefficients, and easiness in the reconstruction of the signal. Since there are different types of basis functions such as Haar, Symmetric and Daubechie, selecting an optimal basis is an important factor in WT. For instance, Haar wavelet is not suitable for the description of smooth functions; instead we used Daubechie-4 in our evaluations with the most commonly used WT, which is Discrete Wavelet Transform (DWT) [42].

Let x be a signal (e.g. a spectra) from X of length m. First we apply zero padding (which may sometime cause a considerable edge effect which linear padding minimizes [2]) to extend the array of $n = 847$ to 1024, which is the nearest 2^{10} format of our dataset (to apply WT, the signal length must be of the form 2^b, where $b \in \mathbb{Z}^+$). DWT was applied on a Daubechies-4 wavelet basis for different number of scales where the maximum was 10. The elements which are less than a selected threshold (λ) were regarded as noise (insignificant information) and removed from the transformed signal. According to [2], there are many thresholding methods such as universal, hard, and soft, but we used the soft thresholding approach. We then obtained our compressed domain data matrix $C_{n\times r}$. The indexes of the selected components are required to reconstruct the original data. In reconstructing the original signal x, we replaced all the removed positions with zero and applied the Inverse DWT for the same numbers of q. We used Multi Resolution Analysis (MRA) [43], which is a simple, fast and easily illustratable DWT method.

In general, MIR spectra contain high and low frequency signal components. The signals that have frequencies above a certain level are considered as noise components. We used the R package *wavelets 0.3 in our MRA based DWT* [6]. Wavelet transform was applied on X and the coefficients were retained, which has the dimensions of 712×1024. The number of SCs and WCs are shown at the 4^{th} scale in Fig. 5 for a single spectrum. We used a threshold (λ) of 0.01 to compress the spectrum at this level discarding insignificant components. According to this threshold, scaling and wavelet coefficients of 53 and 74, respectively can be selected. These components need to be stored as key-value pairs at the compression stage to use in compressed learning.

The optimal number of scale levels (q) and threshold (λ) are the main parameters required for selecting the most significant WCs in WT. Therefore, the behaviour of the number of significant WCs were experimented with

[6] http://CRAN.R-project.org/package=wavelets

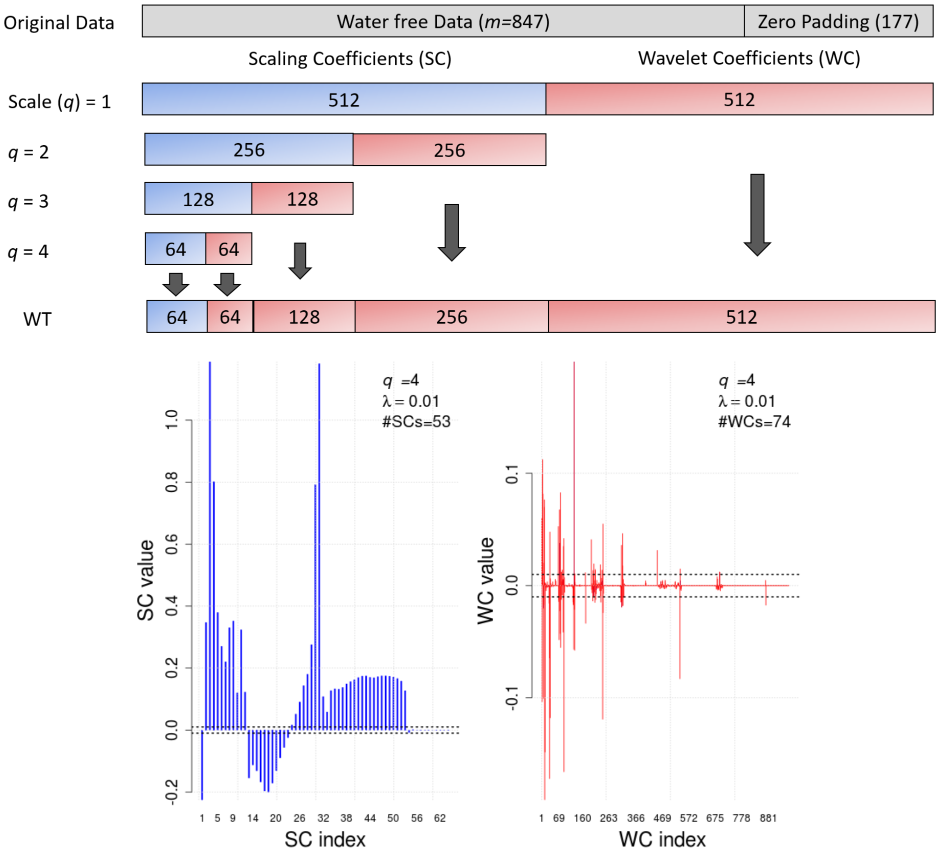


Fig. 5. Distribution of SCs and WCs at the 4^{th} scaling ($q = 4$) of a single water-free spectrum of our MIRS data using 'Daubechies-4'. Threshold (λ) of 0.01 indicate that the spectrum compresses to 127 components.

by changing the values of q and λ. The first graph of Fig. 6 shows the variability in the number of WCs at different scales under different threshold values (exponentially selected between 0.0012 and 0.02). For simplicity, we refer to the number of WCs as the sum of both scaling and wavelet coefficients at a certain threshold in the rest of the paper. According to Fig. 6, the number of coefficients is high (but with lower reconstruction error) for small thresholds. Increasing q up to the maximum possible scale is not required. Fig. 6 shows a saturation behaviour at the number of WCs after the 6^{th} scale, which we will use in Section 5. Therefore, WT does not capture any high or low frequencies of spectra after this point in our MIRS data. For example, at a threshold of 0.01, our data can be compressed to 127 coefficients with a 1.9×10^{-3} reconstruction error. To select an optimal value for r, the reconstruction error was computed for different WCs using the scale and threshold values of 6 and 0.0025. The second graph in Fig. 6 represents the behavior of reconstruction error. The reconstruction error of X was almost saturated (the error change was less than 10^{-3}) after 200 WCs. Therefore, the optimal r value was selected as 200 with the reconstruction error of 8.2×10^{-4} and a compression ratio of 71.91%.

Our MIR spectra can be considerably compressed while keeping most of the critical information and discarding most of the unnecessary information both using PCA or WT techniques. This concludes therefore that spectra can be transformed into their compressed domain and can be recovered with minimal error, if necessary. However, our results show that PCA required a fewer number of components than the required number of coefficients in WT to achieve a similar reconstruction error. The next section will investigate the impact of our compression on the PLS prediction accuracy of four different milk traits and hence derive our generalized/harmonized compression parameters (l and r) for compressed learning.

5. IMPACT ON PREDICTION ACCURACY BY COMPRESSED LEARNING

This section investigates the impact of compression parameters on compressed learning performances. First, we study the impact of l and r on the learning performances derived from a supervised compressed learning approach and second, we select optimal parameter values based on their impact on the learning performances. We apply PLS, which is commonly used for analyzing MIRS data [1, 27, 10], on the compressed MIRS data (i.e. PCA scores $G_{n\times l}$ and Wavelet-transformed data $C_{n\times r}$) to quantify how much the predictive accuracy is impacted by PCA and WT based compressions. At different compression levels (i.e., varying l and r), prediction performance in model calibration and external validation using compressed data is compared with the data in the uncompressed mea-

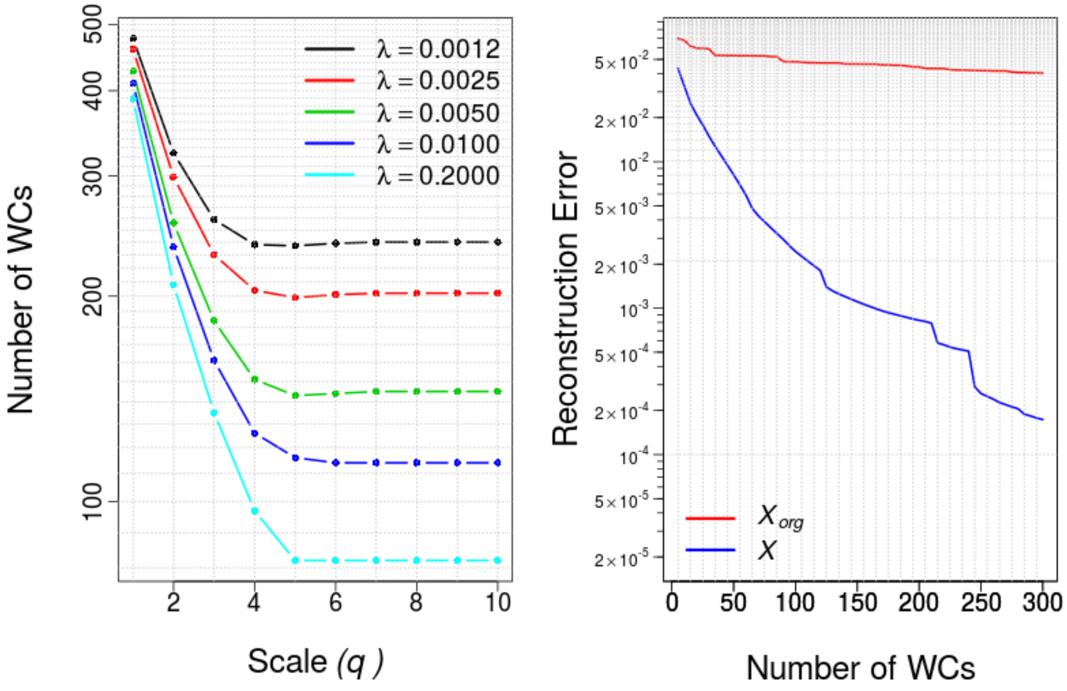


Fig. 6. Number of significant WCs at different thresholds (λ) and different scales (q). The number of coefficients saturate at scale 6. Reconstruction errors in WT are higher when compared to PCA.

surement domain (X). The following indexes for the regression model have been used to evaluate the compressed learning performances.

The root mean-square error ($RMSE$) quantifies the standard deviation of the residuals (between the real and the predicted response variable Y) and is shown in the units of absorbance. The coefficient of determination (R^2) depicts the proportion of variance in the response variable Y explained by the predictor variables in X. These measures are computed by the following two equations (subscript i - real response value and p - predicted value).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - y_p)^2}{N-1}}, \quad R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - y_p)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

The Ratio Performance Deviation (RPD) represents the practical utility of the model, and is calculated as $(1 - R^2)^{-1/2}$. As a rule of thumb, if $RPD > 3$, then the model can be used for practical analytical purposes.

All performance indexes are calculated for both the calibration (c) and external validation (prediction) (p) data segments of our gold standard MIRS data. Based on these evaluations, near lossless compression parameters l and r for PCA and WT, respectively are derived for each milk trait. We have selected four of the most used milk quality parameters: lactose, fat, protein and urea, all derived from milk MIRS [27].

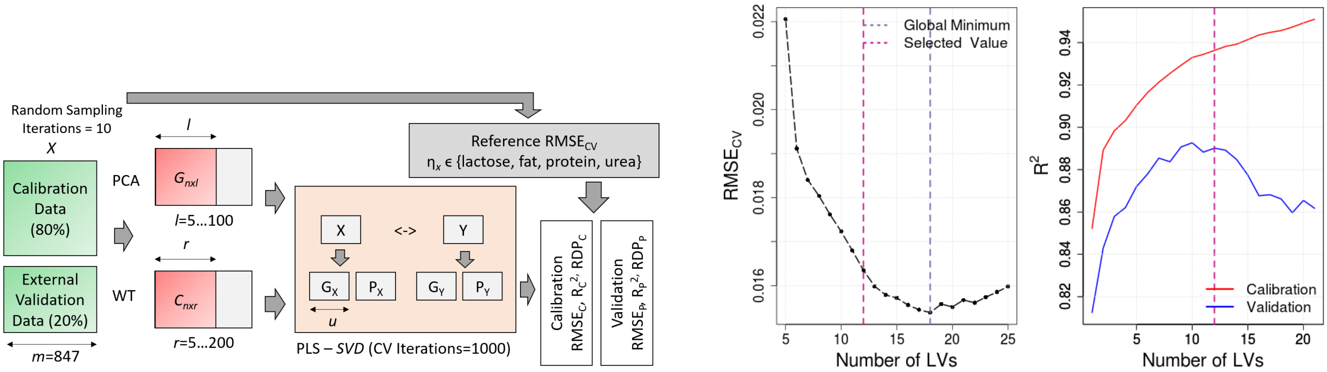
5.1. *Partial Least Squares (PLS) Regression*

PLS is a projection method that models the relationship between the predictors X and responses Y (a.k.a. Projection on Latent Structures) [9]. The PLS method considers not only the correlations among the predictor variables in X, but also the correlations each predictor in X and the response in Y. The general procedure of PLS is somewhat similar to when dimensionality reduction of PCA is combined with Least Squares Regression (LSQ), which is called as PCR. However, PLS and PCR differs mainly in the methods used in extracting factor scores. PCR produces a loading matrix P reflecting the covariance structure among the predictor variables. PLS produces a loading matrix P reflecting the covariance structure between the predictor and the response variables [9]. The set of significant components in PLS is called the Latent Variables (LV). PLS decomposes both X and Y using SVD.

Fig. 7(a) shows the logical overview of the essential steps that we have followed in this section to derive l and r. The sub-sampled training dataset (model calibration) is selected randomly having 80% of the total n samples. The remaining set of samples is used for testing the model (external validation). To increase the validity of model performance, we repeat the above process for 10 different data selections while keeping the same ratio for training and test data partitions. We have selected samples randomly (from $n = 712$) under each iteration and the average of performance measures has been calculated.

5.2. *PLS Accuracy using uncompressed MIRS Data*

First, we calculated PLS accuracy with the MIRS data in the measurement domain (X). This accuracy ($RMSE_{CV} = \eta$) was used as the reference to estimate our near lossless compression parameters. Different compression param-



(a) Overview of the different stages when PCA and WT compressed data is applied with the PLS regression (Green and pink colours respectively represent the data in the measurement and the compressed domains)

(b) Determination of the optimal number of LVs at an optimal $RMSE_{CV}$ for lactose using uncompressed data. The optimal value is selected not to exceed $RMSE_{CV}$ of 0.001 from the absolute minimum

Fig. 7. Overview of the PLS learning procedure from PCA and WT and selection of LVs from PLS calibration for building predictive models

eters; l_x and r_x where $x \in \{lactose, fat, protein, urea\}$, were derived by fitting a PLS model on the compressed data for the four selected milk parameters. We achieved the prediction performance using compressed data to be comparative with the reference model performance (η_x). The selection procedure of the compression parameters is explained only for lactose but the same procedure was followed for fat, protein and urea and the summary is given.

First, a PLS regression model was fitted to the training data in the uncompressed domain (X) and we obtained the cross-validated mean of $RMSE_{CV}$ by changing the number of LVs in the PLS model to select the minimum error at an optimal number of LVs (u). However, in this process, u was selected as the LV corresponding to the $RMSE_{CV}$, which did not make a considerable difference (p-value ≤ 0.001) to the global minimum of $RMSE_{CV}$ (i.e., LV corresponds to the selected $RMSE_{CV}$ such that the difference between the selected and the global minimum $RMSE_{CV}$ is not greater than the p-value). u has been selected (as explained above) in our evaluations according to a permutation model explained in [29] and a 10-fold cross validation, followed 1000 iterations, for selecting each LV.

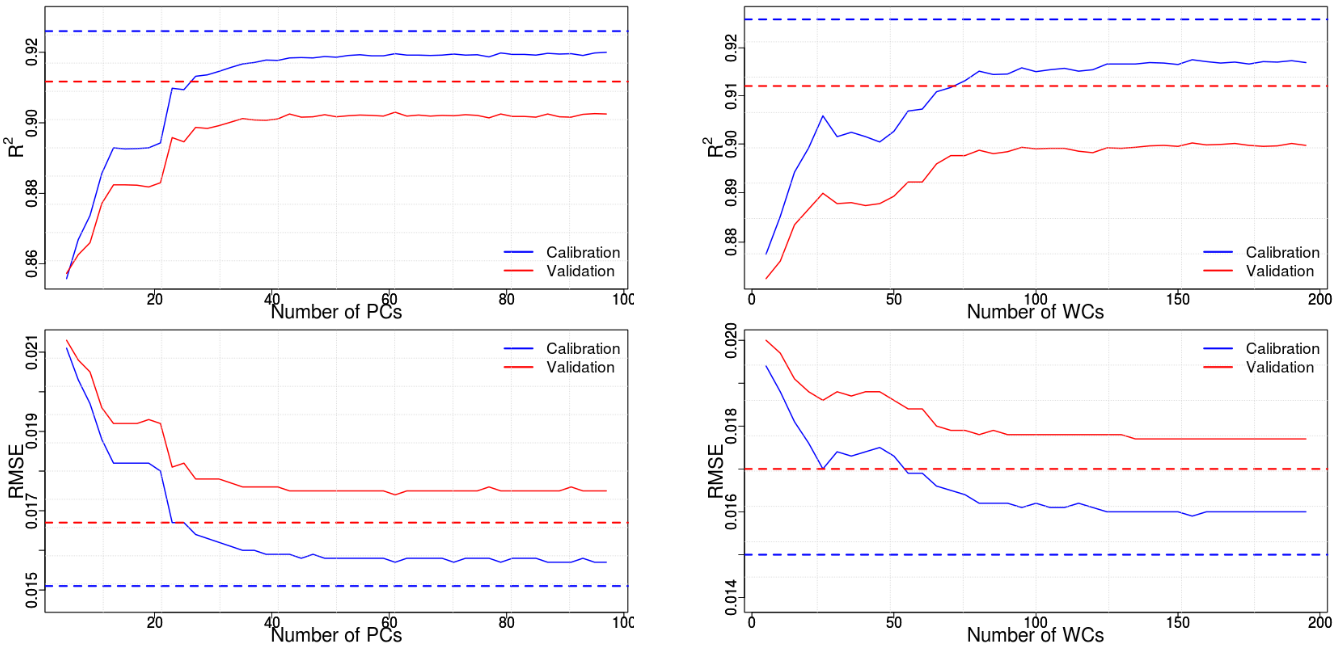
According to Fig. 7(b), the optimal $RMSE_{CV}$ of 0.0154, is achieved with 12 LVs for lactose ($u_{lactose} = 12$), when the water-removed spectra (X) were used. Twelve LVs were selected as the optimal number of LVs even though the absolute minimum of $RMSE_{CV}$ occurred at 17 LVs. The graph on the right of Fig. 7(b) provides performance statistics of the PLS model with a comparison of calibration and external validation statistics based on R^2. In Fig 7(b), the performance indexes of the calibration and external validation values do not change much beyond the selected optimal LV point (after the dashed line). Therefore, the optimal PLS model can reliably be derived with 12 LVs for lactose.

Table 2 presents the optimal $RMSE_{CV}$ and minimum LVs we can achieve with the measurement domain data for all the four different milk traits we have selected. Statistics in Table 2 indicate a well performed regression models for lactose and fat, because the R^2 were $> 87\%$ for both models and the RPD was ≥ 3 for lactose and close to three for fat in both the calibration and validation. The regression models of protein and urea content were not as good as the lactose and fat regression models, because R^2 and RPD values were only $> 73\%$ and ≥ 2 in the both the calibration and external validation.

Table 2 also shows that we can achieve a 12.7% improvement in $RMSE_{CV}$ concurrent with a 20.1% compression by just removing the water-related wavelengths of the spectrum during the pre-processing stage of our compressed learning. Then we conducted the same PLS procedure using PCA and WT compressed data by changing compressed dimension parameters l and r .

Table 2
PLS model performance on the original (X). Our near lossless PCA and WT compressions find optimum number for l and r according to these reference values.

	Reference Values		Calibration			External Validation		
Milk Trait	#LVs (u)	$RMSE_{CV}$ (η)	$RMSE_c$	R_c^2	RPD_c	$RMSE_p$	R_p^2	RPD_p
Lactose (X_{org})	5		0.0173	90.25	3.2152	0.0190	88.53	3.0379
PLS model performance for water removed Spectra (X)								
Lactose (X)	12	0.0154	0.0151	92.60	3.6929	0.0167	91.17	3.4578
Fat	5	0.0892	0.0865	88.49	2.9540	0.0919	87.35	2.8607
Protein	4	0.0601	0.0574	76.12	2.0570	0.0625	73.61	2.0461
Urea	15	0.3443	0.3098	81.55	2.3350	0.3523	77.64	2.1428



(a) PCA

(b) WT

Fig. 8. Compressed domain PLS performance at different number of PCs and WTs for lactose. The dashed line represents the optimal PCA (a) and WT (b)

5.3. *Impact on PLS Accuracy with PCA compression*

The performances of the PLS model were computed by changing the number of PCs, $l = 5, \cdots, 100$ with a step of 5 PCs. The results on predicting lactose are given in Fig. 8(a). During the evaluations, the score matrix ($G_{n \times l}$) at different selected number of PCs was applied as the compressed domain input to the PLS model. $RMSE_{CV}$ of cross-validation were compared with the reference accuracy of lactose ($\eta_{lactose}$), which is given in the Table 2.

PLS calibration and validation accuracies using PCA compressed data decreased as the number of PCs increased. With 45 PCs, it shows a similar minimum $RMSE_{CV}$ compared to the reference PLS accuracy of 0.0154. Adding more PCs after 45 PCs into to the model did not make a significant contribution to improve the model performance (i.e. the impact of l on lactose predictive model is up to a certain value only). Thus, the results reveal that the PCA compression with at least 45 PCs is stable. Therefore, we conclude that the optimal compression level can be achieved with 45 PCs for lactose prediction ($l_{lactose} = 45$) resulting in a compression ratio of 94.7%.

Results in Table 3 shows the optimal number of PCs required to predict all the milk traits using PLS. These models were derived in the similar way to as described for lactose. Moreover, different milk traits have their own optimum number of PCs; $l_{lactose} = 45, l_{fat} = 30, l_{protein} = 37$ and $l_{urea} = 65$. Therefore, with respect to each trait, the water-removed spectra can be compressed by $94.7\%, 96.5\%, 95.6\%$ and 92.3% for lactose, fat, protein and urea, respectively using PCA at the compression stage.

5.4. *PLS Accuracy with WT compressed Data*

The same procedure of PLS regression as explained in the previous section for PCA compression was applied for the WT compressed data. In this case, the PLS was applied to the WT coefficient matrix $C_{n \times r}$ by changing the number of WCs, $r = 5, \cdots, 200$ with a step of 5 WCs. Fig. 8(b) shows PLS prediction performance for lactose and the regression model with 100 WCs indicates an $RMSE_{CV}$ close to the data domain accuracy of $\eta_{lactose}$. Therefore, the optimal compression was achieved using 100 WCs for the prediction of lactose ($r_{lactose} = 70$). In addition, the behaviour of the impact of r was also similar to the behaviour which was obtained with PCs in Fig. 8(a).

Table 3
PLS model accuracies for the selected milk traits at optimal PCA compressed points. Optimal number of PCs has been selected based on 0.01 $RMSE_{CV}$ threshold from the absolute minimum. Optimal $RMSE_{CV}$ has been tallied to reference η.

Milk Trait	#PCs (l)	#LVs (u)	Calibration			External Validation			Reconstruction Error
			$RMSE_c$	R_c^2	RPD_c	$RMSE_p$	R_p^2	RPD_p	
Lactose	45	12	0.1580	91.85	3.5152	0.0175	90.16	3.3018	1.266×10^{-4}
Fat	30	5	0.0871	88.30	2.9386	0.0918	87.41	2.8537	1.860×10^{-4}
Protein	37	4	0.0577	75.91	2.0474	0.0627	73.50	2.0448	1.554×10^{-4}
Urea	65	15	0.3334	78.63	2.1687	0.3705	75.07	2.0351	0.928×10^{-4}

Table 4
PLS model performance for different milk traits for Wavelet compressed data. Optimal number of PCs has been selected based on 0.01 $RESE_{CV}$ threshold from the absolute minimum. Optimal $RMSE_{CV}$ has been tallied to reference η.

Milk Trait	#WCs (r)	#LVs (u)	Calibration			External Validation			Reconstruction Error
			$RMSE_c$	R^2_c	RPD_c	$RMSE_p$	R^2_p	RPD_p	
Lactose	70	12	0.1650	91.31	3.3803	0.0178	89.86	3.2284	4.8×10^{-3}
Fat	40	5	0.0864	88.49	2.9682	0.0920	87.37	2.8455	4.3×10^{-3}
Protein	45	5	0.0575	76.08	2.0549	0.0625	73.72	2.0432	9.5×10^{-3}
Urea	75	15	0.3333	78.64	2.1700	0.3730	74.95	2.0239	3.9×10^{-3}

Table 4 shows the prediction performance in the WT compressed domain for all the selected milk traits. Different milk traits had their own optimum number of WCs; $r_{lactose} = 70, r_{fat} = 40, r_{protein} = 45$ and $r_{urea} = 75$. WT can compress MIRS data by 91.7%, 95.3%, 94.7% and 91.1% for lactose, fat, protein and urea, respectively.

PLS regression models focused on finding an optimum level of compression (optimal l and r) for our MIRS data based on either PCA or WT. We validated the near lossless compression using its impact on PLS regression-based learning accuracies for the different milk traits. Therefore, transformed data can be used to learn in their compressed domain. Both PCA and WT compressions had similar compression performance. Based on the four milk quality traits we selected, the number of PCs in a general PCA compression (l) and the number of WCs in a general WT compression (r) should have at least $l = 65$ and $r = 75$ components (i.e., 92.3% and 91.1% compression can be achieved from PCA and WT, respectively). Therefore, selection of the largest number of PCs and WCs is the requirement to preserve the predictability of urea without losing any information on the investigated milk traits.

6. DISCUSSION

6.1. *Sample size selection of PCA and WT*

Real-time data transfer always consumes greater energy and is not used in many agricultural infrastructures. Instead delay-tolerant networks and data logging systems are mostly used [20]. Therefore, the MIRS source can collect a certain number of spectra before data compression and transmission takes place (e. g. in robotic milking cows are milked in every 7-10 minutes by a single machine). If the delay is large, some extra memory space is needed to store the spectra until data are compressed and later transmitted. However, there can also be cases where in-situ milk quality (online) monitoring is used by the dairy industry. In this case, time becomes a critical factor and WT should be used for compression instead of PCA.

The sample size (n) plays an important role in PCA-based compression since fewer samples create instability in PCA. The general understanding is that the larger the sample size, the better the stability. Selecting an adequate sample size for our MIRS data is a compromise for timeliness of decision making. There are no simple rules to determine the appropriate sample size for PCA. The variability in reconstruction error with respect to sample size was examined with PCA and WT compressions for our MIRS data X. According to Fig. 9(a), WT using our recommended number of WCs, does not improve reconstruction error as the number of spectra available increase. PCA using the recommended number of components can improve reconstruction error by increasing the number of samples. At a certain point beyond 190 samples, PCA has less reconstruction error than WT.

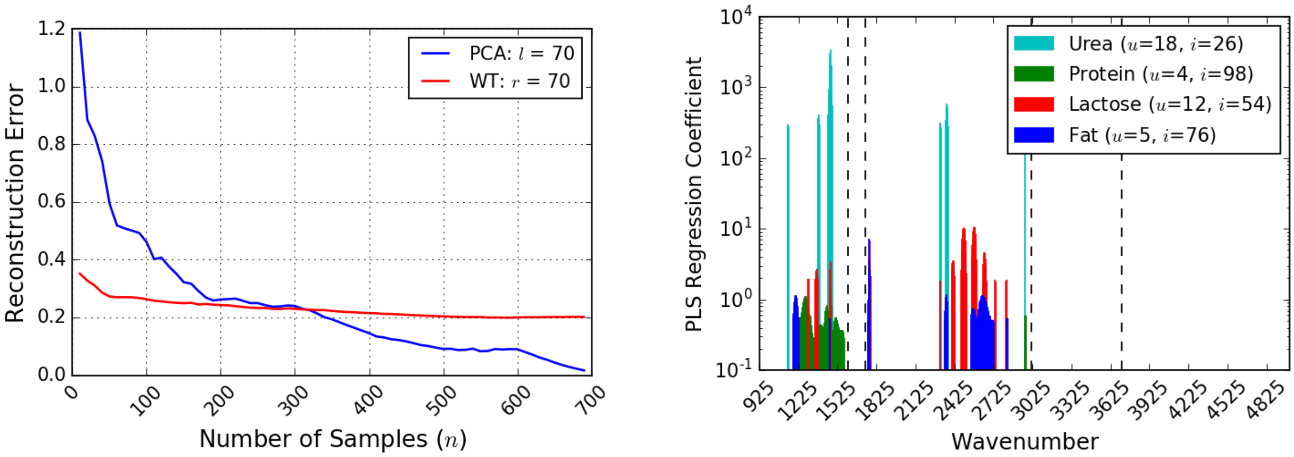
6.2. *Customization using supervised compression*

Standard PCA does not know what portion of variance in each variable is important and should be preserved. Sufficient application knowledge with intended milk traits (supervisory learning) can further optimize our compression performance. PLS can be used in supervised compression only using LVs (compressed domain) or using significant wave indexes in a linear model (measurement domain). We analyzed the composition of each milk quality trait within the spectrum using PLS (Fig. 9(b)). The results show that a customized approach can be applied at the compression stage or on top of our generally compressed data using PCA or WT to further improve our compression performance.

As an example, a farm decision support tool may need to identify only the fat and protein content of certain milk samples [28] to quantify cow-level energy balance in the herd. Such a customized system can further compress MIRS data beyond our unsupervised compression techniques, when data are transferred between the Fog nodes or into the big data systems.

6.3. *Impact on advanced analytics*

Linear PCA assumes that the original data can be converted into a single scale and the relationship between the orthogonal PCs are linear. However, these assumptions are not always true with real data. For instance, categorical data consists of ordinal and nominal variables, which is not easy to convert onto a single scale. Hence, PCA compression could possibly lose significant information, due to multi-scale data with non-linear behaviour and correlations. If data have non-linear behaviours, linear PCA may inadequately capture significant variances.



(a) Batch-based compression of PCA and WT selecting different number of samples ($n = 10 \cdots 700$)

(b) Significant PLS coefficients (higher than the stand deviation) at the optimal number of LVs.

Fig. 9. Compressed Learning with sample-size sensitivity and supervised compression

According to [24], non-linear PCA overcomes not only these issues, but also facilitates the application of PCA without changing the existing scales. Even though some PCs capture very little variance from the data, those PCs may represent substantial information. Therefore, PCA variants such as kernel PCA may solve some of these difficulties in linear PCA, where compressed learning with MIRS needs further investigation.

PCR and PLS predictive methods are commonly used for statistical learning processes in spectrometric analytics. However, these methods fit a linear regression model. If compressed domain data presents a non-linear behaviour, those linear models would not contribute to derive best fit stable predictive models. Use of linear models may create a negative impact on the robustness and accuracy of the learning process. Therefore, advanced methods such as Support Vector Machine (SVM) ([39]) and Artificial Neural Network (ANN) ([22]) are available (with the improvement of pervasive computational capabilities) and can be used to address non-linear behaviours in MIRS data subjected to the fact that we have preserved non-linearity in the compressed domain data.

6.4. A Comparison with State-of-the-art Techniques

We have compared PCA and WT compressed learning performances with deep auto-encoder (DEA) [48], LeNet-5, Vgg-19 , GoogLeNet, and ResNet [12], [40], using our MIRS data, all of which are emerging deep learning techniques. The LeNet-5, Vgg-19, ResNet, and GoogLeNet can be considered as the extensions of the LeNet model. These techniques are un-supervised and different forms of convolution neural network (CNN) models, which can also be considered as lossy compression techniques yet differ from the engineered compression techniques (e. g. JPEG, LZW). The PCA and WT are faster, simpler, and require less computational power, but considered only linear properties in the data. Whereas the deep learning techniques are much flexible and able to achieve more precise outcomes than PCA and WT based learning approaches by accounting for the non-linearity in the data. However, for instance, higher model complexity and computational requirements are the main implementation constraints in the deep learning approach. To overcome these issues, more advanced versions of CNN approaches have been proposed and the techniques mentioned above are a few of them.

The water-removed MIRS data was used for deep learning. Three encoding layers were used in the deep auto-encoder (DAE) model. The number of decoding layers was same as the number of encoding layers. PLS-based learning procedure was followed to quantify compressed learning performances as in Section 5. The compressed dimension was set to 70 as a middle compressed dimension to the highest feature variables (65-Table 3 and 75-Table 4), which were observed from PCA and WT based compression, respectively. Fig. 10 shows the LeNet-5, Vgg-19, ResNet, and GoogLeNet network models, and to apply these models to our data, each sample was re-sized as a $32 \times 32 \times 1$ matrix, applying zero padding. The convolution layers mostly have 1×1 and 3×3 filters and 2×2 Maxpooling filters. The convolution and pooling operations were performed in the intermediate layer (purple color box) and pooling was applied after the convolution. The red color box was removed from the intermediate layer when the same convolution was not repeated. We did not use a dropout layer before making the fully connected layers. Each model has two fully connected layers (second fully connected layer has 70 neurons) and the last dense layer is a regression layer. The network architectures given in [12] were followed to configure the Vgg$-$19 and ResNet models. Although the same convolution was repeated for six times in the ResNet model in [12], we did it only for four times. The solid and dashed lines in the ResNet model represent the shortcut connection with same and increased dimensions, respectively. When the dimension was increased with stride 2, zero padding and 1×1 convolution were used to match dimensions. Three inception modules (the inception

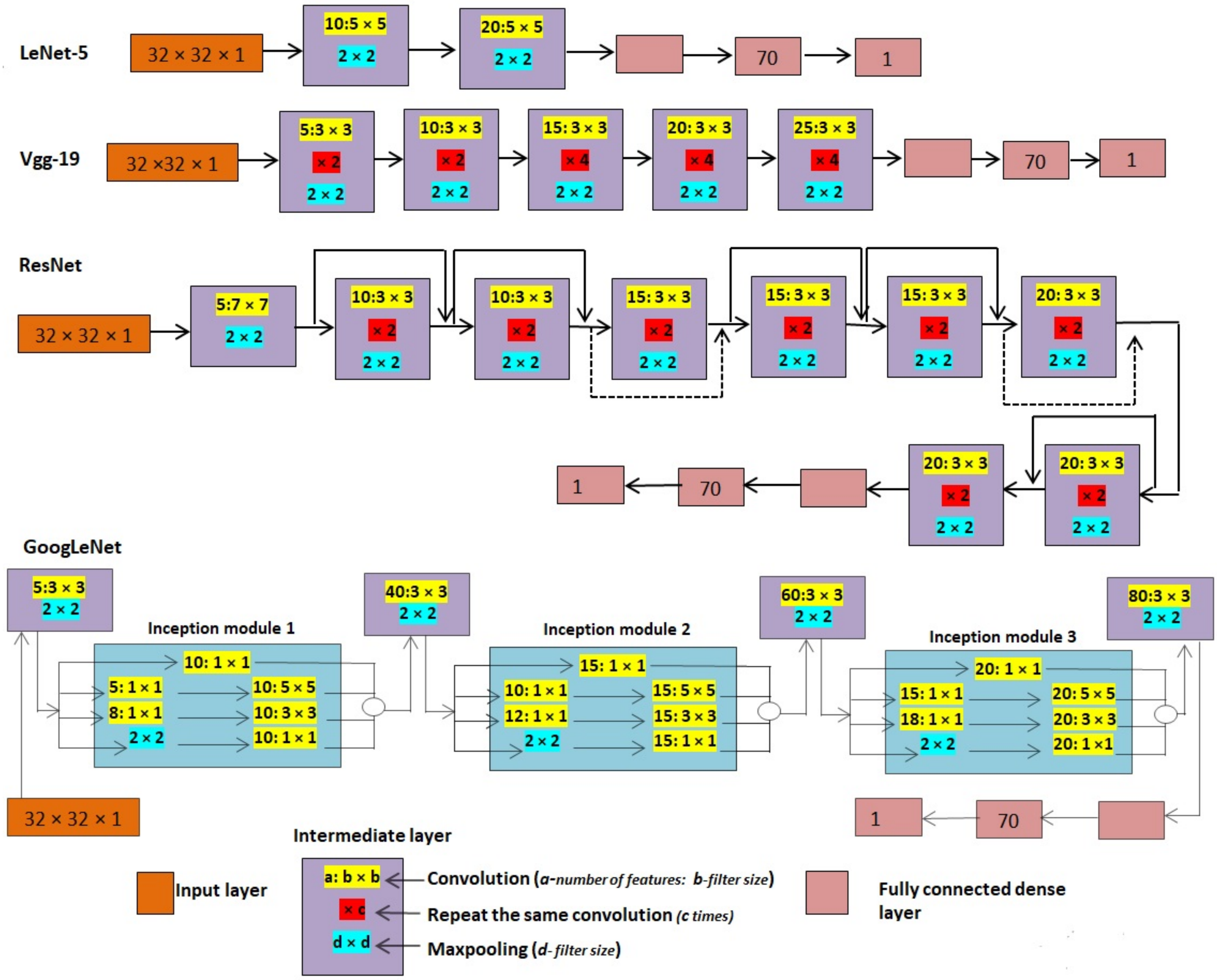


Fig. 10. Network architectures of four state-of-the art deep learning techniques LeNet-5, Vgg-19, GoogLeNet, and ResNet.

module with dimension reduction [40]) were stacked together to form the GoogLeNet model (for more details about these network configurations, please follow [12, 40]). Each model was trained for up to 1,000 iterations using the ADAM optimizer and mean squared error loss function. Also, we used a fixed learning rate of 0.01 and the rectified activation function. Finally, to compare the performance of these deep learning techniques with our compressed learning outcomes, the $RMSE_P$ was computed, applying all these techniques to each milk quality parameter.

The Fig. 11 shows the predictive learning accuracies from each deep learning model, including PCA and WT. The learning performances from all methods were approximately similar for lactose and an improvement was observed for fat, protein, and urea. This can be due to the existence of non-linear associations in the MIRS data, which has an impact on predicting fat, protein, and urea in milk [46]. The predictive accuracy also increased with the increasing depth of network models so that the ResNet model achieved the greatest accuracy. Due to the small data size, selecting a sufficient number of features in convolution, and over-fitting were the major issues when training these models. Therefore, learning performances may even improve further by using larger datasets with a comprehensive study of different factors such as data pre-processing, proper constraints, optimizers, and network design.

These state-of-the-art techniques can also be used for the compressed learning which we have discussed in this study and performed well compared to the traditional methods. However, employing them under some circumstances such as with low complexity and under limited computational resources may not be feasible for applications such as distributed data processing using Fog computing, which is one of our main concern in the smart farming industry. These limitations would be minimized by using the ResNet and GoogLeNet. The GoogLeNet model has the potential to control the computational cost required with deep networks so that, it can be used even with limited resources and low-memory requirements [40]. The ResNet model is easy to optimize and gain accuracy by increasing the depth and width of the network [12]. Although more reliable outcomes can be derived efficiently from these deep learning methods, further studies are essential to study the feasibility in employing these techniques in the smart dairy industry because the resources, such as computational infrastructure, energy, and lack of data might

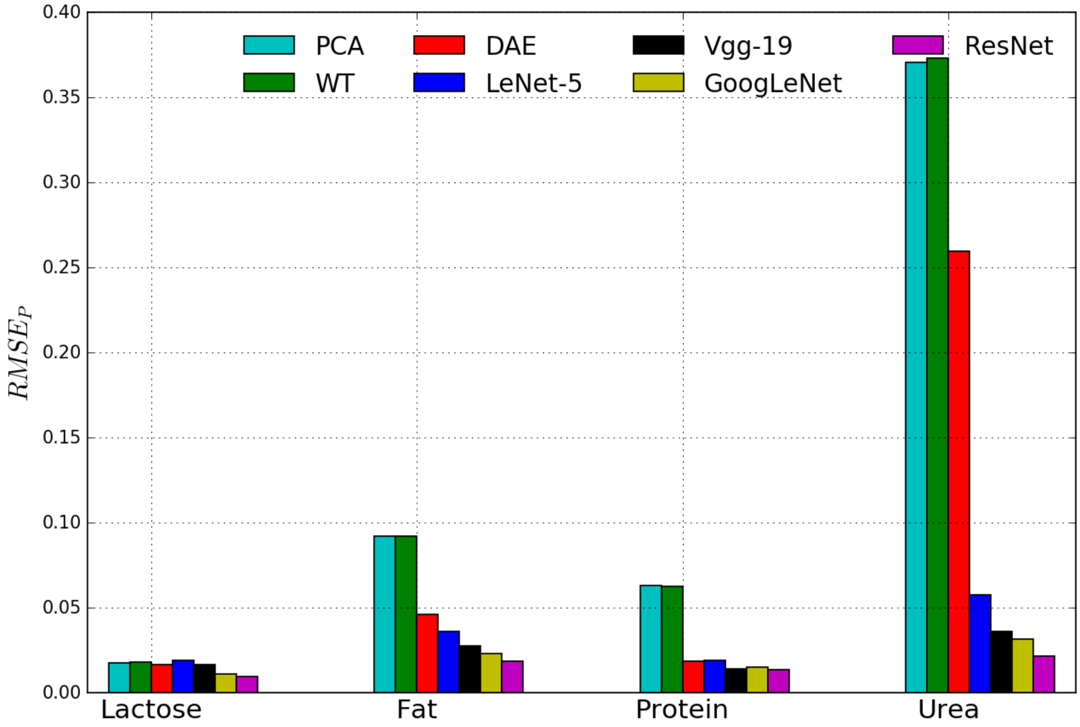


Fig. 11. Comparison of compressed learning performances of PCA and WT with four state-of-the-art deep learning methods LeNet-5, Vgg-19, GoogLeNet, and ResNet.

still be the major constraints to run these advanced algorithms. As we can see in Fig. 11, the learning performances from all methods are approximately similar for lactose, it may not necessary to apply deep learning for deriving a predictive model for lactose. Thus, performing an initial study to get an overall idea about the general characteristics such as non-linear associations in the original data would help to select the most suitable compressed learning approach. Consequently, we can optimize the utilization of available resources and obtain reliable outcomes in resource constraint environment such as Fog Computing.

7. CONCLUSIONS

In this paper, we have shown that MIRS data can be pre-processed and compressed effectively near the data source without impacting the prediction accuracy of most measured milk quality traits. PCA can generally be compressed to 65 principal components and WT can be compressed to 75 wavelet coefficients, which leads to compression ratios of 92.3% and 91.1%, respectively. At these compression levels, PLS using PCA and WT compressed data (i.e. 65 significant scores in PCA and 75 significant coefficients in WT) can achieve the same accuracy, as PLS can achieve using the pre-processed data in the original measurement domain. Therefore, the results show that the compressed learning with MIRS is highly advantageous both in Fog and big data processing, which can preserve communication and computation energy, minimize required memory and storage spaces, reduce application latency and preserve scarce rural network bandwidths.

ACKNOWLEDGEMENT

This research was supported by Teagasc and Science Foundation Ireland (SFI) through the project ”PrecisionDairy-*Using Precision Technologies, Technology Platforms and Computational Biology to increase the Economic and Environmental Sustainability of Pasture-based Production Systems* (ID: 13/1A/1977)”.

References

[1] Alexandratos, N., Bruinsma, J., 2012. World agriculture towards 2030/2050. In: Food and Agriculture Organization in the United Nations, EAS working Paper No 12-03. pp. 12–21.

[2] Artime, C. E. C., 2016. On-line estimation of fresh milk composition by means of vsi-nir spectrometry and partial least squares method (pls). In: IEEE Instrumentation and Measurement Technology Conference. pp. 1471–1475.

[3] Buyya, R., Mahapatra, C., Leung, V., Chen, M., Sahu, P., 2016. Fog computing: Internet of things realize its potential. IEEE Computer Magazine 49 (8), 12–116.

[4] Calrebank, R., Jafarpor, S., Schapier, R., 2009. Compressed learning: Universal dimensionality reduction and learning in the measurement domain.

[5] Cands, E. J., Wakin, M. B., 2008. Introduction to compressive sampling. IEEE Signal Processing Magazine 25 (2), 21–30.

[6] Donoho, D. L., April 2006. Compressed sensing. IEEE Information Theory 52 (4), 1289–1306.

[7] Duarte, M. F., Eldar, Y. C., 2011. Structured compress sensing: From theory to application. IEEE Transactions on Signal Processing 59 (9), 4053–4085.

[8] Elgohary, A., 2016. Compressed linear algebra for large-scale machine learning. VLDB Endowment 9 (12), 960–971.

[9] Garthwaite, P. H., 1994. An interpretation of partial least squares. American Statistical Association 89 (425), 122–127.

[10] Gunrdeniz, G., Ozen, B., 2009. Detection of adulteration of extra-virgin olive oil by chemometric analysis of mid-infrared spectral data. Food Chemistry 116 (2), 519–525.

[11] Guo, J., Song, B., Jian, F., Qin, H., 2015. Texture classification with cross-covariance matrices in compressive measurement domain. Signal, Image and Video Processing 10 (8), 1377–1384.

[12] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

[13] Jian, M., Lam, K.-M., Dong, J., 2011. Image retrieval using wavelet-based salient regions. Imaging Science Journal 59 (4).

[14] Jian, M., Lam, K.-M., Dong, J., 2013. A novel face-hallucination scheme based on singular value decomposition. Pattern Recognition 46 (11), 3091–3102.

[15] Jian, M., Lam, K.-M., Dong, J., 2014. Face-image retrieval based on singular values and potential-field representation. Signal Processing 100, 9–15.

[16] Jian, M., Lam, K.-M., Dong, J., 2014. Facial-feature detection and localization based on a hierarchical scheme. Information Sciences 262, 1–14.

[17] Jian, M., Lam, K.-M., Dong, J., 2014. Illumination-insensitive texture discrimination based on illumination compensation and enhancement. Information Sciences 269, 60–72.

[18] Jian, M., Lam, K.-M., Dong, J., 2015. Visual-patch-attention-aware saliency detection. IEEE tran. on Cybernetics 45 (8).

[19] Klema, V., 1980. The singular value decomposition: Its computation and some applications. IEEE Transactions on Automatic Control 25 (2), 164–176.

[20] Kulatunga, C., 2017. Opportunistic wireless networking for smart dairy farming. IEEE IT Professional Magazine 19 (2), 16–23.

[21] Kulatunga, C., Bhargava, K., Vimalajeewa, D., Ivanov, S., October 2017. Cooperative in-network computation in energy harvesting device clouds. Elsevier Journal on Sustainable Computing.

[22] Lancashire, L. J., Lemetre, C., Ball, G. R., 2008. An introduction to artificial neural networks in bioinformatics - application to complex microarray and mass spectrometry datasets in cancer studies. Briefing in Bioinformatics 10 (3), 315–329.

[23] Li, X., Luo, L., He, Y., Xu., N., 2013. Determination of dry matter content of tea by near and middle infrared spectroscopy coupled with wavelet-based data mining algorithms. Computers and Electronics in Agriculture 98, 46–53.

[24] Linting, M., 2007. Nonlinear principal components analysis: Introduction and application. J. Psychological Methods 12 (3), 336–358.

[25] Loung, H. V., 2017. Incorporating prior information in compressive online robust principal component analysis. In: arXiv:1701.06852.

[26] Lu, J., Verma, N., Jha, N. K., 2016. Compressed signal processing on nyquist-sampled signals. IEEE Transactions on Computers 65 (11), 3293–3303.

[27] Marchi, M. D., Toffanin, V., Cassandro, M., 2014. Invited review: Mid-infrared spectroscopy as phenotyping for milk quality traits. Dairy Science 97 (3), 1171–1186.

[28] McParland, S., Berry, D. P., 2016. The potential of fourier transform infrared spectroscopy of milk samples to predict energy intake and efficiency in dairy cows. Dairy Science 99 (5), 40564070.

[29] Mevik, B., Wehrens, R., 2016. Introduction to the pls package. In: R Project.
URL `http://cran.r-project.org`

[30] Notsu, A., 2012. Information compression effect based on pca for reinforcement learning agents' communication. In: International Symposium on Advanced Intelligent Systems.

[31] Paskov, H. S., West, R., Mitchell, J. C., Hastie, T. J., 2013. Compressive feature learning. In: Advances in Neural Information Processing Systems 26 (NIPS 2013).

[32] Perera, L. P., Mo, S., 2016. Machine intelligence for energy efficient ships: A big data solution. In: International Conference on Maritime Technology and Engineering. pp. 143–150.

[33] Qaisar, S., Bilal, R. M., Iqbal, W., Naureen, M., Lee, S., October 2013. Compressive sensing: From theory to applications, a survey. COMMUNICATIONS AND NETWORKS, 15 (5), 443–456.

[34] Razali, N. M., Yah, Y. B., 2011. Power comparison of shapiro-wilk, kolmogorow-smirnov, lilliefors and anderson-darling test. Statistical Modelling of Analytics 2 (1), 23–33.

[35] S. Han, H. M., Dally, W. J., 2016. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In: International Conference on Learning Representations.
URL `arXiv:1510.00149v5`

[36] Shoabi, M., Jha, N. K., Verma, N., 2015. Signal processing with direct computations on compressively sensed data. IEEE Transactions Very Early Scale Integration (VLSI) Systems 23 (1), 30–43.

[37] Steenveld, J., Hogeveen, H., 2015. Characterization of dutch dairy farms using sensor systems for cow management. Dairy Science 98 (1), 709–717.

[38] Strohmer, T., December 2012. Measure what should be measured : progress and challenges in compressive sensing. IEEE Singal Processing Letters 19 (12).

[39] Subasi, A., Gurosy, M. I., 2010. Eeg signal classification using pca, ica, lda and support vector machine. Expert Systems with Applications 37 (12), 8659–8666.

[40] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., S. Reed, D. A., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[41] Tang, J., Deng, C., Hung, G., Zhao, B., 2015. Compressed-domain ship detection an spaceborne optical image using deep neural network and extreme learning machine. IEEE Transactions on GeoScience and Remote Sensing 53 (3), 1174–1185.

[42] Trygg, J., Kettaneh-Wold, N., Wallba, L., 2001. 2d wavelet analysis and compression of on-line industrial process data. Chemometrics 15 (4), 299–319.

[43] Trygg, J., Wold, S., 1998. Pls regression on wavelet compressed nir spectra. Chemometrics and Intelligent Laboratory Systems 42 (1-2), 209–220.

[44] Valencia, D., Salazar, J., Valencia, J., 2016. Comparison analysis between rigrsure, sqtwolog, heursure and minimaxi techniques using hard and soft thresholding methods. In: XXI Symposium on Signal Processing, Images and Artificial Vision (STSIVA). pp. 1–5.

[45] Vang, X., Palival, K. K., 2013. Feature extraction and dimensionality reduction algorithms and their applications. Pattern Recognition 30 (10), 2429–2439.

[46] Vimalajeewa, D., Robson, E., Berry, D. P., Kulatunga, C., 2017. Evaluation of non-linearity in mir spectroscopic data for compressed learning. In: IEEE International Conference in Data Mining.

[47] Wolfert, S., Verdouw, C., Bogaard, M. J., 2017. Big data in smart farming - a review. Agricultural Systems 153 (5), 69–80.

[48] Xing, C., Ma, L., , Yang, X., 2016. Stacked denoise autoencoder based feature extraction and classification for hyperspectral images. Journal of Sensors 2016.

[49] Zhang, J., Wang, M., Li, Z., 2015. Parallel and distributed dimensionality reduction of hyperspectral data on cloud computing architectures. IEEE Selected Topics on Applied Earth Observations and Remote Sensing 9 (6), 2270–2278.

[50] Zheng, S., Kulkarni, S. R., Poor, H. V., 2011. Attribute-distributed learning: Models, limits, and algorithms. IEEE Transactions on Signal Processing 59 (1), 386–398.

Dixon Vimalajeewa is a PhD student at Telecommunications Software and Systems Group (TSSG) at Waterford Institute of Technology (WIT). His research interests include data analytics, sensor-based animal phenotypes and distributed learning algorithms (dvimalajeewa@tssg.org)

Chamil Kulatunga is a postdoctoral researcher in the Telecommunications Software and Systems Group (TSSG) at Waterford Institute of Technology (WIT). His research interests include distributed analytics, fog computing and smart agriculture (ckulatunga@tssg.org).

Donagh Berry is a quantitative geneticist at the Animal and Grassland Research and Innovation Centre at Teagasc. His research interests include genomic analysis, predictive modelling, chemometrics, breeding objectives and production indexes, decision support tools (donagh.berry@teagasc.ie).