VCI-LSTM: Vector Choquet Integral-based Long Short-Term Memory

Mikel Ferrero-Jaurrieta, Zdenko Takáč, Javier Fernández, *Member, IEEE*, Ľubomíra Horanská, Graçaliz Pereira Dimuro, Susana Montes, Irene Díaz and Humberto Bustince, *Fellow, IEEE.*

***Abstract*—Choquet integral is a widely used aggregation operator on one-dimensional and interval-valued information, since it is able to take into account the possible interaction among data. However, there are many cases where the information taken into account is vectorial, such as Long Short-Term Memories (LSTM). LSTM units are a kind of Recurrent Neural Networks that have become one of the most powerful tools to deal with sequential information since they have the power of controlling the information flow. In this paper, we first generalize the standard Choquet integral to admit an input composed by n-dimensional vectors, which produces an n-dimensional vector output. We study several properties and construction methods of vector Choquet integrals. Then, we use this integral in the place of the summation operator, introducing in this way the new VCI-LSTM architecture. Finally, we use the proposed VCI-LSTM to deal with two problems: sequential image classification and text classification.**

***Index Terms*—Choquet Integral, Aggregation Functions, Vector Choquet Integral, Recurrent Neural Networks, LSTM.**

I. Introduction

INFORMATION aggregation process is a fundamental procedure when combining or aggregating different information structures into a single one [1]. Its use is usual in several fields, such as: multi-criteria decision making, economics and finance [2], statistics, image processing [3], machine learning [4], etc. Recently it has also been applied in deep learning, for example in the pooling layers of convolutional neural networks [5], [6].

Fuzzy integrals [7] are aggregation operators based on fuzzy measures [8], [9] that are capable of modelling the possible coalition among data [4]. In particular, the discrete Choquet integral [10] and its generalizations [11]–[13] are frequently used in several applications, such as classification [4], [14]–[16], multi-criteria decision making [17]–[19] and brain-computer interfaces [20].

Often the data considered for aggregation are multi-valued, i.e. they have a dimensionality greater than one. Examples of this are vectors, which store n-dimensional information. In this paper, we focus on vectors as the source of information to be aggregated.

An example of application where vectorial information is used are Long Short-Term Memories (LSTM) [21] which are a kind recurrent neural networks and a powerful tool to model sequential data, such as time series [22]–[24] and natural language [25]–[28]. These networks perform an information fusion process in order to calculate vectorial values, where the input data, hidden memory and bias information are fused. Traditionally, since summation is used to fuse the data, taking the information as independent of each other. However, these parameters may have interaction among them.

Until now different approaches of the discrete Choquet integrals have been introduced for aggregating different structures, such as intervals, with respect to admissible orders [29] and with respect to admissible permutations [30]. However, since the Choquet integral is defined for punctual or interval-valued entries, it can not be used as the aggregation operator admitting n-dimensional vector inputs, as required, for example, by the LSTM model.

To do so, in this paper we present the Vector Choquet Integral (VCI): an n-dimensional extension of the discrete Choquet-like integral, such that the inputs are n-dimensional vectors, and recovering an n-dimensional vector as output. First, the main definition and construction methods are presented and after the main properties of the Vector Choquet Integral are studied.

Regarding the application of the new introduced theory, we present a new LSTM architecture based on the VCI, called VCI-LSTM, as an example of an application where our theoretical developments can be applied. For that, in the vector-wise parameter data fusion process, we replace the summation by the new vector discrete Choquet integral. To check out the performance of the VCI-LSTM, we have tested it in two simple architectures, in order to solve two kind of problems: sequential image classification and text classification. Concerning text classification, spam detection, sentiment analysis and question classification datasets are considered. Three different datasets have been used to evaluate the achievement in each one. Results are compared with standard fusion method and also with an order statistic, like the maximum function. We show, validated by statistical analysis, that in most cases

Mikel Ferrero-Jaurrieta, Javier Fernández and Humberto Bustince are with the Departament of Statistics, Computer Science and Mathematics, Public University of Navarra, 31006 Pamplona, Spain (e-mails: {mikel.ferrero, fcojavier.fernandez, bustince}@unavarra.es).

Graçaliz Pereira Dimuro is with the Departament of Statistics, Computer Science and Mathematics, Public University of Navarra, 31006 Pamplona, Spain, and also with the Centro de Ciências Computacionais, Universidade Federal do Rio Grande, Rio Grande, 96077540, Brazil (e-mails: gracaliz.pereira@unavarra.es, gracalizdimuro@furg.br).

Zdenko Takáč and Ľubomíra Horanská are with the Institute of Information Engineering, Automation and Mathematics, Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Bratislava, Radlinskeho, 9, Bratislava, Slovakia (e-mail: {zdenko.takac, lubomira.horanska}@stuba.sk).

Susana Montes is with the Department of Statistics and Operational Research, University of Oviedo, Spain (e-mail: montes@uniovi.es).

Irene Díaz is with the Department of Computer Science, University of Oviedo, Spain (e-mail: sirene@uniovi.es).

VCI-LSTM-based models outperform standard LSTM ones.

The structure of this work is as follows. In Section II, aggregation functions and LSTM concepts are reminded. In Section III, the new vector Choquet integral is introduced, and its properties are studied in Section IV. In Section V, the new VCI-LSTM unit is proposed. In Section VI, experimental results are presented. Finally, in Section VII, conclusion and future work are explained.

II. Preliminaries

In this section, we recall basic notions necessary for the work. We present, on the one hand, the basic theoretical definitions of an aggregation function, a fuzzy measure and a Choquet integral, and on the other, the explanation of the performance of an RNN-LSTM network.

A. Aggregation functions and Choquet integral

Let us consider the lattice $(L, \leq)$ where $L = [0, 1]$ and $\leq$ is the natural order on the real numbers. The elements in L^n are in boldface, as $\mathbf{x} = (x_1, \ldots, x_n) \in L^n$, for $m > 0$ and $[0] = 0$.

We denote $\mathbf{0} = (0, \ldots, 0) \in L^n$, $\mathbf{1} = (1, \ldots, 1) \in L^n$ and $[m] = \{1, \ldots, m\}$.

Two vectors $(x_1, \ldots, x_m), (y_1, \ldots, y_m) \in L^m$ are comonotone if and only if there exists a permutation $\pi : [m] \to [m]$ such that $x_{\pi(1)} \leq \ldots \leq x_{\pi(m)}$ and $y_{\pi(1)} \leq \ldots \leq y_{\pi(m)}$.

Definition II.1. *[31] Let m be a positive integer. A function $M : L^m \to L$ is called an m-ary aggregation function if (i) $M(\mathbf{0}) = 0$ and $M(\mathbf{1}) = 1$; and (ii) is non-decreasing in each variable, i.e., for all $\mathbf{x} = (x_1, \ldots, x_m), \mathbf{y} = (y_1, \ldots, y_m) \in L^m$, $M(\mathbf{x}) \leq M(\mathbf{y})$ if $x_1 \leq y_1, \ldots, x_m \leq y_m$.*

A function $F : L^m \to L$ is called:

- Symmetric if $F(x_{\pi(1)}, \ldots, x_{\pi(m)}) = F(x_1, \ldots, x_m)$, for all $x_1, \ldots, x_m \in L$ and any permutation $\pi : [m] \to [m]$
- Idempotent if $F(x, \ldots, x) = x$, for all $x \in L$;
- Self-dual if $F(x_1, \ldots, x_m) = 1 - F(1 - x_1, \ldots, 1 - x_m)$, for all $x_1, \ldots, x_m \in L$;
- Shift-invariant if $F(x_1 + y, \ldots, x_m + y) = y + F(x_1, \ldots, x_m)$, for all $y, x_1, \ldots, x_m \in L$ such that $x_1 + y, \ldots, x_m + y \in L$;
- Positively homogeneous if $F(px_1, \ldots, px_m) = pF(x_1, \ldots, x_m)$, for all $p, x_1, \ldots, x_m \in L$
- Averaging if $\min(x_1, \ldots, x_m) \leq F(x_1, \ldots, x_m) \leq \max(x_1, \ldots, x_m)$, for all $x_1, \ldots, x_m \in L$;
- Comonotone additive if $F(x_1 + y_1, \ldots, x_m + y_m) = F(x_1, \ldots, x_m) + F(y_1, \ldots, y_m)$, for all comonotone vectors $(x_1, \ldots, x_m), (y_1, \ldots, y_m) \in L^m$ such that $(x_1 + y_1, \ldots, x_m + y_m) \in L^m$.

There is a partial order $\leq_P$ induced by $\leq$ given as follows: $\mathbf{x} \leq_P \mathbf{y}$ if and only if $x_i \leq y_i$, for all $i \in [n]$.
In fact, we can verify that $(L^n, \leq)$ is a lattice with operations $\mathbf{x} \wedge \mathbf{y} = (\min(x_1, y_1), \ldots, \min(x_n, y_n))$ and $\mathbf{x} \vee \mathbf{y} = (\max(x_1, y_1), \ldots, \max(x_n, y_n))$, having the minimum element $\mathbf{0}$ and the maximum $\mathbf{1}$. However, we intend to use also the total orders on L^n introduced in [32].

Definition II.2. *[32] Let n be a positive integer. A linear order $\leq_{Adm}$ in L^n is called admissible if, for all $\mathbf{x}, \mathbf{y} \in L^n$: $\mathbf{x} \leq_{Adm} \mathbf{y}$ whenever $x_i \leq y_i$, for all $i \in [n]$.*

In general, an order on L^n, no matter if partial or admissible, is denoted by $\leq_L$.

Definition II.3. *[9] A function $\nu : 2^{[m]} \to L$ is called a fuzzy measure on $[m]$ if (i) $\nu(\emptyset) = 0$ and $\nu([m]) = 1$ and (ii) $\nu(\mathcal{A}) \leq \nu(\mathcal{B})$, for all $\mathcal{A} \subseteq \mathcal{B} \subseteq [m]$*

By $Card(\mathcal{A})$ it is denoted the cardinality of the set $\mathcal{A}$. A fuzzy measure $\nu : 2^{[m]} \to L$ is called additive if $\nu(\mathcal{A} \cup \mathcal{B}) = \nu(\mathcal{A}) + \nu(\mathcal{B})$ for all $\mathcal{A}, \mathcal{B} \subseteq [m]$ such that $\mathcal{A} \cap \mathcal{B} = \emptyset$ and symmetric if $\nu(\mathcal{A}) = \nu(\mathcal{B})$, for all $\mathcal{A}, \mathcal{B} \subseteq [m]$ such that $Card(\mathcal{A}) = Card(\mathcal{B})$. Also, ν is called subadditive if, for $A, B \subseteq [m]$ such that $\mathcal{A} \cap \mathcal{B} = \emptyset$ it holds that $\nu(\mathcal{A} \cup \mathcal{B}) \leq \nu(\mathcal{A}) + \nu(\mathcal{B})$ and ν is called superadditive if $\nu(\mathcal{A} \cup \mathcal{B}) \geq \nu(\mathcal{A}) + \nu(\mathcal{B})$. Note that this properties are also referred to as submodularity and supermodularity, respectively, in the literature [1].

This implies that if a fuzzy measure is superadditive there is a positive correlation between the data, i.e. the data are redundant with each other. On the other hand, if a fuzzy measure is subadditive there is a negative correlation between the data, and therefore the data are complementary to each other.

Example II.4. *A fuzzy measure considered in this work is the power measure, which is a symmetric measure. It is defined, for all $\mathcal{A} \subseteq [m]$, as:*

$$\nu_q(\mathcal{A}) = \left(\frac{Card(\mathcal{A})}{m} \right)^q \tag{1}$$

where $q \in (0, \infty)$.

Definition II.5. *[10] The discrete Choquet integral on L with respect to the fuzzy measure ν is defined as a map $C_\nu : L^m \to L$ such that*

$$C_\nu(\mathbf{x}) = \sum_{i=1}^{m} \left(x_{\pi(i)} - x_{\pi(i-1)} \right) \nu \left(\mathcal{A}_{\pi(i)} \right)$$

where $\mathbf{x} = (x_1, \ldots, x_m) \in L^m$, $\nu : 2^{[m]} \to L$ is a fuzzy measure on $[m]$, $\pi : [m] \to [m]$ is a permutation, with $x_{\pi(1)} \leq \ldots \leq x_{\pi(m)}$ with the convention $x_{\pi(0)} = 0$, $\mathcal{A}_{\pi(i)} := \{\pi(i), \ldots, \pi(m)\}$ is the subset of the indices corresponding to the $m - i + 1$ greatest elements of $\mathbf{x}$, for all $i \in [m]$.

Remark II.6. *[1] The discrete Choquet integral C_ν fulfills the following properties: idempotence, self-duality, shift-invariance, positive homogeneity, averaging, monotonicity and commonotone additivity. C_ν is symmetric if the corresponding fuzzy measure is symmetric. In a similar way, C_ν is additive if the corresponding fuzzy measure is additive.*

Definition II.7. *[32] Let n be a positive integer and $\tilde{\mathbf{M}} = (M_1, \ldots, M_n)$ be a sequence of n-ary aggregation functions $M_i : L^n \to L$. Given $\mathbf{x}, \mathbf{y} \in L^n$, we define an order $\leq_M$ induced by $\tilde{\mathbf{M}}$ as follows:*

- $\mathbf{x} <_M \mathbf{y}$ *if and only if there exists $k \in [n]$ such that $M_j(\mathbf{x}) = M_j(\mathbf{y})$ for all $j \in [k-1]$ and $M_k(\mathbf{x}) < M_k(\mathbf{y})$;*

- $\mathbf{x} \leq_M \mathbf{y}$ *if and only if* $\mathbf{x} <_M \mathbf{y}$ *or* $\mathbf{x} = \mathbf{y}$.

Proposition II.8. *[32] Let n be a positive integer and $\tilde{\mathbf{M}} = (M_1, \ldots, M_n)$ be a sequence of n-ary aggregation functions $M_i : L^n \to L$. The order $\leq_M$ induced by $\tilde{\mathbf{M}}$ is an admissible order in L^n if and only if the aggregation functions M_i satisfy: $\mathbf{x} = \mathbf{y}$ if and only if $M_i(\mathbf{x}) = M_i(\mathbf{y})$ for all $i \in [n]$.*

Remark II.9. *Let $M_i(\mathbf{x}) = x_i$ for all $i \in [n]$ and all $\mathbf{x} = (x_1, \ldots, x_n) \in L^n$. Then the admissible order $\leq_M$ induced by the sequence $\tilde{\mathbf{M}} = (M_1, \ldots, M_n)$ is the lexicographical order, that we denote as $\leq_{Lex}$.*

In the following, we adapt the concept of admissible order discussed in [32]–[34] for our context:

Definition II.10. *Let n, m be positive integers and $\leq_L$ be a partial or admissible order on L^n. A function $\mathbf{M} : (L^n)^m \to L^n$ is a vector m-dimensional aggregation function if it satisfies (i) $\mathbf{M}(\mathbf{0}, \ldots, \mathbf{0}) = \mathbf{0}$ and $\mathbf{M}(\mathbf{1}, \ldots, \mathbf{1}) = \mathbf{1}$; and (ii) for all $\mathbf{x_1}, \ldots, \mathbf{x_m}, \mathbf{y_1}, \ldots, \mathbf{y_m} \in L^n$ it holds $\mathbf{x_1} \leq_L \mathbf{y_1}, \ldots, \mathbf{x_m} \leq_L \mathbf{y_m}$, then $\mathbf{M}(\mathbf{x_1}, \ldots, \mathbf{x_m}) \leq_L \mathbf{M}(\mathbf{y_1}, \ldots, \mathbf{y_m})$.*

B. Recurrent Neural Networks: Long Short-Term Memory

Recurrent Neural Networks (RNN) were introduced with the objective of modeling data with temporal dependence. However, since learning algorithms for RNNs are usually based on the backpropagation through time and gradient descent methods, they incur in the problem of vanishing gradient [35]. This problem consists in the recurrent decrease of the value of a variable in the output of the neural network. This is an especially serious problem when trying to train networks with long dependencies or time sequences.

In order to solve this problem, Long Short-Term Memory (LSTM) [36] was introduced, representing a radical change [35] in the training of recurrent networks since it avoids the continuous decrease of the parameters. This artificial neuron architecture [36] generates a state that allows the store of knowledge that is used in later time instants. In this way, special multiplicative units called gates are introduced in this new architecture (Figure 1).

LSTM neurons have had various modifications in the literature, but in this work we consider one of the most widely used [21]. In Figure 1 we can observe in detail the structure of an LSTM. It is important to remark the following elements [37]:

- Forget gate (f). Introduced by F. Gers in 2000 [38], it decides about which part of the long-term information should be discarded and what part of the long-term information is important and should be retained and moved on to the next time step.
- Input gate (i). It makes it possible for a part of the current state to be transmitted and reflected in long-term memory. It selects which part of the input should be removed while the others carry over to the long-term state corresponding to the next step.
- Output gate (o). It calculates the outputs and also decides about which part of the long-term information is going to the next step.

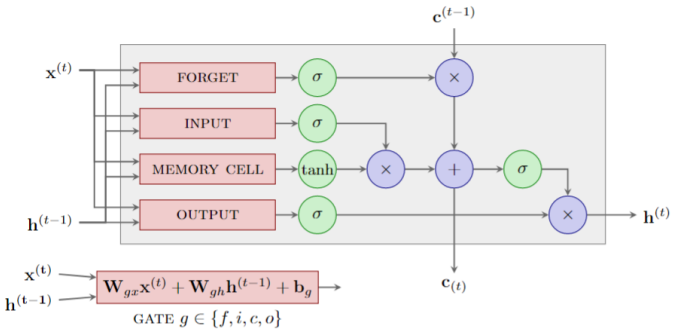


Fig. 1: LSTM unit representation

Next, we are going to explain the operation of the LSTM unit. Let N be the input sequence length, H the hidden size of the LSTM unit and T the number of timesteps. Then we get the following weights [21] for the matrices and vectors associated with the gates and the candidate cell ($g \in \{f, i, c, o\}$):

- Input weight matrices: $\mathbf{W}_{gx} \in \mathbb{R}^{H \times N}$
- Recurrent weight matrices: $\mathbf{W}_{gh} \in \mathbb{R}^{H \times H}$
- Bias weight vectors: $\mathbf{b}_g \in \mathbb{R}^H$

The operations description for each timestep $t \in \{1, \ldots, T\}$ is the following:

i. The input values $\mathbf{x}^{(t)}$ and $\mathbf{h}^{(t-1)}$ enter to the gates f (Eq. 2), i (Eq. 3), $\tilde{c}$ (Eq. 4) and o (Eq. 6). In each of them, the value of $\mathbf{x}^{(t)}$ is multiplied by each of the input weight matrices ($\mathbf{W}_{gx}$, depending on the gate g). The same occurs with the values of $\mathbf{h}^{(t-1)}$ and the recurrent weight matrices. The H-dimensional vectors obtained from these multiplications with the corresponding bias $\mathbf{b}_g$ for each gate g are fused summing them.

ii. As activation function non-linear functions are used. As gate activation function the sigmoid logistic function, $\sigma(x) = (1+e^{-x})^{-1}$ is considered. As activation function of the candidate cell the hyperbolic tangent $\tanh(x)$ is taken. Both of these functions are defined on $\mathbb{R}$. When acting over vectors, they are applied coordenate-wise.

iii. The previous timestep long-term memory vector ($\mathbf{c}^{(t-1)}$) and the candidate cell one ($\tilde{\mathbf{c}}^{(t)}$) are combined in this step. The Hadamard or element-wise product ($\circ$) is calculated between the values of the forget gate and input gate respectively (Eq. 5). Both values are added obtaining the current timestep value of the long-term vector ($\mathbf{c}^{(t)}$).

iv. Finally, the short-time memory vector ($\mathbf{h}^{(t)}$) is calculated. First, the long-term information ($\mathbf{c}^{(t)}$) is evaluated by the $\tanh(x)$ activation function. Subsequently, the Hadamard product is calculated between the value of the output gate ($\mathbf{o}^{(t)}$) and the information obtained from the last activation function, obtaining the value of the short-term memory vector ($\mathbf{h}^{(t)}$).

The process describing equations are the following (Eq. (2)-(7)):

$$\mathbf{f}^{(t)} = \sigma(\mathbf{W}_{fx}\mathbf{x}^{(t)} + \mathbf{W}_{fh}\mathbf{h}^{(t-1)} + \mathbf{b}_f) \tag{2}$$

$$\mathbf{i}^{(t)} = \sigma(\mathbf{W}_{ix}\mathbf{x}^{(t)} + \mathbf{W}_{ih}\mathbf{h}^{(t-1)} + \mathbf{b}_i) \tag{3}$$

$$\tilde{\mathbf{c}}^{(t)} = \tanh(\mathbf{W}_{cx}\mathbf{x}^{(t)} + \mathbf{W}_{ch}\mathbf{h}^{(t-1)} + \mathbf{b}_c) \tag{4}$$

$$\mathbf{c}^{(t)} = \mathbf{f}^{(t)} \circ \mathbf{c}^{(t-1)} + \mathbf{i}^{(t)} \circ \tilde{\mathbf{c}}^{(t)} \tag{5}$$

$$\mathbf{o}^{(t)} = \sigma(\mathbf{W}_{ox}\mathbf{x}^{(t)} + \mathbf{W}_{oh}\mathbf{h}^{(t-1)} + \mathbf{b}_o) \tag{6}$$

$$\mathbf{h}^{(t)} = \mathbf{o}^{(t)} \circ \tanh(\mathbf{c}^{(t)}) \tag{7}$$

From Eq. (2)-(7) is easy to see that $\mathbf{f}^{(t)}, \mathbf{i}^{(t)}, \mathbf{o}^{(t)} \in [0,1]^H$, $\tilde{\mathbf{c}}^{(t)}, \mathbf{h}^{(t)} \in [-1,1]^H$ and $\mathbf{c}^{(t)} \in \mathbb{R}^H$.

III. Vector Choquet integral (VCI)

In this section, we introduce the concept of vector Choquet integral, studying its properties.

Let $\mathbf{x_1} = (x_{11}, \ldots, x_{1n}), \ldots, \mathbf{x_m} = (x_{m1}, \ldots, x_{mn})$ be m vectors in L^n, and $\mathbf{x^1} = (x_{11}, \ldots, x_{m1}), \ldots, \mathbf{x^n} = (x_{1n}, \ldots, x_{mn})$ be n vectors in L^m obtained as follows: Define the $m \times n$ matrix on L, where the rows are given by $\mathbf{x_1}, \ldots, \mathbf{x_m}$. Then, the columns are $\mathbf{x^1}, \ldots, \mathbf{x^n}$.

$$\begin{array}{c} \\ \mathbf{x_1} \\ \mathbf{x_2} \\ \vdots \\ \mathbf{x_m} \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{x^1} & \mathbf{x^2} & \ldots & \mathbf{x^n} \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \ldots & x_{mn} \end{pmatrix} \end{array}$$

Definition III.1. *Let n, m be positive integers, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$ be a sequence of fuzzy measures on $[m]$ and $C_{\nu_i} : L^m \to L$ be discrete Choquet integrals on L^m with respect to ν_i, for all $i \in [n]$. A function $\mathbf{C_\nu} : (L^n)^m \to L^n$, given, for all $\mathbf{x_1}, \ldots, \mathbf{x_m} \in L^n$, by*

$$\mathbf{C_\nu}(\mathbf{x_1}, \ldots, \mathbf{x_m}) = \left(C_{\nu_1}(\mathbf{x^1}), \ldots, C_{\nu_n}(\mathbf{x^n})\right), \tag{8}$$

is called a discrete vector Choquet integral (VCI) on $(L^n)^m$ with respect to $\boldsymbol{\nu}$ and order $\leq_L$.

The VCI $\mathbf{C_\nu}$ is said to be "representable" since it is obtained by using n Choquet integrals on L^m separately for each component:

$$\mathbf{C_\nu}(\mathbf{x_1}, \ldots, \mathbf{x_m}) = (C_{\nu_1}(x_{11}, \ldots, x_{m1}), \ldots, C_{\nu_n}(x_{1n}, \ldots, x_{mn})). \tag{9}$$

This expression is a generalization of the standard Choquet integral on L in the sense that if all the inputs are n-tuples with the same coordinates, i.e. $\mathbf{x} = (x, \ldots, x)$ and $\nu_1 = \ldots = \nu_n = \nu$, then the output is an n-tuple with the same coordinates equal to the output of C_ν.

Proposition III.2. *Under the assumption of Definition III.1, consider $\nu_1 = \ldots = \nu_n = \nu$. Then:*

$$\mathbf{C_\nu}(\mathbf{x_1}, \ldots, \mathbf{x_m}) = (C_\nu(x_1, \ldots, x_m), \ldots, C_\nu(x_1, \ldots, x_m)), \tag{10}$$

for all $\mathbf{x_i} = (x_i, \ldots, x_i) \in L^n$, $i \in [m]$.

Definition III.3. *Let m, n be positive integers, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$ be a sequence of fuzzy measures on $[m]$ and $C_{\nu_i} : L^m \to L$ be Choquet integrals on L with respect to ν_i, for all $i \in [n]$. Let $\tilde{\mathbf{M}}_{in} = (M_1, \ldots, M_n)$ be a sequence of n-ary aggregation functions $M_1, \ldots, M_n : L^n \longrightarrow L$. A function $\mathbf{C_\nu^{\tilde{M}_{in}}} : (L^n)^m \to L^n$ given by*

$$\mathbf{C_\nu^{\tilde{M}_{in}}}(\mathbf{x_1}, \ldots, \mathbf{x_m}) = (C_{\nu_1}(M_1(\mathbf{x_1}), \ldots, M_1(\mathbf{x_m})), \ldots, C_{\nu_n}(M_n(\mathbf{x_1}), \ldots, M_n(\mathbf{x_m}))), \tag{11}$$

for all $\mathbf{x_1}, \ldots, \mathbf{x_m} \in L^n$, is called a vector $\tilde{\mathbf{M}}_{in}$-Choquet integral (VCI-M_{in}) on L^n with respect to $\boldsymbol{\nu}$, $\tilde{\mathbf{M}}_{in}$ and the order $\leq_L$.

Definition III.4. *Let m, n be positive integers, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$ be a sequence of fuzzy measures on $[m]$ and $C_{\nu_i} : L^m \to L$ be discrete Choquet integrals on L^m with respect to ν_i, for all $i \in [n]$. Let $\tilde{\mathbf{M}}_{out} = (M_1, \ldots, M_n)$ be a sequence of n-ary aggregation functions $M_1, \ldots, M_n : L^n \longrightarrow L$. A function $\mathbf{C_\nu^{\tilde{M}_{out}}} : (L^n)^m \to L^n$, given by*

$$\mathbf{C_\nu^{\tilde{M}_{out}}}(\mathbf{x_1}, \ldots, \mathbf{x_m}) = (M_1(C_{\nu_1}(\mathbf{x^1}), \ldots, C_{\nu_n}(\mathbf{x^n})), \ldots, M_n(C_{\nu_1}(\mathbf{x^1}), \ldots, C_{\nu_n}(\mathbf{x^n}))), \tag{12}$$

for all $\mathbf{x_1}, \ldots, \mathbf{x_m} \in L^n$, is called a vector $\tilde{\mathbf{M}}_{out}$-Choquet integral (VCI-M_{out}) on L^n with respect to $\boldsymbol{\nu}$, $\tilde{\mathbf{M}}_{out}$, and the order $\leq_L$.

Remark III.5. *Both $\tilde{\mathbf{M}}_{in}$-Choquet integral and $\tilde{\mathbf{M}}_{out}$-Choquet integral can be expressed in terms of vector Choquet integrals as follows:*

$$\mathbf{C_\nu^{\tilde{M}_{in}}}(\mathbf{x_1}, \ldots, \mathbf{x_m}) = \mathbf{C_\nu}((M_1(\mathbf{x_1}), \ldots, M_n(\mathbf{x_1})), \ldots, (M_1(\mathbf{x_m}), \ldots, M_n(\mathbf{x_m})))$$

$$\mathbf{C_\nu^{\tilde{M}_{out}}}(\mathbf{x_1}, \ldots, \mathbf{x_m}) = (M_1(\mathbf{C_\nu}(\mathbf{x_1}, \ldots, \mathbf{x_m})), \ldots, M_n(\mathbf{C_\nu}(\mathbf{x_1}, \ldots, \mathbf{x_m}))),$$

for all $\mathbf{x_1}, \ldots, \mathbf{x_m} \in L^n$.

The following relation between the three types of vector Choquet integrals is immediate.

Proposition III.6. *Under the assumptions of Definition III.3, if $M_i(\mathbf{y}) = y_i$, for all $\mathbf{y} = (y_1, \ldots, y_n) \in L^n$, $i \in [n]$, then $\mathbf{C_\nu}(\mathbf{x_1}, \ldots, \mathbf{x_m}) = \mathbf{C_\nu^{\tilde{M}_{out}}}(\mathbf{x_1}, \ldots, \mathbf{x_m}) = \mathbf{C_\nu^{\tilde{M}_{in}}}(\mathbf{x_1}, \ldots, \mathbf{x_m})$, for all $\mathbf{x_1}, \ldots, \mathbf{x_m} \in L^n$.*

Note that, by Proposition III.2 and Proposition III.6, all the three introduced vector Choquet integrals, namely, $\mathbf{C_\nu}, \mathbf{C_\nu^{\tilde{M}_{out}}}, \mathbf{C_\nu^{\tilde{M}_{in}}}$, generalize the standard Choquet integral, i.e., they recover standard Choquet integral for inputs being n-tuples with the same coordinates.

IV. Properties of vector Choquet integrals

By Remark II.6, the discrete Choquet integral C_ν fulfills the following properties: idempotence, self-duality, shift-invariance, positive homogeneity, averaging, monotonicity and commonotone additivity. It is also symmetric/additive if the fuzzy measure is symmetric/additive. These properties are studied for the three vector Choquet integrals $\mathbf{C_\nu}, \mathbf{C_\nu^{\tilde{M}_{in}}}, \mathbf{C_\nu^{\tilde{M}_{out}}}$.

A. Symmetry, boundary conditions and idempotency

Lemma IV.1. *Let $\mathbf{C_\nu^{\tilde{M}_{in}}} : (L^n)^m \to L^n$ be a vector $\tilde{\mathbf{M}}_{in}$-Choquet integral given by Definition III.3 and $\mathbf{C_\nu^{\tilde{M}_{out}}} : (L^n)^m \to L^n$ be a vector $\tilde{\mathbf{M}}_{out}$-Choquet integral given by*

Definition III.4. Then, $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x},\dots,\mathbf{x}) = \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x},\dots,\mathbf{x})$ *for all* $\mathbf{x} \in L^n$.

Proof. $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(\mathbf{x},\dots,\mathbf{x}) = (C_{\nu_1}(M_1(\mathbf{x}),\dots,M_1(\mathbf{x})),\dots,$
$C_{\nu_n}(M_n(\mathbf{x}),\dots,M_n(\mathbf{x}))) = (M_1(\mathbf{x}),\dots,M_n(\mathbf{x}))$
and $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{out}}}(\mathbf{x},\dots,\mathbf{x}) = (M_1(C_{\nu_1}(x_1,\dots,x_1),\dots,$
$C_{\nu_n}(x_n,\dots,x_n)),\dots,M_n(C_{\nu_1}(x_1,\dots,x_1),\dots,$
$C_{\nu_n}(x_n,\dots,x_n))) = (M_1(\mathbf{x}),\dots,M_n(\mathbf{x}))$

□

Theorem IV.2. *Let* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}} : (L^n)^m \to L^n$ *be a vector* $\tilde{\mathbf{M}}_{in}$*-Choquet integral given by Definition III.3 and* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}} : (L^n)^m \to L^n$ *be a vector* $\tilde{\mathbf{M}}_{out}$*-Choquet integral given by Definition III.4. Then:*

- *(i) Let* ν_i *be symmetric for each* $i \in [n]$*. Then* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x_1},\dots,\mathbf{x_m}) = \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x}_{\pi(\mathbf{1})},\dots,\mathbf{x}_{\pi(\mathbf{m})})$ *and* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x_1},\dots,\mathbf{x_m}) = \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x}_{\pi(\mathbf{1})},\dots,\mathbf{x}_{\pi(\mathbf{m})})$*, for all* $\mathbf{x_1},\dots,\mathbf{x_m} \in L^m$ *and for any permutation* $\pi : [m] \to [m]$*;*
- *(ii)* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{0},\dots,\mathbf{0}) = \mathbf{0}, \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{1},\dots,\mathbf{1}) = \mathbf{1}$ *and* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{0},\dots,\mathbf{0}) = \mathbf{0}$*,* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{1},\dots,\mathbf{1}) = \mathbf{1}$*, for all* $\mathbf{x_1},\dots,\mathbf{x_m} \in L^m$*;*
- *(iii)* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x},\dots,\mathbf{x}) = \mathbf{x}$*, for all* $\mathbf{x} \in L^n$ *and for any sequence* $\boldsymbol{\nu} = (\nu_1,\dots,\nu_n)$ *of fuzzy measures on* $[m]$*, if and only if* $M_i(\mathbf{x}) = x_i$*, for all* $\mathbf{x} \in L^n$*,* $i \in [n]$*;*
- *(iv)* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x},\dots,\mathbf{x}) = \mathbf{x}$*, for all* $\mathbf{x} \in L^n$ *and for any sequence* $\boldsymbol{\nu} = (\nu_1,\dots,\nu_n)$ *of fuzzy measures on* $[m]$*, if and only if* $M_i(\mathbf{x}) = x_i$*, for all* $\mathbf{x} \in L^n$*,* $i \in [n]$*.*

Proof. The proof of *(i)* follows from the symmetry of C_{ν_i}, $i \in [n]$. The proof of *(ii)* follows from the boundary conditions of $M_i, C_{\nu_i}, i \in [n]$, and the proof of *(iii)* from Lemma IV.1, in the particular case $M_i(\mathbf{x}) = x_i$, where $(M_1(\mathbf{x}),\dots,M_n(\mathbf{x})) = (x_1,\dots,x_n) = \mathbf{x}$.The proof of *(iv)* follows from *(iii)* and Lemma IV.1. □

B. Self-duality

Theorem IV.3. *Let* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}} : (L^n)^m \to L^n$ *be a vector* $\tilde{\mathbf{M}}_{in}$*-Choquet integral given by Definition III.3 and* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}} : (L^n)^m \to L^n$ *be a vector* $\tilde{\mathbf{M}}_{out}$*-Choquet integral given by Definition III.4. Then:*

- *(i) For any sequence* $\boldsymbol{\nu} = (\nu_1,\dots,\nu_n)$ *of fuzzy measures on* $[m]$ *and for all* $\mathbf{x_1},\dots,\mathbf{x_m} \in L^n$*, it holds that:* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x_1},\dots,\mathbf{x_m}) = \mathbf{1} - \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{1}-\mathbf{x_1},\dots,\mathbf{1}-\mathbf{x_m})$ *if and only if* M_i *is self-dual, for all* $i \in [n]$*;*
- *(ii) For any sequence* $\boldsymbol{\nu} = (\nu_1,\dots,\nu_n)$ *of fuzzy measures on* $[m]$ *and for all* $\mathbf{x_1},\dots,\mathbf{x_m} \in L^n$*, it holds that:* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x_1},\dots,\mathbf{x_m}) = \mathbf{1} - \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{1}-\mathbf{x_1},\dots,\mathbf{1}-\mathbf{x_m})$ *if and only if* M_i *is self-dual for all* $i \in [n]$*.*

Proof. For *(i)* sufficiency, the proof follows from the self-duality of M_i and C_{ν_i}, $i \in [n]$. Necessity: Consider that there exist $k \in [n]$ and $\mathbf{x} \in L^n$ such that $M_k(\mathbf{1}-\mathbf{x}) \neq 1 - M_k(\mathbf{x})$. Then: $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(\mathbf{1}-\mathbf{x},\dots,\mathbf{1}-\mathbf{x}) =$
$(C_{\nu_1}(M_1(\mathbf{1}-\mathbf{x}),\dots,M_1(\mathbf{1}-\mathbf{x})),\dots,C_{\nu_n}(M_n(\mathbf{1}-\mathbf{x}),$
$\dots,M_n(\mathbf{1}-\mathbf{x}))) = (M_1(\mathbf{1}-\mathbf{x}),\dots,M_n(\mathbf{1}-\mathbf{x}))$
by idempotency $\neq (1 - M_1(\mathbf{x}),\dots,1 - M_n(\mathbf{x}))$
$= \mathbf{1} - \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(\mathbf{x},\dots,\mathbf{x})$. The proof of *(ii)* is similar to the proof of *(i)*, but considering IV.1 for the necessity. □

C. Shift-invariance

For $y \in L$ and $\mathbf{x} = (x_1,\dots,x_n) \in L^n$, we denote $\mathbf{x} + y = (x_1 + y,\dots,x_n + y)$.

Theorem IV.4. *Let* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}} : (L^n)^m \to L^n$ *be a vector* $\tilde{\mathbf{M}}_{in}$*-Choquet integral given by Definition III.3 and* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}} : (L^n)^m \to L^n$ *be a vector* $\tilde{\mathbf{M}}_{out}$*-Choquet integral given by Definition III.4. Then:*

- *(i) For all* $\mathbf{x_1},\dots,\mathbf{x_m},\mathbf{y} \in L^n$*,* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x_1}+\mathbf{y},\dots,\mathbf{x_m}+\mathbf{y}) = \mathbf{y} + \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x_1},\dots,\mathbf{x_m})$*, whenever* $M_i(\mathbf{x}) = x_i$*, for all* $\mathbf{x} \in L^n$*,* $i \in [n]$*;*
- *(ii) For any sequence* $\boldsymbol{\nu} = (\nu_1,\dots,\nu_n)$ *of fuzzy measures on* $[m]$ *and for all* $\mathbf{x_1},\dots,\mathbf{x_m} \in L^n$*,* $y \in L$*, it holds that:* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x_1}+y,\dots,\mathbf{x_m}+y) = y + \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x_1},\dots,\mathbf{x_m})$ *if and only if* M_i *is shift-invariant for all* $i \in [n]$*;*
- *(iii) For all* $\mathbf{x_1},\dots,\mathbf{x_m},\mathbf{y} \in L^n$*,* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x_1}+\mathbf{y},\dots,\mathbf{x_m}+\mathbf{y}) = \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x_1},\dots,\mathbf{x_m})$ *whenever* $M_i(\mathbf{x}) = x_i$*, for all* $\mathbf{x} \in L^n$*,* $i \in [n]$*;*
- *(iv) For any sequence* $\boldsymbol{\nu} = (\nu_1,\dots,\nu_n)$ *of fuzzy measures on* $[m]$ *and for all* $\mathbf{x_1},\dots,\mathbf{x_m} \in L^n$*,* $y \in L$*, it holds that:* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x_1}+y,\dots,\mathbf{x_m}+y) = y + \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x_1},\dots,\mathbf{x_m})$ *if and only if* M_i *is shift-invariant for all* $i \in [n]$*.*

Proof. For *(i)*, the proof follows from the shift-invariance of C_{ν_i}, $i \in [n]$, and the observation that $M_i(\mathbf{y} + \mathbf{x_j}) = y_i + M_i(\mathbf{x_j})$, for all $i \in [n], j \in [m]$. For *(ii)* sufficiency, the proof follows from the shift-invariance of M_i and C_{ν_i}, $i \in [n]$. Necessity: Consider that there exist $k \in [n]$ and $\mathbf{x} \in L^n, y \in L$, such that $M_k(\mathbf{x}+y) \neq y + M_k(\mathbf{x})$. Then:
$\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(\mathbf{x}+y,\dots,\mathbf{x}+y) = (C_{\nu_1}(M_1(\mathbf{x}+y),\dots,$
$M_1(\mathbf{x}+y)),\dots,C_{\nu_n}(M_n(\mathbf{x}+y),\dots,M_n(\mathbf{x}+y))) = (M_1(\mathbf{x}+$
$y),\dots,M_n(\mathbf{x}+y)) \neq (y + M_1(\mathbf{x},\dots,y + M_n(\mathbf{x})) = y +$
$\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(\mathbf{x},\dots,\mathbf{x})$. The proofs of *(iii)* and *(iv)* are similar to the proofs of *(i)* and *(ii)*, considering Lemma IV.1 for the necessity in *(iv)*. □

D. Positive homogeneity

Theorem IV.5. *Let* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}} : (L^n)^m \to L^n$ *be a vector* $\tilde{\mathbf{M}}_{in}$*-Choquet integral given by Definition III.3 and* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}} : (L^n)^m \to L^n$ *be a vector* $\tilde{\mathbf{M}}_{out}$*-Choquet integral given by Definition III.4. Then:*

- *(i) For any sequence* $\boldsymbol{\nu} = (\nu_1,\dots,\nu_n)$ *of fuzzy measures on* $[m]$ *and for all* $\mathbf{x_1},\dots,\mathbf{x_m} \in L^n$*,* $p \in L$*, it holds that:*$\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(p\mathbf{x_1},\dots,p\mathbf{x_m}) = p\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x_1},\dots,\mathbf{x_m})$ *if and only if* M_i *is positively homogeneous, for all* $i \in [n]$*;*
- *(ii) For any sequence* $\boldsymbol{\nu} = (\nu_1,\dots,\nu_n)$ *of fuzzy measures on* $[m]$ *and for all* $\mathbf{x_1},\dots,\mathbf{x_m} \in L^n$*,* $p \in L$*, it holds that:* $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(p\mathbf{x_1},\dots,p\mathbf{x_m}) = p\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x_1},\dots,\mathbf{x_m}$ *if and only if* M_i *is positively homogeneous, for all* $i \in [n]$*.*

Proof. For *(i)* sufficiency, the proof follows from the positive homogeneity of M_i and C_{ν_i}, for all $i \in [n]$. Necessity: Consider that there exist $k \in [n]$ and $\mathbf{x} \in L^n$, $p \in L$ such that $M_k(p\mathbf{x}) \neq pM_k(\mathbf{x})$. Then:
$\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(p\mathbf{x},\dots,p\mathbf{x}) = (C_{\nu_1}(M_1(p\mathbf{x}),\dots,M_1(p\mathbf{x})),\dots,$
$C_{\nu_n}(M_n(p\mathbf{x}),\dots,M_n(p\mathbf{x}))) = (M_1(p\mathbf{x}),\dots,M_n(p\mathbf{x}))$
$\neq (pM_1(\mathbf{x},\dots,pM_n(\mathbf{x})) = p\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(\mathbf{x},\dots,\mathbf{x})$.

The proof of *(ii)* is similar, considering Lemma IV.1 for the necessity. □

E. Averageness

Now, a result with respect to averageness considering the partial order $\leq_P$ on L^n is given.

Theorem IV.6. *Let $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}} : (L^n)^m \to L^n$ be a vector $\tilde{\mathbf{M}}_{in}$-Choquet integral given by Definition III.3 and $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}} : (L^n)^m \to L^n$ be a vector $\tilde{\mathbf{M}}_{out}$-Choquet integral given by Definition III.4. Then:*

(i) For any sequence $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$ of fuzzy measures on $[m]$ and for all $\mathbf{x_1}, \ldots, \mathbf{x_m} \in L^n$, it holds that: $\min(\mathbf{x_1}, \ldots, \mathbf{x_m}) \leq_P \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x_1}, \ldots, \mathbf{x_m}) \leq_P \max(\mathbf{x_1}, \ldots, \mathbf{x_m})$ if and only if $M_i(\mathbf{x}) = x_i$, for all $\mathbf{x} \in L^n$, $i \in [n]$;

(ii) For any sequence $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$ of fuzzy measures on $[m]$ and for all $\mathbf{x_1}, \ldots, \mathbf{x_m} \in L^n$, it holds that: $\min(\mathbf{x_1}, \ldots, \mathbf{x_m}) \leq_P \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x_1}, \ldots, \mathbf{x_m}) \leq_P \max(\mathbf{x_1}, \ldots, \mathbf{x_m})$ if and only if $M_i(\mathbf{x}) = x_i$, for all $\mathbf{x} \in L^n$, $i \in [n]$.

Proof. For *(i)* sufficiency, the proof directly follows from $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(\mathbf{x_1}, \ldots, \mathbf{x_m}) = (C_{\nu_1}(x_{11}, \ldots, x_{m1}), \ldots, C_{\nu_n}(x_{1n}, \ldots, x_{mn}))$ and the averagingness of C_{ν_i}, $i \in [n]$. Necessity: Consider that there exist $k \in [n]$ and $\mathbf{x} \in L^n$ such that $M_k(\mathbf{x}) > x_k$. Then: $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{out}}}(\mathbf{x}, \ldots, \mathbf{x}) = (M_1(\mathbf{x}), \ldots, M_n(\mathbf{x})) \not\leq_P \mathbf{x}$.
In a similar way we obtain $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{out}}}(\mathbf{x}, \ldots, \mathbf{x}) \not\geq_P \mathbf{x}$ for $M_k(\mathbf{x}) < x_k$. For *(ii)*, the proof follows from *(i)* considering Proposition III.6 for sufficiency and Lemma IV.1 for necessity. □

With respect to admissible orders, there is a weaker result than the previous Theorem.

Corollary IV.7. *Let $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}} : (L^n)^m \to L^n$ be a vector $\tilde{\mathbf{M}}_{in}$-Choquet integral given by Definition III.3 and $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}} : (L^n)^m \to L^n$ be a vector $\tilde{\mathbf{M}}_{out}$-Choquet integral given by Definition III.4. Then:*

(i) For all $\mathbf{x_1}, \ldots, \mathbf{x_m} \in L^n$ such that $\mathbf{x_j} = (x_{j1}, \ldots, x_{jn})$, for all $j \in [m]$, one has that:
$((\min(x_{11}, \ldots, x_{m1}), \ldots, \min(x_{1n}, \ldots, x_{mn})) \leq_{Adm} \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x_1}, \ldots, \mathbf{x_m})$
$\leq_{Adm} ((\max(x_{11}, \ldots, x_{m1}), \ldots, \max(x_{1n}, \ldots, x_{mn}))$
whenever $M_i(\mathbf{x}) = x_i$, for all $\mathbf{x} \in L^n$, $i \in [n]$;

(ii) For all $\mathbf{x_1}, \ldots, \mathbf{x_m} \in L^n$ such that $\mathbf{x_j} = (x_{j1}, \ldots, x_{jn})$ for all $j \in [m]$, one has that
$((\min(x_{11}, \ldots, x_{m1}), \ldots, \min(x_{1n}, \ldots, x_{mn}))$
$\leq_{Adm} \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x_1}, \ldots, \mathbf{x_m})$
$\leq_{Adm} ((\max(x_{11}, \ldots, x_{m1}), \ldots, \max(x_{1n}, \ldots, x_{mn}))$
whenever $M_i(\mathbf{x}) = x_i$, for all $\mathbf{x} \in L^n$, $i \in [n]$.

F. Monotonicity

The $\tilde{\mathbf{M}}_{in}$-Choquet integral $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}$ is monotone, increasing in each component, with respect to the partial order $\leq_P$. However, with respect to admissible orders, it is monotone only under a specific condition. First, a direct result with respect to the partial order $\leq_P$ is given.

Theorem IV.8. *Let $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}} : (L^n)^m \to L^n$ be a vector $\tilde{\mathbf{M}}_{in}$-Choquet integral given by Definition III.3 and $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}} : (L^n)^m \to L^n$ be a vector $\tilde{\mathbf{M}}_{out}$-Choquet integral given by Definition III.4. Then:*

(i) For all $\mathbf{x_1}, \ldots, \mathbf{x_m}, \mathbf{y} \in L^n$, one has that: $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x_1}, \ldots, \mathbf{x_m}) \leq_P \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x_1}, \ldots, \mathbf{x_{k-1}}, \mathbf{y}, \mathbf{x_{k+1}}, \ldots \mathbf{x_m})$ whenever $\mathbf{x_k} \leq_P \mathbf{y}$, for some $k \in [m]$;

(ii) For all $\mathbf{x_1}, \ldots, \mathbf{x_m}, \mathbf{y} \in L^n$, one has that $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x_1}, \ldots, \mathbf{x_m}) \leq_P \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x_1}, \ldots, \mathbf{x_{k-1}}, \mathbf{y}, \mathbf{x_{k+1}}, \ldots \mathbf{x_m})$ whenever $\mathbf{x_k} \leq_P \mathbf{y}$, for some $k \in [m]$.

Proof. It directly follows from the monotonicity of M_i and C_{ν_i}, for all $i \in [n]$. □

The situation with respect to admissible orders is not so straightfoward. The monotonicity is preserved only under rather restrictive assumptions. First, the condition for a fuzzy measure ν under which a standard Choquet integral with respect to ν is strictly increasing is given in the following proposition.

Proposition IV.9. *Let $C_\nu : L^m \to L$ be a discrete Choquet-like integral on L with respect to fuzzy measure ν. If $\nu(\mathcal{A}) < \nu(\mathcal{B})$, for all $\mathcal{A}, \mathcal{B} \subseteq [m]$ such that $Card(\mathcal{A}) < Card(\mathcal{B})$, then C_ν is strictly increasing.*

Recall that, according to the Remark II.9, the admissible order $\leq_{Adm}$ induced by the sequence $\tilde{\mathbf{M}}_{in} = (M_1, \ldots, M_n)$ such that $M_i(\mathbf{x}) = x_i$, for all $\mathbf{x} = (x_1, \ldots, x_n) \in L^n$, is the lexicographical order $\leq_{Lex}$. So the following theorem gives us the conditions under which an $\tilde{\mathbf{M}}_{in}$-Choquet integral on L^n is monotone with respect to $\leq_{Lex}$.

Theorem IV.10. *Let $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}} : (L^n)^m \to L^n$ be a vector $\tilde{\mathbf{M}}_{in}$-Choquet integral given by Definition III.3 and $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}} : (L^n)^m \to L^n$ be a vector $\tilde{\mathbf{M}}_{out}$-Choquet integral given by Definition III.4. For all $i \in [n]$, let C_{ν_i} be strictly increasing and $M_i(\mathbf{x}) = x_i$, for all $\mathbf{x} = (x_1, \ldots, x_n) \in L^n$. Let $\leq_{Adm}$ be the admissible order induced by the sequence $\tilde{\mathbf{M}} = (M_1, \ldots, M_n)$, according to the Proposition II.8 ($\leq_{Adm}$ is $\leq_{Lex}$). Then:*

(i) For all $\mathbf{x_1}, \ldots, \mathbf{x_m}, \mathbf{y} \in L^n$, one has that: $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x_1}, \ldots, \mathbf{x_m}) \leq_{Adm} \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x_1}, \ldots, \mathbf{x_{k-1}}, \mathbf{y}, \mathbf{x_{k+1}}, \ldots \mathbf{x_m})$ whenever $\mathbf{x_k} \leq_P \mathbf{y}$, for some $k \in [m]$;

(ii) For all $\mathbf{x_1}, \ldots, \mathbf{x_m}, \mathbf{y} \in L^n$, one has that: $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x_1}, \ldots, \mathbf{x_m}) \leq_{Adm} \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}(\mathbf{x_1}, \ldots, \mathbf{x_{k-1}}, \mathbf{y}, \mathbf{x_{k+1}}, \ldots \mathbf{x_m})$ whenever $\mathbf{x_k} \leq_P \mathbf{y}$, for some $k \in [m]$.

Proof. (i) Consider $\mathbf{x_k} \leq_{Adm} \mathbf{y}$. Then there exists $p \in [n]$ such that $M_i(\mathbf{x_k}) = M_i(\mathbf{y})$, for all $i \in [p-1]$ and $M_p(\mathbf{x_k}) < M_p(\mathbf{y})$, that is, $x_{k1} = y_1, \ldots, x_{k,p-1} = y_{p-1}$ and $x_{kp} < y_p$. Hence, for all $i \in \{1, \ldots, p-1\}$, we have that: $C_{\nu_i}(x_{1i}, \ldots, x_{mi}) = C_{\nu_i}(x_{1i}, \ldots, x_{k-1,i}, y_i, x_{k+1,i}, \ldots x_{mi})$ and $C_{\nu_p}(x_{1p}, \ldots, x_{mp}) <$

$C_{\nu_p}(x_{1p},\dots,x_{k-1,p},y_p,x_{k+1,p},\dots x_{mp})$, from which it follows that $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(\mathbf{x_1},\dots,\mathbf{x_m}) \leq_{Adm} \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(\mathbf{x_1},\dots,\mathbf{x_{k-1}},\mathbf{y},\mathbf{x_{k+1}},\dots\mathbf{x_m})$. *(ii)* The proof follows from that of item *(i)*, considering Proposition III.6. □

G. Comonotone additivity

First, we study the comonotone additivity with respect to the partial order $\leq_P$. We have the following result for $\tilde{\mathbf{M}}_{in}$-Choquet integrals.

Theorem IV.11. *Let $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}:(L^n)^m\to L^n$ be a vector $\tilde{\mathbf{M}}_{in}$-Choquet integral given by Definition III.4, where M_i are additive for all $i\in[n]$. Then, for all $\mathbf{x_1},\dots,\mathbf{x_m},\mathbf{y_1},\dots,\mathbf{y_m}\in L^n$ such that there exists a permutation $\pi:[m]\to[m]$ with $\mathbf{x}_{\pi(1)}\leq_P\dots\leq_P\mathbf{x}_{\pi(m)}$ and $\mathbf{y}_{\pi(1)}\leq_P\dots\leq_P\mathbf{y}_{\pi(m)}$, it holds that: $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x_1}+\mathbf{y_1},\dots,\mathbf{x_m}+\mathbf{y_m}) = \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{x_1},\dots,\mathbf{x_m}) + \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}(\mathbf{y_1},\dots,\mathbf{y_m})$*

Proof. By the additivity of M_i and the observation that, for all $i\in[n]$, the vectors $(M_i(\mathbf{x_1}),\dots,M_i(\mathbf{x_m}))$, $(M_i(\mathbf{y_1}),\dots,M_i(\mathbf{y_m}))$ are comonotone whenever $(\mathbf{x_1},\dots,\mathbf{x_m})$, $(\mathbf{y_1},\dots,\mathbf{y_m})$ are comonotone with respect to $\leq_P$, we have that:

$$
\begin{aligned}
&\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(\mathbf{x_1}+\mathbf{y_1},\dots,\mathbf{x_m}+\mathbf{y_m}) =\\
&\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(C_{\nu_1}(M_1(\mathbf{x_1}+\mathbf{y_1}),\dots,M_1(\mathbf{x_m}+\mathbf{y_m})),\dots,\\
&\qquad C_{\nu_n}(M_n(\mathbf{x_1}+\mathbf{y_1}),\dots,M_n(\mathbf{x_m}+\mathbf{y_m}))) =\\
&\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(C_{\nu_1}(M_1(\mathbf{x_1})+M_1(\mathbf{y_1}),\dots,M_1(\mathbf{x_m})+M_1(\mathbf{y_m})),\dots\\
&\qquad C_{\nu_n}(M_n(\mathbf{x_1})+M_n(\mathbf{y_1}),\dots,M_n(\mathbf{x_m}))+M_n(\mathbf{y_m}))) =\\
&\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(C_{\nu_1}(M_1(\mathbf{x_1}),\dots,M_1(\mathbf{x_m}))+\\
&\qquad C_{\nu_1}(M_1(\mathbf{y_1}),\dots,M_1(\mathbf{y_m})),\\
&\qquad C_{\nu_n}(M_n(\mathbf{x_1}),\dots,M_n(\mathbf{x_m}))+\\
&\qquad C_{\nu_n}(M_n(\mathbf{y_1}),\dots,M_n(\mathbf{y_m}))) =\\
&\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(\mathbf{x_1},\dots,\mathbf{x_m})+\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(\mathbf{y_1},\dots,\mathbf{y_m}).
\end{aligned}
$$

□

Next, an immediate consequence with respect to the additivity of $\tilde{\mathbf{M}}_{in}$-Choquet integral integrals is given.

Corollary IV.12. *Under the assumptions of Theorem IV.11, if ν_i are additive for all $i\in[n]$, then $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{in}}$ is additive.*

However, since the comonotonicity of the vectors $(\mathbf{x_1},\dots,\mathbf{x_m})$, $(\mathbf{y_1},\dots,\mathbf{y_m})$ does not imply the comonotonicity of the vectors $(\mathbf{x^1},\dots,\mathbf{x^n})$, $(\mathbf{y^1},\dots,\mathbf{y^n})$, we only have the following result about the additivity of $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{out}}}$.

Theorem IV.13. *Let $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}:(L^n)^m\to L^n$ be a $\tilde{\mathbf{M}}_{out}$-Choquet integral given by Definition III.4 where M_i are additive, for all $i\in[n]$. If ν_i are additive for all $i\in[n]$, then $\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{out}}$ is additive.*

Proof. By the additivity of M_i and ν_i, we have that:

$$
\begin{aligned}
&\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{out}}}(\mathbf{x_1}+\mathbf{y_1},\dots,\mathbf{x_m}+\mathbf{y_m}) =\\
&\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{out}}}(M_1(C_{\nu_1}(\mathbf{x^1}+\mathbf{y^1}),\dots,C_{\nu_n}(\mathbf{x^n}+\mathbf{y^n})),\dots,\\
&\qquad M_n(C_{\nu_1}(\mathbf{x^1}+\mathbf{y^1}),\dots,C_{\nu_n}(\mathbf{x_m}+\mathbf{y_m}))) =\\
&\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(M_1(C_{\nu_1}(\mathbf{x^1})+C_{\nu_1}(\mathbf{y^1}),\dots,C_{\nu_n}(\mathbf{x^n})+C_{\nu_n}(\mathbf{y^n})),\dots\\
&\qquad M_n(C_{\nu_1}(\mathbf{x^1})+C_{\nu_1}(\mathbf{y^1}),\dots,C_{\nu_n}(\mathbf{x^n}))+C_{\nu_n}(\mathbf{y^n})))\\
&\quad = \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{in}}}(M_1(C_{\nu_1}(\mathbf{x^1}),\dots,C_{\nu_n}(\mathbf{x^n}))+\\
&\qquad M_1(C_{\nu_1}(\mathbf{y^1}),\dots,C_{\nu_n}(\mathbf{y^n})),\dots,\\
&\qquad M_n(C_{\nu_1}(\mathbf{x^1}),\dots,C_{\nu_n}(\mathbf{x^n}))+\\
&\qquad M_n(C_{\nu_1}(\mathbf{y^1}),\dots,C_{\nu_n}(\mathbf{y^n})))\\
&\quad = \mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{out}}}(\mathbf{x_1},\dots,\mathbf{x_m})+\mathbf{C}_{\boldsymbol{\nu}}^{\tilde{\mathbf{M}}_{\text{out}}}(\mathbf{y_1},\dots,\mathbf{y_m}).
\end{aligned}
$$

□

Remark IV.14. *Since there is not a relation between the comonotonicity of the vectors $(M_i(\mathbf{x_1}),\dots,M_i(\mathbf{x_m}))$, $(M_i(\mathbf{y_1}),\dots,M_i(\mathbf{y_m}))$ and the comonotonicity of the vectors $(\mathbf{x_1},\dots,\mathbf{x_m})$, $(\mathbf{y_1},\dots,\mathbf{y_m})$ with respect to $\leq_{Adm}$, an $\tilde{\mathbf{M}}_{in}$-Choquet integral on L^n is not comonotone additive with respect to admissible orders $\leq_{Adm}$. The same holds for a $\tilde{\mathbf{M}}_{out}$-Choquet integral on L^n.*

H. Properties of the Vector Choquet Integral

Remark IV.15. *According to Proposition III.6 and the results in Section IV, a vector Choquet integral on L^n defined by Definition III.1, $\mathbf{C}_{\boldsymbol{\nu}}:(L^n)^m\to L^n$ is:*

- *symmetric (if ν_i is symmetric for each $i\in[n]$);*
- *satisfies the boundary conditions;*
- *idempotent;*
- *self-dual;*
- *shift-invariant;*
- *positively homogeneous;*
- *averaging with respect to $\leq_P$, and in the sense of Corollary IV.7, also with respect to $\leq_{Adm}$;*
- *monotone with respect to $\leq_P$, and monotone with respect to $\leq_{Lex}$ if C_{ν_i} are strictly increasing for all $i\in[n]$;*

V. VCI-LSTM: LSTM UNIT ARCHITECTURE MODIFICATION BASED ON THE VECTOR CHOQUET INTEGRAL

This section explains the Choquet integral-based LSTM (VCI-LSTM), that is, the introduction of the definitions set out in the previous sections in the architecture of this kind of recurrent neural networks.

The modification is based in two steps:

1) The normalization of sequential data vectors using the sigmoid function. Instead of applying the sigmoid function after aggregating the data, we first normalise the data to $[0,1]^H$ element-wise with the sigmoid function.
2) The replacement of the classical operator of the LSTM network (sum of vectors) by the aggregation of vectorial data using an n-dimensional aggregation function

$$\mathbf{M}:([0,1]^H)^3\to[0,1]^H.$$

This article has been accepted for publication in IEEE Transactions on Fuzzy Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2022.3222035

In this case $\mathbf{M}$ (Fig. 2) is the Vector Choquet Integral ($\mathbf{C}_\nu$, Definition III.1). The new equations that configure the VCI-LSTM performance generated by modifying the aggregation function are represented in Figure 2, where it is shown the modified part of Figure 1 (standard LSTM).

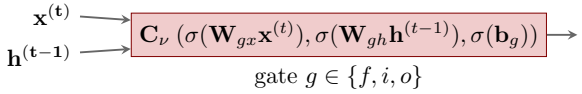


Fig. 2: VCI-LSTM parameter fusion.

In this work, two different VCI-LSTM are considered, depending on the type of fuzzy measure used for the Choquet-like integral:

- Fixed fuzzy measure-based VCI-LSTM. We adopt the power measure (Example II.4, Eq. (1)), fixing $q = 2$ in the exponent of the power fuzzy measure $\nu_2 : 2^{[3]} \to [0, 1]$. We use a superadditive fuzzy measure (setting $q = 2$ in the exponent of the power fuzzy measure), in which the elements have a negative correlation between the data and therefore analyse what is the behaviour of the integral based on such a measure in which complementarity between the data is assumed. However, as we show in Section VI, we also consider the possibility of a positive correlation between the data.
  Therefore, the fuzzy measures sequence used for the vector Choquet-like discrete integral is given by $\boldsymbol{\nu_2} = (\nu_2, \ldots, \nu_2)$. This aggregation operator is denoted by VCI_2.
- Parameter-learned fuzzy measure-based VCI-LSTM. In this case, the power measure (Example II.4, Eq. (1)) is used, but with $q \in (0, \infty)$ as a trainable parameter of the model. Then, in the same way as the recurrent neural network model learns weights matrices, it also learns this parameter. The used fuzzy measure is denoted by ν_q and the sequence of fuzzy measures is $\boldsymbol{\nu_q} = (\nu_q, \ldots, \nu_q)$. This aggregation operator is denoted by VCI_q.

To compare the LSTM (*Sum*) and the two types of VCI-LSTM (VCI_2 and VCI_q), we also use a statistical order, the maximum (*Max*), which is also an extreme concrete case of the Choquet integral.

VI. Experimental Study

The present section presents the experimental study done to test the architecture modification raised in Section V. The section is divided into three subsections. The first two ones correspond to the two completed experimental blocks: sequential image classification and text classification. In each of them, the datasets used, the specific architecture and the experimental results obtained are explained. The third one corresponds to the realized analysis about the fuzzy measure on the modelling context.

A. Experiment 1: Sequential image classification

Although the use of images in a recurrent neural network may be unusual due to the sequential dependence of the input information, in many recurrent network architectures images are used as benchmarks [39], [40]. This is because images offer a large amount of information in each pixel. This information in this case is interpreted as sequential information [41].

1) Experimental framework: In this experiment three standard image datasets are used. In Table I the most important attributes for each of them are displayed. All of them are balanced datasets.

TABLE I: First experiment dataset descriptions

Name	*Description*	*Train*	*Test*	*Dims*	*# Classes*
Fashion-MNIST [42]	clothing items	60.000	10.000	28×28	10
MNIST [43]	handwriting digits	60.000	10.000	28×28	10
EMNIST [44]	handwriting letters	88.800	14.800	28×28	27

Regarding the used neural architecture, as we can see in Figure 3, we set up an structure in which the images are taken as sequential data. In each time step $t \in \{1, \ldots, T\}$ a row of the image is taken as input data $\mathbf{x}^{(t)} \in [0, 1]^N$. In the case of these three datasets in particular, $T = N = 28$.

The used architecture consists in two layers. The first one, an LSTM unit (Section II-B) with $H = 128$ hidden size layer. Second, a dense layer, that connects the H nodes from the LSTM with N_C nodes of the dense layer, giving a probability value in $[0, 1]$ to each of them. If $(\varphi_1, \ldots, \varphi_{N_C})$ is the vector of probabilities extracted from the dense layer, being N_C the number of classes, it is classified in the class number corresponding to the maximum probability value of the softmax vector ($\mathbf{S} = (S_1, \ldots, S_{N_C})$). That is, $S_j = \arg\max \frac{\exp(\varphi_j)}{\sum_{k=1}^{N_C} \exp(\varphi_k)}$ for $j \in \{1, \ldots, N_C\}$.

In this experiment, for each dataset and for each aggregation, 10 independent runs of 40 epochs each were performed. The learning rate set for the experiment is $\alpha = 0.1$ and the optimization method used for learning has been the stochastic gradient descent (SGD) [45]. The optimization method used is the Cross Entropy Loss.

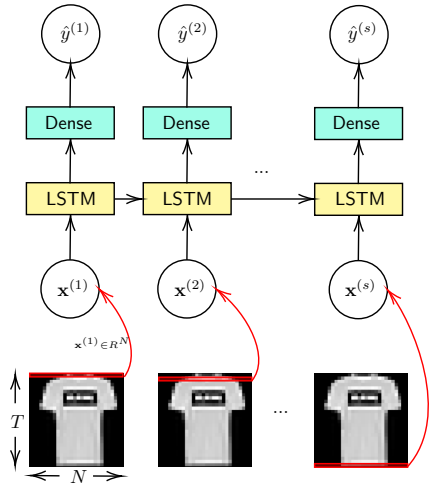


Fig. 3: Graphical representation of the used architecture.

2) Experimental results: We show the results obtained after calculating the arithmetic mean and standard deviation of 10

independent runs for each of the aggregation functions (Table II) and the summation. The result with the greatest average accuracy is highlighted in bold.

First, regarding the results obtained with a single aggregation function (Table II), in two of the three cases, the best average result is obtained when we aggregate the values using the vector Choquet integral, but when the exponent $q \in (0, \infty)$ it is learned by the recurrent neural network itself. This means that using the vector Choquet integral and learning this parameter, the algorithm adjusts better the weighting of the data as well as the interaction modeling between them.

The main difference we see in the case of the dataset *Fashion-MNIST*, where the data fusion with the vector Choquet integral improves 0.33 points in comparison with the classical form of data fusion in this architecture, the sum. In the case of the other datasets, in *MNIST* does not improve and *EMNIST* improves very little in comparison with the sum. The contrast on the results between the first and second and third datasets can also be seen in the obtained p-values after aplying statistical test (Table III). *Fashion-MNIST* dataset results with VCI-LSTM are statistically different in comparison with the standard one, but this is not the case for *EMNIST*. This difference between the results is justified with the different distribution of the pixels in the images of three datasets. Whereas *Fashion-MNIST* dataset has more distributed information, *MNIST* and *EMNIST* are datasets with less distributed information, where less correlation and interaction between them may be.

TABLE II: Different aggregation functions and summation mean accuracy for the first architecture

Aggr. Fn.	*Fashion-MNIST*	*MNIST*	*EMNIST*
Max	86.45 ± 0.69	98.88 ± 0.10	92.76 ± 0.10
VCI_2	87.22 ± 0.28	98.08 ± 0.30	91.65 ± 0.42
VCI_q	$\mathbf{89.33} \pm 0.17$	98.68 ± 0.10	$\mathbf{92.80} \pm 0.10$
Sum	89.00 ± 0.24	$\mathbf{98.90} \pm 0.08$	92.78 ± 0.13

TABLE III: Mann-Whitney U test p-values for best accuracy aggregation comparison against Sum

Aggr. Fn.	*Fashion-MNIST*	*MNIST*	*EMNIST*
Sum/VCI_q	.002	.025	.248

B. Experiment 2: Text classification - Sentiment analysis

Second, we use natural language processing datasets to evaluate the VCI-LSTM. In this case, we propose an architecture different from the previous case, since for language processing and text classification a deep recurrent network is necessary. For it, we will chain more than one LSTM units.

TABLE IV: Second experiment dataset descriptions

Name	*Description*	*Train*	*Test*	*# Classes*
IMDb [25]	film reviews	25.000	25.000	2
TREC [46]	questions classif.	5.500	500	6
sms_spam [26]	spam filtering	3.344	2.240	2

1) Experimental framework: As we can see in Table IV, the datasets are different from each other, TREC has little more than 20% of data volume of *IMDb* dataset, and also has the triple of classes. Also, *IMDb* dataset is balanced, *TREC* is not enough balanced for all the classes and *sms_spam* dataset So, we see the operation of the architecture in very different datasets.

Before entering in the architecture used in this second experiment, we performed a pre-processing of the data, based on the "tokenization" and the construction of the vocabulary. Tokenization consists of separating words using the spaces between them as separators. Building vocabulary is a larger process. First, a positive integer P is fixed (in this experiment we fix it at $25,000$). Subsequently, the P most frequent words are encoded with numbers from the set $\{1, \ldots, P\}$. Words that are not among the most frequent P are encoded with the value $P+1$, which means unknown. To keep the same length in each sentence within a batch, the value $P+2$ is used representing the *padding*. In this way, each example of the data set consists of a vector $\mathbf{x} \in \{1, \ldots P+2\}^s$.

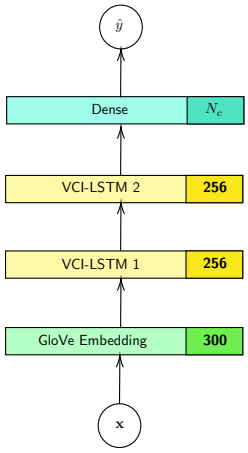


Fig. 4: Graphical representation of the architecture used in the second experiment

Regarding the used architecture (Figure 4), we establish a structure in which at each timestep $t \in \{1, \ldots, T\}$ an element of the input vector $\mathbf{x}^{(t)} \in \{1, \ldots, P+2\}$ is taken, which represents a word.

The architecture consists in four layers: embedding layer, double stacked LSTM units (which includes two layers) and a dense layer:

- *Embedding* [47]. It is the way to reduce the input space (in this case, a sentence) to a smaller dimension. It consists of a simple search matrix that encodes the input words in vectors, taking values depending on how the words are related in the input texts. In this sense, words with close representations have a greater relationship. The relationship between words is defined by subtraction of vectors. For example, if we subtract the vector that represents *Ireland* from the vector that represents *Dublin* and add the vector that represents *England*, we would obtain a vector close to the one that represents *London* [48]. In this case, for obtaining vector representation

of words we use the GloVe (Global Vectors for Word Representation) [49] unsupervised learning algorithm, with 6B-sized tokens and 50 or 300 dimensions vectors, depending on the dataset.

- *Double stacked LSTM units*. Two LSTM memory units, joined together; each of them has the size of $H = 256$ nodes in their hidden layer. The first memory unit, LSTM 1, connects 50 nodes from the *embedding* with 256 from the first cell. The second, LSTM 2, joins the 256 nodes of LSTM 1 with this second cell.
- *Dense layer*. Finally, a fully connected or dense layer is used. It connects the 256 nodes of the LSTM unit with 10 nodes, assigning a probability value of $[0, 1]$ to each of them. The operation of this layer is the same as in the previous experiment.

In this experiment, for each dataset and for each function combination, 10 independent runs of 15 epochs each were performed (25 epochs for the last one). The learning rate set for the experiment is $\alpha = 1 \times 10^{-3}$ and the optimization method used for learning has been the Adam algorithm [50].

2) Experimental results: Next, we show the arithmetic mean and standard deviation of the results obtained for 10 independent runs in this second experiment. As in the previous experiment, for each data set, the aggregation with the highest mean accuracy is highlighted in bold. First, we will analyze the results obtained assigning the same function for both stacked LSTM units. As we can see in Table V, when we use the same unit for both double stacked units, VCI-LSTM (with LSTM-learned fuzzy measure parameter) generally obtains better results. Regarding to *TREC* set, there is a difference of 2.87 points in comparison with classical unit. Bigger average difference is observed in *sms_spam* one, where is a gaining of 6.83 points, which also justifies that VCI-LSTM regularizes better the learning in comparison with the LSTM. This means that, in several runs with another fusion operators, it is more difficult to reach a global minimum of the loss function, and recurrently it ends up sinking to a local minimum, or stagnating at a saddle point. Regarding to *IMDb* set, although the mean accuracy value is lower than the baseline, the difference in comparison with summation is smaller, but the number of epochs that the algorithm needs to reach the convergence is smaller (Figure 5). Also, we have evaluated our results applying Mann-Whitney U statistical test [51], where we have upholded that all of them are statistically significant.

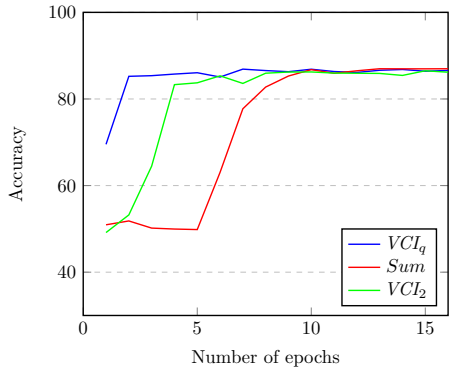
TABLE V: Different aggregation functions and summation mean accuracy for double stacked units with same function for each one

Aggr. Fn.	*IMDb*	*TREC*	*sms_spam*
Max	85.69 ± 0.68	81.80 ± 4.02	94.83 ± 4.64
VCI_2	77.47 ± 2.02	79.80 ± 2.50	94.17 ± 4.07
VCI_q	85.93 ± 0.29	$\mathbf{83.38} \pm 1.54$	$\mathbf{97.94} \pm 0.68$
Sum	$\mathbf{86.04} \pm 0.19$	80.51 ± 4.61	91.11 ± 4.45

TABLE VI: Mann-Whitney U test p-values for best accuracy aggregation comparison against Sum

Aggr. Fn.	*IMDb*	*TREC*	*sms_spam*
Sum/VCI_q	.158	.035	$< .001$

Fig. 5: Comparison of the evolution of the mean accuracy for the validation set, for VCI-LSTM (VCI_q and VCI_2) and LSTM (Sum) for *IMDb* dataset.



In order to try all possible combinations, the second part of this experiment consists in using different units (different functions) in the double stacked units. In this sense, for each dataset, we run all unit combinations between the three previous aggregation functions and the summation. As in the previous cases, we take the arithmetic mean and standard deviation of 10 independent runs (Table VII). The results for same aggregations (diagonals) are also shown for comparing.

TABLE VII: Different aggregation functions and summation mean accuracy for double stacked units with different functions for each one

		LSTM 1				
		Aggr	*Max*	VCI_2	VCI_q	*Sum*
LSTM 2	*IMDb*	*Max*	85.69 ± 0.68	86.28 ± 0.31	85.72 ± 0.33	$\mathbf{86.83} \pm 0.44$
		VCI_2	85.98 ± 0.49	77.47 ± 2.02	85.15 ± 0.19	75.01 ± 9.31
		VCI_q	86.02 ± 0.46	$\mathbf{86.49} \pm 0.29$	85.93 ± 0.29	86.31 ± 0.39
		Sum	85.30 ± 0.55	82.31 ± 3.54	86.04 ± 0.09	86.04 ± 0.04
	TREC	*Max*	81.80 ± 4.20	75.91 ± 6.32	79.73 ± 4.01	82.52 ± 1.52
		VCI_2	81.03 ± 2.85	79.80 ± 2.50	79.62 ± 4.28	79.45 ± 3.66
		VCI_q	$\mathbf{83.22} \pm 2.24$	81.07 ± 2.18	$\mathbf{83.38} \pm 1.54$	82.75 ± 1.60
		Sum	82.17 ± 2.24	77.05 ± 4.05	82.64 ± 2.23	80.51 ± 4.61
	sms_spam	*Max*	94.83 ± 4.46	95.31 ± 4.33	97.26 ± 0.63	91.75 ± 4.90
		VCI_2	93.96 ± 3.96	94.17 ± 4.07	94.26 ± 3.76	92.13 ± 4.92
		VCI_q	96.60 ± 0.78	94.03 ± 4.61	$\mathbf{97.94} \pm 0.68$	94.82 ± 3.81
		Sum	95.14 ± 4.52	88.49 ± 3.66	$\mathbf{97.34} \pm 1.32$	91.11 ± 4.41

Regarding to *IMDb* set, the best result is obtained with Sum-Max combination. Nevertheless, 4 of the 5 best results are when any kind of VCI-LSTM is used. Although the best mean accuracy only gains 0.69 percentage points over the sum, three best results give statistical significance p-values in comparison with summation (Table VIII).

For *TREC* dataset, over all combinations, best results are obtained when parameter-learned VCI-LSTM is used. Concretely, the three best results are VCI-LSTM-based Choquet integral and maximum combinations. The best result gains more than 3 points against the sum. After evaluating with the p-value (Table VIII), although the best three results gain some points over the sum, there are only statistically significant differences for the two best ones.

Finally, in the *sms_spam* dataset, we can observe a big difference in mean accuracy between the cases where self-

TABLE VIII: Three best combinations of aggregation functions in LSTM and Mann-Whitney U test p-values for best accuracy aggregation comparison against *Sum*

IMDB	*LSTM 1*	*LSTM 2*	Mean acc.	p-value
Best comb.	*Sum*	*Max*	86.83 ± 0.44	$< .001$
2nd Best comb.	VCI_2	VCI_q	86.49 ± 0.29	$< .001$
3rd Best comb.	*Sum*	VCI_q	86.31 ± 0.39	.008
TREC	*LSTM 1*	*LSTM 2*	Mean acc.	p-value
Best comb.	VCI_q	VCI_q	83.38 ± 1.54	.042
2nd Best comb.	*Max*	VCI_q	83.22 ± 2.24	.059
3rd Best comb.	*Sum*	VCI_q	82.75 ± 1.60	.153
sms_spam	*LSTM 1*	*LSTM 2*	Mean acc.	p-value
Best comb.	VCI_q	VCI_q	97.94 ± 0.68	$< .001$
2nd Best comb.	VCI_q	*Sum*	97.34 ± 1.32	$< .001$
3rd Best comb.	*Max*	VCI_q	97.26 ± 0.63	$< .001$

learnt VCI-LSTM unit is used and otherwise. All the VCI-LSTM-based model results are statistically significant (Table VIII shows for the three best ones). In this case we have also observed that when we use VCI-LSTM with VCI_q the model does not sinks on saddle points, in comparison with other aggregation-based units. This also explains the difference on standard deviations between different models.

C. Analysis of the fuzzy measure on the modelling context

In this experimentation we have used the Vector Choquet Integral based on a fixed fuzzy measure and a model-learned fuzzy measure. Concretely, the fixed one is a superadditive fuzzy measure in order to evaluate how the Choquet Integral behaves with a negative correlation among data. Regarding the model-learned fuzzy measure, we can study which fuzzy measure obtains better results based on gradient descent learning. On the modeling context, we can observe that depending on the fuzzy measure, the correlation between data is different, so this adaptative method of VCI can improve some model results. To do so, we have extracted the arithmetic mean and standard deviation of the value of the parameter $q > 0$ for all experiments in which we have learned the fuzzy VCI measure in all LSTM cells.

TABLE IX: Values of q parameter for each dataset in the first experiment

$q > 0$	*Fashion-MNIST*	*MNIST*	*EMNIST*
VCI-LSTM	0.3575 ± 0.0180	0.0796 ± 0.0174	0.1947 ± 0.0171

TABLE X: Values of q parameter for each dataset in the second experiment

$q > 0$	*IMDb*	*TREC*	*sms_spam*
VCI-LSTM 1	0.1145 ± 0.0208	0.0261 ± 0.0113	0.0358 ± 0.0148
VCI-LSTM 2	0.0988 ± 0.0221	0.0380 ± 0.0269	0.0417 ± 0.0120

On Tables IX and X we can observe that the exponent of the average parameter of the power fuzzy measure is a small number, always less than 0.4, and with the exception of one dataset, less than 0.2. The standard deviation of the parameter shows little difference between cases independently, where it has no relation to the average value. This implies that in the case where the mean value is similar, the width of the parameter range is larger, while when the mean value is larger, the width is smaller.

With the results about the average values of the parameter of the fuzzy measure learned by the neural network, we can observe that the fuzzy measure is always subadditive, i.e., that there is a large positive correlation between the data.. That is, two of the three vectors that are added in this process are almost redundant. This means that the maximum value in each case contains information that is also contained in the other vector components. However, we can observe that although the maximum function by itself obtains good results, in most cases it needs to be complemented with other values. That is, in terms of averaging aggregation functions, the best performing functions are close to the maximum.

In addition, we have checked in each of the cases what the maximum value is, on the understanding that if it were always the same, there would be no point in training some parameters. However, the maximum value has a similar distribution among the values to be aggregated, which makes it necessary to use this method of aggregating information.

VII. Conclusion

In this paper we have presented the Vector Choquet Integral, as a vectorial extension of the classical Choquet-like discrete integral, and we have studied some of its properties. Also we have applied it in order to introduce the VCI-LSTM unit: a modification of a LSTM recurrent neural network based on the replacement of the summation by the VCI. We have tested this unit in different scenarios to solve different problems.

Regarding the future work, in the theoretical side, intend to consider new forms of n-dimensional information fusion. In the applied side, we want to use these modifications which are capable of n-dimensional modeling, such as, Generalized Extended Bonferroni Means in newer and more complex architectures, in order to optimize the information fusion methods on recurrent neural network models based on fuzzy measures, in order to improve the performing of more concrete problems dealing with sequential information.

Acknowledgments

The Department of Migration Policies and Justice of the Government of Navarre has promoted this project together with the Public University of Navarra, with the collaboration of the public company TRACASA. Based on the conviction that artificial intelligence provides valuable elements of judgment for the development of public policies, the aim is to offer the Judiciary, the Prosecutor's Office, the Bar and social work tools so that they can carry out their work in an increasingly efficient and effective way, both in relation to people who at some point in their lives have problems with justice, as well as those who are victims of a crime. In the context of today's digital society, public agencies must ensure that criminal enforcement policies are based on data, and that they do so in a permanently updated way, improving the efficiency of

these policies with each new case. Grant PID2019-108392GB-I00 funded by MCIN/AEI/10.13039/501100011033, by CNPq (Proc. 301618/2019-4), FAPERGS (Proc. 19/2551-0001660), by Tracasa Instrumental and the Immigration Policy and Justice Department of the Government of Navarre and by the project VEGA 1/0267/21.

REFERENCES

[1] G. Beliakov, H. Bustince, and T. Calvo, "A practical guide to averaging functions," in *Studies in Fuzziness and Soft Computing*, 2016.

[2] M. Grabisch, J.-L. Marichal, R. Mesiar, and E. Pap, *Aggregation Functions*, ser. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2009.

[3] D. Paternain, J. Fernandez, H. Bustince, R. Mesiar, and G. Beliakov, "Construction of image reduction operators using averaging aggregation functions," *Fuzzy Sets and Systems*, vol. 261, pp. 87–111, 2015, theme: Aggregation operators. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165011414001122

[4] G. Lucca, J. A. Sanz, G. P. Dimuro, B. Bedregal, M. J. Asiain, M. Elkano, and H. Bustince, "Cc-integrals: Choquet-like copula-based aggregation functions and its application in fuzzy rule-based classification systems," *Knowledge-Based Systems*, vol. 119, pp. 32–43, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705116304889

[5] C. A. Dias, J. C. S. Bueno, E. N. Borges, S. Botelho, G. Dimuro, G. Lucca, J. Fernández, H. Bustince, and P. Drews, "Using the choquet integral in the pooling layer in deep learning networks," in *NAFIPS*, 2018.

[6] I. Rodriguez-Martinez, J. Lafuente, R. H. Santiago, G. Pereira Dimuro, F. Herrera, and H. Bustince, "Replacing pooling functions in Convolutional Neural Networks by linear combinations of increasing functions," *Neural Networks (submitted)*.

[7] M. Grabisch and C. Labreuche, "A decade of application of the choquet and sugeno integrals in multi-criteria decision aid," *Annals of Operations Research*, vol. 175, pp. 247–286, 2008.

[8] G. Lucca, J. A. Sanz, G. P. Dimuro, E. N. Borges, H. Santos, and H. Bustince, "Analyzing the performance of different fuzzy measures with generalizations of the choquet integral in classification problems," in *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2019, pp. 1–6.

[9] T. Murofushi, M. Sugeno, and M. Machida, "Non-monotonic fuzzy measures and the choquet integral," *Fuzzy Sets and Systems*, vol. 64, no. 1, pp. 73–86, 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0165011494900086

[10] G. Choquet, "Theory of capacities," *Annales de l'Institut Fourier*, vol. 5, pp. 131–295, 1954. [Online]. Available: http://www.numdam.org/articles/10.5802/aif.53/

[11] G. P. Dimuro, J. Fernández, B. Bedregal, R. Mesiar, J. A. Sanz, G. Lucca, and H. Bustince, "The state-of-art of the generalizations of the Choquet integral: From aggregation and pre-aggregation to ordered directionally monotone functions," *Information Fusion*, vol. 57, pp. 27 – 43, 2020.

[12] G. Lucca, J. Sanz, G. Pereira Dimuro, B. Bedregal, R. Mesiar, A. Kolesárová, and H. Bustince Sola, "Pre-aggregation functions: construction and an application," *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 2, pp. 260–272, April 2016.

[13] H. Bustince, R. Mesiar, J. Fernandez, M. Galar, D. Paternain, A. Altalhi, G. Dimuro, B. Bedregal, and Z. Takáč, "d-choquet integrals: Choquet integrals based on dissimilarities," *Fuzzy Sets and Systems*, vol. 414, pp. 1–27, 2021, aggregation Functions. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165011420301032

[14] G. P. Dimuro, G. Lucca, J. A. Sanz, H. Bustince, and B. Bedregal, "Cmin-integral: A choquet-like aggregation function based on the minimum t-norm for applications to fuzzy rule-based classification systems," in *Aggregation Functions in Theory and in Practice*, V. Torra, R. Mesiar, and B. D. Baets, Eds. Cham: Springer International Publishing, 2018, pp. 83–95.

[15] G. Lucca, J. Antonio Sanz, G. P. Dimuro, B. Bedregal, H. Bustince, and R. Mesiar, "Cf-integrals: A new family of pre-aggregation functions with application to fuzzy rule-based classification systems," *Information Sciences*, vol. 435, pp. 94–110, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025517311507

[16] G. Lucca, G. P. Dimuro, J. Fernandez, H. Bustince, B. Bedregal, and J. A. Sanz, "Improving the performance of fuzzy rule-based classification systems based on a nonaveraging generalization of CC-integrals named $C_{F_1F_2}$-integrals," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 1, pp. 124–134, Jan 2019.

[17] R. Lourenzutti, R. A. Krohling, and M. Z. Reformat, "Choquet based topsis and todim for dynamic and heterogeneous decision making with criteria interaction," *Information Sciences*, vol. 408, pp. 41 – 69, 2017.

[18] J. C. Wieczynski, G. P. Dimuro, E. N. Borges, H. S. Santos, G. Lucca, R. Lourenzutti, and H. Bustince, "Generalizing the GMC-RTOPSIS method using CT-integral pre-aggregation functions," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2020, pp. 1–8.

[19] P. Liu, S.-M. Chen, and G. Tang, "Multicriteria decision making with incomplete weights based on 2-d uncertain linguistic choquet integral operators," *IEEE Transactions on Cybernetics*, vol. 51, no. 4, pp. 1860–1874, 2021.

[20] J. Fumanal-Idocin, Y.-K. Wang, C.-T. Lin, J. Fernández, J. A. Sanz, and H. Bustince, "Motor-imagery-based brain-computer interface using signal derivation and aggregation functions," *IEEE Transactions on Cybernetics*, pp. 1–12, 2021.

[21] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

[22] A. Graves, A. rahman Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649, 2013.

[23] K. Choi, G. Fazekas, and M. Sandler, "Text-based lstm networks for automatic music composition," *ArXiv*, vol. abs/1604.05358, 2016.

[24] J. Du, C.-M. Vong, and C. L. P. Chen, "Novel efficient rnn and lstm-like architectures: Recurrent and gated broad learning systems and their applications for text classification," *IEEE Transactions on Cybernetics*, vol. 51, no. 3, pp. 1586–1597, 2021.

[25] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 142–150. [Online]. Available: https://www.aclweb.org/anthology/P11-1015

[26] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the study of sms spam filtering: New collection and results," in *Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG'11)*, 2011.

[27] Y. Meng, A. Rumshisky, and A. Romanov, "Temporal information extraction for question answering using syntactic dependencies in an lstm-based architecture," *CoRR*, vol. abs/1703.05851, 2017. [Online]. Available: http://arxiv.org/abs/1703.05851

[28] Q. dao-er-ji Ren, Y. Su, and W. Liu, "Research on the lstm mongolian and chinese machine translation based on morpheme encoding," *Neural Computing and Applications*, vol. 32, pp. 41–49, 2018.

[29] H. Bustince, M. Galar, B. Bedregal, A. Kolesárová, and R. Mesiar, "A new approach to interval-valued choquet integrals and the problem of ordering in interval-valued fuzzy set applications," *IEEE Transactions on Fuzzy Systems*, vol. 21, no. 6, pp. 1150–1162, 2013.

[30] D. Paternain, L. De Miguel, G. Ochoa, I. Lizasoain, R. Mesiar, and H. Bustince, "The interval-valued choquet integral based on admissible permutations," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 8, pp. 1638–1647, 2019.

[31] G. Beliakov, H. B. Sola, and T. C. Snchez, *A Practical Guide to Averaging Functions*, 1st ed. Springer Publishing Company, Incorporated, 2015.

[32] L. De Miguel, M. Sesma-Sara, M. Elkano, M. Asiain, and H. Bustince, "An algorithm for group decision making using n-dimensional fuzzy sets, admissible orders and owa operators," *Information Fusion*, vol. 37, pp. 126–131, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253517300684

[33] B. Bedregal, G. Beliakov, H. Bustince, T. Calvo, R. Mesiar, and D. Paternain, "A class of fuzzy multisets with a fixed number of memberships," *Information Sciences*, vol. 189, pp. 1–17, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025511006256

[34] T. Milfont, I. Mezzomo, B. Bedregal, E. Mansilla, and H. Bustince, "Aggregation functions on n-dimensional ordered vectors equipped with an admissible order and an application in multi-criteria group decision-making," *International Journal of Approximate Reasoning*, vol. 137, pp. 34–50, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888613X2100089X

[35] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," in *A Field Guide to Dynamical Recurrent Neural Networks*, S. C. Kremer and J. F. Kolen, Eds. IEEE Press, 2001.

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] T. Pradhan, P. Kumar, and S. Pal, "Claver: An integrated framework of convolutional layer, bidirectional lstm with attention mechanism based scholarly venue recommendation," *Information Sciences*, vol. 559, pp. 212–235, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025520311890

[38] F. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with lstm," in *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, vol. 2, 1999, pp. 850–855 vol.2.

[39] T. M. Breuel, "Benchmarking of LSTM Networks," *arXiv e-prints*, p. arXiv:1508.02774, Aug. 2015.

[40] O. Kaziha and T. Bonny, "A comparison of quantized convolutional and lstm recurrent neural network models using mnist," in *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, 2019, pp. 1–5.

[41] Q. V. Le, N. Jaitly, and G. E. Hinton, "A simple way to initialize recurrent networks of rectified linear units," *CoRR*, vol. abs/1504.00941, 2015.

[42] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *ArXiv*, vol. abs/1708.07747, 2017.

[43] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[44] G. Cohen, S. Afshar, J. Tapson, and A. V. Schaik, "Emnist: Extending mnist to handwritten letters," *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926, 2017.

[45] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. NIPS Foundation (http://books.nips.cc), 2008, pp. 161–168. [Online]. Available: http://leon.bottou.org/papers/bottou-bousquet-2008

[46] X. Li and D. Roth, "Learning question classifiers," in *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, ser. COLING '02. USA: Association for Computational Linguistics, 2002, p. 1–7. [Online]. Available: https://doi.org/10.3115/1072228.1072378

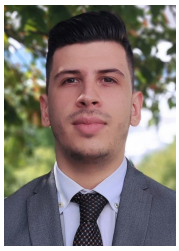
[47] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR*, 2013.

[48] P. Pérez-Núñez, O. Luaces, A. Bahamonde, and J. Díez, "Improving recommender systems by encoding items and user profiles considering the order in their consumption history," *Prog. Artif. Intell.*, vol. 9, no. 1, pp. 67–75, 2020. [Online]. Available: https://doi.org/10.1007/s13748-019-00199-7

[49] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[51] H. B. Mann and D. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18, pp. 50–60, 1947.



Mikel Ferrero-Jaurrieta received his B.Sc. degree in Computer Science from the Public University of Navarre (Pamplona, Spain) in 2018. He received his M.Sc. in Mathematical Modelling from the Public University of Navarre, University of Zaragoza and Univesity of the Basque Country in 2020. He is currently a PhD candidate at the Department of Statistics, Computer Science and Mathematics, Public University of Navarre in the field of Computer Science and Artificial Intelligence researching about recurrent neural network architectures and its applications in the Justice Administration. His research interests include Sequence Models, Natural Language Processing, Fuzzy Logic and Aggregation functions.

Zdenko Takáč received the Graduate degree in teaching mathematics and physics from the Faculty of Mathematics and Physics, Bratislava, Slovakia, in 1998 and the Ph.D. degree in teaching mathematics with the thesis Analysis of mathematical proof from Pavol Jozef Šafárik University, Košice, Slovakia, in 2007.

Since 1999, he has been a member of the Department of Mathematics, Faculty of Education, Catholic University in Ružomberok, Ružomberok, Slovakia, and since 2010, a member of the Department of Mathematics, Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Bratislava, Slovakia. His research interests include uncertainty modeling, fuzzy sets and fuzzy logic, aggregation operators, interval-valued, and type-2 fuzzy sets.

Javier Fernández (Member, IEEE) received the M.Sc. degree in mathematics from the University of Zaragoza, Zaragoza, Spain, in 1999, and the Ph.D. degree in mathematics from the University of the Basque Country, Leioa, Spain, in 2003. He is currently an Associate Lecturer with the Department of Statistics, Computer Science and Mathematics, Public University of Navarre, Pamplona, Spain. He has authored or coauthored approximately 45 original articles and is involved with teaching artificial intelligence and computational mathematics for students of the computer sciences and data science. His research interests include fuzzy techniques for image processing, fuzzy sets theory, interval-valued fuzzy sets theory, aggregation functions, fuzzy measures, deep learning, stability, evolution equation, and unique continuation

Ľubomíra Horanská received the Graduate degree in mathematics, and the Ph.D. degree in geometry and topology, from the Faculty of Mathematics and Physics, Comenius University, Bratislava, Slovakia, in 1993 and 2001, respectively. Since 1993, she is with the Institute of Information Engineering, Automation and Mathematics, Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Slovakia. Her research interests include uncertainty modeling, aggregation functions, with a special stress to copulas, measures and integrals and algebraic and differetial topology.

Graçaliz Pereira Dimuro Graçaliz Pereira Dimuro received the M.Sc. and Ph.D. degrees 1991 and 1998, respectively, both from the Instituto de Informatica of Universidade Federal do Rio Grande do Sul, Brazil. In 2015, she had a Pos-Doc post-doctorate grant from the Science Without Borders Program from the Brazilian Research Funding Agency CNPq, to join GIARA research group at Universidad Publica de Navarra (UPNA), and, in 2017, she had a talent grant at the Institute of Smart Cities of Universidad Publica de Navarra, Spain. Currently, she is a full professor with Universidade Federal do Rio Grande, Brazil, a Researcher level 1 of CNPq, and a visitant professor with UPNA.

Susana Montes received the M.Sc degree in mathematics, option statistics and operational research, from the University of Valladolid, Valladolid, Spain, in 1993, and the Ph.D. (cum laude) degree from the University of Oviedo, Gijón, Spain, in 1998.

She is currently a Full Professor with the Department of Statistics and Operational Research, University of Oviedo, where she is the Leader of the research group UNIMODE. She has several publications in international journals and communications in international conferences, and she is participating in several national and international projects at the moment, some of them led by her. Dr. Montes received the Best Mathematics Ph.D. Thesis Award from the University of Oviedo. She is currently the Secretary of EUSFLAT and Vice-President of IFSA.

Irene Díaz received the M.Sc degree in mathematics, option applied mathematics and computation, from the University of Oviedo, Oviedo, Spain, in 1995, and the Ph.D. (cum laude) degree from the University Carlos III of Madrid, Spain, in 2001.

She is currently a Full Professor of Computer Science and Artificial Intelligence at the Department of Computer Science of the University of Oviedo, where she belongs to the research group UNIMODE. She has several publications in international journals and communications in international conferences, and she is participating in national and international projects at the moment, some of them led by her.

Humberto Bustince received the Graduate degree in physics from the University of Salamanca in 1983 and Ph.D. in mathematics from the Public University of Navarra, Pamplona, Spain, in 1994. He is a Full Professor of Computer Science and Artificial Intelligence in the Public University of Navarra, Pamplona, Spain where he is the main researcher of the Artificial Intelligence and Approximate Reasoning group, whose main research lines are both theoretical (aggregation functions, information and comparison measures, fuzzy sets, and extensions) and applied (image processing, classification, machine learning, data mining, and big data). He has led 11 I+D public-funded research projects, at a national and at a regional level. He is currently the main researcher of a project in the Spanish Science Program and of a scientific network about fuzzy logic and soft computing. He has been in charge of research projects collaborating with private companies. He has taken part in two international research projects. He has authored more than 210 works, according to Web of Science, in conferences and international journals, with around 110 of them in journals of the first quartile of JCR. Moreover, five of these works are also among the highly cited papers of the last ten years, according to Science Essential Indicators of Web of Science. Dr. Bustince is the Editor-in-Chief of the online magazine Mathware & Soft Computing of the European Society for Fuzzy Logic and technologies and of the Axioms journal. He is an Associated Editor of the IEEE Transactions on Fuzzy Systems Journal and a member of the editorial board of the Journals Fuzzy Sets and Systems, Information Fusion, International Journal of Computational Intelligence Systems and Journal of Intelligent & Fuzzy Systems. He is the coauthor of a monography about averaging functions and coeditor of several books. He has organized some renowned international conferences such as EUROFUSE 2009 and AGOP. Honorary Professor at the University of Nottingham, National Spanish Computer Science Award in 2019 and EUSFLAT Excellence Research Award in 2019.