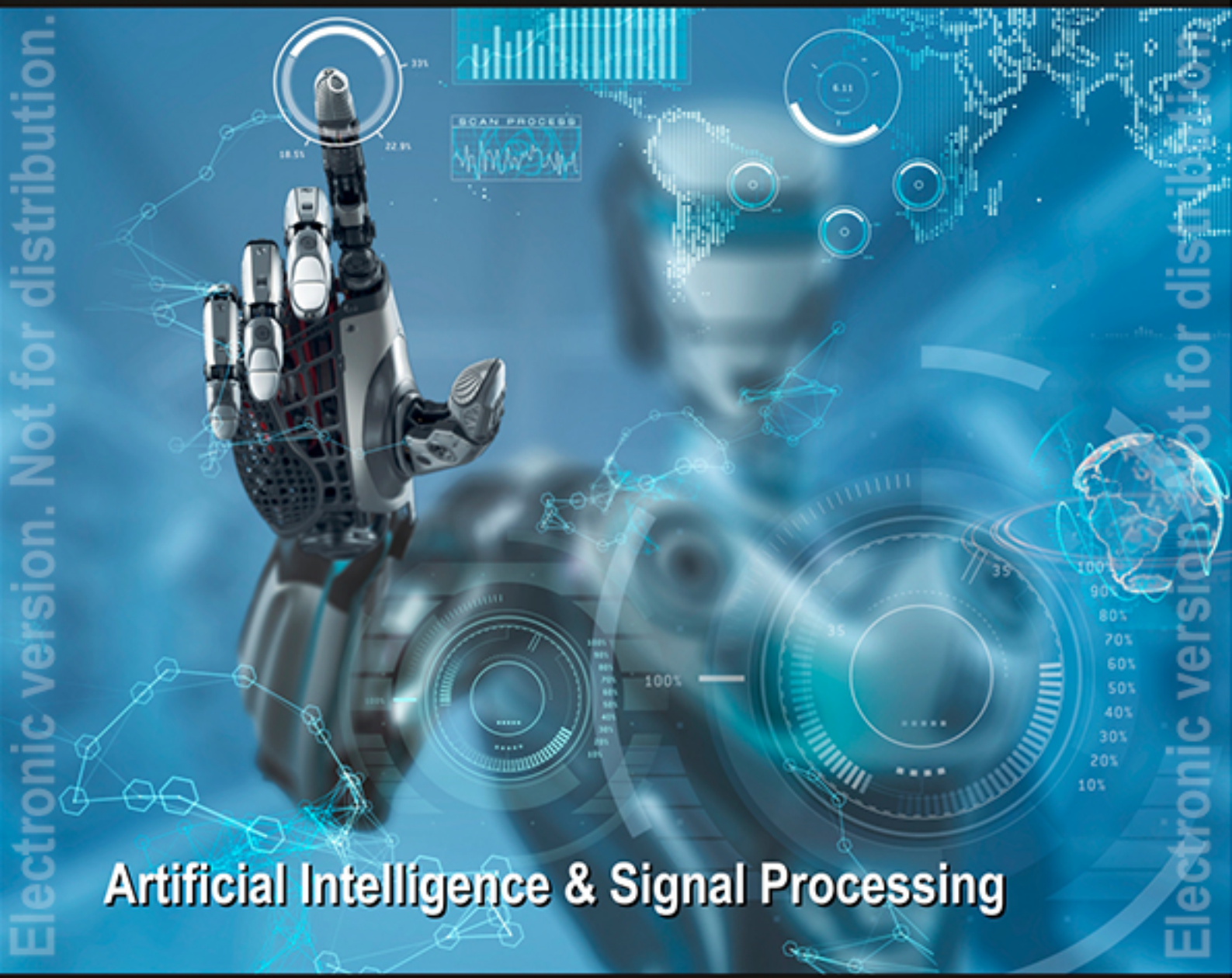
ISSN 1726-5479

SENSORS & TRANSDUCERS

vol. 256
2/22



Artificial Intelligence & Signal Processing

International Frequency Sensor Association Publishing

IFSA

Sensors & Transducers

International Official Open Access Journal of the International Frequency Sensor Association (IFSA) Devoted to Research and Development of Sensors and Transducers

Volume 256, Issue 2, March 2022

Editor-in-Chief
Prof., Dr. Sergey Y. YURISH

IFSA Publishing: Barcelona • Toronto

Published by International Frequency Sensor Association (IFSA) Publishing. Printed in the USA.

Sensors & Transducers

Volume 256, Issue 2
March 2022

www.sensorsportal.com

e-ISSN 1726-5479
ISSN 2306-8515

Sensors & Transducers journal is an open access, peer review international journal published monthly by International Frequency Sensor Association (IFSA).
Available in both: print and electronic (printable pdf) formats.

Sensors & Transducers Journal

Contents

Volume 256
Issue 2, March 2022

www.sensorsportal.com

ISSN 2306-8515
e-ISSN 1726-5479

Research Articles

Authors are encouraged to submit article in MS Word (doc) and Acrobat (pdf) formats
by e-mail: editor@sensorsportal.com. Please visit journal's webpage with preparation instructions:
http://www.sensorsportal.com/HTML/DIGEST/Submition.htm

International Frequency Sensor Association (IFSA).

Sensors & Transducers, Vol. 256, Issue 2, March 2022, pp. 1-11

Sensors & Transducers

Published by IFSA Publishing, S. L., 2022
http://www.sensorsportal.com

Learning Binary Data Representation for Optical Processing Units

[1, *] Bogdan Kozyrskiy, [1] Maurizio Filippone, [2] Iacopo Poli, [2, 3] Ruben Ohana, [2] Laurent Daudet and [2] Igor Carron

[1] Department of Data Science, EURECOM, 450 Route des Chappes, 06410 Biot, France
[2] LightOn, 2 rue de la Bourse, F-75002 Paris, France
[3] Laboratoire de Physique, Ecole Normale Supérieure, 24 rue Lhomond, 75005 Paris, France
[1] Tel.: +33 4 93 00 81 00
[*] E-mail: Bogdan.Kozyrskiy@eurecom.fr

Received: 1 February 2022 /Accepted: 4 March 2022 /Published: 31 March 2022

Abstract: Optical Processing Units (OPUs) are computing devices that perform random projections of input data by exploiting the physical phenomenon of scattering a light source through a diffusive medium. Random projections calculated by OPUs have been used successfully for approximating kernel ridge regression for large datasets with low power consumption and at high speed. However, OPUs require the input data to be binary. In this paper, we propose to use shallow and deep neural networks (NN) as binary encoders to perform input data binarization. The difficulty in developing a binarization strategy which is learned in an end-to-end fashion along with kernel ridge regression parameters, is due to the non-differentiability of the operation performed by the OPU. We overcome this difficulty by considering OPUs as a black-box and by employing the REINFORCE gradient estimator, which allows us to calculate the gradient of the loss function with respect to the weights of the binarization encoder and to optimize these together with the parameters of kernel ridge regression with gradient-based optimization.

Through our experimental campaign on a variety of tasks and datasets, we show that our method outperforms alternative unsupervised and supervised binarization techniques.

Keywords: Optimization, Random features, Linear regression, Optical processing unit.

1. Introduction

Statistical models based on kernel methods offer powerful and theoretically well-understood tools for complex data modeling problems. The limitation of employing these kernel-based models in practice is that a naive implementation scales poorly with the size of the data set, and there has been a tremendous amount of work in the direction of mitigating this issue by introducing approximations.

In this context, Nyström approximations [1] and random features [2] are very popular techniques to scale kernel methods virtually to any number of data, thanks to mini-batch formulations [3, 4].

The focus of this work is on random feature approximations, where by kernel-based models are "linearized" by an equivalent linear model with a set of suitably constructed random basis functions. The motivation behind this work is to considerably accelerate the construction of random features, while reducing power consumption, by resorting to a dedicated hardware, which we refer to Optical Processing Units (OPUs).

OPUs are computing devices which perform random projections of input vectors by exploiting the physical phenomenon of scattering a light source through a diffusive medium [5]. The random projection is then followed by a nonlinear operation,

making the whole pipeline of computation exactly what is needed to construct random features to approximate kernel-based models. Crucially, OPUs offer the possibility to operate with a number of random features at the speed of light and with low-power consumption, representing a unique solution to further improve scalability of kernel machines. As an example, OPU-based random feature approximations have successfully been proposed to carry out approximate kernel ridge regression in [6, 7].

One limitation associated with working with OPUs is that, because of the hardware setup, input vectors need to be binarized. In addition, the random projection matrix characterizing the device is unknown, and can only be retrieved through an expensive calibration procedure.

In this paper, we propose a novel binarization strategy for OPUs which is learned along with the regression/classification task in an end-to-end manner, meaning that the parameters of the binarization part are learned along with the kernel-based model parameters. In order to achieve this, we overcome the limitation that OPU projection matrices are unknown by employing the so-called REINFORCE gradient estimator, which allows us to treat the OPU as a black-box. Through experiments on several UCI classification/regression problems, we show that our proposal outperforms alternative unsupervised and supervised binarization techniques. This paper is an extended version of [8]; compared to the shorter version, we expand on the methods by analyzing the bounds on the objective functions of the proposed approaches, and we expand on the experiment by considering a larger class of kernels and image-based classification problems.

2. Related Work

In neural networks, binarization is generally targeting intermediate layer activations, and it may also stem from binarization of model parameters; in these cases, binarization is mostly introduced to reduce computational cost and memory consumption [9]. Neural networks with binary hidden layers find applications in binary autoencoders for hashing [10], data compression [11], and hard attention mechanism [12]. The binarization of layer activations is obtained by a suitable choice of activation functions; for instance, the sign or Heaviside functions for the deterministic case, or the sigmoid or $tanh$ functions combined with the Bernoulli distribution for the stochastic case [13, 14]. The most popular technique to propagate gradients through such activation functions is the so called straight-through estimator (STE) [15]. More recently, there have been proposals to replace the STE with another estimator through a relaxation technique, also known as the Gumbel Softmax-trick [16]. Also, different kinds of target propagation are used to learn suitable targets for each binary layer and then train the associated parameters with relaxation techniques or combinatorial optimization [17-19].

Focusing on OPUs, currently the standard approach to binarize data makes use of a binary autoencoder [11]. Such a binary autoencoder is trained independently from the OPU device, and it gives the possibility to perform the binarization operation by means of its encoder part. The autoencoder consists of a fully-connected encoder and decoder. The hidden layer has a Heaviside activation function, so its output is binary. The training procedure updates the weights of the decoder with backpropagation and weights of the encoder are forced to be equal to the weights of the decoder in order to be able to reproduce the input.

In this work, we aim to develop a supervised binarization model which is learned together with the supervised learning task. That is, we aim to provide a training procedure for the heterogeneous model consisting of the kernel ridge regression model approximated with random features and the binarization encoder before the OPU. In this context, a general-purpose framework called Method of Auxiliary Coordinates (MAC) was proposed in [19] with examples of application in [10] and [20]. The authors propose to introduce auxiliary variables into a deep neural network. These auxiliary variables are assigned the role of pre-activations for each layer, and they get replaced during the forward pass. The first step of the optimization targets the auxiliary variables, and, after this step, the parameters of each layer are optimized to regress on these variables, which take the role of layer-specific labels. This is very beneficial when some layers are discrete and vanilla backpropagation is not applicable. In [20], this approach is used to train a fully connected network with binary activation functions, using a STE to propagate a learning signal through the non-differentiable parts. Reference [10] is especially interesting because authors illustrate, how discrete binary layers can be optimized within larger, non-binary model.

While splitting the optimization of the binarization and the model is a viable option, we still need a way to training each part individually. There is a wide variety of ways to obtain a solution for kernel ridge regression with the random feature approximation, so the most difficult point is how to optimize the part consisting of the binary encoder and the OPU, because it combines a non-differentiable function with an implicit random projection. These make the STE from [20] inapplicable. Also, we found that the combinatorial approach used in [10] and [17] is inapplicable for our case for two reasons. First, it is suitable only when the binary dimension is relatively small, which might be a limitation for a general solution. Second, the combinatorial approach combined with MAC converges in one iteration to poor local optima, and this happens because of the model setup which is different from the ones in [10] and [17].

From a different point of view, it is possible to view our problem through the lenses of reinforcement

learning, where it is necessary to propagate binary codes through the OPU instead of discrete actions through the black-box environment. Instead of maximizing the reward from the environment, we are trying to minimize the loss function. The classical algorithm to solve this problem is REINFORCE [1]. This allows one to calculate gradients of the reward with respect to parameters of the policy that generates actions. The applicability of this method to other settings with black-box elements was shown in [21]. There are various versions of this algorithm intended to reduce variance of the gradient of the parameters. Very frequently they are based on relaxations of the non-differentiable sampling procedure [22], or approximation of the black-box part of the model [23]. It also worth noting that there exist competitive alternatives to REINFORCE, such as the one in [24], later extended with variance reduction [25] or relaxation [26].

3. Background

3.1. Kernel Ridge Regression

In this paper, we focus on kernel ridge regression for supervised learning tasks. Let $\mathbf{X} = \mathbf{x}_1, \dots, \mathbf{x}_n$ be a set of input vectors $\mathbf{x} \in \mathbb{R}^d$ and let $\mathbf{y} = y_1, \dots y_n$ be a set of labels associated with the input vectors.

The labels y_i can be continuous or binary depending on whether the task is regression or classification. Kernel ridge regression is a statistical model which constructs a functional relationship between the inputs and the labels which belongs to the so-called Reproducing Kernel Hilbert Space (RKHS). The properties of such functions, such as smoothness, are characterized by the choice of a so-called kernel function $k(\cdot,\cdot)\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ [27], which is a positive semi-definite function of pairs of input points returning a scalar. The reproducing property of kernel functions is $\langle k(\mathbf{x},\cdot), k(\mathbf{y},\cdot)\rangle = k(\mathbf{x},\mathbf{y})$. Positive definiteness of kernel functions implies that we can express $k(\mathbf{x_i},\mathbf{x_j}) = \varphi(\mathbf{x_i})^T\varphi(\mathbf{x_j})$ for some set of (possibly infinite) basis functions $\varphi(\cdot)$.

In order to derive the conventional formulation of kernel ridge regression, it is useful to start from linear regression, where a set of model parameters $\mathbf{w}$ is introduced to express a linear relationship between input and labels. Then, one introduces the following optimization problem:

$$\hat{\mathbf{w}} = \underset{w}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n}(y - \mathbf{w}^\top \mathbf{x})^2 - \frac{\lambda}{2}||\mathbf{w}||_2^2 \qquad (1)$$

The objective function contains two terms; the first is a model fitting term, while the second is a regularization term, which prevents the weights to become too large. The solution to this optimization problem is available in closed form, given that the objective is quadratic with respect to the parameters, yielding:

$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{y} \qquad (2)$$

Using standard algebraic manipulations involving the Woodbury identity, we can re-express the solution as:

$$\hat{\mathbf{w}} = \mathbf{X}^T(\mathbf{X}\,\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{y} \qquad (3)$$

While this is costly than the previous expression in the common case where $d < n$ (inversion of a $n \times n$ matrix rather than a $d \times d$ matrix), this formulation is useful to derive kernel ridge regression.

Imagining to introduce basis functions $\phi(\cdot) = \left(\phi_1(\cdot), \dots, \phi_D(\cdot)\right)^\top$, we can solve this new optimization problem

$$\hat{\mathbf{w}} = \underset{W}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i)\right)^2 + \frac{\lambda}{2}||\mathbf{w}||^2 \qquad (4)$$

with solution

$$\hat{\mathbf{w}} = \Phi^\top(\Phi\,\Phi^\top + \lambda\mathbf{I})^{-1}\mathbf{y} \qquad (5)$$

Evaluating the model at a given input $\mathbf{x}_*$ yields:

$$\phi(\mathbf{x}_*)^\top\hat{\mathbf{w}} = \phi(\mathbf{x}_*)^\top\mathbf{\Phi}^\top(\mathbf{\Phi}\mathbf{\Phi}^\top + \lambda\mathbf{I})^{-1}\mathbf{y} \qquad (6)$$

In this expression, we recognize the scalar product of vectors of basis functions. What we can do then, is to express these scalar products as a kernel function and obtain:

$$\phi(\mathbf{x})^\top\hat{\mathbf{w}} = \mathbf{k}_*(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \qquad (7)$$

where $\mathbf{k}_* = \left(k(\mathbf{x}_1,\mathbf{x}_*), \dots, k(\mathbf{x}_n,\mathbf{x}_*)\right)^\top$ and $K_{ij} = k(\mathbf{x}_i,\mathbf{x}_j)$. In practice, one first chooses a kernel function, and this induces a set of basis function; the beauty of this formulation is that one never explicitly works with the set of basis functions and all we need to use this model in practice is the evaluation of kernel functions among inputs.

3.2. Random Feature Approximation

One of main limitations of kernel methods is scalability to large datasets. The problem arises from the need to evaluate and perform algebraic operations with the so-called Gram matrix $\mathbf{K}$. Because $\mathbf{K}$ is an $n \times n$ matrix, evaluating and storing $\mathbf{K}$ requires $\mathcal{O}(n^2)$ computations and storage, while any algebraic operations, such as factorizations to handle the inverse of $\mathbf{K} + \lambda\mathbf{I}$, requires $\mathcal{O}(n^3)$ operations. These prevent the applicability of kernel methods in their exact form to datasets of size beyond a few thousand. It is worth noting that some approaches have been proposed to solve algebraic operations in an iterative fashion and without the need to store $\mathbf{K}$ [28, 29, 30], but they still require $\mathcal{O}(n^2)$ computations for each iteration of their solvers. Furthermore, while the number of iterations of

the solvers is much lower than n in practice, in the worst case it can be $\mathcal{O}(n)$, leading to a worst-case complexity of $\mathcal{O}(n^3)$.

The literature offers a number of solutions to scale kernel methods to large data linearly in the number of data, such as Nyström approximations [31] and random features [2]. In this work we focus in particular on random feature approximations, given that these have a practical implementation in hardware in the optical processing units that we consider in this work.

The random feature approximations form a class of approximations which attempt to construct a finite set of basis functions $\phi(\cdot) \in \mathbb{R}^D$ such that

$$k(\mathbf{x}_i, \mathbf{x}_j) \approx \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j) \tag{8}$$

There are different ways to construct such sets of basis functions, depending on the kernel. For example, so-called random Fourier features are commonly employed to approximate the Gaussian kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right) \tag{9}$$

Appealing to Bochner's theorem [2], this kernel, which is shift-invariant due to dependence on $\boldsymbol{\tau} = \mathbf{x}_i - \mathbf{x}_j$, admits an alternative expression as:

$$k(\boldsymbol{\tau}) = \int p(\boldsymbol{\omega}) \exp(\mathrm{i}2\pi\boldsymbol{\omega}\boldsymbol{\tau})\, d\boldsymbol{\omega} \tag{10}$$

where $p(\boldsymbol{\omega})$ is a proper density function and $\mathrm{i} = \sqrt{-1}$. Interpreting this as an expectation under $p(\boldsymbol{\omega})$, it is possible to approximate the integral as an expectation using Monte Carlo.

$$k(\boldsymbol{\tau}) = \frac{1}{D}\sum_r \exp\left(\mathrm{i}2\pi\boldsymbol{\omega}^{(r)}\boldsymbol{\tau}\right) \tag{11}$$

with $\boldsymbol{\omega}^{(r)} \sim p(\boldsymbol{\omega})$. Furthermore, it is possible to use simple trigonometric identities to verify that the complex exponential can be broken down as a scalar product with terms depending on $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively

$$k(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{D}\phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j) \tag{12}$$

with

$$\phi_r(\mathbf{x}) = \left(\sin\left(\mathbf{x}^\top\boldsymbol{\omega}^{(r)}\right), \cos\left(\mathbf{x}^\top\boldsymbol{\omega}^{(r)}\right)\right) \tag{13}$$

We refer the reader to [2, 32, 3, 33] for random features derived from alternative integral representation to the Fourier transform.

3.3. Random Features on Optical Processing Units

In this section we discuss Optical Processing Units (OPUs) in the context of random features. In the previous section we discussed random features as a way to approximate models involving kernels; for OPUs, instead, the device produces random features (fast and with little power consumption) and the question that we aim to address here is how to use these to implement approximate kernel machines.

OPU are computing devices which exploit the physical process of scattering of light to perform a random projection operation of a given vector. In particular, given a binary vector $\mathbf{x}_\mathrm{i} \in \mathbb{R}^d$, OPUs perform a multiplication by a random matrix $\mathbf{R}$ and apply the nonlinear activation function $||\cdot||^2$. In other words,

$$\phi(\mathbf{x}) = \frac{1}{\sqrt{D}}\|\mathbf{R}\mathbf{x}\|^2 \tag{14}$$

The matrix $\mathbf{R} \in \mathcal{C}^{D\times d}$ is a complex Gaussian matrix with elements $R_{ij} \sim \mathcal{CN}(0,1)$. Previous works have established that in the limit of an infinite number of random features, the equivalent kernel is the following [6]:

$$k(\mathbf{x},\mathbf{y}) \approx \phi(\mathbf{x})\phi(\mathbf{y}) \overset{D\to\infty}{=} ||\mathbf{x}||^2||\mathbf{y}||^2 + (\mathbf{x}^T\mathbf{y})^2 \tag{15}$$

Therefore, when using OPUs for kernel ridge regression, we are implicitly working with this polynomial kernel.

Recently a new version of OPUs has been proposed and developed in [34], which allows one to perform linear random feature projections

$$\psi(\mathbf{x}) = C\mathbf{R}\mathbf{x} \tag{16}$$

where C is the fixed constant.

This novel type of OPU opens to the possibility to approximate a wide variety of kernels by choosing an appropriate activation function [2, 33]. For example, it is possible to apply trigonometric activation functions to the outputs of the OPU:

$$\psi'(\mathbf{x}) = \begin{bmatrix}\sin(\psi(\mathbf{x}))\\ \cos(\psi(\mathbf{x}))\end{bmatrix} \tag{17}$$

This type of random features is called Random Fourier Features (RFF). It was proven in [2] that this kind of random features allows to approximate RBF kernels.

$$\begin{aligned}&\frac{1}{D}\sum_{i=1}^{D}\psi(\mathbf{x})^\top\psi(\mathbf{y})\\ &= \frac{1}{D}\sum_{i=1}^{D}\left(\begin{bmatrix}\sin(\psi(\mathbf{x}))\\ \cos(\psi(\mathbf{x}))\end{bmatrix}^\top\begin{bmatrix}\sin(\psi(\mathbf{y}))\\ \cos(\psi(\mathbf{y}))\end{bmatrix}\right) =\\ &= \mathbb{E}_\omega[cos(\omega(\mathrm{x}-\mathrm{y}))] = k_{RBF}(\mathrm{x},\mathrm{y})\end{aligned} \tag{18}$$

As mentioned before, an important aspect of OPUs is that their input should be binary; this paper proposes a novel way to carry out a binarization of its input along with the kernel ridge regression task in an end-to-end fashion.

4. Methods

4.1. REINFORCE for Kernel Ridge Regression with Binarized Inputs

In order to be able to implement kernel ridge regression on OPUs we need to binarize the inputs $\mathbf{x}_i$, and we propose to do so by employing an encoder, implemented as a neural network, parameterized by a set of weights $\mathbf{W}_{\text{enc}}$. The encoder transforms the inputs to kernel-based models $\mathbf{x}_i$ and turns them into a set of Bernoulli-distributed binary random variables $\mathbf{z}_i$.

In particular, we denote by $f_k(\mathbf{x}, \mathbf{W}_{\text{enc}})$ the function implemented by the encoder which parameterizes the Bernoulli distribution associated with the *k*th element of the output, that is z_k. Recalling the random feature formulation of linear regression of **Section 3**, we propose the following approach to construct an approximate kernel-based model with binary inputs:

$$\tilde{y} = \mathbb{E}_{\mathbf{z}}\left[\mathbf{w}_{\text{regr}}^{\top}\phi(\mathbf{z})\right] + \varepsilon \tag{19}$$

where $\mathbf{z} \sim \text{Bernoulli}\big(f(\mathbf{x}, \mathbf{W}_{\text{enc}})\big)$ and $\mathbf{w}_{\text{regr}}$ are parameters of the linearized regression model. Note how in this formulation the binary vectors $\mathbf{z}$ are treated stochastically due to the expectation under the Bernoulli distribution induced by the encoder. The reason for this is that it allows us to employ the so-called REINFORCE gradient estimator, as we discuss next.

REINFORCE, also known as the log-derivative trick or score function estimator, offers a way to estimate the gradient of the expectation of a non-differentiable function $f(z)$ under the distribution of the input random vector variables z:

$$\begin{aligned}\nabla_\theta E_{p(z;\theta)} f(z) &= \nabla_\theta \int p(z;\theta) f(z) dz = \\ &\int \nabla_\theta p(z;\theta) f(z) dz = \\ &\int p(z;\theta) \frac{\nabla_\theta p(z;\theta)}{p(z;\theta)} f(z) dz \\ &= \mathbb{E}_{p(z;\theta)} \nabla_\theta \log p(z;\theta) f(z) \\ &\approx \frac{1}{M} \sum_{i=1}^{M} \nabla_\theta \log p(z;\theta) f(z)\end{aligned} \tag{20}$$

where M is the number of samples drown from $p(z, \theta)$. Applying REINFORCE to our approximate kernel-based model yields the following optimization objective:

$$\begin{aligned}\min_{\mathbf{w}_{\text{regr}}, \mathbf{W}_{\text{enc}}} \mathbb{E}_{\mathbf{z}\sim\text{Bernoulli}(f(\mathbf{x},\mathbf{W}_{\text{enc}}))}\left[\mathcal{L}\left(\mathbf{y}, \phi(\mathbf{Z})\mathbf{w}_{\text{regr}}\right)\right] \\ +\lambda_{\text{enc}}\|\mathbf{W}_{\text{enc}}\|^2 + \lambda_{\text{regr}}\left\|\mathbf{w}_{\text{regr}}\right\|^2\end{aligned} \tag{21}$$

In this expression, we denoted by $\mathcal{L}(\mathbf{y}, \tilde{\mathbf{y}})$ the loss function associated with the task at hand and by $\mathbf{Z}$ the matrix that contains binary encoded variables for the whole training set $\mathbf{X}$. We can optimize this objective by means of gradient-based techniques; for this we require that we are able to compute the gradient of the objective with respect to all parameters. The gradient of the first term of the objective with respect to $\mathbf{W}_{\text{enc}}$, which is the most involved part, is:

$$\begin{aligned}&\nabla_{\mathbf{W}_{\text{enc}}} \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\mathcal{L}\left(y, \mathbf{w}_{\text{regr}}^{\top}\phi(\mathbf{z})\right)\right] \approx \\ &\approx \frac{1}{M}\sum_{i=1}^{M} \mathcal{L}\left(y, \mathbf{w}_{\text{regr}}^{\top}\phi(\mathbf{z})\right) \nabla_{\mathbf{W}_{\text{enc}}} \log q(\mathbf{z})\end{aligned} \tag{22}$$

while the derivatives of the other terms are straightforward to compute. With this derivation, we observe that it is then possible to jointly optimize all parameters, leading to what it is commonly referred to as an *end-to-end* approach. In the remainder of this paper, we refer to this method as End-to-End SE, where SE stands for Supervised Encoder.

4.2. Variance Reduction

REINFORCE is known to suffer from large variance of the gradients. In order to reduce the variance of this estimator, we employ control variates [25]. In this approach, we add a set of random variables to the estimator, such that these variables have zero mean, so they do not alter the expectation of the gradient. The aim is to construct such variables so as to reduce the overall variance of the estimator:

$$\begin{aligned}&\nabla_{\mathbf{W}_{\text{enc}}} \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\mathcal{L}\left(y, \mathbf{w}_{\text{regr}}^{\top}\phi(\mathbf{z})\right)\right] \approx \\ &\frac{1}{M}\sum_{i=1}^{M} \nabla_{\mathbf{W}_{\text{enc}}} \log q(\mathbf{z})\left(\mathcal{L}\left(y, \mathbf{w}_{\text{regr}}^{\top}\phi(\mathbf{z})\right) - \mathbf{v}\right) \\ &\text{where } \mathbf{v} = \frac{1}{M-1}\sum_{i\neq j} \mathcal{L}\left(y, \mathbf{w}_{\text{regr}}^{\top}\phi(\mathbf{z})\right)\end{aligned} \tag{23}$$

4.3. Lowering the Cost of REINFORCE

The estimation of the gradient of the End-to-End SE with respect to $\mathbf{W}_{\text{enc}}$ can be expensive when the number of random features is large. This is due to the fact that this requires multiple samples to be passed from the encoder through the random projection and the approximate kernel ridge regression model. In this section we propose a strategy to reduce the complexity of REINFORCE applied to our model, whereby we average set of basis functions under the resampling of the binary variables as follows:

$$\tilde{y} = \mathbf{w}_{\text{regr}}^{\top}\mathbb{E}_{\mathbf{z}}\left[\phi(\mathbf{z})\right] + \varepsilon \tag{24}$$

where $\mathbf{z} \sim \text{Bernoulli}\big(f(\mathbf{x}, \mathbf{W}_{\text{enc}})\big)$

With this new modeling assumption, the training is based on a modified optimization problem as follows:

$$\min_{\mathbf{w}_{\text{regr}}\mathbf{W}_{\text{enc}}} \mathcal{L}\left(\mathbf{y}, \mathbb{E}_{\mathbf{z}\sim\text{Bernoulli}(f(\mathbf{x},\mathbf{W}_{\text{enc}}))}[\phi(\mathbf{Z})]\mathbf{w}_{\text{regr}}\right) \tag{25}$$

$$+\lambda_{\text{enc}}\|\mathbf{W}_{\text{enc}}\|^2 + \lambda_{\text{regr}}\left\|\mathbf{w}_{\text{regr}}\right\|^2$$

Again, we can perform gradient-based optimization. Focusing on the first term, which is the nontrivial one to differentiate in the objective, we obtain

$$\nabla_{\mathbf{W}_{\text{enc}}}\mathcal{L} = \frac{d\mathcal{L}}{d\left(\mathbb{E}\phi(\mathbf{z})\right)}\nabla_{\mathbf{W}_{\text{enc}}}\mathbb{E}\left(\phi(\mathbf{z})\right) \tag{26}$$

where $\nabla_{\mathbf{W}_{\text{enc}}}\mathbb{E}\left(\phi(\mathbf{z})\right)$ is calculated with the REINFORCE estimator. In the remainder of the paper, we will refer to this method as Isolated Supervised Encoder (SE).

Regarding the comparison of the End-to-End SE and Isolated SE, we can note the following relationship between these models in the case of regression problems. In the data term of the optimization objective (25) we can put an expectation over the whole matrix product of the random features map and the regression weights instead of an expectation over the random features only. Then we can move the expectation in such a way that it is taken over the whole term within the squared norm. We can do this because within one gradient step iteration, neither $\mathbf{y}$ nor $\mathbf{W}_{\text{enc}}$ are considered as random variables. In this case the data term looks as follows:

$$\begin{aligned}\left\|\mathbf{y} - \mathbb{E}[\phi(\mathbf{Z})]\mathbf{w}_{\text{regr}}\right\|^2 &= \left\|\mathbf{y} - \mathbb{E}\left[\phi(\mathbf{Z})\mathbf{w}_{\text{regr}}\right]\right\|^2 \\ &= \left\|\mathbb{E}\left[\mathbf{y} - \phi(\mathbf{Z})\mathbf{w}_{\text{regr}}\right]\right\|^2\end{aligned} \tag{27}$$

In turn, End-to-End SE has a following data term as part of its optimization objective (21):

$$\mathbb{E}\left[\left\|\mathbf{y} - \phi(\mathbf{Z})\mathbf{w}_{\text{regr}}\right\|^2\right] \tag{28}$$

We can note that the squared loss is convex function. Thus, we can apply Jensen's inequality to obtain the following expression:

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{y} - \phi(\mathbf{Z})\mathbf{w}_{\text{regr}}\right\|^2\right] &\geq \left\|\mathbb{E}\left[\mathbf{y} - \phi(\mathbf{Z})\mathbf{w}_{\text{regr}}\right]\right\|^2 \\ &= \left\|\mathbf{y} - \mathbb{E}\left[\phi(\mathbf{Z})\mathbf{w}_{\text{regr}}\right]\right\|^2\end{aligned} \tag{29}$$

As a result, End-to-End SE optimizes upper bound of the Isolated SE objective.

5. Results

5.1. Experiments on the UCI Datasets

We compared the performance of the proposed approaches for a non-linear OPU (End-to-End SE and Isolated SE) and a linear OPU that uses trigonometric activations (End-to-End SE with RFF) against a model based on unsupervised autoencoder proposed in [11], an encoder trained with direct feedback alignment (DFA) [35] and a Kernel Ridge Regression (KRR) based on a Radial Based Function kernel (RBF). Results are reported in Fig. 1 for several UCI regression and classification problems [36]. We want to emphasize that the main competitors of the proposed methods are the ones based on unsupervised autoencoder and encoder trained by DFA, because kernel ridge regression is unable to work with large datasets, and OPU-based regression just approximates this method and is intended to replace it on large datasets.

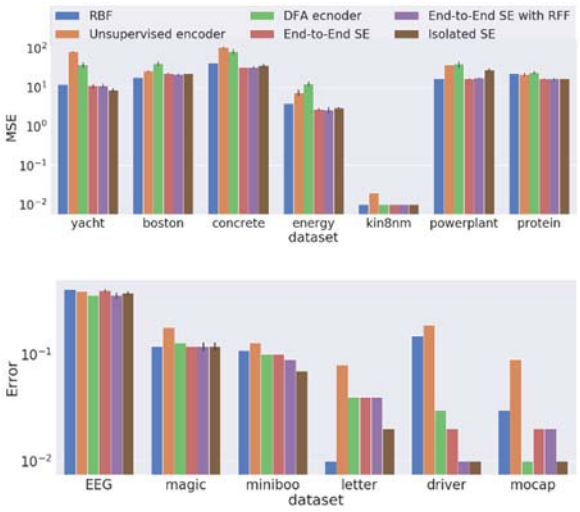


Fig. 1. Mean squared error (MSE) for regression (top) and negative error on classification (bottom) datasets comparison.

For KRR experiments we used Mean Squared Error (MSE) as a loss function. To apply KRR to the classification problems we replaced 0 and 1 in class labels with -1, 1 and solved a classification problem as a regression one using MSE loss as an optimization objective. For all other models we used MSE loss for the regression problems and Cross Entropy (CE) loss for the classification problems. We used a logistic activation function on the last layer for the classification tasks.

For Isolated SE and End-to-End SE as an encoding function $f(\mathbf{x}, \mathbf{W}_{\text{enc}})$ providing parameters for the Bernoulli distribution, we chose a single linear layer with a sigmoid activation.

$$f(\mathbf{x}, \mathbf{W}_{\text{enc}}) = \sigma\left(\mathbf{W}_{\text{enc}}^{\mathrm{T}}\,\mathbf{x}\right) \tag{30}$$

All hyperparameters for the DFA encoder, End-to-End SE and Isolated SE models (size of binary embedding, learning rate, l2 regularization for the encoder and the regression layer) were chosen with a random search during cross-validation. Kernel parameters of KRR were tuned by random search with cross-validation. This poses computational challenges for the large datasets (MiniBoo, MoCap), so we resort to random Fourier feature approximations for these cases.

For the models involving random features (both Fourier and OPU-generated ones) we have tuned the

variance of the distribution that generates these random features. Concretely, assuming that the elements of the **R** matrix generating the random projections are distributed through the standard Normal distribution, we can obtain a new random matrix $\mathbf{R}'$ by multiplying **R** by any variance, for instance:

$$\phi'(\mathbf{x}) = c|\mathbf{R}'\mathbf{x}|^2 = c|\frac{\mathbf{R}}{\alpha}\mathbf{x}|^2 = c\frac{1}{\alpha^2}|\mathbf{R}\mathbf{x}|^2 \qquad (31)$$

It is enough to multiply the output of the OPU by an additional parameter γ, such that $\gamma^2 = \frac{1}{\alpha^2}$, and optimize them with standard gradient descent. The parameter γ is not equivalent to the lengthscale parameter of the RBF kernel. In practice, it has an effect of outputscale parameter of RBF kernel, as it has simply a scaling effect on the kernel.

On the regression problems, both proposed methods outperformed their main competitors. On the classification problems, the DFA-based approach was better only on one dataset, and on all other datasets the proposed methods performed better or equally well. Regarding the type of a kernel approximated by the OPU, the experiments show that the linear OPU with trigonometric activations performs as good as OPU kernel for most of the datasets. It gives performance gains only for some classification problems. Considering the comparison between the proposed methods, we see that End-to-End SE is more stable and requires a significantly fewer number of samples from the encoder, although Isolated SE showed slightly better results on classification problems.

We considered including results obtained by running these models on the real OPU (Fig. 2). Unfortunately, the regression problems required such a large number of epochs that we could not perform the experiments in a reasonable amount of time.

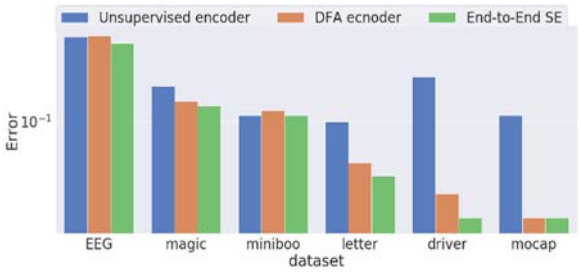


Fig. 2. Error comparison on classification (bottom) datasets for experiments on a real hardware.

We also tested the performance of our approach with respect to the number of samples required to employ REINFORCE. We found that End-to-End SE can achieve good results with a small amount of samples from the encoder, and the increase of amount of samples does not seem to improve performance.

Finally, we evaluated the effect of variance reduction on convergence speed and performance for End-to-End SE model. In Fig. 3 we report results for one classification and one regression problem. The convergence curves indicate that the convergence speed is benefits from the gradient variance reduction.

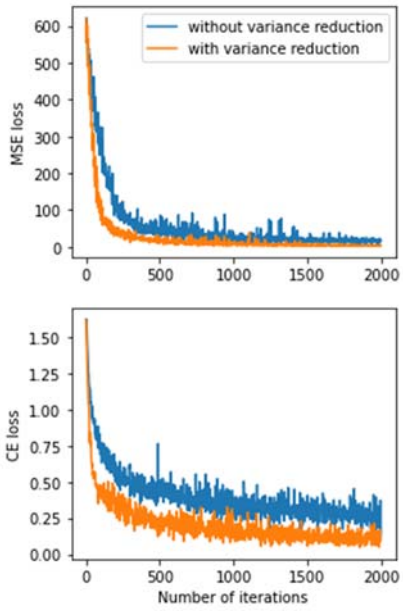


Fig. 3. Convergence of the training procedure on classification problem: mocap dataset (bottom) and regression problem: boston dataset (top).

5.2. Experiments on Image Data

In this section we evaluate an optical random feature regression approach for image classification tasks with several different binarization techniques including the proposed methods.

The kernel generated by the OPU (15) is an example of a polynomial kernel. Polynomial kernels, unlike more popular RBF kernels, take into account interaction between different feature dimensions. This property is especially important for image data because a relative alignment of pixels is crucial for image classification. Of course, when we are working with OPUs, kernel takes into account an alignment of different dimensions of the binary embedding of an image instead of the pixels. The relative alignment of different dimensions of the binary embedding most probably does not contain exactly the same information as the mutual alignment of pixels. But until the binary encoder does not have disentanglement properties, the mutual interaction of dimensions of the binary embedding have to contain an additional information about the image. That is why it is still important to use a kernel that is capable to take into account these relationships.

For the experiments on image data, we used two convolutional architectures of the binary encoder. First architecture was inspired by the LeNet model (Table 1).

Table 1. LeNet-based binary encoder architecture.

Layer	Dimensions
Conv2D	5×5, 6 filters
MaxPooling	2×2
Conv2D	5×5, 16 filters
MaxPooling	2×2
Linear	576×512
Linear	512×d_{binary}
Binarization layer	d_{binary}

We performed experiments on three classical image classification datasets (Table 2). We compared End-to-End SE with a model that used autoencoder to train the encoder (AE) and a model that used direct feedback alignment (DFA) for the same purpose. The results are shown in Table 3.

Tab. 2. Image datasets used in the experiments.

Dataset	Train/test size	Classes	Dimension
MNIST	60000/10000	10	1×28×28
F-MNIST	60000/10000	10	1×28×28
CIFAR10	50000/10000	10	3×32×32

Tab. 3. Classification error obtained by the model with the LeNet encoder.

Dataset	AE	DFA	End-to-End SE
MNIST	0.06±0.02	0.31±0.03	0.01±0.00
F-MNIST	0.20±0.01	0.47±0.01	0.09±0.00
CIFAR10	0.55±0.01	0.81±0.02	0.32±0.01

We used the same encoder architecture with the same hyperparameters for each binarization method. The size of the binary embedding was set to 400, except of the unsupervised AE method. The reason why we trained the unsupervised AE differently for these experiments is because we were using complex convolutional models and it was hard to adapt the method proposed in [11]. This binarization approach requires to use exactly the same values of weights both for the encoder and for the decoder models. It means that this method requires to build a decoder model that is symmetrical to the encoder model. Achieving this property for convolutional neural networks is difficult because for the decoder it is difficult to pick equivalent symmetric operations for convolutional and pooling layers in the encoder. Transposed convolutions and interpolation operations that are used in decoders for image data, are suitable for training of the autoencoder by end-to-end backpropagation. They learn operations that are not symmetric to convolutions and pooling layers of the encoder. The method proposed in [11] assumes that only the decoder is trained and the encoder copies weights from it, that is impossible to do for asymmetric operations in decoder and encoder. That is why we trained the autoencoder model in a different way. The encoder of the AE model used a $tanh$ activation function at the output. We used a parameter β that controls steepness of the $tanh$ function. We slowly increased the value of this parameter from $\beta = 1$ to $\beta = 100$ in the process of training. At the end of the training, the function $tanh$ with a high value of the parameter β is almost equivalent to a shifted and scaled Heaviside function.

$$2h(x) - 1 \approx tanh(\beta x),\ \beta \to \infty \tag{32}$$

The results of the DFA approach signify that this type of gradient updates is not suitable for convolutional models. This observation is supported by other researchers [37, 38].

Unsupervised AE was able to provide acceptable accuracy for the MNIST dataset, but on CIFAR-10 its performance dropped significantly. A possible explanation is that simple convolutional AE is unable to extract reasonable binary representation of complex images. The AE model used in the experiments was able to reconstruct simple images from MNIST dataset. But the reconstruction quality of the same model was much worse for the CIFAR-10 dataset. When the dimensionality of the binary embedding was equal to 400, the AE model was unable to generate any sensible images. Thus, we had to increase the size of the binary embedding to 1024. But the reconstructed images were very blurry even with this modification.

Because of the poor performance of the unsupervised AE baseline, we decided to add another baseline to the comparison. For this experiment we used a RESNET-based convolutional network as the encoder. LBAE approach proposed in [39] implements an autoencoder with a binary latent space. The training procedure of this method is based on straight-through gradient estimator. The architecture of the binary encoder is represented in Table 4 and Table 5. This encoder used leaky ReLU as an activation function. Batch-normalization was used before each activation function.

Tab. 4. Architecture of the RESNET-based binary encoder.

Layer	Dimensions
Conv2D	3×3, 64 filters
Conv2D	4×4, 64 filters
Residual Block	3×3, 64 filters
Conv2D	4×4, 64 filters
Residual Block	3×3, 64 filters
Conv2D	4×4, 128 filters
Linear	4096×d_{binary}
Binarization layer	d_{binary}

Tab. 5. Residual block structure. The number of filters is specified in Tab. 4.

Layer	Dimensions
Conv2d	3×3
Conv2d	3×3

Table 6 contains the results of the comparison between the LBAE-based and End-to-End SE-based encoders in terms of classification error.

Tab. 6. Classification error for models with the RESNET encoder.

Dataset	LBAE	End-to-End SE
MNIST	0.15±0.01	0.01 ±0.00
F-MNIST	0.24±0.02	0.06±0.01
CIFAR-10	0.63±0.02	0.17±0.01

Both binarization approaches used the same RESNET-based architecture of the encoder network. We dropped the DFA approach from the comparison because of its poor performance.

The results of this experiment showed interesting property of the unsupervised approach for training of the binary encoder. The LBAE-based encoder with a deeper network performed worse than the simpler autoencoder with $tanh$ annealing in terms of classification error, while the LBAE approach was better in image reconstruction task. It seems, that the binary latent projection of image data, that is suitable for image reconstruction, is unsuitable for image classification.

As with the UCI data we evaluated the effect of variance reduction.

As we can see, variance reduction plays a crucial role for image classification (Fig. 4). Without this technique the proposed method is unable to train the model for CIFAR-10. We assume, that it happens because in deeper models variance has a multiplicative effect when the number of layers increases.

6. Conclusion

Recent advances in alternatives to transistor-based hardware are bringing a new wave of sustainable computing solutions for machine learning [40]. This paper focuses on optical-based computing through OPUs [5], which perform randomized projections of binary input vectors at the speed of light with low energy consumption. In this paper, we considered these randomized computations to implement kernel-based models for regression and classification tasks through random feature approximations. In particular, we proposed a novel strategy to binarize the inputs of the given task so as to be able to employ OPUs inspired by reinforcement learning. The proposed strategy uses an encoder to map the inputs to a set of binary variables and employs the REINFORCE gradient estimator to estimate its parameters jointly with the parameters of the kernel-based model. We also explored ways to reduce the variance of the gradient estimator and accelerate convergence, which is key in a number of challenging modeling tasks such as image classification. Through a series of experiments, we showed that our proposal outperforms competitors based on unsupervised binarization and those that do not employ gradient information.

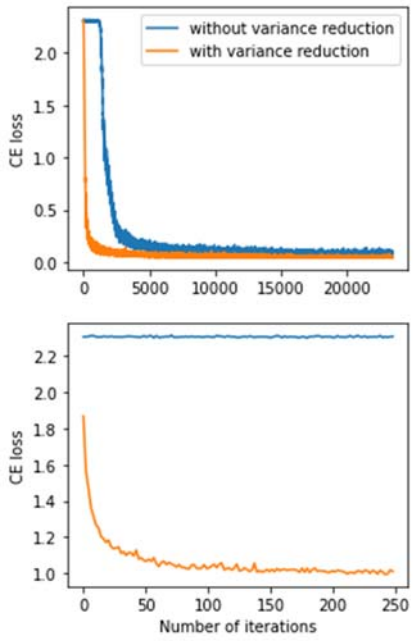


Fig. 4. Training loss with and without variance reduction.

We are currently investigating our approach in the context of other kernel-based models, such as Gaussian processes [41], and their extension to deep models, such as Deep Kernel Learning [42] and Deep Gaussian processes [3].

References

[1]. R. J. Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, *Machine Learning*, Vol. 8, 1992, p. 229–256.

[2]. A. Rahimi and B. Recht, Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning, in *Proceedings of the Advances in Neural Information Processing Systems 21 (NIPS 2008)*, 2009.

[3]. K. Cutajar, E. V. Bonilla, P. Michiardi and M. Filippone, Random feature expansions for deep Gaussian processes, in *Proceedings of the International Conference on Machine Learning*, Vol. 70, August 2017, pp. 884–893.

[4]. J. Hensman, N. Fusi and N. D. Lawrence, Gaussian Processes for Big Data, in *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, Arlington, 2013, pp. 282–290.

[5]. A. Saade, F. Caltagirone, I. Carron, L. Daudet, A. Drémeau, S. Gigan and F. Krzakala, Random projections through multiple optical scattering: Approximating kernels at the speed of light, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6215–6219.

[6]. R. Ohana, J. Wacker, J. Dong, S. Marmin, F. Krzakala, M. Filippone and L. Daudet, Kernel computations from large-scale random features obtained by optical processing units, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP' 2020)*, 2020, pp. 9294-9298.

[7]. A. Cappelli, R. Ohana, J. Launay, L. Meunier, I. Poli and F. Krzakala, Adversarial Robustness by Design through Analog Computing and Synthetic Gradients, arXiv preprint arXiv:2101.02115, 2021.

[8]. B. Kozyrskiy, I. Poli, R. Ohana, L. Daudet, I. Carron and M. Filippone, Binarization for Optical Processing Units via REINFORCE, in *Proceedings of the 3rd International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI' 2021)*, Porto, Portugal, 17-19 November 2021, pp. 23-27.

[9]. H. Qin, R. Gong, X. Liu, X. Bai, J. Song and N. Sebe, Binary neural networks: A survey, *Pattern Recognition*, Vol. 105, 2020, p. 107281.

[10]. M. A. Carreira-Perpinán and R. Raziperchikolaei, Hashing with binary autoencoders, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 557-566.

[11]. J. Tissier, C. Gravier and A. Habrard, Near-lossless binarization of word embeddings, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 7104–7111.

[12]. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel and Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 2048-2057.

[13]. M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv and Y. Bengio, Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1, arXiv preprint arXiv:1602.02830, 2016.

[14]. J. W. T. Peters and M. Welling, Probabilistic binary neural networks, arXiv preprint arXiv:1809.03368, 2018.

[15]. Y. Bengio, N. Léonard and A. Courville, Estimating or propagating gradients through stochastic neurons for conditional computation, arXiv preprint arXiv:1308.3432, 2013.

[16]. E. Jang, S. Gu and B. Poole, Categorical reparameterization with gumbel-softmax, arXiv preprint arXiv:1611.01144, 2016.

[17]. A. L. Friesen and P. Domingos, Deep learning as a mixed convex-combinatorial optimization problem, arXiv preprint arXiv:1710.11573, 2017.

[18]. D.-H. Lee, S. Zhang, A. Fischer and Y. Bengio, Difference target propagation, in ECML PKDD 2015: Machine Learning and Knowledge Discovery in Databases, *Lecture Notes in Computer Science*, Vol. 2015, pp. 498–515.

[19]. M. Carreira-Perpinan and W. Wang, Distributed optimization of deeply nested systems, in *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, Reykjavik, 2014, pp. 10-19.

[20]. A. Choromanska, B. Cowen, S. Kumaravel, R. Luss, M. Rigotti, I. Rish, P. Diachille, V. Gurev, B. Kingsbury, R. Tejwani and others, Beyond backprop: Online alternating minimization with auxiliary variables, in *Proceedings of the International Conference on Machine Learning*, 2019.

[21]. R. Ranganath, S. Gerrish and D. Blei, Black box variational inference, in *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Reykjavik, Iceland, 2014, pp. 814-822.

[22]. G. Tucker, A. Mnih, C. J. Maddison, D. Lawson and J. Sohl-Dickstein, Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models, arXiv preprint arXiv:1703.07370, 2017.

[23]. W. Grathwohl, D. Choi, Y. Wu, G. Roeder and D. Duvenaud, Backpropagation through the void: Optimizing control variates for black-box gradient estimation, arXiv preprint arXiv:1711.00123, 2017.

[24]. M. Yin and M. Zhou, ARM: Augment-REINFORCE-merge gradient for stochastic binary networks, arXiv preprint arXiv:1807.11143, 2018.

[25]. W. Kool, H. van Hoof and M. Welling, Buy 4 REINFORCE Samples, Get a Baseline for Free!, *Deep Reinforcement Learning Meets Structured Prediction Workshop at the International Conference on Learning Representations*, 2019.

[26]. Z. Dong, A. Mnih and G. Tucker, DisARM: An antithetic gradient estimator for binary latent variables, arXiv preprint arXiv:2006.10680, 2020.

[27]. K. P. Murphy, Machine learning: a probabilistic perspective, *MIT Press*, 2012.

[28]. M. Filippone and R. Engler, Enabling scalable stochastic gradient-based inference for Gaussian processes by employing the Unbiased LInear System SolvEr (ULISSE), in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, 2015, pp. 1015-1024.

[29]. K. Cutajar, M. A. Osborne, J. P. Cunningham and M. Filippone, Preconditioning Kernel Matrices, in *Proceedings of the 33rd International Conference on International Conference on Machine Learning* - Volume 48, New York, NY, USA, 2016, arXiv:1602.06693.

[30]. K. Wang, G. Pleiss, J. Gardner, S. Tyree, K. Q. Weinberger and A. G. Wilson, Exact Gaussian Processes on a Million Data Points, in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS' 2019)*, 2019.

[31]. C. Fowlkes, S. Belongie and J. Malik, Efficient spatiotemporal grouping using the nystrom method, in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR' 2001)*, 2001.

[32]. Y. Cho and L. Saul, Kernel methods for deep learning, *Advances in Neural Information Processing Systems*, vol. 22, 2009.

[33]. J. Wacker, M. Kanagawa and M. Filippone, Improved Random Features for Dot Product Kernels, arXiv preprint arXiv:2201.08712, 2022.

[34]. I. Poli, J. Launay, K. Müller, G. Pariente, I. Carron, L. Daudet, R. Ohana and D. Hesslow, Method and system for machine learning using optical data. *USA Patent US 2021/0287079 A1*, 16 September 2021.

[35]. A. Nøkland, Direct feedback alignment provides learning in deep neural networks, arXiv preprint arXiv:1609.01596, 2016.
[36]. D. Dua and C. Graff, UCI Machine Learning Repository, 2017.
[37]. S. Bartunov, A. Santoro, B. Richards, L. Marris, G. E. Hinton and T. Lillicrap, Assessing the scalability of biologically-motivated deep learning algorithms and architectures, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*, 2018, pp. 9390–9400.
[38]. J. Launay, I. Poli and F. Krzakala, Principled training of neural networks with direct feedback alignment, arXiv preprint arXiv:1906.04554, 2019.
[39]. J. Fajtl, V. Argyriou, D. Monekosso and P. Remagnino, Latent Bernoulli Autoencoder, in *Proceedings of the 37th International Conference on Machine Learning (PMLR'20)*, 2020, pp. 2964-2974..
[40]. K. Berggren, Q. Xia and K. Likharev, Roadmap on emerging hardware and technology for machine learning, *Nanotechnology*, Vol. 31, No. 1, 2020, p. 012002.
[41]. C. E. Rasmussen, Gaussian processes in machine learning, *Summer School on Machine Learning*, 2003.
[42]. A. G. Wilson, Z. Hu, R. R. Salakhutdinov and E. P. Xing, Stochastic variational deep kernel learning, in *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*, 2016, pp. 2594–2602.

Sensors & Transducers, Vol. 256, Issue 2, March 2022, pp. 12-18

www.sensorsportal.com

Sensors & Transducers

Published by IFSA Publishing, S. L., 2022
http://www.sensorsportal.com

Symmetry-based Method for Water Level Prediction using Sentinel 2 Data

[1, *] David JESENKO, [2] Lukáš HRUDA, [2] Ivana KOLINGEROVÁ, [1] Borut ŽALIK and [1] David PODGORELEC

[1] University of Maribor, Faculty of Electrical Engineering and Computer Science, Koroška cesta 46, SI-2000 Maribor, Slovenia
[2] University of West Bohemia, Faculty of Applied Sciences, Department of Computer Science and Engineering, Technická 8, 301 00, Plzeň, Czech Republic
[*] Tel.: +386-2-220-7476, fax: +386-2-220-7272
[*] E-mail: david.jesenko@um.si

Received: 31 January 2022 /Accepted: 1 March 2022 /Published: 31 March 2022

Abstract: The Sentinel satellite constellation series, developed and operated by the European Space Agency, represents a dedicated space component of the European Copernicus Programme, committed to long-term operational services in the environment, climate and security. A huge amount of acquired data allow us different surveys. The paper considers the detection of changes in water levels in Lake Cerknica. The multispectral index has been calculated from Sentinel-2 data and transformed to a 3D point cloud. As shown by the results, symmetry measures of 3D point clouds could be used for the detection of water levels. Prediction functions using a genetic algorithm have been fitted, and the best result achieved was RMSE = 0.9824.

Keywords: Sentinel-2, Lake Cerknica, Remote sensing, Water monitoring, Symmetry, Symmetry measure, Genetic algorithm.

1. Introduction

Monitoring open water bodies accurately is an important task and one of the basic applications in Remote sensing. They are a significant part of the Earth's water cycle, and water bodies such as rivers, lakes and reservoirs are irreplaceable for the global climate system and the ecosystem. Remote sensing has become a conventional approach for monitoring water bodies, as it is often real-time, dynamic and cost-effective [1]. The measuring and monitoring of surface water using Remote sensing technology is, therefore, an essential topic [2]. In particular, the use of freely available high-spatial resolution optical satellite data is relevant [3]. Such data include the images obtained by the Landsat series [4-6], Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) [7-8], and Sentinel-2 [1, 9] multispectral imagery. A high extraction accuracy has been achieved in the mapping of surface water bodies, including lakes [10], rivers [11], coastlines [12] and water bodies in rural areas [13].

Among all the existing water body mapping methods, the calculation of a multispectral index is the most reliable, as it is user-friendly, efficient and has low computational cost [14]. The use of the water index is currently accepted to enhance the differences between water and non-water bodies, based on combinations of two or more spectral bands using various algebraic operations [15]. The well-known Normalised Difference Water Index (NDWI) [13] is sensitive to built-up lands, and frequently results in the

overestimation of water bodies in urban areas [16]. The modified NDWI (MNDWI) [17] is used mostly in urban scenes to improve the separability of the built-up areas. The Automated Water Extraction Index (AWEI) [18] highlights the water bodies in urban areas against shadowed and dark surfaces. It consists of two separate indices, one for areas where shadow is not important, and a second one, where shadow is significant.

A number of methods exist that are designed to seek symmetries in 3D objects. Some aim to detect symmetries of rather general types, including reflectional ones such as [19] or the well-known [20] and its newer modification [21]. However, when trying to find specifically reflectional symmetries it is usually more convenient to employ methods that specialise in detecting this type of symmetry. Since reflectional symmetry is probably the most occurring symmetry type in real world objects, there are quite a few such methods, such as [22-25] among the older ones. The more recent ones include [26-31]. For our purpose we chose among the newer methods and selected the method of Hruda et al. [31] because it is recent, appears to be robust, fast and only requires a set of points on the input which is not the case with some other methods. A detailed description of the method [31] follows in the next Section.

The observed water unit in the presented paper is Lake Cerknica. It is one of the largest intermittent lakes in Europe. Lake Cerknica is located in the southwestern part of Slovenia, caught between the Javorniki hills and the Bloke plateau on one side, and Mount Slivnica on the other. It appears every year on the karst plain, and is present for about eight months of the year. Water usually spreads over a surface of 20 km2, but, at its fullest, the lake covers a surface of about 26 km2. The height above sea level is in the range from 546 m to 551 m, with the maximum depth about 10 m. When full it is the largest lake in Slovenia [32]. The lake is an important wildlife resort, especially as a nesting place for many bird species. During the dry season the lake disappears, which enables hiking and grass mowing, but, on the other hand, while present, allows for paddling and fishing. For these reasons it is crucial to detect water levels at different times of the year.

The methodology used in the presented paper is described in Section 2. The results are given in Section 3, while Section 4 concludes the paper.

2. Methodology

This Section presents the methodology used in this paper to determine the water levels in Lake Cerknica using Sentinel-2 data and symmetry measure [33]. Firstly, the Sentinel-2 data were acquired from the European Space Agency's (ESA) official website, then the multispectral index was calculated, followed by extraction of the area of Lake Cerknica from the multispectral index and calculation of the symmetry measure of the lake. The water level of Lake Cerknica is predicted finally. Each of these steps is explained additionally in the continuation. A flowchart of the proposed method is represented in Fig. 1.

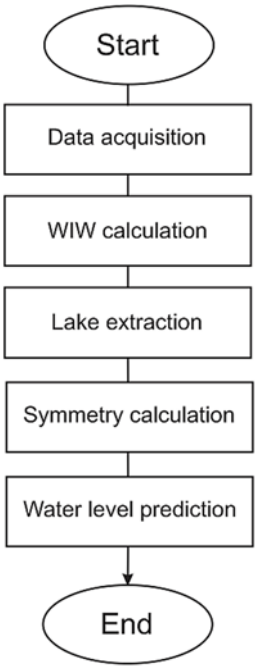


Fig. 1. Flowchart of the proposed approach.

Sentinel-2 data are composed of 13 spectral bands, that range from the visible range to the shortwave infrared. Data are freely available on the ESA website, accessible at https://scihub.copernicus.eu. The multispectral index, named the Water In Wetlands index (WIW), was calculated from the acquired data [33]. The WIW (Eq. 1) is defined by:

$$WIW = B8A \leq 0.1804 \;\&\&\; B12 \leq 0.1131, \quad (1)$$

where B8A represents a narrow Near Infra-Red (NIR) band, and B12 is a Short Wave Infra-Red (SWIR) band. In other words, flooded areas could be distinguished from dry areas when the pixels satisfy the conditions in Eq. 1 [34].

The area of Lake Cerknica was extracted using a mask, which, in the detail, describes the entire area of the lake. The extracted pixels from the WIW index were converted from 2D to 3D in such a way that a pixels with x and y coordinates, contained in the area of the flooded lake, store the intensity (coordinate z) of the WIW index.

2.1. Symmetry Estimation

The symmetry was calculated using the method presented in [31]. The method is designed to find the plane of reflectional symmetry of a 3D point set, and is based on maximizing an objective function called symmetry measure. Having a set of points

$X = \{x_1, x_2, \ldots, x_n\}$, $x_i \in E^3$, $i = 1, \ldots, n$, and a plane $P: ax + by + cz + d = 0$ represented by a vector $p = [a, b, c, d]^T$ the symmetry measure is defined as follows (Eq. 2):

$$s_X(p) = \sum_{i=1}^{n} \sum_{j=1}^{n} \varphi\left(\left\| r(p, x_i) - x_j \right\|\right), \quad (2)$$

where $r(p, x) \in E^3$ is a function that reflects a point $x \in E^3$ over the plane represented by p, and $\varphi(l)$ is a locally supported differentiable Wendland's function [35] which resembles the Gaussian for $l \geq 0$, and basically measures the similarity of two points based on their distance. Each point of the point set X is reflected over the plane represented by p and its similarity is computed to all points. All these similarities are summed together, providing the symmetry measure value. The idea behind the method is that maximizing the symmetry measure for p should maximise the similarity of the point set to its reflected version, and therefore maximize the symmetry over the plane represented by p. Since φ is locally supported, the distances only need to be computed for nearby points, and the computation of the symmetry measure can easily be accelerated using a 3D grid. Also, the symmetry measure is differentiable with regard to p, so any gradient-based optimization technique can be employed to locate its maximum.

Having the input point set X, two simplified versions are created, one denoted X_{simp} with approximately 1,000 points intended for the symmetry measure computation, and one denoted X_{cand} with roughly 100 points for candidate creation. The number of symmetry plane candidates are created by pairing points of X_{cand}, and each candidate p_i is evaluated by the symmetry measure $s_{X_{simp}}(p_i)$. Subsequently, S candidates with the largest value of the symmetry measure are selected, and local numerical optimization using the L-BFGS method [36] is started from all of them. This provides S local maxima of the symmetry measure $s_{X_{simp}}$, and selecting the one with the largest measure value results in the final symmetry plane.

2.2. Genetic Algorithm

This subsection presents an evolutionary algorithm, more precisely a genetic algorithm, for finding the most suitable predictive function [37, 38]. The function will predict the water level in Lake Cerknica based on estimated symmetry measure in the previous subsection. The task of the genetic algorithm is estimation of the coefficients for the predictive function. In a genetic algorithm the coefficients are represented as individuals. The process of estimation of coefficients consists of a few main steps, which are, generation of the initial population, selection, crossover, mutation, evaluation of individuals and the stopping criteria. A flowchart and the main steps can be seen in Fig. 2. In the continuation each of the main steps will be explained in detail.

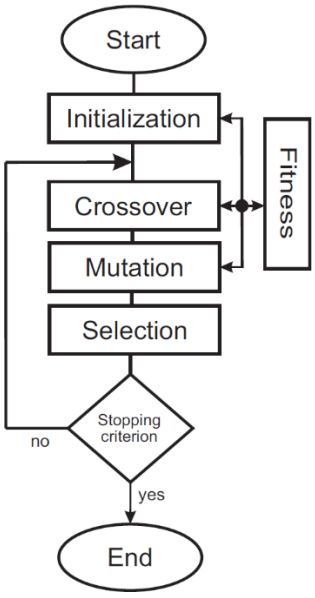


Fig. 2. Flowchart of the proposed genetic algorithm.

As mentioned, the following steps are in charge of finding the most suitable solution:

- **The initial population** was generated first. It consists of a set of randomly generated individuals. Each individual consists of chromosomes, and each chromosome represents a completely randomly generated coefficient for the predictive function.
- **Selection** of the individuals is a very important step towards finding the most suitable solution (predictive function). The tournament selection of size two is used, which means, that two individuals from the population are selected and compared, and, finally, the one with the best fitness, in our case the lowest Root-Mean-Square Error (RMSE), continues into the next generation.
- **Crossover** is used to generate new individuals and develop the population by combining the fittest individuals (chromosomes) from the previous generation. Two parents were selected randomly from the population, and a new individual was generated by copying the chromosomes from the first or the second parent. A two-point crossover was applied, which means that two crossover points are picked randomly from the parents and the chromosomes were swapped between the two points.
- **Mutation** is performed over randomly chosen individuals from the population, to allow new genetic material to enter, and, thus, keep the population diverse. Chromosomes from individuals are also selected randomly, and modified by a small positive or negative value.
- **Evaluation of individuals,** also called a fitness function, is performed by measuring the efficiency of the forecast made by each individual, which serves as the predictive function. Efficiency is measured by the already mentioned RMSE.

- **Stopping criteria** terminates the execution of the genetic algorithm. Three different criteria have been used. The first one, which is almost never achieved in practice, is when RMSE is equal to 0, the second one occurs when the minimal RMSE is not changed in five consecutive generations. The last one is triggered when 4,000 individuals are evaluated. If the stopping criteria are not satisfied, the proposed genetic algorithm repeats the basic steps (crossover, mutation, and selection).

3. Results

3.1. Parameter Sensitivity

The proposed method (genetic algorithm) for prediction relies on five input parameters. These parameters are the size of the population, the probability of crossover, the probability of mutation, the coefficients' change range, and the number of individual evaluations. In order to examine their influences on the method performance and estimate their optimal values, a systematic sensitivity analysis was performed, as proposed by Chang and Delleur [39, 40]. The well-known dataset Iris [41] in machine learning was used for this purpose. The ranges of the input parameter values were first defined during the sensitivity analysis. The size of the population was from the range [10, 200], while the coefficients' change range was on the interval [-1, 1]. The probability of mutation and of crossover were from the range [0.1, 1], and, finally, the number of individual evaluations from the interval [100, 10,000]. Taking these restrictions into account, 3,000 sets of parameter values were generated randomly, and the proposed genetic algorithm was evaluated accordingly. The obtained results were classified as acceptable if the run finished with a score of correctly predicted samples of more than 95%; otherwise the results were regarded as unacceptable. The cumulative frequencies were plotted of acceptable and unacceptable results. Insensitive parameters are characterized by similar cumulative frequencies of acceptable and unacceptable results throughout the whole range of parameter values. The following parameters were classified as insensitive: The probability of crossover, the probability of mutation, the coefficients' change range and the number of individual evaluations. On the other hand, the size of population was recognized as sensitive. In that case, a significantly larger number (approximately 30.2%) of acceptable results was achieved when its value was equal to 40 than in any other case. While influences of the sensitive parameters on the method's performance could be recognized straightforwardly, the optimal values of the insensitive parameters have to be tuned accordingly. In our case, a model-based approach with the linear regression was applied, as proposed in [42]. The selected optimal values for the genetic algorithm's parameters can be seen in Table 1.

Table 1. Parameters and their optimal values.

Parameter	Value
Probability of crossover	70 %
Probability of mutation	40 %
Size of the population	40
Number of individual evaluations	4,000
Coefficients' change range	[-0,2. 0.2]

3.2. Data Processing

Sentinel-2 data processing is a demanding and computer intensive process. An AMD Ryzen 5 4500U with 32 GB of main memory on Windows 10 was used for these reasons. In Fig. 3 we can see the exact location of Lake Cerknica in Slovenia. Fig. 4 shows the whole area of Lake Cerknica after the calculated WIW index and applied extraction mask. The Sentinel-2 data used for the calculation of WIW presented in Fig. 4 were acquired on 24 February, 2021. At that time, most of the lake was flooded with water (the blue pixels in Fig. 4), and only the east part and a small portion of the centre of the lake were dry.

Fig. 3. Location of Lake Cerknica in Slovenia.

Fig. 4. Area of Lake Cerknica, where blue pixels represent the flooded area of the lake.

Fig. 5 represents only the flooded parts of the lake. Recently, the flooded parts were converted from 2D to 3D, and, in that way, prepared for the calculation of the symmetry measure, which was calculated finally. The black line in Fig. 5 represents the plane of the 3D object of the flooded Lake Cerknica where the largest symmetry measure value was achieved. The calculated symmetry measure for the flooded parts in Fig. 5 was 1016.89. Then, the several different flooded areas of the lake were tested (see Fig. 6).

When the whole area of Lake Cerknica was flooded, the symmetry measure equaled 1,588.41, but, at the same time, when the lake was flooded less than presented in Fig. 5, the symmetry measures were also lower than in Fig. 5. An example of this can be seen in Fig. 6, where the symmetry measure was 983.77. This indicates that the lower the symmetry value is, the lower is the water level in Lake Cerknica. For the completely dry lake we assumed that the symmetry measure was equal to 0 and the surface elevation (above the sea level) was 541 m, while, at the surface level of 551 m, the symmetry measure was 1,588.41 (the whole lake was flooded). All the other calculated symmetry measures were in the mentioned range (from 0 to 1,588.41).

3.3. Interpretation of Prediction Model

The obtained prediction functions using the genetic algorithm presented in subsection 2.2 can be seen in Table 2.

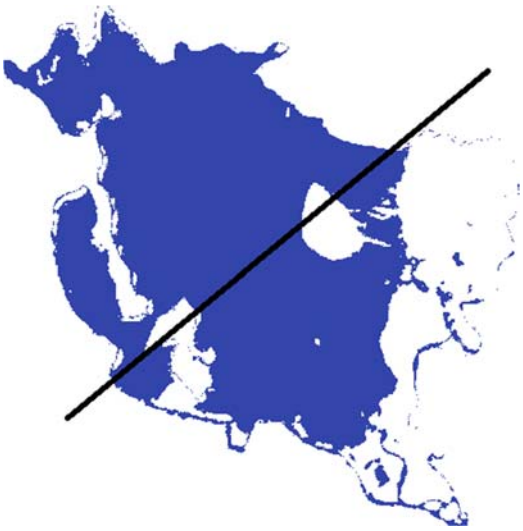


Fig. 5. Only the flooded area of Lake Cerknica, where the black line represents the largest symmetry value.

Fig. 6. Only a small part of the flooded area of Lake Cerknica, where the symmetry measure was 983.77.

Table 2. Prediction functions and their achieved RMSE.

ID	Prediction function	RMSE
1	$SE = -0.0042\ SM^3 + 4.1911\ SM^2 + 0.7824\ SM + 0.6433\ \sqrt{SM} + 0.2341$	0.9824
2	$SE = -0.0011\ SM^3 + 1.1708\ SM^2 - 4.6062\ SM - 4.5357\ \sqrt{SM} + 0.1523$	0.9963
3	$SE = -0.0012\ SM^3 + 1.1293\ SM^2 - 7.0471\ SM - 0.3752\ \sqrt{SM} + 0.2911$	1.0574
4	$SE = -0.0129\ SM^3 + 13.1534\ SM^2 - 1.4888\ SM + 0.4396\ \sqrt{SM} + 0.1472$	1.0641
5	$SE = -0.0028\ SM^3 + 2.9462\ SM^2 - 7.3371\ SM - 0.6071\ \sqrt{SM} + 0.0852$	1.0783
6	$SE = -0.0024\ SM^3 + 2.4943\ SM^2 + 9.0621\ SM + 3.2507\ \sqrt{SM} + 0.0513$	1.0969
7	$SE = -0.0018\ SM^3 + 1.8412\ SM^2 - 5.9313\ SM - 1.1612\ \sqrt{SM} + 0.0974$	1.1045
8	$SE = -0.0056\ SM^3 + 5.6911\ SM^2 - 5.9312\ SM - 1.1614\ \sqrt{SM} + 0.1692$	1.1103
9	$SE = -0.0034\ SM^3 + 3.4957\ SM^2 + 2.2993\ SM - 1.9101\ \sqrt{SM} + 0.1396$	1.1163
10	$SE = -0.0065\ SM^3 + 6.5582\ SM^2 - 4.7865\ SM - 2.4498\ \sqrt{SM} + 0.1623$	1.1201
11	$SE = -0.0043\ SM^3 + 4.8641\ SM^2 - 1.9744\ SM + 1.2902\ \sqrt{SM} + 0.4910$	1.1300
12	$SE = -0.0017\ SM^3 + 9.7071\ SM^2 + 2.3065\ SM - 0.0994\ \sqrt{SM} + 0.2901$	1.1309
13	$SE = -0.0052\ SM^3 + 3.8514\ SM^2 + 1.0087\ SM - 0.1851\ \sqrt{SM} + 0.1493$	1.1312
14	$SE = -0.0029\ SM^3 + 7.9763\ SM^2 - 2.6295\ SM + 0.9652\ \sqrt{SM} + 0.2475$	1.1331
15	$SE = -0.0038\ SM^3 + 1.5629\ SM^2 + 5.1702\ SM + 2.5801\ \sqrt{SM} + 0.6029$	1.1349

In all prediction functions SE represents the predicted Surface Elevation, and SM is the calculated Symmetry Measure. The performance of fitted functions was measured using RMSE. Using that type (predictive functions) of representation of a prediction model not only achieves very good results, but also allows us the interpretation of the learned functions. From the obtained functions it can be observed that all symmetry measures that were cubed (SM^3), have negative coefficients, which means that they have a negative influence on the final result. On the other hand, all squared (SM^2) symmetry measures have a positive influence due to their positive coefficients. In the case of symmetry measure (SM) and rooted symmetry measure ($\sqrt{SM}$) the influences are hard to explain, as some have negative, and some have positive coefficients. Finally, all the free coefficients have positive values and have positive influence on the final result. All the coefficients at SM^3 are very small and have small influence. There could be two main reasons: The first is that they do not have much influence, and the second, more likely, is due to the very large numbers that occur in the case of cubed values. Fifteen prediction functions with the highest achieved accuracies can be seen in Table 2. The smallest RMSE was equal to 0.9824, which means that the prediction error, on average, was less than one metre. As the difference between the lowest and highest surface elevations (above the sea level) is about 10 m, the achieved results show an error less than 10%, which is, if we take everything into account (Sentinel-2 resolution, shadows, clouds), a very good result.

4. Conclusion

The methodology for detection of water levels in the intermittent Lake Cerknica using a multispectral index acquired from Sentinel-2 and symmetry measure is presented in this paper. The results presented in the previous Section show that the symmetry measure of the 3D point cloud generated from the WIW index could be used for the prediction of water levels. The learned prediction functions achieved good results, and can be used for the prediction of surface levels. The use of ground truth data of the measured surface elevations of Lake Cerknica (e.g. sending experts to measure the exact elevation) and a bigger learning set (a combination of measured surface elevations and calculated symmetry measures), will be the main topic of the future work. The influence of other factors, such as shadows and clouds, on the quality of the WIW index will also be part of our future research work.

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency (Research Core Funding No. P2-0041 and Project No. N2-0181). This work was also supported by the Czech Science Foundation, Project 21-08009K Generalized Symmetries and Equivalences of Geometric Data.

References

[1]. Du Y., Zhang Y., Ling F., Wang Q., Li W., and Li X., Water bodies' mapping from Sentinel-2 imagery with modified normalized difference water index at 10-m spatial resolution produced by sharpening the SWIR band, *Remote Sensing*, Vol. 8, No. 4, 2016, p. 354.

[2]. Du N., Ottens H., and Sliuzas R., Spatial impact of urban expansion on surface water bodies: A case study of Wuhan, China, *Landscape and Urban Planning*, Vol. 94, No. 3, 2010, pp. 175-185.

[3]. Pekel J. F., Cottam A., Gorelick N., and Belward A. S., High-resolution mapping of global surface water and its long-term changes, *Nature*, Vol. 540, No. 7633, 2016, pp. 418-422.

[4]. Tulbure M. G. and Broich M., Spatiotemporal dynamic of surface water bodies using Landsat time-series data from 1999 to 2011, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 79, No. 1, 2013, pp. 44-52.

[5]. Singh K., Ghosh M., and Sharma S. R., WSB-DA: Water surface boundary detection algorithm using Landsat 8 OLI data, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* Vol. 9, No. 1, 2015, pp. 363-368.

[6]. Acharya T. D., Lee D. H., Yang I. T., and Lee J. K., Identification of water bodies in a Landsat 8 OLI image using a J48 decision tree, *Sensors*, Vol. 16, No. 7, 2016, pp. 1075-1091.

[7]. Sivanpillai R. and Miller S. N., Improvements in mapping water bodies using ASTER data, *Ecological Informatics,* Vol. 5, No. 1, 2010, pp. 73-78.

[8]. Zhou Y., Luo J., Shen Z., Hu X., and Yang H., Multiscale water body extraction in urban environments from satellite images, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 7, No. 10, 2014, pp. 4301-4312.

[9]. Yang X., Zhao S., Qin X., Zhao N., and Liang L., Mapping of urban surface water bodies from Sentinel-2 MSI imagery at 10 m resolution via NDWI-based image sharpening, *Remote Sensing*, Vol. 9, No. 6, 2017, pp. 596-615.

[10]. Bhardwaj A., and Singh M. K., Joshi P. K., Singh S., Sam L., Gupta R. D., and Kumar R., A lake detection algorithm (LDA) using Landsat 8 data: A comparative approach in glacial environment, *International Journal of Applied Earth Observation and Geoinformation*, Vol. 28, No. 1, 2015, pp. 150-163.

[11]. Jiang H., Feng M., Zhu Y., Lu N., Huang J., and Xiao T., An automated method for extracting rivers and lakes from Landsat imagery, *Remote Sensing*, Vol. 6, No. 6, 2014, pp. 5067-5089.

[12]. Li W. and Gong P., Continuous monitoring of coastline dynamics in western Florida with a 30-year time series of Landsat imagery, *Remote Sensing of Environment*, Vol. 179, No. 1, 2016, pp. 196-209.

[13]. McFeeters S. K., The use of the normalized difference water index (NDWI) in the delineation of open water features, *International Journal of Remote Sensing*, Vol. 17, No. 7, 1996, pp. 1425-1432.

[14]. Ryu J.-H., Won J.-S., and Min K. D., Waterline extraction from Landsat TM data in a tidal flat: a case

study in Gomso Bay, Korea, *Remote Sensing of Environment*, Vol. 83, No. 3, 2002, pp. 442-456.
[15]. Yang X., Qin Q., Grussenmeyer P., and Koehl M., Urban surface water body detection with suppressed built-up noise based on water indices from Sentinel-2 MSI imagery, *Remote Sensing of Environment*, Vol. 219, No. 1, 2018, pp. 259-270.
[16]. Huang X., Xie C., Fang X., and Zhang L., Combining pixel-and object-based machine learning for identification of water-body types from urban high-resolution remote-sensing imagery, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 8, No. 5, 2015, pp. 2097-2110.
[17]. Xu H., Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery, *International Journal of Remote Sensing,* Vol. 27, No. 14, 2006, pp. 3025-3033.
[18]. Feyisa G. L., Meilby H., Fensholt R., and Proud S. R., Automated water extraction index: a new technique for surface water mapping using Landsat imagery, *Remote Sensing of Environment*, Vol. 140, No. 1, 2014, pp. 23-35.
[19]. Lipman Y., Chen X., Daubechies I., and Funkhouser T., Symmetry factored embedding and distance, *ACM Transactions on Graphics*, 29, 4, 2010, p. 103.
[20]. Mitra N. J., Guibas L. J., and Pauly M., Partial and approximate symmetry detection for 3D geometry, *ACM Transactions on Graphics (TOG),* Vol. 25, No. 3, 2006, pp. 560-568.
[21]. Shi Z., Alliez P., Desbrun M. Bao H., and Huang J., Symmetry and orbit detection via lie-algebra voting, *Computer Graphics Forum*, Vol. 35, No. 5, 2016, pp. 217-227.
[22]. Sun C. and Sherrah J., 3D symmetry detection using the extended Gaussian image, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 2, 1997, pp. 164-168.
[23]. Podolak J., Shilane P., Golovinskiy A., Rusinkiewicz S., and Funkhouser T., A planar-reflective symmetry transform for 3D shapes, *ACM SIGGRAPH 2006 Papers*, Vol. 1, No. 1, 2006, pp. 549-559.
[24]. Simari P., Kalogerakis E., and Singh K., Folding meshes: Hierarchical mesh segmentation based on planar symmetry, in *Proceedings of the Fourth Eurographics Symposium on Geometry Processing,* Cagliari, Sardinia, Italy, June 26-28, 2006, pp. 111-119.
[25]. Combes B., Hennessy R., Waddington J., Roberts N., and Prima S., Automatic symmetry plane estimation of bilateral objects in point clouds, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1-8.
[26]. Sipiran I., Gregor R., and Schreck T., Approximate symmetry detection in partial 3D meshes, *Computer Graphics Forum*, Vol. 33, No. 7, 2014, pp. 131-140.
[27]. Li B., Johan H., Ye Y., and Lu Y., Efficient 3D reflection symmetry detection: A view-based approach, *Graphical Models*, Vol. 83, No. 1, 2016, pp. 2-14.
[28]. Speciale P., Oswald M. R., Cohen A., and Pollefeys M., A symmetry prior for convex variational 3D reconstruction, in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 313-328.
[29]. Ecins A., Fermuller C., and Aloimonos Y., Detecting reflectional symmetries in 3D data through symmetrical fitting, in *Proceedings of the IEEE International Conference on Computer Vision Workshops,* 2017, pp. 1779-1783.
[30]. Nagar R. and Raman S., 3dsymm: robust and accurate 3D reflection symmetry detection, *Pattern Recognition,* Vol. 107, No. 1, 2020, p. 107483.
[31]. Hruda L., Kolingerova I., and Vaša L., Robust, fast and flexible symmetry plane detection based on differentiable symmetry measure, *The Visual Computer*, 2021, pp. 1-17.
[32]. Šmid H., Golež G., Podjed D., Kladnik D., Erhartič B., Pavlin P., and Jerele I., Encyclopedia of Natural and Cultural Heritage in Slovenia, Ljubljana, 2012.
[33]. Jesenko D., Hruda L., Kolingerova I., Žalik B., and Podgorelec D., Detection of water levels in Lake Cerknica using Sentinel-2 data and symmetry, in *Proceedings of the 3rd International Conference on Advances in Signal Processing and Artificial Intelligence,* 2021, pp. 65-69.
[34]. Lefebvre G., Davranche A., Willm L., Campagna J., Redmond, L., Merle C., Guelmami A., and Poulin B., Introducing WIW for detecting the presence of water in wetlands with landsat and sentinel satellites, *Remote Sensing*, Vol. 11, No. 19, 2019, p. 2210.
[35]. Wendland H., Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree, *Advances in Computational Mathematics*, Vol. 4, No. 1, 1995, pp. 389-396.
[36]. Liu D. C. and Nocedal J., On the limited memory BFGS method for large scale optimization, *Mathematical Programming*, Vol. 45, No. 1, 1989, pp. 503-528.
[37]. Lešnik U, Mongus D., and Jesenko D., Predictive analytics of PM10 concentration levels using detailed traffic data, Transportation research. Part D, *Transport and Environment,* Vol. 67, No. 1, 2019, pp. 131-141.
[38]. Mongus D., Vilhar U., Skudnik M., Žalik, B., and Jesenko D., Predictive analytics of tree growth based on complex networks of tree competition, *Forest Ecology and Management,* Vol. 425, No. 1, 2018, pp. 164-176.
[39]. Chang F. J. and Delleur J. W., Systematic parameter estimation of watershed acidification model, *Hydological Processes*, Vol. 6, No. 1, 1992, pp. 29-44.
[40]. Jesenko D., Mernik M., Žalik B., and Mongus D., Two-level evolutionary algorithm for discovering relations between nodes' features in a complex network, *Applied Soft Computing*, Vol. 56, No. 1, 2017, pp. 82-93.
[41]. Blake C. L. and Merz C. J., UCI repository of machine learning databases, *University of California*, 1998, pp. 1-2.
[42]. Czarn A., MacNish C., Vijayan K., Turlach B., and Gupta R., Statistical exploratoryanalysis of genetic algorithms, *IEEE Transactions on Evolutionary Computation,* Vol. 8, No. 4, 2004, pp. 405-421.

Sensors & Transducers, Vol. 256, Issue 2, March 2022, pp. 19-26

www.sensorsportal.com

Sensors & Transducers

Published by IFSA Publishing, S. L., 2022
http://www.sensorsportal.com

A Description of Data Handling Practices and Software Tools Developed by the Boston Children's Hospital Echo Core Laboratory

[1 *] Edward Marcus, [2] Meena Nathan, [1] Patrick McGeoghegan and [1] Kevin Friedman

[1] Department of Cardiology, [2] Department of Cardiac Surgery,
Boston Children's Hospital, Boston, Massachusetts
9 Hope St., Waltham, Massachusetts, USA
Tel: (001) 781-216-2532
E-mail: ed.marcus@cardio.chboston.org

Received: 25 January 2022 /Accepted: 28 February 2022 /Published: 31 March 2022

Abstract: Ultrasound 'Echocardiogram (echo)' images generated by commercial scanners are non-invasive, economical, and have therefore become an important clinical and research tool for cardiovascular disease. The Boston Children's Hospital (BCH, USA) cardiology department has developed echo image handling processes and software tools to simplify and manage the flow of echo data from sites participating in multi-center research studies. With frequent needs to correlate echocardiographic findings with surgical outcomes and more recently, better understand the acute and long-term effects of the SARS-CoV-2 pandemic on the heart in a multi-institutional collaborative [1], the need for a central corelab for image review and reporting has spurred the development of new software tools that seamlessly prepare, transmit and measure echocardiographic images and generate reports formatted to merge easily with clinical databases. In this paper we provide a descriptive guide to these software tools, and offer suggested approaches to data preparations that can simplify corelab processing.

Keywords: Echo Corelab, Health insurance portability and accountability act (HIPAA), Protected health information (PHI), Ultrasound images, Structured reporting.

1. Introduction: Echocardiographic Corelab

The advent of the echocardiogram (echo) as an economical and non-invasive imaging modality has provided new opportunities for multicenter research in cardiovascular diseases. In the early phase of echocardiographic imaging (pre 2000) most echo laboratories initially utilized video cassette tapes [2] to record and review echo exams that ranged in duration from 15 minutes to one hour. More recently, echo device manufacturers have adopted the Digital Imaging and Communication in Medicine (DICOM) standard [3] to enable digital communications across various devices and to allow for uniformity in imaging data interpretation using software that is specifically designed around this standard.

Fundamentally, an echo corelab is organized as a centralized facility to receive DICOM echo images performed by participating sites of a research network. Boston Children's Hospital (BCH) corelab has the capacity to handle data from 30 or more sites.

In the 20 years BCH has served as an echo corelab, nearly two dozen research protocols have been implemented, to generate publications on topics ranging from anatomy and pathophysiological effects of congenital heart disease [4], efficacy of pharmaceutical interventions [5], inter observer variability of measurement [6], and comparisons of outcomes from differing disease management strategies [7].

Multi-site research protocols can generate thousands of echocardiographic studies for management by the corelab. Each of these studies undergoes progressive stages of processing from initial scanning, to transmission and eventual reading by the corelab's echo technicians and echocardiographers. Each study received by the corelab is also checked and stored for easy access from a reading list. From a design perspective, at each of these preparation steps, implemented software should have the capacity to assign and recognize study identifiers (ID), ensure patient identifiers are anonymized, and to seamlessly transmit de-identified study echos for the corelab review [8] .

A standard corelab research project begins with the definition of research aims, followed by the enrollment of participating sites [9], development of subject inclusion/exclusion criteria, preparation of training materials, budget forecasting, and specification of software tools to be implemented at progressive stages of the research data pipeline [8]. These challenges require expertise from physicians with experience in the diseases being studied, as well as the supporting efforts by administrators, data managers, information technology (IT) professionals, and software engineers to implement needed processes and tools.

The echo corelab also works in close collaboration with a centrally administered data coordinating center (DCC). The DCC has responsibility for monitoring patient enrollments, developing maintained clinical databases, and managing the centralized channels of communication between participating sites of the research network [10]. The corelab's partnership with DCC includes creating training materials for image acquisitions and transfers. Additionally, corelab reports of measurement data are provided for upload to the study specific database at the DCC. This requires appropriate formatting of all corelab data for merging into databases maintained at the DCC.

2. Methods

2.1. Study Identifiers

Before a corelab protocol is implemented, it is helpful to have some estimate of: 1) The number of participating research sites; 2) The number of subjects to be enrolled, and 3) The number of echo studies to be performed on each enrolled subject. These preliminary estimates will determine the required storage capacity for the corelab, and allow for appropriate formatting of each study's identifier (Study ID).

Study IDs are designed to uniquely identify each recorded image set. The encoding of the study ID is vital to identify echo data through each stage of image processing, review, and reporting. BCH study IDs are encoded with these identifying components (Fig. 1):

Protocol ID : Protocol IDs are the component of the Study ID needed to link the dataset with the appropriate research project.

Site ID : Site ID's are a component of the Study ID needed to indicate the site from which the study was received.

Subject ID : Subject IDs are proxy references that can link to mapped patient medical records maintained by sites. For the corelab, the subject ID therefore provides a means to communicate with the site about a particular subject, without having to access the subject's private identifying information.

Time Point ID: Corelab protocols frequently consist of multiple follow-up echocardiograms for a given subject. Starting with a baseline diagnosis scan (scan 1), the subject might undergo multiple follow-up echocardiograms over a period of days, months or years. Each scan's placement in this follow up sequence is specified by the Time Point ID.

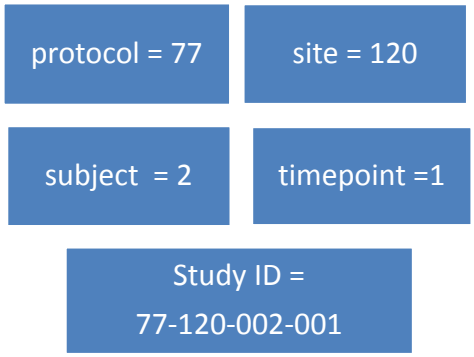


Fig. 1. The component factors, which uniquely identify each echo study include the protocol, site, subject, and time point.

The planning of anticipated site and subject participations is a key component of the Study ID's format to be implemented for a research protocol. For example, if participating site counts will not exceed 99, the format of the Study Id's site component can be limited to two digits. Similarly, if subject participation counts will not exceed 9999, then four digits of the Study ID's subject component will suffice to identify each subject.

2.2. Protected Health Information (PHI)

Since the introduction of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) [11], site research data provided to outside entities

needs to be 'scrubbed' of protected health information (PHI) that can reveal a subject's identity. With an echo data set, there are two basic types of PHI that need be removed at the site before the images can be transmitted.

The first type of PHI is found in the embedded pixels of each echo image, which may include textual information of the research subject's name or medical record number. With most image formats, this identifying information is visualized at the very top regions of the echo image, and is readily accessed for image redactions by software (Fig. 2).

A second type of PHI is found in DICOM tag elements which contain the metadata descriptions of each image file. These data elements include information such as the research subject's name, medical record number, address, telephone, or date of birth. Our laboratory has implemented a tool, which applies open source DICOM methods [12] to reconfigure PHI sensitive data elements by automatically de-identifying these elements with substitute or blank values.

In practice, techniques for PHI removal may be site specific and some participating sites may apply preferred software packages for the de-identification. Different sites may also have hospital specific PHI regulations related to types of information that can be released. However, to ensure compliance with HIPPA, the corelab always performs a final check to confirm that all PHI elements are redacted. An ideal solution for image preparations in the future would be a single software package suitable for all sites, irrespective of the echo platform used, standardizing the method of

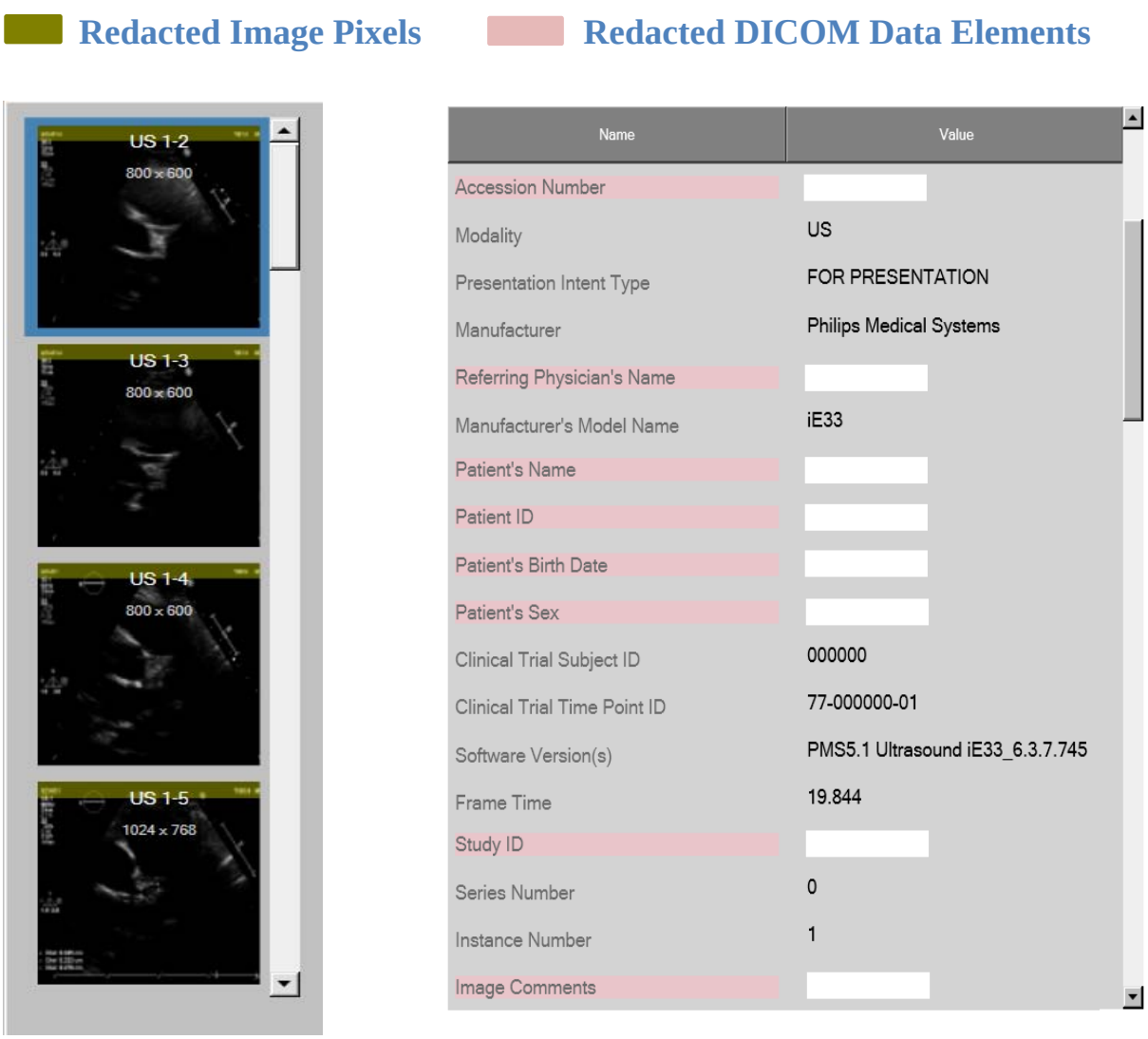


Fig. 2. Protected health information (PHI) is embedded as pixels in the echo images, or as the values contained within each image file's DICOM data elements. Pixels are redacted by manually identifying image regions containing PHI, and DICOM data elements are cleared by a configured automatic process to remove elements with potentially sensitive PHI.

image de-identifications by sites prior to the corelab transfer.

2.3. Patient Characteristics

In addition to removing PHI, the site will also provide metadata information for purposes of characterizing the study subject, and for inclusion of the data as part of defined subject categories. Information on age, sex, height, weight, blood pressure, or other vital data will allow for measures at the corelab to be normalized for statistical analysis. In a pediatric study there will be a spectrum of different ages and stages of developmental growth; thus, the normal ranges for a cardiac measure depend on factors such as age, height, weight, or body surface area. Standardization allows for the assessments of whether a measurement is within or outside of a defined normal range. Mathematically, these normative values are based on standard deviations from normal population mean, and are collectively known as measurement 'z scores' [13, 14].

2.4. Data Uploads From Sites

Once PHI data is removed, the sites upload images to a central and secure data server. For the images to upload, BCH has implemented a system for synchronization between echo data folders on site computers and objects of archiving systems contained within the Amazon Web Service (AWS) [15, 16].

The AWS archiving system is convenient for several reasons. First, the AWS upload occurs rapidly and reliably. As AWS is open source, needed configuration changes such as new user accounts, can be made by the corelab directly in the study specific space of the AWS system. Finally, for large studies the AWS storage system is cost-effective, and provides for a tiered storage pricing based on frequency of data access.

2.5. Image Transfers to the Corelab

From the AWS study archives, the corelab can download and review each study. However, prior to the downloaded study being released for review by the corelab reader, a quality assurance check is performed. When images are first received, preliminary checking procedures will compare image file counts with a manifest of expected files from a DICOM directory. If missing images are identified, then the site can be informed, and a data resend request might be activated.

After all study content is confirmed as received, each image is then tested to ensure that it is readable. Although the DICOM standard dictates a generic image format, DICOM accepts a wide variety of specific image formats, some of which may not have been previously encountered by the corelab's reading software. When a new format arrives that cannot be rendered by the existing software version, the software is modified and redesigned to accommodate the new image format.

Another important factor to be checked is to ensure that there has not been inadvertent 'mixing', which can occur if images from two or more study subjects (or study time points) have been combined into the same study data folder. To prevent study mixing, the DICOM standard's unique identifiers and study dates are assessed for each folder. If more than one study is indicated to be in the folder, the mixing indications are flagged to be further assessed and to consider that a resend might be appropriate.

The final check of received data is the visual inspection of images for PHI redactions that may have been inadvertently missed by the site's de-identifications. A typical echo study can have 100 or more images, and checking all the images can be labor intensive. To mitigate this, we have developed a system to automatically identify 'key images' containing PHI. A key image is a representative image in a format that is shared by other images in the study. Using this method, the redaction of pixels identified in the key image, can be automatically applied to the pixels of other image files with the same format as the key image.

2.6. Corelab Accession Numbers

When a study is checked and cleared for echo reading, the data are assigned with an accession number to specify the order of the study's arrival to the corelab. Starting with '1' (the first received study) the subsequent arrival of new studies from different sites will lead to progressively incremented accession numbers.

Accession numbers serve two important purposes. First, the accession number is an implicit connection to the counts of all studies received to date. The second feature of accession numbers is the simplicity the numbers provide to reference the study in ways that are more easily communicated verbally than the more complex study ID. The accession number becomes an effective pointer to a particular study at the corelab, while the Study ID contains supplemental data components with further details about the study, such as the originating site, subject number, and echo scanning time point.

There are two basic techniques for study accessions. An unstructured accession process will accept studies in the order in which they are received. A structured accession process will order study placements into the reading list in a specific way that might be guided by study protocols. For example, if a subject is to have three scanning time points and these scans are to be reviewed in sequence, then the studies are assigned three consecutive accession numbers. While structured accessions save time during the echo read, they are harder to implement.

2.7. Reading List Activation

The final step to prepare a study for the reading list is to activate a flag that will serve as an indication of the study's progress through the reading pipeline. At our corelab, these enumerated flags show the study's progress through the different stages of the review:

1) *Active Flag*: the study is in the process of being measured.
2) *Signed Flag*: the study measures are completed and ready for sending to the DCC.
3) *Sent Flag*: the study measures have been successfully transmitted to the DCC, and are locked from additional edits.
4) *Retired Flag*: the study measures are not to be sent to the DCC.

When the study is first placed into the reading list, the flag is set to 'active'. When all measures are completed, the study is 'signed' by the reviewer. Signed results are sent to the DCC and the 'sent' flag is applied. Studies indicated for withholding of data from the DCC are designated to be 'retired'.

3. Reviewing, Measurement and Reporting

3.1. The Echo Report

Corelab physicians and echo technologists apply a custom designed software program to render images and trace geometric tissue measurements as image overlays. These measures are presented to the reviewer as a checklist of items that form the content of each study's structured echo report. This report checklist will have been specified by the group of principal investigators with knowledge about the measures considered to be important and essential to fulfill the specific aims of the research. All needed measurements are then organized into the report as sections of a tree data structure that can be completed by reviewers as they read the echo.

Content of the echo report generally includes items that are classified into three basic categories: 1) Traced Measures 2) Calculations and 3) Coded Assessments. Each of these item types is configured in a similar fashion. First, the item of interest is selected from a dictionary of coded concepts [17] for inclusion into the report. The selected dictionary item is then characterized by its unique code to identify what will be reported. The item's presentation in the report's interface may also be customized and presented to the reviewer with a helpful guide to the intended measurement's anatomical imaging location. Measured content also has the specification of desired units, and display precisions. Finally, each item of the report is exported to the DCC with a unique database field code that is mapped directly to the DCC's master database.

'Traced' items of the echo report are based on the reviewer generated outlines drawn as image overlays. These measures might include point-to-point distance assessments (Fig. 3), or a measure of the heart chamber's cross-sectional area visualized from a 2-dimensional echo view [18]. Velocity of blood flows or tissue motions can also be quantified from the echo scanner's application of Doppler processing [19].

Calculated report items are automatically assessed by computing derivations based on the inputs provided from the tracing measures. Reported calculations might consist of the heart chamber's volume reconstruction [20], a ratio between two traced measures, or the statistical identification of a measurement outlier from an expected normal range [13].

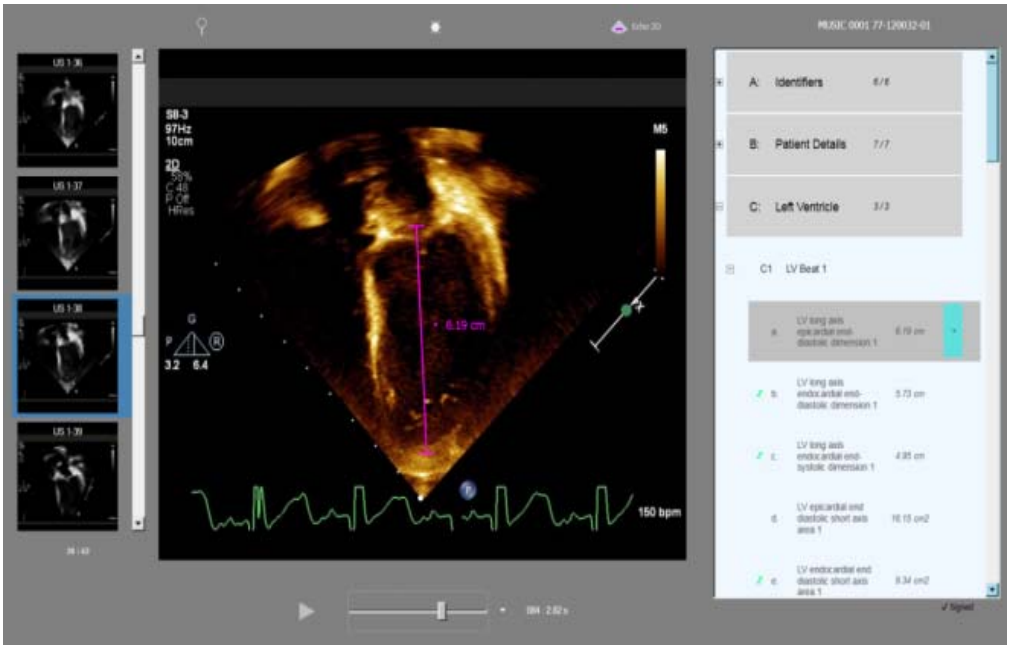


Fig 3. The dimension of a traced left ventricle long-axis is stored as one item of the echo report.

Coded report items are not measures per se, but are assessed as selectable options from a list. An example of a coded report item may be a reviewer's indication for the severity of a disease condition as either 'mild, moderate, or severe'. Each selectable option from the list will have a unique identifying code formatted for compatibility with database structures maintained at the DCC.

An important feature of each report item is the option to indicate if the item's measurement or assessment cannot be made. Generally, there are two reasons that a particular measurement or assessment might be incomplete. First, if the image of the heart structure to be visualized for the report item is not available from received images, the report item is said to be 'Missing', because the image is missing. A second cause of incomplete measurement may arise from poor image quality. In such cases, the reporting item result can be indicated as 'Indeterminate', i.e. the image is present, but a reliable assessment is nevertheless not possible.

3.2. Signatures and Exports

After measures are completed, the report is finalized with a signature from the reviewer. The study findings are then considered ready for inclusion with results from other completed studies into a dataset, which is exported to the DCC on a periodic basis. The exported data will constitute one key component for the analysis of study outcomes, and will form the comprehensive statistical basis of the research protocol's clinical recommendations. The DCC's statisticians and corelab personnel are then convened into study committees that review results and develop manuscripts for scientific publication.

4. Discussion

4.1. Designing Perspectives

Some aspects of the software features specifically designed for a corelab are very different from the features of a clinical software design. A clinical software program undergoes a rigorous Food and Drug Administration (FDA) approval for well-defined functional requirements in patient care scenarios. With a corelab application, the design and implementation of the software is more fluid and versatile, and has an ability to respond to needs that may not be anticipated at the initiation of the research protocol. Several changes to the software may be indicated after the corelab's review of initial image datasets. The software is then designed to rapidly accommodate any alterations that might be suggested by the new knowledge gained as the research project progresses. Some of these changes might include the addition of a measurement for echo reporting, or a modification to the imaging aims which can alter the process at the sites for imaging and data upload preparations.

With the goal to achieve software flexibility, it is useful to specify some of the critical functional details that will be the primary features shared by all corelab protocols. These critical features can then define the planning for effective and rigorous testing of the software. From the designer's perspective, the critical features also become more convenient to maintain, when the features can be partitioned from the other software elements that will be less critical for the research project.

To effectively manage corelab software, it is also important to understand which software features are to be hard coded (i.e. fixed and cannot be manipulated), and which features are to be user configurable for mid-stream protocol changes. Fixed features include the rendering of true image pixel colors, cine-image playback synchronization to a clock, and assured accuracy of traced measures when checked against calibrated standards. These basic and critical functionalities represent the essential components of the software needs for all research protocols. The more flexible aspects of designed software are ideally configurable in order to assist with protocol specific changes anticipated throughout the lifetime of the research project. A configurable item might include the addition of a new reporting measure, or some change to the Study ID format to handle growing patient enrollments and study counts.

Another practical and important design factor is to ensure all implemented software will be compatible with the various computer operating systems (OS) at the different sites of the corelab network. As some sites will have computers with a more recent version of the Microsoft Windows OS, it is prudent to choose which developed software features can be relied upon to perform for the various OS environments.

4.2. Process Considerations

In parallel to these considerations for design, the corelab's research output takes into account the linked processing between site image preparations, corelab reviewer measures, and the DCC's maintenance of research data. To facilitate a seamless integration between these stages of the corelab network, it is vital to provide some training of site personnel, and also test the corelab software capacities as early as possible with preliminary data validations. Early stage 'qualifying study' tests can then be applied to limited study counts and identify process or software changes to be made before large study influxes can be received. A test study process has now become standard practice for all stages of corelab data flow, including site image preparations, receiving area data handling, and structuring of echo measures forming the content of the study report.

4.2. Merged and Split Storage

The different types of files handled by the corelab include the image files of DICOM, together with echo

report files containing results of image measurement data. A typical echo study consists of approximately 100 DICOM images, and may consume over 2 gigabytes (GB) of disk storage space. By contrast, an echo measurement file is formatted in the Extensible Markup Language (XML), and is relatively compact at approximately 100 kilobytes (KB) in size. For most corelab protocols, the DICOM images and the XML report file of the study will be stored in the same data folder. This is a 'merged' method of data storage that is simple to maintain, and is preferred for simple corelab protocols with one read per study.

Some corelab protocols will have images that might undergo multiple reads by different reviewers. The measures between the reviewers will then be assessed for inter-observer variability. When multiple sets of measures are to be generated for the same image files, it becomes useful to consider 'split' data storage configurations, with reports generated by each reviewer stored into separate folders, which are linked to the shared location of the centrally accessible images. A split storage configuration is more complex to implement, but facilitates the review and export of measures to be generated by different reviewers.

4.3. Other Data Pipeline Models

The traditional model for a corelab network follows a data flow based on having sites provide images to a central reading facility. An implicit aspect of the centralized reading method therefore involves the de-identification of images by site data coordinators before the images can be uploaded. At the corelab, there is also a need to maintain significant resources to handle the gradual accumulation of all received site images which eventually reside on the corelab servers.

With the recent trends for economical 'cloud-based' data sharing, new possibilities are now introduced to simplify and alleviate the corelab's storage resource requirements. A cloud-based storage can allow for on-demand images to be downloaded for each initiated corelab study review session. When the study reads of the session are completed, the downloaded images can be purged from the corelab's server, greatly reducing the storage costs associated with a permanent local archive of the study. As more corelab projects are completed, a long term archiving of all image data (from all projects) is also economically achieved from a cloud-based storage.

Another corelab data exchange method to be considered, is to provide each site with reporting software designed to generate measurement results for a direct upload. In this 'distributed' study reading model, instead of receiving and reviewing images the corelab effectively functions to organize all the measurements generated by the site reviewers. The participating sites are then responsible for maintaining local image data, reviewing the images, and reporting measurement results with a periodic data send. With this envisioned data flow, the corelab only requires a minimal storage capacity for the received measures. Additionally, the distributed reads alleviate the need for PHI redactions prior to image transfers, because all images remain at the site location.

The choice between a centralized or distributed data flow will ultimately depend on the complexity and specific aims of the research study. Centralized reads will benefit from the quality assurance procedures that can be easily applied by the corelab's limited pool of readers. Alternatively, distributed reading methods provide for the assessments of measurement variation between sites and the distributed model is implemented at a reduced cost. It is also important to note that with the distributed review process, the results from the different site reviewers may be influenced by each site's image reading practices. Subsequently, some variation factors can be anticipated in the statistics of measures that would be gathered from different sites.

Summary

In this paper, we have provided descriptive guides to effective corelab planning, software component design, and data handling practices. For the purposes of planning and implementation, we suggest the initial steps to be considered for a new corelab protocol to include: (a) specification of the echo image views and measures needed for the research aims (b) preparation of training materials, (c) configuration of software for image preparations, transmissions, quality assurance checking, and reporting, and (d) a process to prepare measurements and related data in formats compatible with DCC records. The success of these linked data stages naturally benefits from engaged channels of communication maintained between the corelab, research sites, and the DCC.

As more sites from the US and other countries participate with corelab research, data pipeline handling capacities will need to be scaled and expanded to meet increasing image preparation, transmission, storage, and measurement activity needs. These factors will naturally spur the development of new software tools to distribute some of the corelab's current workload activity to the sites, where trained data managers and reviewers can independently apply image quality assurances and perhaps also perform site based measures. This envisioned distribution of corelab activity will be facilitated by improved software developments, and lead to cost effective and timely responses to arising public health research challenges.

Acknowledgements

This manuscript is an extension of materials presented during the 3nd International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI' 2021) [21].

References

[1]. T. Truong *et al.*, The NHLBI study on long-term outcomes after the multisystem inflammatory syndrome in children (MUSIC): Design and objectives: Design and rationale of the MUSIC Study, *Am Heart J*, 243, 2022, pp. 43-53.

[2]. H. Feigenbaum, Evolution of echocardiography, *Circulation*, Vol. 93, No. 7, April 1 1996, pp. 1321-1327.

[3]. W. Bidgood, Jr., S. Horii, F. Prior, and D Van Syckle, Understanding and using DICOM, the data interchange standard for biomedical imaging, *J Am Med Inform Assoc*, Vol. 4, No. 3, 1997, pp. 199-212.

[4]. K. Burns, V. Pemberton, and G. Pearson, The pediatric heart network: meeting the challenges to multicenter studies in pediatric heart disease, *Curr Opin Pediatr*, Vol. 27, No. 5, Oct 2015, pp. 548-554.

[5]. R. Lacro, H. Dietz, and L. Mahony, Atenolol versus losartan in Marfan's syndrome, *N Engl J Med*, Vol. 372, No. 10, 2015, pp. 980-981.

[6]. S. Colan *et al.*, The ventricular volume variability study of the Pediatric Heart Network: study design and impact of beat averaging and variable type on the reproducibility of echocardiographic measurements in children with chronic dilated cardiomyopathy, *J Am Soc Echocardiogr*, Vol. 25, No. 8, 2012, pp. 842-854.

[7]. M. Nathan *et al.*, Impact of major residual lesions on outcomes after surgery for congenital heart disease, *J Am Coll Cardiol*, Vol. 77, No. 19, 2021, pp. 2382-2394.

[8]. E. Marcus, M. Nathan, P. McGeoghegan, and K. Friedman, Data handling practices of the cardiac echo core laboratory at Boston Children's Hospital, in *Proceedings of the 3rd International Conference on Advances in Signal Processing and Artificial Intelligence*, Porto, Portugal, 2021, pp. 32-34.

[9]. National Institutes of Health, 2001-Present, Pediatric Heart Network (PHN), https://www.nhlbi.nih.gov/science/pediatric-heart-network-phn

[10]. National Institutes of Health, 2011, Compendium of Best Practices for Data Coordinating Centers, https://www.nhlbi.nih.gov/events/2011/compendium-best-practices-data-coordinating-centers

[11]. W. Moore and S. Frye, Review of HIPAA, part 1: History, protected health information, and privacy and security rules, *J Nucl Med Technol*, Vol. 47, No. 4, Dec 2019, pp. 269-272.

[12]. Open Source Contributors, 2022, *Fellow Oak DICOM*, https://github.com/fo-dicom.

[13]. S. Colan, The why and how of Z scores, *J Am Soc Echocardiogr*, Vol. 26, No. 1, Jan 2013, pp. 38-40.

[14]. L. Lopez *et al.*, Pediatric heart network echocardiographic z scores: Comparison with other published models, *J Am Soc Echocardiogr*, Vol. 34, No. 2, Feb 2021, pp. 185-192.

[15]. Amazon Corporation, 2022, *Amazon Web Services*, https://aws.amazon.com/

[16]. S. Subramanian, 2022, *Boston Children's Hospital Heart Center Vault*, https://hcbchvault.org

[17]. National Electrical Manufacturers Association, 2011, *Digital imaging and communications in medicine (DICOM) part 16: Content mapping resource*, http://medical.nema.org/

[18]. H. Feigenbaum, Digital echocardiography, *Am J Cardiol*, Vol. 86, No. 4A, 2000, pp. 2G-3G.

[19]. W. Armstrong and T. Ryan, Feigenbaum's Echocardiography, Eighth Edition, *Wolter's Kluwer Health*, Philadelphia, 2018, p. 980.

[20]. R. M. Lang *et al.*, Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging, *J Am Soc Echocardiogr*, Vol. 28, No. 1, January 2015, pp. 1-39 e14.

[21]. E. Marcus, M. Nathan, P. McGeoghegan, K. Friedman, Data Handling Practices of the Cardiac Echo Core Laboratory at Boston Children's Hospital, in Proceedings of the 3nd International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI' 2021), Porto, Portugal, 17-18 November 2021, pp. 32-34.

Sensors & Transducers, Vol. 256, Issue 2, March 2022, pp. 27-35

www.sensorsportal.com

Sensors & Transducers

Published by IFSA Publishing, S. L., 2022
http://www.sensorsportal.com

Opportunities and Challenges of IoT Security Using Distributed Ledger Technology

Chérif DIALLO
Laboratoire d'Algèbre, de Cryptographie, Codes et Applications (LACCA)
Dept. Informatique, UFR Sciences Appliquées et Technologie (UFR SAT)
Université Gaston Berger (UGB), BP 234, Saint-Louis, Sénégal
E-mail: cherif.diallo@ugb.edu.sn

Received: 21 February 2022 /Accepted: 22 March 2022 /Published: 31 March 2022

Abstract: The Internet of Things (IoT) has finally made its way into people's daily lives. Multiple connected objects are present in many markets. Their rapid adoption is due to the fact that IoT applications are developing rapidly and are easily deployed. However, these applications present several security challenges, although they are based on existing network technologies, which are already more or less secure. These challenges can be broadly classified as identification, authentication, jamming, cloning, privacy, etc. To face these challenges and in the perspective of ensuring end-to-end IoT networks security, several encryption mechanisms have been adapted and widely used including public key infrastructure (PKI), AES standard and ECC (Elliptic Curve Cryptography). To correct the weaknesses and vulnerabilities that are still present, other technologies such as DLTs (Distributed Ledger Technology) have been recently adopted by the Internet of Things. This adoption is still in its infancy as there are no standards nor methods yet indicating how to manage the security of the IoT through DLTs. In this article, we review the IoT-DLT paradigm in order to identify the benefits and major challenges of the convergence of these two technologies.

Keywords: IoT, DLT, Blockchain, Security, Cloud.

1. Introduction

IoT (Internet of Things) systems have developed rapidly. Many people interact with the IoT on a daily basis without even realizing it. It may be present in a water supply system, in an electricity network, in transport or in healthcare providing system. IoT systems are connected to the internet, which presents a high risk of threats. If it is not sufficiently protected, the IoT could be could be the door open to many attacks with adverse effects. However, IoT security is a real challenge considering the limitations that characterize connected objects.

The use of cloud computing has proven to be advantageous because it allows secure storage, availability, accessibility and interoperability for IoT applications. However, in IoT-Cloud architecture, the cloud is a point of failure, and therefore availability and accessibility issues appear in case of an attack. This is why other solutions, such as DLTs, would be explored and seem promising. DLTs eliminate the trusted third party from the IoT-Cloud model. DLT based IoT architecture could provide solutions to accessibility and availability issues. Nevertheless, many challenges and questions remain to be explored for a good convergence of these two technologies. In

this paper, we present some key points on the DLT-IoT paradigm, and in particular on the use of DLT to secure the IoT environment.

The rest of the paper is organized as follows: in Section 2, we discuss IoT security issues. In Section 3, we present the fundamentals of DLTs, and then review recent works on IoT security using DLTs in Section 4. Finally, in Section 5, we describe the problems, challenges and future research directions before concluding in Section 6.

2. IoT Security Issues

The fairly widespread use of the Internet of Things make life easier. For example, the electricity supply chain, the water supply system, the smart farming and smart home applications are good illustrations. However, such IoT systems present hacking risks that could have harmful consequences [1, 2].

An IoT based water supply system in a city could be a vector for a terrorist attack consisting of corrupting connected objects in order to insert bacteria, viruses, or even worse, poisons without being detected. The consequences of such an attack could be very harmful for the whole population using this supply system.

A smart home has many connected devices such as fridges, TVs, locks, cameras, etc. As connected locks are designed to communicate with smartphones, the use of a specialized digital lock-picking application could allow hackers to make a physical intrusion into the premises.

Criminals could think of attacking a business through a connected electricity meter installed within this business in order to black out it. In another way, they could carry out larger attacks such as forcing a power plant to shut down through its computer system to cause a complete blackout.

Some attacks exploit connected devices as a means of accessing larger networks such as the Internet. The addition of a connected object to a secure infrastructure can create a new attack surface. The connected object can be used as a relay for a large scale attack.

Attacks [1, 2] such as Man-in-the-middle (MITM), DoS/DDoS, Sybil attack, which are very common in the internet network, are also inherited by IoT systems, (Fig. 1). So, IoT attacks are often intended to data or hardware theft, espionage, service compromise, system destabilization, etc. All these attacks could be carried out at different levels of the Internet of Things architecture (Fig. 1).

Therefore, securing an IoT System would mean securing the different levels of its architecture, which is not a trivial matter. Researchers are therefore working to provide security services such as authentication, availability, integrity, confidentiality, non-repudiation, scalability, transparency, reliable data storage, etc.

The encryption algorithms used for a long time and the security mechanisms used to ensure authentication and security of communication on the Internet have proved to be very energy consuming for IoT connected objects, which do not have much energy reserve.

.

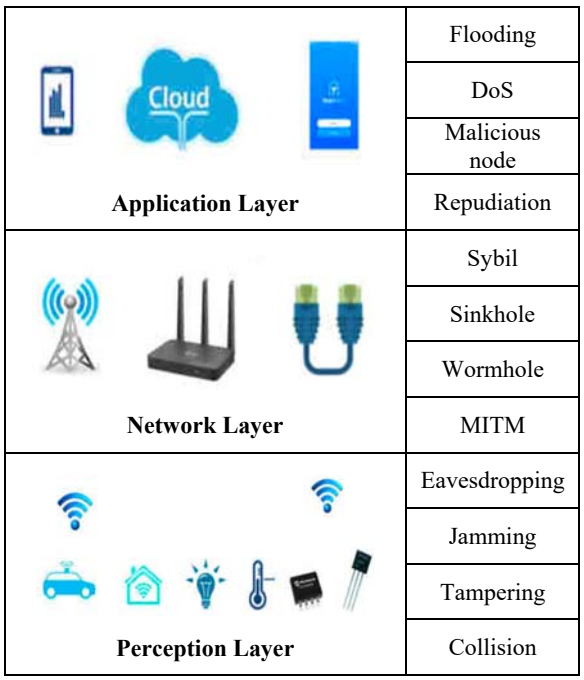


Fig. 1. IoT attacks.

On the other hand, IoT connected objects memory storage and computing power requirements do not facilitate the adoption of such algorithms. Thereby, a readjustment of these mechanisms is absolutely necessary. Similarly, the exploration of other efficient technologies in convergence with the IoT system has also its advantages.

Thus, cloud technology has been used to propel the IoT. The cloud has greatly contributed to the adoption of the Internet of Things in various domains (Table 1). It has 3 models:

- **Public cloud:** In the public cloud, resources are offered as a service, usually over an Internet connection, for a fee. Users can tailor their use on demand and do not need to buy hardware to use the service. In the public cloud, computing resources are open to a shared use [3].
- **Private cloud:** The aim is not to offer cloud services to the general public, but to use it within the organization. A private cloud is hosted in a company's data center and provides its services only to users within that company or its partners. A private cloud offers more security than the public cloud, and cost savings in case it would otherwise use unused capacity in an existing data center [3]. It is a model where all resources are reserved for the exclusive use of a company.
- **Hybrid cloud:** this is a mixture of public and private clouds. Hybrid clouds are more complex than the other deployment models, since they involve a composition of two or more clouds

(private or public). Each member remains a unique entity, but is bound to others through standardized or proprietary technology that enables application and data portability among them [3].

The cloud offers storage space, low-cost processing capacity (the customer only pays for what he uses), high availability, data accessibility and interoperability between IoT solutions [4]. However, the client-server model of the IoT-Cloud paradigm has many short comings. The cloud is a point of failure in the IoT-Cloud system.

A poorly configured public cloud (configuration flaws) could expose the data it stores on the internet. In a private cloud, an API to which objects will connect to send data and receive instructions, if it has a vulnerability, could be attacked in order to bounce into the company's network. This intrusion could then be used to steal industrial secrets. And if the API has an SQL injection vulnerability, the attacker will be able to use it to access the database to which the API is connected and thus gain access to private data. An attack on the cloud could have a direct impact on the IoT system.

The cloud has enabled the IoT to be easily usable in many cases, bringing many benefits. Apart from its security vulnerabilities, the cloud is opaque because participants do not know how their data is used.

Table. 1. IoT-Cloud paradigm benefits and deficiencies

IoT-Cloud paradigm	
Benefits	**Deficiencies**
Accessibility	Security
Connectivity	
Interoperability	Aupacity
Data analysis	
Real-time data translation	Trusted third party
Cost effective delivery	

To address the security deficit, other technologies such as DLTs are being explored. The combination of innovative technologies such as cloud computing and IoT has proven invaluable. DLTs are expected to further revolutionize the IoT by providing more security and establishing a reliable information sharing service that ensures that data is immutable [5, 6].

3. Fundamentals of Distributed Ledger Technologies (DLTs)

Distributed ledger technology is the generic term to describe any system that distributes data across multiple sites. A distributed registry is a decentralized database distributed across multiple computers or nodes. Each node will maintain the registry and if data changes, the registry will be updated. Each update to the system is recorded individually and then updated by the node itself. Thus, each node participating in the network operates independently of the other nodes. The blockchain is currently the best known distributed ledger technology (DLT). How this data is distributed, structured and agreed upon (consensus) will dictate the type of DLT (Fig. 2).

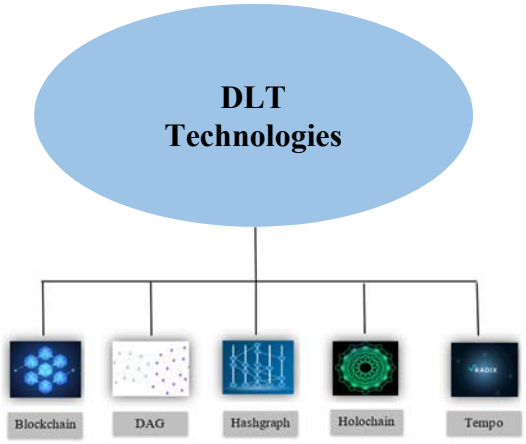


Fig. 2. Type of DLT Technologies

3.1. Blockchain

Blockchain is one of the most popular types of DLT. Blockchain is a type of DLT where transaction records are kept in the ledger in the form of a blockchain. In a blockchain, each block (Fig. 3) is linked to the previous one thanks to the digital signature called here block hash. Each block consists of three important elements: the previous block hash, the timestamp and the transaction data usually presented by a Merkle tree.

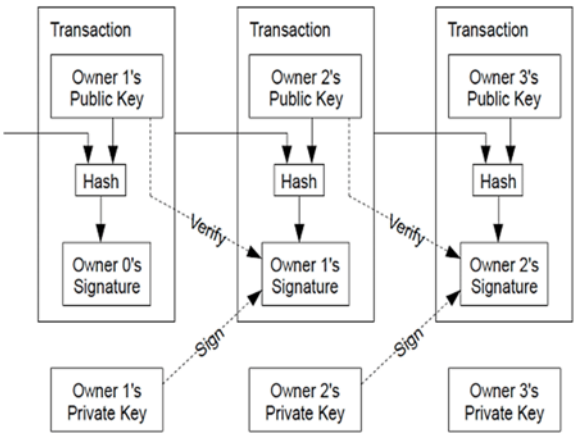


Fig. 3. Block structure.

When there is a new transaction, it is sent to all nodes of the network. The miner nodes authenticate

the transaction using their signature, if the transaction is valid, they mine it in a securely encrypted block. The miner who first succeeds in solving the underlying consensus problem, will then be able to publish his block which will be added to the blockchain. Since then, the transactions contained in this block can no longer be modified. This makes the blockchain a secure system because it is difficult to corrupt. There are several types of blockchain. A blockchain can be private, public or consortium. It can also allow or not the implementation of smart contract.

3.2. DAG

Another ambitious addition to the distributed ledger without the blockchain family is the DAG (Directed Acyclic Graph). A DAG is a particular data structure that allows transactions to be validated and time stamped by interconnecting them to each other unlike the blockchain which groups them together. Even though it is an alternative, the structure of this ledger is really different. One of the main benefits of implementing the DAG distributed ledger is the ability to offer nano transactions at no cost. This is because scalability improves as the network grows. In other words, the more transactions there are on the network, the faster it can settle them. A DAG is much more scalable than a traditional blockchain.

DAG takes a different route to consensus. The distributed ledger system stores transaction processes on nodes. Here, each member of the network is called a "node" just like the blockchain. All nodes in the network validate transactions on the ledger and are also represented by validated transactions. Any node can initiate transactions. However, in order to validate them, it must verify at least two of the previous transactions in the ledger. Once it has validated them, its transaction is confirmed. The more a node validates, the more valid its transactions become in the distributed ledger database.

Thus, if a transaction (Fig. 4) has a longer branch of previously validated transactions, it will have the most weight in the ledger. However, an algorithm will randomly select the two previous transactions for each member to validate. Because if not, members will only validate their transactions and leave another one. This is actually a new form of consensus to achieve greater scalability. Because of this nature of the distributed ledger implementation, companies that need a higher volume of transactions every second should use it.

IOTA [7] is the most popular example of an open source DAG distributed ledger. It is intuitive and scalable, designed to support friction less data and value transfer. IOTA uses quantum robust signatures to protect the network from next generation computing power. It is based on the Tangle [5] witch is a public registry replicated on all nodes. All data in the Tangle are stored in objects called transactions. Once a transaction is attached to the Tangle, it is immutable. All transactions in the Tangle are attached to two others to form a Directed Acyclic Graph (DAG). Each transaction in the graph is represented by a box and each attachment is represented by a line. When a new transaction is attached to the Tangle, it is attached to two previous transactions, adding two new lines to the graph.

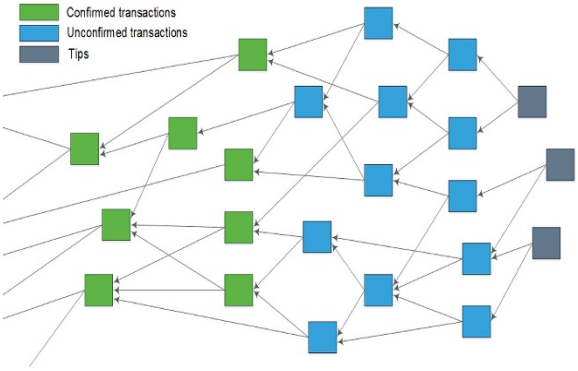


Fig. 4. IOTA transaction [7].

3.3. Hedera Hashgraph

Hedera is a project born in August 2018. It doesn't use a blockchain structure but a particular DAG structure called hashgraph. In the hashgraph [5], all transactions are stored in a parallel structure (Fig. 5).

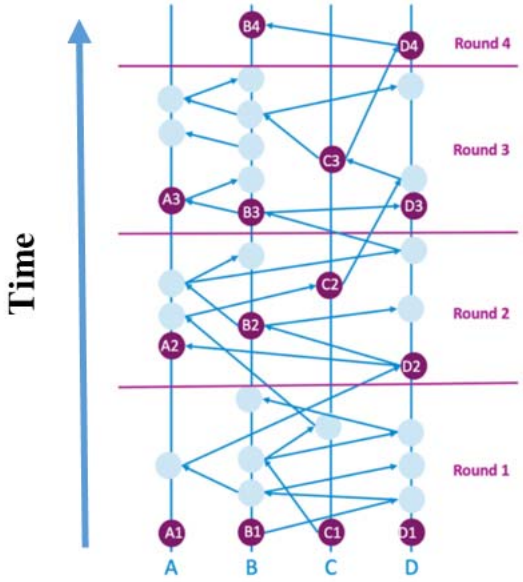


Fig. 5. Hashgraph vote.

Each circle represents a ledger record called an event, which is analogous to a block on a bockchain. An event contains: the timestamp, the transactions, and the hash of the two previous events witch they are joins. All events are therefore linked together irreversibly until the first event of the graph. Hashgraph uses a gossip protocol, the nodes are constantly trying to contact each other in a random way. The gossip algorithm makes the information flow

at a high speed on the network. As all transactions occur on the network, within seconds, everyone on the network will know where the transaction would be placed in the ledger. In addition, everyone on the network will know that the entire network is aware of the existence of the transaction and, therefore, will make changes accordingly.

To validate its transactions, nodes use a virtual voting system. The hashgraph is first divided into several rounds and in each round the nodes will vote to find a consensus on the validity of a transaction and on the timestamp and therefore their order of inscription on the ledger.

3.4. Holochain

In holochain [8], each node or agent maintains a single source string of their transactions, associated with a shared space implemented as a validating, monotonic, sharded Distributed Hash Table (DHT), where each node applies validation rules to the data in the DHT and provides the provenance of the data in the source strings from which it came. A Holochain (Fig. 6) is the set of these three elements: the shared database of public data (DHT), the local hash-chain of private data to an agent (local hash-chain), and a set of common rules (Nucleus).

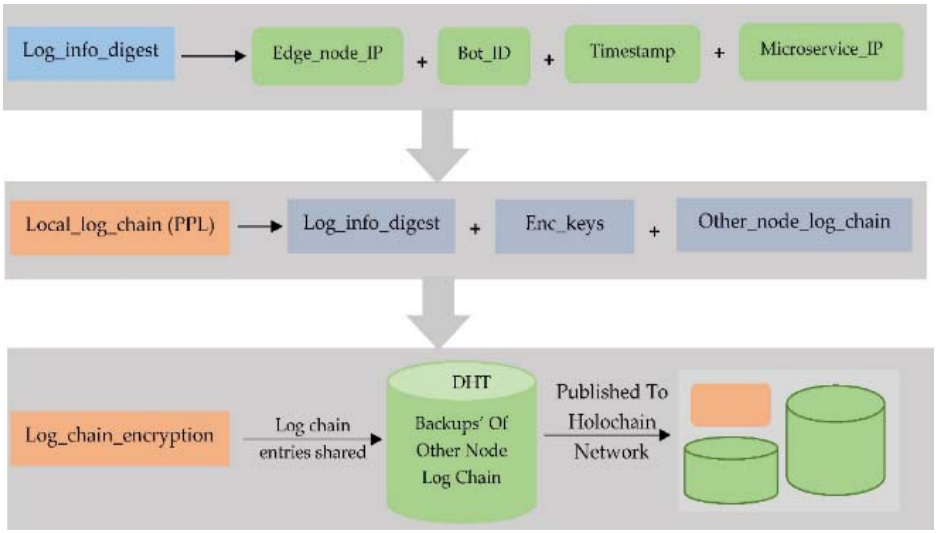


Fig. 6. Holochain's components [8].

Each agent has the common rules of operation: the source code of the distributed application, called Nucleus. By consulting its copy of the Nucleus, the node can check the validity of the information it receives from the other nodes (the signals). Therefore, each node has its own local hash-chain of private data. A node can allow, at its discretion, another node to view its local hash-chain. Also, when a node decides to modify the shared database (DHT), it must first write the information to its local hash-chain and then have it audited by other nodes to see if the addition meets the common rules. Each node decides which private data it wants to share on the common database, so it has sovereignty over its data. Holochain consumes very little energy and will run really well on a phone or small 45 W computer.

3.5. Tempo

The tempo radix distributed ledger database works on three major principles:

- Have a networked cluster of nodes.
- Distributed global ledger among the cluster of nodes. Special algorithms for timestamp events in the ledger.
- Each instance of this distributed ledger database is known as a universe. In the universe, each event is called an "atom".

This is a bit different from other distributed ledger databases on the market. Any node can choose to carry a subset of the entire global ledger with it. The subset of the ledger is called fragments, and each node carrying a fragment will receive a unique identifier for its subset of the ledger. Thus, nodes are not required to carry the burden of the global ledger on the network. This ensures that the network can handle a larger amount of load, increasing scalability. When a node wants to validate transactions, it uses logical clocks to do so. The usual distributed ledger database timestamp is not capable of achieving consensus on its own. This is because of the perspective of time changes from one person to another.

4. Some IoT Security Solutions using DLTs

The Table 2 gives a comparative analysis between the different DLTs, whereas Table 3 lists challenges and opportunities due to certain keys factors.

Table. 2. DLTs comparison table.

DLTs	Applications	Transaction/s	Consensus algorithm	Storage	Scalability	Security	Smart Contract
Blockchain	Bitcoin	About 3	PoW	Important	↓	High	No
	Ethereum	7-15	PoW	Important	↓	High	Yes
	Hyperledger	3000-20000	pBFT	Important	↓	High	Yes
DAG	IOTA	More than 1000	DAG-based consensus	Low	↑	Medium	Yes
	Hashgraph	About 10000	Gossip + virtual vote	Low	↑	Medium	Yes
Holochain	Holochain	About 56000	No general consensus	Low	↑	Medium	Yes
Tempo	Radix	About 1000000	Gossip	Low	↑	Medium	Yes

Table 3. IoT-DLT challenges and opportunities due to certain keys factors.

Key factors	Challenges/Opportunities	
Immature technology, Implementation cost	Absence of widespread adoption Challenges to businesses	**Challenges**
Lack of integration with existing solutions		
Strong encryption mechanisms		
Lack of adaptation of existing regulatory frameworks	Changes needed for adoption across sectors	
Nascent technology	Lack of governance framework	
Multiple non-interoperable implementations	Fragmented ecosystem	
Security vulnerabilities	Challenges for personal data protection	
Distributed systems	High energy consumption	
Need for increased computing power	Associated additional costs	
Legal enforceability of smart contracts	Smart contracts implementation	
Automating processes	Efficiency gains	**Opportunities**
Reducing the need for third-party intermediaries	Cost savings	
Technology adoption	New revenue sources for businesses	
Growth of the ecosystem	Creation of novel business	
Decentralized nature of the technology	Facilitate transactional systems	
Users control their own information	Improve users' trust in carrying out transactions	
Transactions immutability	Reducing the propensity for fraud	
Lack of a central point of failure	More resilient and secure	
Use of public key cryptography	Efficient and cost-effective management of digital identity	
Smart contracts implementation	Smart auditing capabilities	

In [9], the authors propose a data exchange model between owners and consumers based on smart contracts. This model eliminates the trusted third party of traditional centralized systems. Their general data exchange process is performed in eight steps ranging from sharing available information and data of IoT objects via the smart contract, to the payment of rental fees of these data by the consumer. In their proposed model, data owners have control over their data and can authorize access to the data and they can track all transactions since the entire transaction process is transparent and traceable. Their security model meets the criteria of confidentiality, integrity and availability.

The authors of [10] present a holochain-based security and privacy framework for IoT health systems. They point out major setbacks to blockchain related to increasing storage and network size requirements. In the proposed system, patients, doctors, staff, are considered as agents who can actively participate in the network and transfer information. Considering the storage space limitations of IoT objects, they use the cloud to process and store the transaction of a holochain network. The proposed scheme has been analyzed showing the higher efficiency of the holochain compared to rival blockchain systems. They highlight a significant reduction in time and complexity.

In [11], the authors identified the need for fast transaction, IoT network privacy. They have highlighted the main areas of vulnerabilities in the private and public sectors in Dubai in order to propose a sanitary framework. Thus, this one presents an approach using IOTA and Tangle technologies. They discussed the creation of two applications on the Tangle: a smart electricity meter and a smart toll system.

In [12], the authors propose a Lightweight Scalable Blockchain (LSB) optimized for IoT requirements while providing end-to-end security. DTC requires that each head cluster waits a random time before mining, i.e., adding a block instead of solving a computationally demanding headache, which significantly reduces the mining processing overhead. It also offers a distributed debit management (DTM) mechanism to dynamically adjust certain system parameters to ensure that the throughput of the public blockchain does not deviate from the transaction load in the public blockchain network.

DTM ensures that network is self scaling, i.e., as the network grows in size, more transactions can be appended to the public blockchain, thus increasing the throughput. Qualitative analyses have shown that their approach is resistant to several types of attacks [13]. Moreover, simulations have shown that packet overhead and delay are reduced, as the scalability of the blockchain has increased.

In [14], the authors design a new protocol to improve the security of distributed IoT devices. Their solution is essentially based on the hardening of IoT devices. This approach enables secure deployment of IoT devices by reducing the threat of surveillance and theft, while improving operator accountability and trust in IoT technology.

5. Challenges of DLT-IoT Integration

Table 2 gives a comparative analysis between the different DLTs studied here. In this table, pBFT (Practical Byzantine Fault Tolerance) is a consensus algorithm for enterprise consortiums where members are partially trusted. Whereas, Proof-of-work (PoW) is a well-known decentralized consensus mechanism. As we can see, this table gives an overview of some important characteristics of DLTs.

The security aspect of DLTs, and in particular of blockchain, justifies the attraction of researchers to address the security challenges of the Internet of Things. With the presence of smart contracts, several scenarios can be realized.

The convergence of IoT and a DLT such as blockchain still presents many challenges that need to be addressed in order to take full advantage of the benefits offered by DLTs [15].

There is not yet a reference model for the integration of IoT and DLTs. However, there are primarily two ways of integration:

- **Tight integration:** In this type of integration all IoT devices are considered as peers of the blockchain. All IoT communications are recorded in the blockchain.
- **Loose integration:** In this type of integration, only resource-rich IoT devices are considered peers of the blockchain. Other devices can communicate with the blockchain through these intermediaries.

The choice between these two types of integration depends on the level of security required but also on the characteristics of the connected devices.

The tight integration offers more advantages in terms of security. All exchanges are recorded in the blockchain and therefore verified and approved by all the nodes. This system provides security services such as non-repudiation.

However, this approach increases storage requirements and may not be feasible for some use cases. In contrast, loose integration allows the blockchain to be used to perform only certain interactions. DLT-IoT integration presents some challenges:

- **Storage:** one of the major problems of this convergence is the great difference in storage space requirements of the blockchain and paradoxically the limitations of connected objects.
- **Computing power:** participation in the blockchain consensus requires a certain amount of computing power and most connected objects do not have this computing power.
- **Scalability:** connected objects tend to produce and exchange a lot of data, and with tight integration, the number of transactions in the blockchain can increase rapidly and may cause the blockchain to slow down.

Nevertheless, despite these major challenges, the use of DLT brings many advantages and opportunities as we can see in table 3 which assesses challenges and opportunities due to certain keys factors. Finally, the table 4 summarizes the advantages and challenges of this convergence.

Table. 4. Benefits and Deficiencies DLT-IoT paradigm

<table>
<tr><th colspan="2">DLT-IoT paradigm</th></tr>
<tr><th colspan="2">Main features</th></tr>
<tr><td colspan="2">– An immutable record
– Disintermediation
– A lack of central control by one party
– New opportunities (Table 3)
– Tight integration
– Loose integration</td></tr>
<tr><th>Benefits</th><th>Deficiencies</th></tr>
<tr><td>Security</td><td rowspan="2">Scalability</td></tr>
<tr><td>Decentralization</td></tr>
<tr><td>Improved interoperability</td><td rowspan="2">Storage</td></tr>
<tr><td>Greater resilience</td></tr>
<tr><td>Greater privacy</td><td rowspan="2">Computing Power</td></tr>
<tr><td>Fault tolerance</td></tr>
</table>

6. Conclusion

In this study, we have presented the existing DLTs and made a comparison according to different parameters. This study is intended to provide additional support to help in the choice of DLTs for IoT applications. Many recent works on IoT has shown that the convergence of these two technologies, although very different, provides several solutions to the challenges of IoT, especially for security.

Thus, the key features of DLT-IoT are associated with its decentralized nature. In DLT-IoT, multiple copies of information are held by different connected objects, with data updating process by consensus and without the need for a third party. As a result, DLT-IoT can offer more efficiency, gains in trust and data reconciliation across all participant nodes. This means that DLT-IoT is able to provide:

– An immutable record. Data added is in theory unchangeable and secure, with the agreement of all participants as to the contents.
– Disintermediation. Direct nodes interaction, without the need for a third party.
– A lack of central control by one party. Changes are decided on a consensus basis by multiple connected participants.
– New opportunities for management and sharing of data and many others, such as, more efficiency, costs savings, more resiliency and security with reduction of the propensity for fraud.

However, some of the major problems of this convergence are the great difference in storage and computing power requirements of blockchain and paradoxically the limitations of connected objects.

Finally, in order to fully benefit from the advantages of DLTs, it is needed to find ways to circumvent the limitations of connected objects in order to set up a good integration architecture for these two technologies.

References

[1]. Ore Ndiaye Diedhiou, Cherif Diallo, An IoT mutual authentication scheme based on PUF and blockchain, in *Proceeding of the IEEE International Conference on Computational Science and Computational Intelligence (CSCI),* Las Vegas, USA, 2020, pp. 1034 – 1040.

[2]. Cherif Diallo, Security Issues and Solutions related to Data Aggregation Process in WSN, *International Journal of Computer Science and Network Security,* Vol. 17 No. 4, April 2017, pp. 59-71.

[3]. Goyal, S, Public vs private vs hybrid vs community-cloud computing: a critical review, *International Journal of Computer Network and Information Security,* 6, 3, 2014, pp. 20-29.

[4]. Xiong, Z., Zhang, Y., Luong, N. C., Niyato, D., Wang, P., & Guizani, N. The best of both worlds: A general architecture for data management in blockchain-enabled Internet-of-Things, *IEEE Network,* 34, 1, 2020, pp. 166-173.

[5]. Schueffel, Patrick, Alternative Distributed Ledger Technologies Blockchain vs. Tangle vs. Hashgraph - A HighLevel Overview and Comparison, *SSRN Electronic Journal,* 2018, https://ssrn.com/abstract=3144241

[6]. Farahani, Bahar Firouzi, Farshad Luecking, Marku, The convergence of IoT and distributed ledger technologies (DLT): Opportunities, challenges, and solutions, in *Journal of Network and Computer Applications,* Vol. 177, 2021, p. 102936.

[7]. IOTA Web site (https://www.iota.org).

[8]. Eric Harris Braun, Nicolas Luck, Arthur Brock, Holochain, scalable agent-centric distributed computing DRAFT(ALPHA 1) – 2/15/2018..

[9]. Pham, Hoang Anh Le, Trung Kien Pham, Thi Ngoc My Nguyen, Hoai Quoc Trung Le, Thanh Van, Enhanced Security of IoT Data Sharing Management by Smart Contracts and Blockchain, in *Proceedings of the 19th International Symposium on Communications and Information Technologies (ISCIT' 2019)*, 2019, pp. 398 – 403.

[10]. Zaman, Shakila Khandaker, Muhammad R. A. Khan, Risala T. Tariq, Faisal Wong, Kai-Kit, Thinking Out of the Blocks: Holochain for Distributed Security in IoT Healthcare, *IEEE Internet of Things Journal*, arXiv:2103.01322.

[11]. Shabandri, Bilal Maheshwari, Piyush, Enhancing IoT Security and Privacy Using Distributed Ledgers with IOTA and the Tangle, in *Proceedings of the 6th International Conference on Signal Processing*

and Integrated Networks (SPIN 2019), 2019, pp. 1069 - 1075.
[12]. Dorri, Ali Kanhere, Salil S.Jurdak, Raja Gauravaram, Praveen, LSB: A Lightweight Scalable Blockchain for IoT security and anonymity, *Journal of Parallel and Distributed Computing,* 134, 2019, pp. 180–197.
[13]. Anuska Gupta, Bhumika Gupta, Kamal Kumar Gola, Blockchain technology for security and privacy issues in internet of things, *International Journal of Scientific and Technology Research,* 9, 3, 2020, pp. 377-383.
[14]. John Wickstrom, Magnus Westerlund, Goran Pulkkis, Rethinking IoT Security: A Protocol Based on Blockchain Smart Contracts for Secure and Automated IoT Deployments, *arXiv:2007.02652,* 2020.
[15]. Advait Deshpande, Katherine Stewart, Louise Lepetit, Salil Gunashekar, Distributed Ledger Technologies/Blockchain: Challenges, opportunities and the prospects for standards, Overview Report, *British Standards Institution (BSI),* 2017.

Sensors & Transducers, Vol. 256, Issue 2, March 2022, pp. 36-44

www.sensorsportal.com

Sensors & Transducers

Published by IFSA Publishing, S. L., 2022
http://www.sensorsportal.com

Passive Inter-modulation Sources and Cancellation Methods

Onur YILDIRIM
YUPANATECH, Inc., Turkey
E-mail: info@yupanatech.com, www.yupanatech.com

Received: 25 February 2022 /Accepted: 26 March 2022 /Published: 31 March 2022

Abstract: When two or more signals of different frequencies pass through a nonlinear system, intermodulation distortion (IMD) occurs, resulting in the formation of spurious distortion signals. IMD is most commonly found in active circuits of a radio system, but it can also be found in passive wireless components such as lters, transmission lines, connectors, antennas, attenuators, and so on, especially when transmit power is quite high. Passive intermodulation (PIM) distortion is the name given to the IMD in the latter scenario. With the evolution of radio systems and the scarcity of radio spectrum, PIM interference is being recognized as a potential stumbling block to a radio network's maximum capacity. This article classifies the PIM sources in BS radio systems into two categories, internal and external sources. Internal sources are the radio's passive components such lters, transmission lines, connections, antennas, and so on. External sources, on the other hand, are passive items that are located outside of the BS antenna but inside the RF signal path, such as metallic and rusted objects in the antenna near eld. The high power current flowing through such passive devices can cause nonlinear behavior, resulting in IMD for both types of sources. Also, a review of PIM mitigation techniques is presented in the article.

Keywords: Passive inter-modulation, Internal/external sources, Cancellation methods.

1. Introduction

Both the user equipment (UE) and the base station (BS) are very important for achieving very high data rates. These are expected to be multimode and multiband to support transmission. However, especially on the UE side, there are restrictions in terms of power consumption, cost, and size. Transmission quality depends on signal quality and wireless system factors, so interference issues should be minimized. Defects in the wireless system architecture cause interference and are caused by technical limitations of the components. This is especially problematic for BS, where large amounts of current flow through the structure and affect the linearity of its components. These high performance systems require the signal path to be highly linear. Otherwise, defects will occur due to non-linear operation. These non-linearities can cause defects in the system, such as intermodulation distortion (IMD).

Intermodulation (IM) is an event that appears when one or more transmit (Tx) signals, with single and multiple frequencies or carriers, are given to a nonlinear system. The output is the inverse spurious frequency caused by the compound of input tones. This produces frequency components within and adjacent to the harmful band. These undesired spectral emissions, called spurious emissions, occur in the receiver band (Rx) and can interfere with the required receiver signal. If the nonlinear system that produces

these new frequency elements is a passive linear element of high power systems like transmission lines, connectors, interconnects, or just a metal part, this event is called passive intermodulation distortion (PIM).

An optimal behavior is sought in any RF communication system, which translates into components with linear relationships. Due to the presence of modest intrinsic nonlinearities, it is unfortunately inescapable. Nonlinearities like this cause interference, intermodulation, and harmonic distortion. Intermodulation occurs when a sum of frequencies input signal travels through a nonlinear system, resulting in new frequency content. When the fundamental frequencies are mixed, additional frequency components are created that are integer multiples of the input signal's frequencies [1-3].

When the IM frequencies generated by the circuits fall into the receiver bands near the transmitter signals in the RF spectrum, intermodulation becomes an interference concern. The input signal V_i is made up of two tones with frequency f_1 and f_2, as well as corresponding amplitudes A_1 and A_2 which can be defined as:

$$V_i(s) = A_1 \cos(2\pi f_1 t) + A_2 \cos(2\pi f_2 t) \quad (1)$$

The signal is subsequently passed via a non-ideal and thus nonlinear current-voltage (I-V) system, whose transfer function is represented by an nth-order power series with coefficients K_1, K_2, K_3, ... The nonlinear system's output signal, V_o, can be described as follows:

$$V_o(s) = K_1 V_i + K_2 V_i^2 + K_3 V_i^3 + \ldots \quad (2)$$

The stronger the K^{th} coefficient in the series (2), the more dominating the nonlinear term, i.e., the larger the nonlinear contribution. Additional terms with new frequencies are formed by combining both equations and expanding the series terms using the trigonometric identity and the Binomial theorem [4, 5]. These erroneous frequency components are either harmonics (multiples) of the original signal or the consequence of adding or subtracting the frequencies of the original signal. The IM products, or frequencies that follow the relationship, are the additions and subtractions of the original frequencies f_1 and f_2.

$$f_{IM} = nf_1 \pm mf_2, \quad (3)$$

where the coefficients n and m are integers. The order of the IM product is determined by adding the absolute values of these coefficients. Despite the fact that some of the higher and lower products may be easily filtered out, odd order IM frequencies are of particular concern because they are often near to the original signals (if the original frequencies in the original signal are close, which is common for multicarrier signals). Fig. 1 shows an example of an output spectrum that shows the entire range of this occurrence in frequency. In general, the proposed approach can be expanded to multiple frequency components; for example, if three frequency components mix in a nonlinear system, the corresponding third-order IM products (IM3) would be $f_1 \pm f_2 \pm f_3$.

If the signals are modulated, the bandwidth (Bw) of the IM products formed, not only the center frequency, must be taken into account. The bandwidth of IM products is greater than the bandwidth of the original signals, and it scales with the order of IM. If both input signals are 1 MHz wide, for example, the third-order product will have a bandwidth of 3 MHz, the fifth-order product will have a bandwidth of 5 MHz, and so on [6]. As a result, assuming both original signals have the same bandwidth, the IM product bandwidth can be calculated as the original signal bandwidth multiplied by the IM product order number. The different amplitudes of the IM products are another crucial aspect. If the input power on the signals is low, IM products have modest amplitudes; but, if the input power is large (as in radio systems), the amplitudes will also grow. If the input signal is increased by 1 dB to achieve the necessary output power, the IM3 product amplitude increases by 3 dB, according to the mathematical equations presented in [4, 5, 7]. Finally, IMD is an event defined by the transmitted signals combined in the nonlinear device, such as relative carrier frequencies, bandwidth, and power.

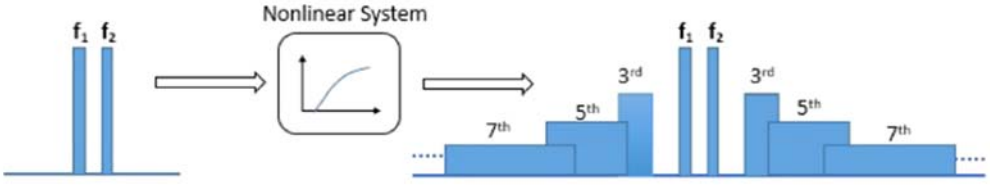


Fig. 1. Inter-modulation of Two Carries.

2. Origins of Intermodulation Distortion

PIM has long been recognized as a potential source of interference in radio communications systems, with the first investigations reaching back to 1989-90 [5, 8]. However, because to the rising saturation of the frequency spectrum and the adoption of wideband multicarrier communications, the PIM problem has resurfaced in recent years. PIM distortion can be

divided into three types: design PIM, assembly PIM, and rusty bolt PIM, according to [9]. This classification is based on several nonlinear trigger mechanisms that cause PIM interference, which necessitates a distinct solution. Design PIM refers to the decisions made by designers while selecting layout members, or picking components based on size, power, rejection, and PIM performance trade-offs. Based on the quality (materials, robustness, stability, interface) of the components and surrounding environment, assembly PIM indicates the interference caused by the degradation of the installed system over time. Rusty bolt PIM is linked to downlink frequency reflections towards the uplink in metallic objects in the beam's propagation path (rusty fence, barn, etc.). Any interference signal that can couple into the radio receiver and has a substantial strength difference from the desired received signal might cause receiver desensitization.

It's helpful to divide PIM sources into two categories: internal and external. PIM sources such as coaxial cables, connectors, waveguides, antennas, filters, and other internal PIM sources [5, 6, 8]. External PIM points out sources outside of the antenna that re-radiate the detrimental spurious emission towards the Rx, such as support structures, towers and masts, wire fences, and surrounding metallic objects.

2.1. Internal PIM Sources

PIM is caused by several physical mechanisms, including electron tunneling and thin dielectric layers between metallic contacts, micro discharge between microcracks and across voids (multipactor discharge), nonlinearities associated with dirt and metal particles on metal surfaces, high current densities, nonlinear resistivity of materials used, nonlinear hysteresis (memory effects) due to ferromagnetic and ferrimagnetic materials, and electro-thermal (ET) cohesion

Until recently, contact mechanisms in RF components were thought to be the primary source of internal PIM in radio system elements such as filters, antennas, and connectors [1, 10-12]. ET research, on the other hand, surfaced and was recognized as another significant contributor to PIM production in internal sources. Dirt particles and corrosion (contamination) are still a significant source of PIM, and they are a concern that must be addressed in radio systems. The description of the triggers will center on contact mechanisms and ET because the physics of contact mechanisms explains why corrosion can also generate PIM.

2.1.1. Contact Nonlinearities

The nonlinearities in metal contacts, as previously indicated, are one of the key mechanisms responsible for PIMP formation. The metal-insulator-metal (MIM) and metal-metal (MM) scenarios are two types of physical contact that can occur. Each of these physical structures, MIM and MM contacts, has numerous nonlinear mechanisms of its own. Electron tunneling, thermionic emission, and corona discharge are all more likely in MIM structures. Due to variations in metal work functions and nonlinear contact resistances caused by thermal processes such as expansion and thermal fluctuation, MM structures can form diode-like junctions [1]. These two contact types can occur in a variety of ways, especially as they are influenced by terrain and pressure on both ends of the contact.

In general, achieving a completely smooth surface on the termination of radio components during the fabrication process is unachievable. When two radio components are coupled, both contact surface topographies have multiple peaks in random positions and a native oxide or sulfide layer covering them on a microscopic level. The thickness of this layer varies depending on the metals used, but it is often in the nanometer range. As a result, contacting two similar surfaces is similar to contacting needles of varying lengths. Based on these findings, one might deduce that the "actual" contact area is only a fraction of the macroscopic contact area, and that it only occurs in peaks making contact. Similarly, MM scenarios only arise at microasperities where the mechanical pressure was significant enough to induce the connection of the peaks due to surface defects. When the mechanical pressure is inadequate to pierce through the thin dielectric layer covering the metal's surface, the comparable MIM condition arises [1]. As a result of the topology of the surface, a contact can be viewed as a combination of both forms of structure. However, as pressure rises, mechanical deformation rises, increasing the size of the "actual" area by pushing more microasperities connections, allowing researchers to evaluate whether MM happens more frequently. Different models exist to characterize the nonlinearities that appear based on the parameters that define the type of structure, such as deterioration, metals used, and cleanliness. The physical situation described above, in which the current is confined to flow via the microasperities, is depicted in Fig. 2. The presence of a thin dielectric layer between the microasperities or the emergence of corrosion in the empty zone determines whether the scenario is MM or MIM. The number of MM and MIM scenarios, however, is dependent on the amount of pressure exerted. The PIM distortion formed at a junction can come from a variety of sources, including MIM and MM, but some contributions are more significant [1]. Because the applied pressure connecting two radio components can decrease in most radio systems, MIM regions are extremely likely to be the main source of IMD, especially since a large number of contacting zones can emerge.

In conclusion, PIM interference can be caused by ferromagnetic materials and dirtiness in contacts. These mechanisms can be found throughout the RF network, including transmission lines, resonant structures, and antennas, in addition to contacts and junctions. Physical conditions incorporated in radio system components, on the other hand, generate PIM

signals that can be accumulated across the system. In high-power BS radio systems, where the downlink signals that contribute to PIM formation are extremely powerful, these interferences become increasingly noticeable.

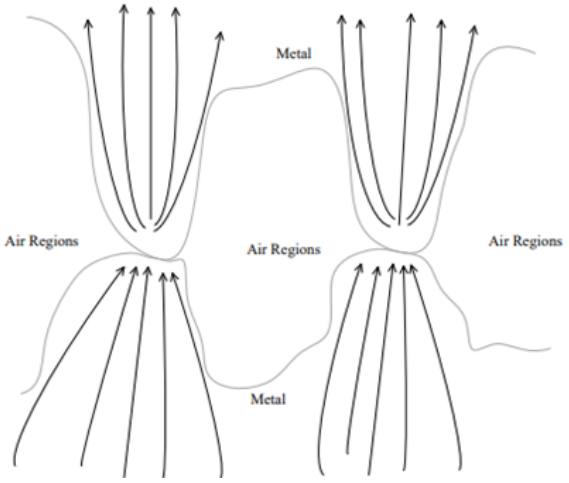


Fig. 2. Constriction of current in the connection between microasperities.

2.1.2. Electro-Thermal PIM Sources

High-power transmit signals might cause electro-thermal effects, which can lead to PIM. When modulated RF signals travel, the continual change in both the thermal and electrical domains causes time variant nonlinear conductivity, which is responsible for PIMP formation [1]. PIM generation due to ET conductivity is also one of the key contributions to exacerbating the problem in internal sources, as previously indicated, and its physics are explored in the following subsections. Electro-thermal effects caused by high-power transmit signals can also cause PIM. When modulated RF signals travel, the continual change in both the thermal and electrical domains causes time variant nonlinear conductivity, which is responsible for PIMP formation [1]. PIM production as a result of ET conductivity is also a major factor in exacerbating the problem in internal sources.

2.1.3. Distributed PIM Sources

In current radio systems, passive nonlinearities are divided into two types: lumped, in which PIM is generated by a single major source, often metal-to-metal contacts, and distributed, in which the sources are dispersed across the infrastructure. In current base stations, weak passive nonlinearities due to ET effects that operate like PIM sources are disseminated throughout the system, similar to the MIM situation. According to [1], the temperature's influence on conductivity causes significant electrical distortion in microwave elements. As a result of the electro-thermal phenomena, distributed PIM distortion is commonly described as a nonlinear transmission line (NTL). This model can be used to describe PIM production in passive components such as coaxial cables and microstrips due to ET conductivity. In a radio system, these elements, along with contact terminations (lumped components), play a critical role in generating PIM. The PIM distortion in the NTL model is caused by singular elements of a nonlinear transmission line. The total of nth order PIM outage power owing to ET effects is the sum of all the impacts reproduced by the cells. The reader is directed to [1] for a thorough explanation of how PIM is formed in each tiny element. To recapitulate, a nonlinear electric field (E) of IM products is generated in a line component by the nonlinear current (J) generated by variable heat dissipation (Q). The PIM signal is the sum of the electric fields that have accumulated down the line as a result of several components. In essence, each line element is a nonlinear generator whose signal power is proportional to its impedance (varies throughout the line). It's worth noting that each point generates two electric fields, one forward and one reverse. However, the latter is minimal due to destructive interference after line length z = /4. PIM can, however, flow backwards by reflecting the forward PIM signal at the line's termination. An illustration of PIM creation from the NTL model is shown in Fig. 3. In conclusion, it is possible that the many resistive parts through which the current travels contribute to PIM distortion due to ET conductivity.

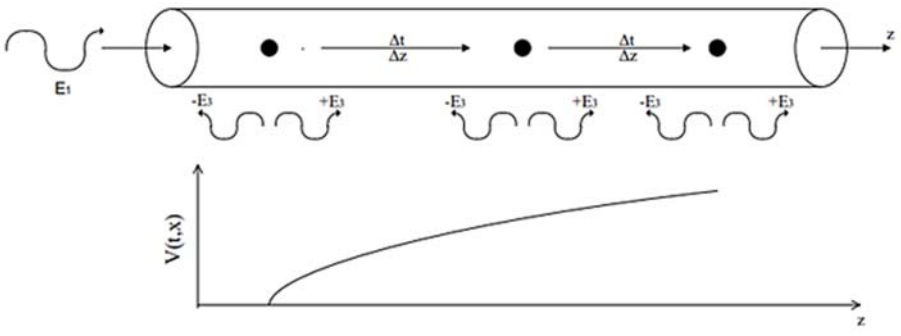


Fig. 3. PIM's signal strength sequential increase due to the generated fields in nonlinear consecutive points in the transmission lines.

2.2. External PIM Sources

The focus of the discussion thus far has been on nonlinear triggers of internal PIM sources; nevertheless, the PIM problem extends beyond the radio system's internal components. External sources of PIM can sometimes be unpredictable and unmanageable. In either situation, indoor or outdoor, resolving site interference concerns reveals the challenges related with PIM from external sources. The author of [13] conducted a study on the challenges that a site faces. It was established that non-linear objects in the RF path, such as sheet metal vents, metal flashing, ceiling tile frames, street lamps, and so on, might generate IM products and re-radiate them into the system. The "Rusty Bolt" effect is a frequent name for this effect. As a result, antenna position and orientation are critical for removing external sources from the system's RF path. The way energy couples with the nonlinear item and how it is received back into the Rx is also affected by antenna polarization. Varying antenna linear polarization (+45 and 45, respectively) leads to different degrees of third order IM products created externally by stacking metal sheets, as illustrated also in [13]. Tx and Rx functions are combined into one antenna in a conventional FDD radio system, while various antennas and bands perform simultaneously in a co-site scenario where multiple radios from the same or separate operators are installed. Intermodulation distortion of signals is a serious challenge in site integration. Internal causes like as contact nonlinearities (described and quoted in 2.1.1), material nonlinearities, and electro-thermal nonlinearities [1, 5, 10] are commonly blamed for PIM interference in antennas. PIM can, however, be generated externally (outside the base stations), in simple metallic components, as discussed in Section 2.2. PIM products can be generated or reflected by simple objects in the RF path, such as rusted junctions or metal structures, and are captured by the antenna as noise [5]. Tx and Rx functions are combined into one antenna in a conventional FDD radio system, while various antennas and bands perform simultaneously in a co-site scenario where multiple radios from the same or separate operators are installed.

2.2.1. External Sources as PIM Antennas

The physics behind IM produced products of reflections in metallic objects, which is based on the TDPO technique, was described in the previous subsections. It works by inducing a nonlinear current that radiates an electric field with the frequency of the IM products back to the BS antenna, which can then couple with the receiver chain. In fact, the TDPO method may be used to determine the scattered EM fields of dielectric substances and huge reflector antennas. However, when a reflector antenna, such as a parabolic reflector, is illuminated by high-power microwave radiation, the principal cause of false signal radiation is electron tunneling in the MIM junctions, which is known as the "Rusty Bolt" phenomenon. Nonlinear currents originating from two or more transmitters are induced in the reflector surface and flow through the object in this situation. MM or MIM junctions are formed when the local current encounters connectors, slits, or cracks during this process. Across tunneling events and corona discharge of the two or more currents running through the junction, nonlinear I-V characteristics emerge, which generate and radiate IM products. The produced IM products are subsequently radiated back at the transmitting antenna, where they may interfere with Rx bands. External sources can be seen as PIM antennas based on this research.

The physics of PIM from internal sources, specifically MIM sjunctions, are the key contributors to radiate PIM towards the Tx antenna, even when the sources are beyond the antenna, e.g., are external to the radio. It is defined by its parameters, just like any other antenna, but these cannot be measured. The strength of the current induced and how it flows in the surface determine the radiation pattern, polarization, directivity, and gain. Furthermore, many MIM junctions and reflections from a single external source might contribute to the radiation of IM products. Dielectric coating and incoming wave polarization are two unpredictable elements that can alter this generation. As a result, it's reasonable to predict that different PIM antennas should be expected based on the external source met by the broadcast tones.

3. Passive Intermodulation Distortion in Radio Systems

PIM is a type of radio interference that has the potential to reduce the efficiency of a cell site, consequently influencing the performance of a radio network directly. Furthermore, the problem of PIM interference is aggravated by the cohabitation of numerous communication systems in the same areas, and it is becoming more of a concern as network complexity and deployments rise. As a result, it is a problem with a growing influence on networks. Intermodulation concerns are exacerbated in broadband radio systems where bandwidth is expanded to obtain higher data rates, especially when CA is used, because PIM interference is a problem in the RF spectrum. However, active narrowband systems such as GSM are susceptible to PIM interference. Because of the nonlinear generating techniques,passive intermodulation is frequently seen as an installation issue, most commonly seen in high-power cell sites. However, this is only a surface consideration,as interference issues are frequently caused by a saturated RF spectrum. PIM has become a volatile concern as a result of the constant updating and expansion of network systems, and its effects on RF systems will continue to worsen. This is especially true in 5G, where the seamless integration of several base station technologies is becoming more common.

3.1. Passive Intermodulation in Broadband Radio Systems

Most literature describe and analyze PIM based on two unmodulated carrier signals. As it is explained in Section 2.1, in this case, the resulting PIM products are also narrow carriers spurs. However, in broadband radio systems like UMTS and LTE, the signal is modulated over a wider bandwidth and can be transmitted through several bands of the RF frequency, especially with the employment of features like CA. Furthermore, the combined modulated signals can be both narrowband and broadband at the same time. Hence, consider a general non-contiguous dual-carrier CA FDD transceiver whose Tx signal is represented as

$$x[n]=x_1e^{jf1}+x_2e^{jf2} \quad (4)$$

This signal is made up of two component carriers (CCs), denoted by x_1 and x_2 and respective frequencies f1 and f_2 ($f_1 < f_2$). Mixing occurs and IM components appear when this non-contiguous Tx signal travels via a nonlinear passive source.

Consider an FDD system in which signals are broadcast from several bands. Consider two channels, B2 and B21 that transmit from band 2 and band 21, respectively. The transmitting carrier has a center frequency of 1940 MHz and a bandwidth of 10 MHz, while the band 2 DL runs from 1930 to 1990 MHz. The transmitting carrier has a central frequency of 2130 MHz and a bandwidth of 10 MHz, while the band 21 DL runs from 2110 MHz to 2155 MHz. PIM products are created during transmission due to a nonlinearity on a passive device. Consequently, because they are focused around 1750 MHz, one of the created spurious 3rd order and fifth order IM products will overlap with the B21 receiver spectrum, which spans 1710 MHz to 1765 MHz. Furthermore, because the B2 carrier's bandwidth extension is big enough to fall within this Rx band, the OOB 5th order product overlaps with B15 UL. Even if the false emission is more detrimental, both possibilities lead to a potential danger of interference. The preceding situation is depicted in Fig. 4.

Spectral regrowth is typical in active broadband base stations, where the carriers merged at the source are from multiple frequency bands. The scenario is known as cross-band PIM when the CCs engaged in PIM generation are from multiple bands, as in the example above. In the same way, if all of the carriers are from the same band, the scenario is known as in-band PIM. The refarming of RF spectrum, such as new DL bands and interband CA, as well as the increasing complexity of the BS, are exacerbating cross-band PIM. Furthermore, because the number of external sources in which the carriers can mix is rising, cross-band PIM is a contemporaneous scenario in modern radio systems. Fig. 5 illustrates this scenario.

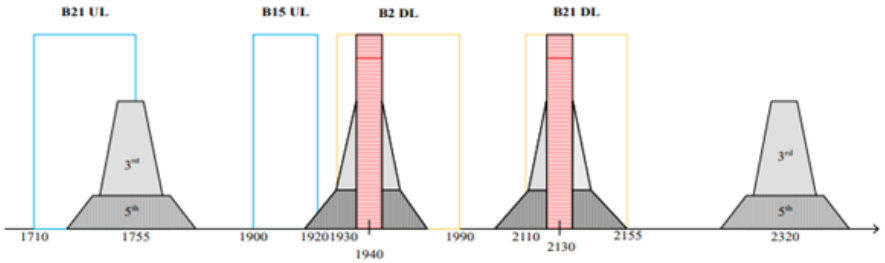


Fig. 4. Spectral regrowth given the spurious and OOB emission due to IM of carriers from B2 and B21 that can lead to interference in B15 and B21s.

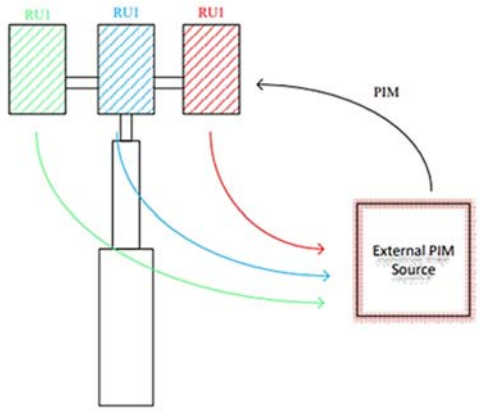


Fig. 5. Cross-PIM scenario involving three radio units, RU1, RU2 and RU3 and an external source close to the BS.

4. Passive Intermodulation Mitigation Techniques

PIM-generated spurious signals interfering in the Rx bands are generally quite unpredictable. The spurious signals are based on the site's environment, RF infrastructure characteristics, external and internal sources, as well as configured bands and used frequency channel, as previously explained. The nonlinear features of the PIM producing source usually determine the strength of interfering signals. Furthermore, there is the uncontrollable factor of dielectric coating in external sources. As a result, dealing with any PIM interference becomes a difficult task. When considering continual radio system enhancements, such as higher performing services,

smaller cells, and radio configuration in the RF route, the complexity of PIM mitigation becomes even more. Nonetheless, PIM interference has been addressed in the past using a variety of mitigation approaches, which are briefly discussed below.

Physical techniques and radio integrated techniques are the two primary kinds of PIM mitigation techniques. Physical mitigation techniques are processes or changes to RF equipment, infrastructure, and the surrounding environment that can be used to correct or minimize the level of PIM interference. These processes are frequently particular countermeasures to the previously described generation mechanisms, and thus necessitate a thorough grasp of the physics involved. To prevent PIM interference, digital mitigation mechanisms are included into the signal's path and use digital signal processing techniques.

4.1. Physical Mitigation of Passive Intermodulation Interference

Because PIM sources can be internal or external, this section should be divided into two subsections to account for the physical activities taken to minimize both sorts of sources.

4.1.1. Guidelines for Mitigation of Internal Sources

Counteracting the effects indicated in Section 2.1 is required to mitigate internal sources of PIM in radio systems. For the past two decades, the following guidelines have generally been followed:

- Minimize metallic contacts and connectors, ensuring that loose contacts and rotating joints are avoided;
- Keep thermal variations in the components of the radio system to a minimum;
- Keep current densities low in the conduction paths by using larger conductors or having bigger contact areas, e.g., MM contact;
- Minimize metallic contacts and connectors, avoiding loose contacts and rotating joints;
- Keep thermal variations in radio system components to a minimum;
- Keep all joints clean and tight, preferably made or coated with materials less prone to oxidation; Although no system is fully free of PIM, careful planning, quality craftsmanship, and a high degree of system maintenance are all necessary to significantly reduce its level [4, 5].

Since electro-thermal effects have been established as the main internal PIM sources, low electronic conductivity materials are currently being used in the production of radio system components. While using high power currents flowing through the RF components of the base station, this practice ensures that thermal changes are maintained to a minimal.

4.1.2. Guidelines for Mitigation of External Sources

These are more harder to deal with than internal sources. Scanning the surroundings, or RF route, for potential nonlinear sources that can cause PIM and then deleting them is a basic and common guideline used when setting up a site. These scanners look for primary test tones that generate IM products in order to pinpoint the source. Objects in the antenna's radiation path, on the other hand, are unexpected and uncontrollable, making mitigation a difficult task. The strength of the distortion generated by these sources is also depending on how far away they are located, and they may even be the primary PIM source in some situations.

Because PIM is inherent in all radio systems, digital signal processing techniques are used to reduce PIM. However, as radio networks become more complicated, so does the complexity of base stations, particularly the number of antennas. As previously explained, MIMO systems are currently in place, with antennas playing a key role in PIM. Because their sensitivity is decreasing, any improvement that enhances the transmission link's power (dB) is important.

Antenna isolation refers to the considerations that co-site base stations make in order to avoid excessive interference (PIM) and improve link quality. The amount of isolation obtained is influenced by a number of elements, including the physical horizontal and vertical separation distance between antennas, antenna polarization, radiation pattern, and the antenna's surrounding environment. In general, the mitigation of IM interference improves as the distance between the antennas and the electrical down tilt (azimuth) angle increase.

4.2. Radio Integrated Mitigation of Passive Intermodulation Interference

Physical PIM mitigation techniques, in general, cannot entirely resolve the PIM interference problem. As a result, digital PIM mitigation measures have lately sparked interest, as evidenced by [4] and others. In most cases, the digital PIM cancellation method is based on the block diagram presented in Fig. 6.

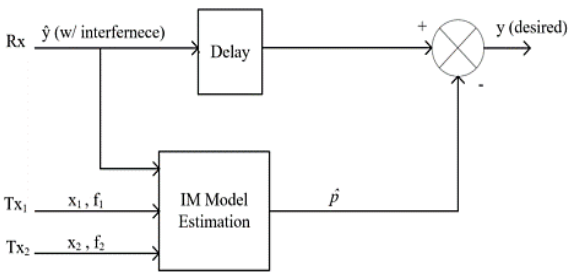


Fig. 6. Diagram block for an adaptive cancellation PIM algorithm.

The fundamental idea behind this model, which is based on adaptive filter theory, is to forecast PIM-induced IM products based on the transmitting signal in order to remove them from the receiving signal chain (after I/Q conversion and duplexer). However, the fundamental issue with digital cancellation is that the amplitudes, phases, and delays of PIM products vary dependent on the RF components employed (PA, duplexer, etc.) and are not constant. Because the parameters must be updated or examined when the transreceiver system is offline, this adds to the complexity of the model used to estimate IM products. The key advantage of this digital cancellation is that it can be utilized to minimize the internal IM products generated after the Tx filter. Frequency or resource planning is another often used method. In this method, a channel spacing rule is used to determine the allocation of available channels for each tier. The channel allocation is made hastily in order to minimize spurious interference signals. The process of superimposing the networks can be accomplished through block assignment or frequency banning.

- The first allocates channel blocks to both networks, ensuring that no potentially IM product interferes with transmission within the cell;
- The second stage consists of three steps:
  - Identifying the channel combinations that produce hazardous IM products in each cell;
  - Avoiding the most popular channels in the combinations; and
  - Using frequency hopping to compensate for the reduction in capacity (number of channels).

Frequency planning, unlike other systems, totally eliminates the PIM problem, ensuring a PIM-free channel. This method, on the other hand, prevents full utilization of existing resources and can result in lower throughput. Unusual strategies can also be used, although they come at a higher cost. Reduced transmit power, for example, can minimize PIM because PIM power is proportional to Tx power. However, this comes at the expense of lesser coverage. When PIM is imminent, another example is to boost Rx sensitivity. However, users on the cell-edge, who will be interacting with BS at a low power level, may have their signals blocked.

5. Conclusions

PIM happens when two or more high-power tones are sent through passive equipment (such as cables, antennae, and so on). The PIM product is created when two (or more) high-power tones mix at nonlinearities in the device, such as dissimilar metal junctions or metal-oxide junctions, such as loose corroded connectors. The effect of the nonlinearities is more obvious at greater signal amplitudes, and the intermodulation is more prominent — even if the system appears to be linear and unable to generate intermodulation at first glance.

When a single antenna is used for both high power transmission and low power receive signals, PIM is a serious challenge in current communication systems. Strength of the passive intermodulation signal is often many orders of magnitude lower than the power of the transmit signal, it is frequently on the same order of magnitude (if not higher) than the power of the receive signal. As a result, passive intermodulation that makes its way into the receive path cannot be filtered or isolated from the receive signal. There are many types of PIM mitigation techniques deployed, including as physical and radio integrated solutions. In terms of physical mitigation approaches, various countermeasures have already been implemented as a result of considerable research into physical generating mechanisms. However, PIM is a persistent and uncontrollable problem as a result of this vast research. As a result, radio integrated mitigation approaches are critical for coping with unpredictability. Until far, the strategy for these types of techniques has been concentrated on anticipating spurious frequency emissions; however, as the spectrum becomes saturated, new tactics are required. Digital cancellation algorithms have emerged as a feasible option in recent years, but they still need to be improved for greater performance. Data collection and analysis are used to make improvements.

6. About Yupana

Yupana is a highly dynamic and fast growing company founded in 2011 in the San Francisco Bay Area. Working closely with the wireless carriers tailored engineering services are developed and products are customized allowing customers to build the networks of tomorrow.

Yupana team of highly skilled technicians and engineering bring an unprecedented level of quality and expertise to all field activities, combined with passionate and ambitious team of UTRAN engineers and project managers, and software developers.

Yupana leverages many years of international experience managing projects and programs in Tier 1 MNOs, along with world class Engineering expertise.

References

[1]. J. R. Wilkerson, Passive intermodulation distortion in radio frequency communication systems, *North Carolina State University*, 2010.

[2]. J. C. Pedro and N. B. Carvalho, Intermodulation distortion in microwave and wireless circuits, *Artech House*, 2003.

[3]. T. Lähteensuo, Linearity requirements in LTE-advanced mobile transmitter, Master's Thesis, *Tampere University of Technology,* 2013.

[4]. J. Jokinen, Passive intermodulation in high-power radio transceivers, Master's thesis, *Tampere University of Technology,* 2016.

[5]. P. L. Lui, Passive intermodulation interference in communication systems, *Electronics Communication*

Engineering Journal, Vol. 2, No. 3, June 1990, pp. 109-118.
[6]. Passive intermodulation (pim), https://www.anritsu.com/en-us/test-measurement/technologies/pim, accessed: 2019-03-18.
[7]. 3GPP, 3gpp tr 37.808 v1.0.0 (2013-09), ARIB, ATIS, CCSA, ETSI, TTA, TTC, Tech. Rep., 2013.
[8]. P. L. Lui and A. D. Rawlins, Passive non-linearities in antenna systems, *IEEE Colloquium on Passive Intermodulation Products in Antennas and Related Structures,* June 1989, pp. 6/1-6/7.
[9]. F. Kearney and Steven Chen, Passive intermodulation (pim) effects in base stations: Understanding the challenges and solutions, Mar 2017. [Online]. Available: https://www.analog.com/en/analog-dialogue/articles/passive-intermodulation-effects-in-base-stations-understanding-the-challenges-and-solutions.
[10]. J. R. Wilkerson, K. G. Gard, and M. B. Steer, Electro-thermal passive intermodulation distortion in microwave attenuators, in *Proceedings of the 2006 European Microwave Conference*, September 2006, pp. 157-160.
[11]. A. Shitvov, A. G. Schuchinsky, M. B. Steer, and J. M. Wetherington, Characterisation of nonlinear distortion and intermodulation in passive devices and antennas, in *Proceedings of the 8th European Conference on Antennas and Propagation (EuCAP 2014),* April 2014, pp. 1454-1458.
[12]. J. R. Wilkerson, P. G. Lam, K. G. Gard, and M. B. Steer, Distributed passive intermodulation distortion on transmission lines, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 59, No. 5, May 2011, pp. 1190-1205.
[13]. T. Bell, Mitigating external sources of Passive Intermodulation, *IET*, 2013.

4th International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI' 2022)

ASPAI CORFU • 2022

19-21 October 2022 • Corfu, Greece

https://aspai-conference.com/

Advances in artificial intelligence (AI) and signal processing are driving the growth of the artificial intelligence market by two major factors: emerging AI technologies and intelligent signal processing. Today, more and more sensor manufacturers are using machine learning to sensors and signal data for analyses. However, the increased number of sensors in devices will inherently generate higher data throughput, which poses a serious challenge in managing and processing the tremendous amount of sensory information. Furthermore, traditional processing techniques in conventional sensing devices are no longer suitable for systematically labeling, processing, and analyzing the exuberant amount of information.

The global artificial intelligence market size was valued at US $ 93.5 billion in 2021 and is projected to expand at a compound annual growth rate (CAGR) of 38.1% from 2022 to 2030. Revenue forecast in 2030 is USD $1,811.8 billion.

The global AI market growth is driven by increasing adoption of cloud-based applications and services growing big data increasing demand for intelligent virtual assistants advances in signal processing. The major restraint for such market is the limited number of AI technology experts. Today, there are approximately 300,000 AI researchers and practitioners worldwide, but the market demand is around millions of roles.

The series of ASPAI conferences have been launched to fill-in this gap and provide a forum for open discussion and development of emerging artificial intelligence and appropriate signal processing technologies focused on real-word implementations by offering Hardware, Software, Services, Technology. The goal of the conference is to provide an interactive environment for establishing collaboration, exchanging ideas, and facilitating discussion between researchers, manufacturers and users.

The topics of interest include (but not limited):

Artificial Intelligence:

- AI Algorithms
- Intelligent System Architectures
- Hybrid Intelligent Systems
- Artificial Neural Networks
- Expert Systems
- Intelligent Networks
- Parallel Processing
- Pattern Recognition
- Pervasive Computing and Ambient Intelligence
- Programming Languages Artificial Intelligence
- Soft Computing and Applications
- Artificial Intelligence Tools & Applications
- CAD Design & Testing
- Computer Vision and Speech Recognition
- Fuzzy Logic, Systems and learning
- Computational Theories, Learning and Intelligence
- Computational Neuroscience
- Soft Computing Theory and Applications
- Software & Hardware Architectures
- Web Intelligence Applications & Search
- Ambient Intelligence
- Artificial Immune Systems
- Autonomous and Ubiquitous Computing
- Bayesian Models and networks
- Data Fusion
- Distributed AI
- Machine Learning
- Deep Learning
- Lerning and Adaptive Sensor Fusion
- Multisensor Data Fusion Based on Neural and Fuzzy Techniques
- Applied Artificial Intelligence
- Self-Organising Networks

Signal Processing:

- Signal Processing Theory and Methods
- Sensor Array and Multichannel Signal Processing
- Multivariable Sensor Systems
- Digital Signal Processing
- Image and Video Processing
- Multichannel Signal Processing
- Biomedical Signal Processing
- Biometrics
- Chaotic and Fractal Systems
- Computer Vision and Pattern Recognition
- Image and Video Processing
- Information Theory and Coding
- Speech Signal Processing
- Signal Processing for Communications and Networking
- Radar and Sonar Signal Processing
- Satellite Signal Processing
- Nonlinear Signal Processing
- Statistical Signal Processing
- Signal Processing for Internet of Things
- Coding
- Wavelet Transformation
- Theory and Application of Filtering
- Spectral Analyze

Deadlines:

Submission (2-page abstract):	1 July 2022
Notification of acceptance:	20 July 2022
Registration:	30 August 2022
Camera ready (4-6 pages paper):	20 September 2022

IFSA World Scientific Connecting Great Minds IOS Press SMC Soft Measurement and Computing Scientific journal

Contribution Types:

- Keynote presentations
- Invited presentations
- Special sessions papers
- Regular papers
- Posters
- Industrial presentations
- Panel discussions and Round Tables
- Exhibition

One event - three different publications !

1. All registered abstracts will be published in the conference proceedings (Flash drive edition of published Proceedings with the ISBN). The conference proceedings will be submitted for indexing in the Web of Science.

2. Authors will be invited to submit full-page extended papers to one of special journals' issues, which are published in both formats: print and electronic:

   - Sensors & Transducers (ISSN 2306-8515, e-ISSN 1726-5479) by IFSA Publishing
   - Integrated Computer-Aided Engineering (ISSN: 1069-2509, e-ISSN: 1875-8835) by IOS Press
   - International Journal of Neural Systems (ISSN: 0129-0657, e-ISSN: 1793-6462) by World Scientific

3. The limited number of full-page papers published in the Sensors & Transducers journal will be selected by the Editorial Board to extend for open access books' chapters for the Book Series on 'Advances in Signal Processing: Reviews', Vol. 3 or 'Advances in Artificial Intelligence: Reviews', Vol. 2, which will be published by IFSA Publishing in 2023.

Special Sessions:

Authors are welcome to propose and manage special sessions during the ASPAI' 2022 conference. Each special session will contain 4-6 papers in a related field as specified above. Session organizers will get:

- Certificate of Appreciation
- Free Registration
- Special Publishing Theme within Conference Proceedings

Chairman:

Prof., Dr. Sergey Y. Yurish (IFSA, Spain)
e-mail: aspai@sensorsportal.com

Advisory Chairman:

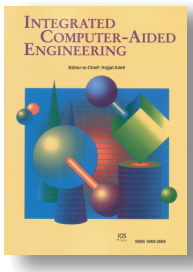
Prof., Dr. Hojjat Adeli (The Ohio State University, USA)

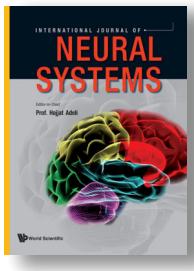
Steering Committee:

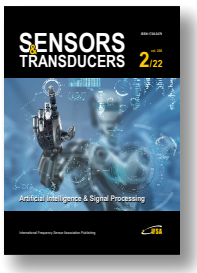
Dr. Mobyen Uddin Ahmed (Mälardalen University, Sweden)
Dr. Sergey Grosman (Siemens PPAL, Germany)
Prof. Svetlana V. Prokopchina (Financial University, Russia)
Prof. Valery B. Tarassov (Bauman Moscow State Technical University, Russia)
Prof. Sandeep Singh Sengar (Cardiff Metropolitan University, UK)

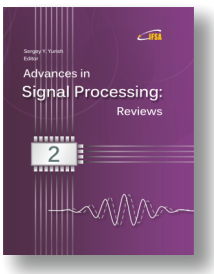
Conference Manager:

Mrs. Tetyana Zakharchenko (IFSA Publishing, S.L., Spain)







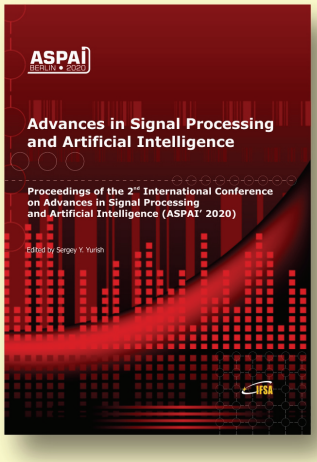


Social Programme

Welcome Cocktail. 18 October 2022 (20:00-21:30). The Welcome Cocktail will take place in the Corfu Holiday Palace Hotel. Do not miss this opportunity to say the first "Hello" to attendees and committee members. The Registration will be opened at the Welcome Cocktail area.
Gala Dinner. 20 October 2022 (20:00-23:00). The Gala Dinner will take place in the Corfu Holiday Palace event hotel, Nausica Ballroom.

Advances in Signal Processing and Artificial Intelligence

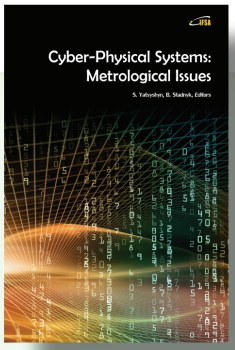
Proceedings of the 2nd ASPAI' 2020 Conference



The proceedings contains all accepted and presented papers of both: oral and poster presentations at ASPAI' 2020 conference of authors from 23 countries. The coverage includes artificial neural networks, emerging trends in machine and deep learnings, knowledge-based soft measuring systems, artificial intelligence, signal, video and image processing.

Formats: hardcover (print book) and PDF (e-book), 264 pages
ISBN: 978-84-09-21931-5, e-ISBN: 978-84-09-21930-8
IFSA Publishing, 2020

https://www.sensorsportal.com/HTML/BOOKSTORE/ASPAI_2020_Proceedings.htm



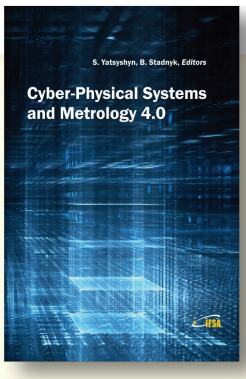
Cyber-Physical Systems: Metrological Issues

S. Yatsyshyn and B. Stadnyk, Editors

This book presents and considers main trends in the branch of metrology of cyber-physical systems, which are becoming a key element of everyday life. First of all it is destined for engineers, lecturers, students, persons who are not acquainted enough with specificity of cyber-physical systems and their metrology but are interested in it. The authors tried to highlight emergence and development of these systems, combined with the study of their metrology provision and support.

Formats: hardcover (print book) and PDF, 326 pages
ISBN: 978-84-608-9962-4, e-ISBN: 978-84-617-6200-2

https://www.sensorsportal.com/HTML/BOOKSTORE/Cyber_Physical_Systems.htm



Cyber-Physical Systems and Metrology 4.0

S. Yatsyshyn and B. Stadnyk, Editors

The book is written by 30 authors whose scientific achievements for the last 5 years cover a significant information technology and measurement science areas.

The purpose of this title is to present and consider the main trends in the field of metrology of Cyber-Physical Systems, which are becoming a key element of everyday life. At the first, the book is intended for engineers, lecturers, students, persons who are not acquainted enough with the specificity of Cyber-Physical Systems and their Metrology, but are interested in it.

Formats: hardcover (print book) and PDF, 332 pages
ISBN: 978-84-09-26899-3, e-ISBN: 978-84-09-26898-6

https://www.sensorsportal.com/HTML/BOOKSTORE/Cyber-Physical_Systems_and_Metrology_4_0.htm

Aims and Scope

Sensors & Transducers is established, international, peer-reviewed, open access journal (print and electronic). It provides the best platform for the researchers and scientist worldwide to exchange their latest findings and results in science and technology of physical, chemical sensors and biosensors. The journal publishes original results of scientific and research works related to strategic and applied studies in all aspects of sensors: reviews, regular research and application specific papers, and short notes. In comparison with other sensors related journals, which are mainly focused on technological aspects and sensing principles, the *Sensors & Transducers* significantly contributes in areas, which are not adequately addressed in other journals, namely: frequency (period), duty-cycle, time-interval, PWM, phase-shift, pulse number output sensors and transducers; sensor systems; digital, smart, intelligent sensors and systems designs; signal processing and ADC in sensor systems; advanced sensor fusion; sensor networks; applications, etc. By this way the journal significantly enriches the appropriate databases of knowledge.

Sensors & Transducers journal has a very high publicity. It is indexed and abstracted very quickly by Chemical Abstracts, EBSCO Publishing, ProQuest Science Journals, CAS Source Index (CASSI), Ulrich's Periodicals Directory, Google Scholar, etc. Since 2011 to 2014 *Sensors & Transducers* journal was covered and indexed by EI Compendex (CPX) index (including a Scopus, Embase, Engineering Village and Reaxys) in Elsevier products. The journal is included in the IFSA List of Recommended Journals (up-dated 9.12.2019), which contains only the best, established sensors related journals.

Topics of Interest

Include, but are not restricted to:

- Physical, chemical sensors and biosensors
- Digital, frequency, period, duty-cycle, time interval, PWM, pulse number output sensors and transducers
- Theory, principles, effects, design, standardization and modelling
- Smart sensors and systems
- Intelligent sensors and systems
- Sensor instrumentation
- Virtual instruments
- Sensors interfaces, buses and networks
- Signal processing, sensor fusion
- Remote sensing
- Frequency (period, duty-cycle)-to-code converters, ADC
- Technologies and materials
- Nanosensors and NEMS
- Microsystems
- Applications

Further information on this journal is available from the Publisher's web site:
http://www.sensorsportal.com/HTML/DIGEST/New_Digest.htm

Free Access and Subscriptions

Sensors & Transducers is the open access journal. All single articles are available for free download (article by article) without any registration and embargo period. For those who want to get the full page journal issue in print and/or pdf format there is an annual subscription. It includes 12 regular issues and some special issues. Annual subscription rates for 2020 are the following: Electronic version (in printable pdf format): 590.00 EUR; Printed with b/w illustrations: 950.00 EUR; 15 % discount is available for IFSA Members.

Further information about subscription is available on the IFSA Publishing's web site: http://www.sensorsportal.com/HTML/DIGEST/Journal_Subscription.htm

Advertising Information

If you are interested in advertising or other commercial opportunities please contact by e-mail: sales@sensorsportal.com

Instructions for Authors

http://www.sensorsportal.com/HTML/DIGEST/Guides_for_Authors.htm

Call for Chapters

IFSA Publishing

Your Chapter may be in the next Volume of the popular Book Series

ADVANCES IN SENSORS: REVIEWS 8

Open Access Book Series

Sinse 2012

- Physical, chemical and biosensors
- Sensor uncertainty, accuracy, calibration
- Smart sensors and systems
- Intelligent sensors and systems
- Artificial Intelligence enabled sensors
- Virtual and soft sensors
- Sensor instrumentation
- Sensor buses and networks
- Signal processing, sensor fusion
- Nanosensors
- Photonic sensors
- Sensor interfacing and signal conditioning
- Materials for sensors
- MEMS, BioMEMS and NEMS
- Remote Sensors and Telemetry
- Sensor packages and reliability

https://www.sensorsportal.com/HTML/IFSA_Publishing.htm

IFSA

International Frequency Sensor Association Publishing

www.sensorsportal.com

ISSN 1726- 5479