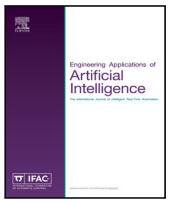
Engineering Applications of Artificial Intelligence 173 (2026) 114442

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Walsh sequences as direct Error Correcting Output Code dichotomies for multiclass problems

Lorena Álvarez-Pérez [a,*], Álvaro Callejas-Ramos [b]

[a] CEIEC Research Institute. Universidad Francisco de Vitoria, Ctra. Pozuelo-Majadahonda KM 1,800, 28223 Pozuelo de Alarcón (Madrid), Spain
[b] Department of Signal Theory and Communications, Avda. de la Universidad, 30, 28911, Leganés, (Madrid), Spain

ARTICLE INFO

Keywords:
Bregman divergence
Multiclass classification
Posterior probability
Walsh sequences

ABSTRACT

Solving multiclass classification problems using Error Correcting Output Code (ECOC) schemes offers flexibility and robustness, but traditional approaches based on Rademacher sequences suffer from an exponential increase in dichotomies with the number of classes. This paper introduces two main contributions. First, we propose the use of Walsh sequences for directly constructing ECOC dichotomies, enabling multiclass problems with C classes to be addressed using only $C - 1$ classifiers, while preserving high pairwise Hamming distances between class codes ($C/2$). Second, we present a principled method to estimate posterior class probabilities from the outputs of discriminative classifiers trained with Bregman divergences. This provides an alternative to Hamming decoding, improves confidence estimation, and avoids matrix inversion due to the orthogonality of Walsh codes. Experimental results on five multiclass datasets—including Fashion-MNIST, Letter Recognition, and Vowel—demonstrate consistent improvements in classification accuracy (e.g., a 5% error reduction on the Letter dataset) and computational efficiency over baseline methods such as OvR and ECOC-Rademacher. Performance was evaluated using standard classification accuracy and confusion matrices, validating the advantages of our approach in both predictive performance and scalability.

1. Introduction

Multiclass classification is a fundamental problem in machine learning, where the goal is to assign a given input to one of several possible categories. It plays a critical role in numerous applications such as cancer detection (Yuan et al., 2018), image recognition (LeCun et al., 1989), EEG analysis (Siuly and Li, 2014), sentiment analysis (Bouazizi and Ohtsuki, 2019), and text categorization (Chaitanya et al., 2017). Recent work has also applied multiclass classification to real-world clinical settings, for instance in predicting knee prosthesis types using deep learning and demographic data (Yoon et al., 2022). While binary classification has been extensively studied, multiclass problems often present additional challenges due to the need to simultaneously discriminate among multiple classes.

Standard discriminative approaches using a softmax output layer face two important limitations. First, selecting from among multiple outputs is inherently more complex than binary decisions. Second, class imbalance complicates training, since a single rebalancing scheme may be insufficient to address disparities across all classes (Rifkin and Klautau, 2004; Krawczyk, 2016). To address these issues, decomposing the multiclass problem into multiple binary problems—so-called dichotomies—has gained significant attention. Popular strategies include One-vs-Rest (OvR) and One-vs-One (OvO) (Rifkin and Klautau, 2004; Duan et al., 2003), but both suffer from limitations: OvO grows quadratically with the number of classes, while OvR can introduce strong class imbalance.

Error Correcting Output Code (ECOC) methods offer an elegant framework to organize and aggregate binary classifiers for multiclass tasks (Dietterich and Bakiri, 1995). Among them, Rademacher-based ECOC designs are attractive for their direct construction and high inter-class Hamming distances. Recent alternatives have explored more scalable ECOC constructions through discrete optimization and graph-coloring formulations (Gupta and Amin, 2022), although they typically require additional problem-specific analysis or search procedures. Other recent work focuses on optimizing the minimum Hamming distance between class codes, such as the Min–Max ECOC strategy proposed in Szűcs (2023), which aims to improve separability at the cost of more complex code design. In parallel, other approaches have proposed generalizing ECOC beyond binary codes using N-ary coding schemes to further enrich the ensemble diversity, such as in Nguyen et al. (2021), although such methods typically involve more complex decoding and training pipelines. Additionally, ECOC has been recently adopted in

* Corresponding author.
E-mail address: lorena.alvarez@ufv.es (L. Álvarez-Pérez).

https://doi.org/10.1016/j.engappai.2026.114442
Received 6 April 2024; Received in revised form 30 May 2025; Accepted 6 March 2026
Available online 10 March 2026

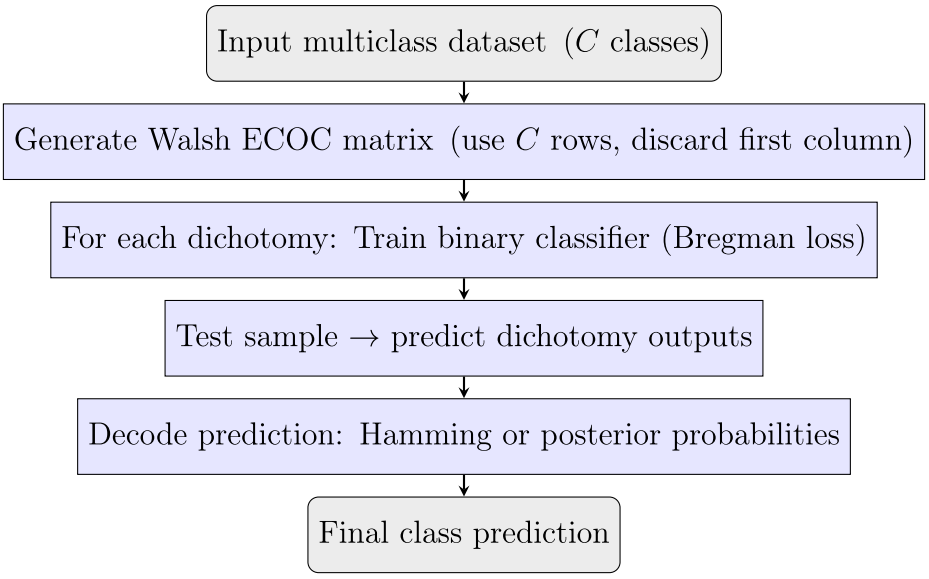


Fig. 1. Flowchart of the proposed Walsh-ECOC classification pipeline.

deep learning frameworks to enhance robustness against weight-level errors and adversarial perturbations, using concepts from the Neural Tangent Kernel (Yu et al., 2024).

However, they require an exponentially increasing number of classifiers as the number of classes grows, making them impractical in large-scale scenarios. These limitations motivate the search for multiclass strategies that are both efficient and robust, without requiring complex problem-specific designs or excessive computational resources. In particular, we are interested in solutions that reduce the number of binary classifiers, preserve discriminative power, and offer reliable estimates of prediction confidence.

The main contributions of this work are summarized as follows:

- We propose the use of Walsh sequences for constructing ECOC dichotomies, which reduces the number of required binary classifiers to $C-1$ for C classes. These codes maintain a high inter-class Hamming distance, and scale efficiently to large numbers of classes.
- We introduce a principled procedure to estimate posterior class probabilities using discriminative classifiers trained with a Bregman divergence. This method outperforms traditional Hamming-based decoding in terms of both accuracy and confidence estimation.
- We show that Walsh-based ECOC matrices are orthogonal, which eliminates the need for matrix inversion when recovering class probabilities, leading to a significant computational advantage.
- We empirically validate our approach on five multiclass datasets, demonstrating improved classification accuracy and scalability over standard techniques such as OvR, Rademacher-based ECOC, and state-of-the-art classifiers including MLP, Random Forests, and XGBoost.

The rest of this paper is structured as follows. Section 2 presents the design of Walsh-based ECOC matrices and the proposed probability estimation framework. Section 3 describes the datasets, experimental setup, and results. Finally, Section 4 summarizes the main findings and discusses possible directions for future research.

2. Methodology and materials

This section describes the proposed methodology, which combines a compact and orthogonal ECOC encoding using Walsh sequences with a principled decoding strategy based on posterior probability estimation. The approach is designed to be scalable to problems with many classes, while maintaining high discriminative capacity and interpretability.

We first review how multiclass classification can be addressed through ECOC decomposition, and then introduce the use of Walsh sequences as a code design strategy. We also describe how binary classifiers trained with Bregman divergences allow us to recover probabilistic outputs for final decision-making.

Fig. 1 illustrates the overall workflow of the proposed method, from Walsh-based ECOC matrix generation to final class prediction using either Hamming decoding or posterior probabilities.

Table 1
The Rademacher sequences of length 8.

Dichotomies							
1	1	1	1	1	1	1	1
1	1	1	1	−1	−1	−1	−1
1	1	−1	−1	1	1	−1	−1
1	−1	1	−1	1	−1	1	−1

2.1. ECOC matrix for solving multiclass problems

Table 1 shows the 4 Rademacher sequences of length 8 (Dietterich and Bakiri, 1995). Their construction is easy: The first row is all ones. The second row consists of 2^{C-2} ones followed by 2^{C-2} ones preceded by the sign minus, where C represents the number of classes ($C = 4$ classes, for the example at hand). The third row consists of 2^{C-3} ones, followed by 2^{C-3} minus ones, followed by 2^{C-3} ones and 2^{C-3} minus ones. In the ith row, there are alternating runs of 2^{C-i} ones and minus ones. The sequences are odd and orthogonal.

To be used as ECOC for multiclass problems, each row represents a class. The first column is omitted, and the rest indicate how to construct the dichotomies: Each dichotomic class is composed of the problem classes with the same value of the Rademacher sequences. For example, the third dichotomic problem (fourth column in Table 1) consists of deciding between classes 1 and 2 and classes 3 and 4. The traditional mode of solving the multiclass problem for a given sample is to construct the sequence of all the winning dichotomic class ± 1 indicators, and to select as the winning class that whose row has the minimal Hamming distance with respect to this syndrome. For example, if the dichotomic winners code is 1 1 1 −1 1 1 1, the sample is assigned to the first class, whose code is at a Hamming distance 1 from that

Table 2
The Walsh sequences of length 8.

Dichotomies							
1	1	1	1	1	1	1	1
1	1	1	1	−1	−1	−1	−1
1	1	−1	−1	−1	−1	1	1
1	1	−1	−1	1	1	−1	−1
1	−1	−1	1	1	−1	−1	1
1	−1	1	−1	−1	1	−1	1
1	−1	−1	−1	−1	−1	−1	−1
1	−1	1	−1	1	−1	1	−1

winners code. Note that the Hamming distance between class codes is 4 ($L/2$, for a code length L).

It is easy to see that the all columns (but the first) of Rademacher functions include all the possible dichotomies: In the case at hand, all the one vs. three and the two vs. two problems. This means that some imbalance is introduced for some dichotomies, but this is not important for the practical margin of application of Rademacher ECOC procedures: A length L means to deal with $L-1$ dichotomies – and the subsequent design of $L-1$ classifiers–, but it permits to deal with only $1+\log_2 L$ classes.

2.2. Walsh sequences and their use as ECOC for solving multiclass problems

Walsh sequences are a complete set of orthogonal binary sequences composed of +1 and −1 elements. For a sequence length $L = 2^n$, where n is a non-negative integer, there exist L Walsh sequences. These sequences can be generated by taking all possible products of Rademacher functions (or Hadamard matrix rows), ordered by the number of zero crossings (sequency order) (Walsh, 1923). Table 2 presents the Walsh sequences of length 8, which correspond to the rows of a 8×8 Walsh matrix.

Let $\mathbf{W}_L$ be the $L \times L$ Walsh matrix of orthogonal rows, where each row corresponds to a Walsh sequence of length L. Given a multiclass problem with C classes, we can define an ECOC coding matrix $\mathbf{M}$ by selecting any C rows from $\mathbf{W}_L$ (typically $L = C$). To ensure a square and orthogonal matrix, we often take the first C rows. The first column (all ones) is typically discarded, as it does not contribute to meaningful dichotomies.

Mathematical proof of the Hamming distance between Walsh codes:

Since the Walsh matrix $\mathbf{W}_L$ is orthogonal, it satisfies the relation:

$$\mathbf{W}_L \mathbf{W}_L^{\top} = L\mathbf{I}_L,$$

where $\mathbf{I}_L$ is the $L \times L$ identity matrix.

This implies that for any two distinct rows (codewords) $\mathbf{c}_i$ and $\mathbf{c}_j$,

$$\mathbf{c}_i \cdot \mathbf{c}_j = \sum_{k=1}^{L} c_{ik} c_{jk} = 0, \quad \text{for } i \neq j,$$

where each element $c_{ik} \in \{-1, +1\}$.

The inner product relates to the Hamming distance $d_H(\mathbf{c}_i, \mathbf{c}_j)$ by the formula:

$$\mathbf{c}_i \cdot \mathbf{c}_j = L - 2d_H(\mathbf{c}_i, \mathbf{c}_j).$$

Setting the inner product to zero gives:

$$0 = L - 2d_H(\mathbf{c}_i, \mathbf{c}_j) \implies d_H(\mathbf{c}_i, \mathbf{c}_j) = \frac{L}{2}.$$

Because we typically choose $L = C$, the number of classes, the minimum Hamming distance between any two Walsh codewords is exactly $C/2$.

Each of the remaining columns defines a binary partition (dichotomy) over the classes: all classes with a +1 value are grouped as the positive class, and those with −1 as the negative. For C classes, this results in $C-1$ dichotomies –each corresponding to a column in the reduced Walsh matrix.

The main benefits of using Walsh sequences for ECOC are:

- Orthogonality: Any pair of rows in $\mathbf{W}_L$ has zero inner product, ensuring maximum separation between class codes (i.e., maximal Hamming distance).
- Completeness: The Walsh basis spans the entire space of binary ±1 sequences of length L, providing flexibility to encode any number of classes (up to L).
- Balanced dichotomies: Many Walsh functions have an equal number of +1 and −1 values, which leads to balanced binary classification problems when classes are evenly distributed.
- Efficiency: Only $C-1$ classifiers are required for C classes, in contrast to the exponential growth in Rademacher ECOC designs. This greatly improves scalability.

Formally, let $\mathbf{c}_i \in \{-1, +1\}^{C-1}$ be the codeword assigned to class i. For an input sample, each trained binary classifier predicts a value $\hat{y}_j \in \{-1, +1\}$ for the jth dichotomy. The class prediction can then be made by selecting the class whose codeword $\mathbf{c}_i$ is closest (e.g., via Hamming distance or estimated posterior) to the predicted code $\hat{\mathbf{y}}$.

When the number of classes is not a power of 2, Walsh sequences can still be used by selecting the first C rows of a longer Walsh matrix (e.g., $L = 2^{\lceil \log_2 C \rceil}$). In this case, some imbalance may appear in a few dichotomies, but the impact is often negligible in practice.

2.3. Bregman divergences and posterior probability estimate with discriminative classifiers

This contribution is completed by introducing a different method from using Hamming distances to solve ECOC multiclass problems. It is based on using a Bregman divergence (Bregman, 1967) as surrogate cost to solve the dichotomies with discriminative machines, such as Multi-Layer Perceptrons (MLPs) or Deep Neural Networks, and even ensembles of them. This is important because these machines offer a very high expressive capacity –to establish input–output correspondence–, and, therefore, are good candidates for classification purposes. Moreover, the use of a Bregman divergence as surrogate cost to train them permits to estimate the posterior probability of each dichotomic class (Cid-Sueiro et al., 1999; Cid-Sueiro and Figueiras-Vidal, 2001; Benítez-Buenache et al., 2019, 2021). These estimates serve to obtain estimates of the posterior probabilities of the problem classes. As it is known, a Bregman divergence $B(o,t)$ is a function that accomplishes:

$$\frac{\partial B(o,t)}{\partial o} = -g(o)(t-o), \;\; g(o) < 0 \tag{1}$$

If it is used as surrogate cost to train a discriminative machine with output o to get a target t, the theoretical objective is to minimize Expression (2):

$$\iint_{\mathbf{x}\;t} B(t,o)p(\mathbf{x},t)\,d\mathbf{x}\,dt = \int_{\mathbf{x}} \left[\int_t B(t,o)p(t|\mathbf{x})\,dt\right] p(\mathbf{x})\,d\mathbf{x} \tag{2}$$

where $\mathbf{x}$ is the machine input and $p(\mathbf{x},t)$ the joint probability density function of $\mathbf{x}$ and t. Obviously, the minimization is carried out by minimizing the inner integral in the second term of (1), and thus leads to:

$$\frac{\partial}{\partial o}\int_t B(t,o)p(t|\mathbf{x})\,dt = -g(o)\int_t (t-o)p(t|\mathbf{x})\,dt = 0 \tag{3}$$

whose solution is obtained by making 0 the integral, from which:

$$o(\mathbf{x}) = \mathrm{E}\{t|\mathbf{x}\} \tag{4}$$

For a binary problem with $t = \pm 1$,

$$\begin{aligned}\mathrm{E}\{t|\mathbf{x}\} &= 1 \cdot \Pr(t=1|\mathbf{x}) - 1 \cdot \Pr(t=-1|\mathbf{x}) \\ &= 2 \cdot \Pr(t=1|\mathbf{x}) - 1\end{aligned} \tag{5}$$

Since the machine has a limited expressive power and it is trained with a finite number of samples, in practice $\Pr(t=1|\mathbf{x})$ in (5) will be an estimate. So, this leads to Expression (6).

$$\widehat{\Pr}(t=1|\mathbf{x}) = \frac{1+o(\mathbf{x})}{2} \tag{6}$$

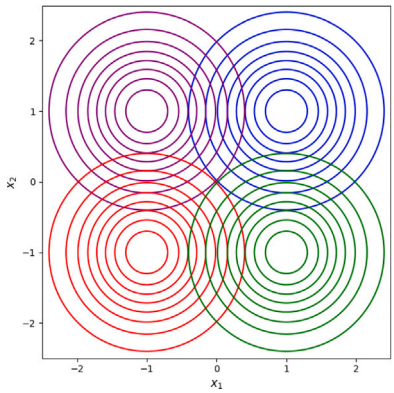


Fig. 2. Illustrative representation of the four 2-dimensional Gaussian distributions of the synthetic dataset.

3. Illustrative experiments

The aim of the results shown in this section is to illustrate the benefits of applying Walsh codes to solve multiclass classification problems with a relatively large number of classes.

3.1. Databases

The presentation of the results will be limited to a few appropriately selected balanced multiclass databases: A synthetic dataset and four real classification problems: Fashion-MNIST (Xiao et al., 2017), Vowel dataset (Renals and Rohwer, 1989), Letter Recognition dataset (Frey and Slate, 1991) and Handwritten numerals dataset using morphological features (Breukelen et al., 1998).

Fashion-MNIST has been taken from Kaggle (Goldbloom and Hammer, 2010); both vowel and handwritten numerals datasets have been obtained from the OpenML Repository of Machine Learning Databases (Vanschoren et al., 2013), and finally, Letter Recognition dataset has been collected from the UCI[1] Machine Learning Repository (Frank and Asuncion, 2010). These databases have been chosen aiming at providing enough diversity in relation to the number of samples, features and properties. They will be briefly described in the next paragraphs.

3.1.1. Synthetic dataset

The synthetic dataset is a simple toy dataset that illustrates the behavior of the Walsh codes in an easy way. It consists of four 2-dimensional Gaussian distributions where each distribution has the same variance but different mean. The variance is $\mathbf{V} = v\mathbf{I}$, where $\mathbf{I}$ is the 2×2 identity matrix and the value of v is $v = 0.45$. As shown in Fig. 2, this value has been set empirically in order to allow samples of each class to interfere with each other to the point where it is neither a perfect classification problem nor an unfeasible classification task.

As it can be observed in the aforementioned figure, the mean vector of each distribution is $\mathbf{m}_1 = [1,1]^\intercal$, $\mathbf{m}_2 = [-1,1]^\intercal$, $\mathbf{m}_3 = [-1,-1]^\intercal$ and $\mathbf{m}_4 = [1,-1]^\intercal$, respectively.

The number of samples belonging to each of the distributions has been set to be 2000, as depicted in Fig. 3(a). For illustrative purposes, it is shown the boundaries of the theoretical MAP test for this classification problem in Fig. 3(b).

3.1.2. Fashion-MNIST

Fashion-MNIST (henceforth, FMNIST) is a set of Zalando's article images, where each image is a 28×28 grayscale sample (784 pixels in total) (Nocentini et al., 2022). Each pixel has a single integer value associated with it (ranging from 0 to 255) indicating how the light or dark is the pixel (the higher the number, the darker the pixel). This dataset consists of a set of 70 000 images (the training set has 60 000 images and the test set has 10 000 images) and each image is associated with a label from 10 classes.

[1] UCI stands for University of California, Irvine.

Table 3
Characteristics of the benchmark problems.

Dataset	Samples	Features	Classes	Distribution
Synthetic	8000	2	4	2000 each
FMNIST	70 000	784	10	7000 each
Vowel	990	10	11	90 each
Letter	20 000	16	26	Varies per class
Mfeat	2000	6	10	200 each

3.1.3. Vowel dataset

This classification problem consists of fifteen individual speakers, each saying the eleven steady state vowels of the British English (Determding, 1990). The vowels are indexed by integers $0-10$ (there are a total of $C = 11$ classes). Each speaker thus yielded six frames of speech from eleven vowels. For each utterance, there are ten floating-point input values.

The speech audio signals were low pass filtered at 4.7 kHz and then digitized to 12 bits with a 10 kHz sampling rate. Twelfth order linear predictive analysis was carried out on six 512 sample Hamming windowed segments from the steady part of the vowel. The reflection coefficients were used to calculate 10 log area parameters, giving a 10 dimensional input space.

3.1.4. Letter recognition dataset

This classification problem was created by David J. Slate in 1991 (Slate, 1991). It consists in identifying each of the 26 capital letters in the English alphabet, which are represented by a large number of black and white rectangular pixels. The images that contained the characters were based on 20 different fonts, and each letter within these 20 fonts was randomly distorted to produce a file of 20 000 unique images. Each image was then converted into a feature vector with 16 numerical values (basically, it is composed of statistical moments and edge counts) which were scaled to fit into a range of integer values from 0 through 15. The dataset consists of 20 000 samples corresponding to 26 capital letters. The number of samples per class is not exactly uniform, but approximately balanced.

3.1.5. Handwritten numerals dataset

This dataset (henceforth, Mfeat) is comprised of morphological features of handwritten numerals $(0-9)$ extracted from a collection of Dutch utility maps. 200 samples per class (for a total of 2000 samples) were digitized (or converted) into binary images. Each of these digits is represented by 6 morphological features (Duin, 1998).

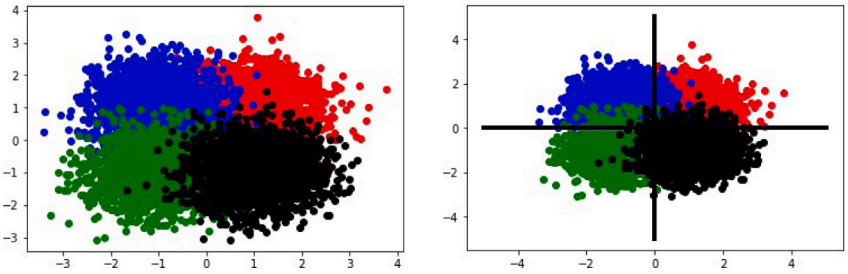
To sum up, Table 3 summarizes the main characteristics of the aforementioned classification problems.

Aiming at carrying out the experiments, since there is not always a pre-established partition of training and test sets (except for FMNIST dataset), the data have been split into two subsets: $75-25\%$ for training and test, respectively (maintaining the proportion of the classes in each subset).

All the results presented in this paper correspond to the test set. It is shown the mean and standard deviation of 10 repetitions, each one with different random initialization.

3.2. Numerical results

The classifier used in each dichotomic problem has been an MLP with one hidden layer (Bishop, 1995). It has been shown that an MLP with only one hidden layer can approximate any continuous function (Funahashi, 1989). In each binary problem, the number of



(a) Graphically representation of the samples belonging to each distribution.

(b) Theoretical MAP classifier (the black lines represent the boundaries).

Fig. 3. Synthetic dataset used in the experiments.

hidden neurons has been obtained by means of cross validation (with a number of folds equal to 5).

Aiming at comparing the results obtained, there have been explored the use of three single classifiers: an MLP-based classifier, the simple but highly effective k-NN algorithm (Cover and Hart, 1967) and a decision tree classifier (easy and widely used classification technique) (Safavian and Landgrebe, 1991). The number of neurons in the hidden layer, the number of neighbors and the maximum number of leaf nodes, respectively, have been determined by using 5-fold cross validation. It has been also studied the use of Random Forest (RF) (Breiman, 2001) and eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016) as a combination or ensemble of multiple decision trees. For all models involving hyperparameters, we performed a grid search with 5-fold cross-validation on the training set to select optimal values. For the MLP, the number of hidden layers was varied between 1 and 3, with each hidden layer size tested at {32, 64, 128, 256} neurons. Activation functions considered were ReLU and tanh. For ensemble methods such as RF and XGBoost, the number of estimators ranged from 50 to 300, maximum tree depth was varied from 3 to 15, and for XGBoost the learning rate was chosen from {0.01, 0.1, 0.2}. The model hyperparameters were selected based on the highest average accuracy obtained on the validation folds. This procedure ensured robust tuning and fair comparison across methods.

The results obtained are also compared with two traditional techniques used in the literature for binarizing multiclass problems: ECOC Rademacher method and OvR, both of them described in Section 2. Please note that the results obtained with ECOC technique is only calculated for the synthetic dataset since it is the only one that accomplishes that the number of classes is between 3 and 7 and, therefore, there exists an exhaustive algorithm to construct the columns of the ECOC matrix.

Table 4 outlines the overall error rates for the datasets explored and the classification algorithms considered. It can be observed the positive effects of using Walsh sequences for solving balanced multiclass problems since it reduces the overall error rate in all the datasets. It is important to highlight the results obtained for the Letter Recognition database (with $C = 26$ classes) in which the error rate is drastically reduced approximately 5% when compared with an MLP-based classifier.

Despite this remarkable improvement and just regarding the above-mentioned error rates, the reader may wonder whether the approach is worthwhile or not. Aiming at answering this question, Table 5 shows the confusion matrices for the FMNIST dataset in which, at a first sight, the overall error rate is not drastically reduced when using Walsh sequences. Having a look at Table 5, it can be noticed that the accuracy rate for the class that achieved the worst performance with an MLP-based classifier or the OvR technique (that is, C_7), has been noticeably improved from 67.7–67.6% to 74.7% when using Walsh sequences.

In order to complete this Section, Table 6 depicts the confusion matrices for the Vowel dataset. In the same line of reasoning, it can be observed that the accuracy rate for all the classes has been improved. It is worth mentioning the considerably increase achieved for C_6.

Table 4
Overall error rates (mean ± standard deviation in %) for the benchmark problems.

	Datasets				
	Synthetic	FMNIST	Vowel	Letter	Mfeat
MLP	13.1 ± 0.6	11.6 ± 0.2	4.8 ± 1.9	7.5 ± 0.4	26.7 ± 0.7
k-NN	13.1 ± 0.7	14.7 ± 0.2	4.9 ± 1.9	5.0 ± 0.1	27.6 ± 1.2
DT	13.3 ± 0.4	18.0 ± 0.2	16.1 ± 1.4	12.7 ± 0.3	27.3 ± 1.8
RF	13.0 ± 0.5	11.5 ± 0.2	3.1 ± 1.3	3.6 ± 0.2	27.3 ± 1.8
XGBoost	14.3 ± 0.5	14.9 ± 0.3	9.5 ± 2.1	3.9 ± 0.2	27.9 ± 2.2
ECOC	12.9 ± 0.8	–	–	–	–
OvR	13.0 ± 0.6	11.4 ± 0.2	4.1 ± 1.7	3.7 ± 0.4	27.5 ± 2.9
Walsh	**12.6 ± 0.5**	**11.1 ± 0.2**	**2.8 ± 1.3**	**2.3 ± 0.2**	**26.1 ± 1.3**

Please note that the confusion matrices for the rest of datasets explored in this work have not been included here due to space limitations.

3.3. Discussion

There are several key aspects that make the proposed methodology appealing for multiclass classification. First, the Walsh-based ECOC design enables a highly compact coding scheme that requires only $C-1$ classifiers for C classes. This drastically reduces training and inference complexity compared to other ECOC methods such as Rademacher codes, which grow exponentially with the number of classes. Second, due to the orthogonality of Walsh sequences, each pair of class codewords maintains maximum Hamming distance (i.e., $C/2$), which directly contributes to higher separability among classes and improved robustness to classification errors.

While we do not claim formal optimality of Hamming distance across all possible coding schemes, it is worth noting that the Walsh matrix guarantees a strong and uniform inter-class separation. The orthogonality property ensures that all codewords are equally distant from one another in Hamming space, which is one of the key desirable traits in ECOC design. Therefore, even without global optimality, Walsh sequences offer a well-structured and theoretically grounded compromise between separability and code compactness.

Another significant advantage lies in the proposed decoding strategy. By training the base classifiers using a Bregman divergence, it is possible to recover posterior probabilities for each dichotomy, and from those, estimate the posterior probability of each class. This probabilistic decoding scheme provides richer information than standard Hamming decoding—particularly valuable when confidence or uncertainty plays a role in downstream tasks or decision-making.

While Eq. (4) derives posterior probabilities under the assumption of unbiased estimates from discriminative classifiers trained with Bregman divergences, practical factors such as finite sample sizes, limited model capacity, and optimization heuristics may lead to deviations from these assumptions. Despite these limitations, our empirical

Table 5
Confusion matrices and overall error rates for FMNIST database with: (a) A 784-500-10 MLP, (b) OvR method and (c) a WALSH ensemble.

(a)

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
C_1	**83.7**	0.2	1.7	2.5	0.6	0.1	10.2	0	1.0	0
C_2	0.3	**97.2**	0.3	1.4	0.3	0	0.4	0	0.1	0
C_3	1.5	0.1	**82.0**	1.0	9.0	0	6.0	0	0.4	0
C_4	2.6	0.9	1	**89.3**	3.0	0	2.9	0	0.3	0
C_5	0.4	0.2	7.9	2.6	**83.2**	0	5.2	0	0.5	0
C_6	0.1	0	0.1	0	0	**94.7**	0	3.0	0.5	1.6
C_7	12.5	0.3	8.5	2.4	7.3	0.1	**67.7**	0	1.2	0
C_8	0	0	0	0	0	1.7	0	**95.1**	0.2	3.0
C_9	0.5	0.1	0.7	0.6	0.6	0.5	1.0	0.3	**95.6**	0.1
C_{10}	0	0	0	0	0	1.1	0	3.1	0.1	**95.7**

Overall error rate: 11.6 ± 0.2

(b)

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
C_1	**84.0**	0.3	1.6	2.7	0.5	0.2	9.9	0	0.7	0.1
C_2	0.1	**97.8**	0.2	1.3	0.2	0	0.3	0	0.1	0
C_3	1.7	0.1	**81.1**	1.2	9.2	0.1	6.3	0	0.3	0
C_4	2.6	0.8	0.8	**90.2**	3.0	0	2.2	0	0.3	0.1
C_5	0.3	0.3	7.9	3.0	**82.7**	0	5.3	0	0.3	0
C_6	0.1	0	0.1	0.1	0	**95.5**	0	2.7	0.2	1.3
C_7	12.3	0.2	8.7	2.7	7.3	0.1	**67.6**	0	1.0	0.1
C_8	0	0	0	0	0	1.7	0	**95.1**	0.2	3.0
C_9	0.5	0.1	0.5	0.5	0.4	0.4	0.8	0.3	**96.4**	0.1
C_{10}	0	0	0	0	0	0.9	0	3.1	0.1	**95.8**

Overall error rate: 11.4% ± 0.2%

(c)

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
C_1	**83.3**	0.2	1.4	2.3	0.2	0	12.1	0	0.5	0
C_2	0.3	**97.9**	0.2	0.8	0.3	0	0.4	0	0.1	0
C_3	1.5	0.2	**81.3**	0.7	7.5	0	8.4	0	0.4	0
C_4	2.6	1.5	1.1	**88.5**	3.1	0.1	2.6	0	0.5	0
C_5	0.4	0.2	8.3	2.8	**81.6**	0	6.2	0	0.5	0
C_6	0.2	0	0.2	0.2	0.2	**95.9**	0.2	1.6	0.4	1.1
C_7	9.6	0.3	6.4	2.1	5.8	0.1	**74.7**	0	0.9	0.1
C_8	0.1	0	0.1	0.1	0.1	2.4	0.2	**93.9**	0.3	2.8
C_9	0.9	0.1	0.6	0.3	0.5	0.3	1.4	0.1	**95.8**	0
C_{10}	0.2	0	0.1	0.1	0.2	1.4	0.2	2.8	0.1	**94.9**

Overall error rate: 11.1% ± 0.2%

Table 6
Confusion matrices and overall error rates for Vowel database with: (a) A 12-242-11 MLP, (b) OvR method and (c) a WALSH ensemble.

(a)

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}
C_1	**99.5**	0	0	0	0	0	0	0	0	0	0.5
C_2	0.8	**97.5**	1.7	0	0	0	0	0	0	0	0
C_3	0	1.4	**97.7**	0	0	0	0	0	0	0	0.9
C_4	0	0	1.3	**95.0**	0	3.7	0	0	0	0	0
C_5	0	0	0	0	**91.0**	7.3	1.3	0	0	0.4	0
C_6	0	0	1.0	1.0	5.2	**89.2**	0	0	0.5	0	3.1
C_7	0	0	0	0	5.6	0	**91.0**	3.0	0.4	0	0
C_8	0	0	0	0	0	0	0.4	**99.1**	0.5	0	0
C_9	0	0	0	0	0	0.5	0	2.4	**90.7**	2.5	3.9
C_{10}	0	0	0	0	0	0	0	0	0	**100**	0
C_{11}	0	0.5	0	0	0	2.8	0	0	0.9	0.5	**95.4**

Overall error rate: 4.8% ± 1.9%

(b)

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}
C_1	**99.5**	0	0	0	0	0	0	0	0	0.5	0
C_2	0.8	**96.8**	0.8	0	0	0	0	0	0	0.4	1.2
C_3	0.4	1.3	**98.3**	0	0	0	0	0	0	0	0
C_4	0	0	1.0	**96.2**	0.9	1.9	0	0	0	0	0
C_5	0	0	0	0	**97.3**	1.8	0.9	0	0	0	0
C_6	0	0	0	1.7	6.2	**88.4**	0	0	0.4	0	3.3
C_7	0	0	0	0	4.2	0.4	**93.5**	1.9	0	0	0
C_8	0	0	0	0	0	0	0	**99.1**	0.9	0	0
C_9	0	0.4	0	0	0	0	0	0.4	**93.1**	3.3	2.8
C_{10}	0	0	0	0	0	0	0	0.9	0.9	**98.2**	0
C_{11}	0.4	0.4	0.4	0	0	2.6	0	0	1.3	0	**94.9**

Overall error rate: 4.1% ± 1.7%

(c)

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}
C_1	**99.5**	0.5	0	0	0	0	0	0	0	0	0
C_2	1.2	**96.7**	1.7	0	0.4	0	0	0	0	0	0
C_3	0	0.9	**98.3**	0.8	0	0	0	0	0	0	0
C_4	0	0	0.5	**98.2**	0.4	0.9	0	0	0	0	0
C_5	0	0	0	0	**97.3**	2.2	0.5	0	0	0	0
C_6	0	0.4	0	2.2	1.8	**93.0**	0	0	0	0	2.6
C_7	0	0	0	0	4.6	0	**92.9**	2.5	0	0	0
C_8	0	0	0	0	0	0	0	**100**	0	0	0
C_9	0	0	0	0	0	0.9	0	0.4	**95.7**	2.6	0.4
C_{10}	0	0	0	0	0	0	0	0	0	**99.6**	0.4
C_{11}	0	0.4	0	0	0	0	0	0	0.5	0	**99.1**

Overall error rate: 2.8% ± 1.2%

results on diverse datasets indicate that the proposed posterior estimation method generally yields well-calibrated and improved predictions compared to traditional Hamming decoding. The use of flexible base classifiers like multilayer perceptrons (MLPs) further aids in approximating the true posterior distributions. Future work will focus on providing formal robustness guarantees and investigating theoretical conditions under which these approximations hold.

It is also worth highlighting the computational benefit of using Walsh sequences. Because the code matrix is orthogonal, there is no need to perform matrix inversion when recovering class probabilities from dichotomy outputs. This makes the approach scalable and computationally efficient, especially for large-scale problems with many classes.

We acknowledge that the datasets used in our evaluation are mostly balanced. This was an intentional decision to ensure that the impact of the ECOC structure and decoding strategy could be analyzed in isolation. Nevertheless, preliminary results (such as those obtained from the Letter dataset) suggest that the method remains stable under mild imbalance. Future work will explore imbalance scenarios more systematically and investigate strategies to integrate class-weighting into the decoding process.

As with any ECOC-based decomposition method, our approach assumes that the multiclass problem can be effectively projected into a set of binary classification tasks where each induced dichotomy exhibits meaningful class separability. This assumption is common across One-vs-Rest, One-vs-One, and other ECOC variants, but may not hold in cases where class distributions are highly overlapping or non-linearly intertwined. Although the use of expressive classifiers like MLPs helps capture complex boundaries, the performance of any coding scheme—Walsh-based or otherwise—ultimately depends on the underlying structure of the data. We acknowledge this as a limitation of the method's generalizability and suggest further exploration in domains with more challenging class geometries.

While orthogonality offers computational and structural advantages—such as uniform Hamming distances and efficient decoding—it does not, by itself, guarantee improved classification performance in all scenarios. The effectiveness of the Walsh ECOC design still depends on how well the induced dichotomies align with the actual separability of the classes in the feature space. In some cases, certain binary partitions may group together classes that are not easily distinguishable, leading to less reliable classifiers for those dichotomies. This is a well-known limitation of fixed ECOC coding strategies and highlights the importance of matching code design with the data distribution when possible.

It is also important to clarify that our approach does not aim to theoretically prove the universal superiority of Walsh sequences over

all other ECOC coding strategies. Instead, we position Walsh-based designs as a principled and efficient alternative, whose orthogonality and compactness offer clear advantages in terms of classifier scalability and robustness. The performance gains observed in our experiments are empirical and task-specific, and we see this work as a contribution to the repertoire of practical and theoretically grounded ECOC schemes, rather than as a claim of absolute optimality.

Finally, we note that this work does not aim to replace or outperform advanced data-driven ECOC techniques, such as those based on clustering, mutual information, or discrete optimization. These approaches typically rely on problem-specific analysis and search procedures to construct highly customized codebooks. In contrast, our goal is to propose a direct and scalable coding scheme with strong mathematical properties—orthogonality, completeness, and compactness—without the need for code tuning. While this makes the Walsh-based method attractive in general-purpose settings or resource-constrained environments, we recognize that adaptive methods may yield superior performance in tailored applications. Comparative studies with such approaches constitute an important avenue for future research.

We also acknowledge that all experimental datasets used in this study are either balanced or only mildly imbalanced. This choice was intentional to allow a clean evaluation of the proposed coding and decoding mechanisms without introducing confounding effects. However, in scenarios with extreme class imbalance, the performance of any ECOC framework—including Walsh-based designs—can be impacted due to the degradation in the accuracy of the underlying binary classifiers. Although the Walsh structure offers robustness in terms of orthogonality and uniform code distribution, it does not inherently compensate for imbalanced data distributions. Addressing such cases would require additional mechanisms such as class weighting, data resampling, or decoding adjustments that account for class priors. We consider this an important direction for future research and plan to explore it systematically in subsequent work.

Moreover, we note that this study did not include formal statistical significance testing to accompany the reported performance differences across methods. Although we provide mean error rates and standard deviations over repeated trials to indicate consistency, we acknowledge that such metrics are not substitutes for statistical validation. Including significance testing, such as Wilcoxon signed-rank or Friedman tests, would offer stronger evidence regarding the reliability of the observed improvements. We consider this a valuable enhancement for future iterations of the work.

4. Conclusions

The experimental work carried out and presented in this paper supports the idea of applying Walsh sequences aiming at designing ECOC dichotomies. The benefits of Walsh sequences schemes are clear. They avoid the moderate increase of the number of dichotomies with the number of classes that appears with Rademacher ECOC designs (without losing a high Hamming distance between codewords or class codes).

Additionally, it is proposed in this paper a principled formulation which permits to get estimates of the posterior class probabilities. These estimates are calculated from the estimates of the posterior probabilities for all the dichotomies that come from the outputs of the classification machines that solve each dichotomy if these machines are trained with a Bregman divergence. This latter entails an additional advantage. The process to estimate posterior dichotomy estimates requires, in general, the (pseudo) inversion of an extended code matrix. However, Walsh sequences are orthogonal. So, it is not necessary to compute the mentioned inversion, which drastically reduces the computation load (especially, when the number of classes is relatively high).

The series of experiments carried out over 5 databases clearly show the increase of the performance of the resulting classification schemes with respect to those coming from a single classifier or an ensemble of classifiers and even more, when compared with traditional binarization schemes for solving multiclass problems. The conclusion is significant because it indicates the highest performance of the methodology proposed and, what is of key importance, they can be applied for multiclass problems with an arbitrary large number of classes.

Extending this procedure to OvO technique is also a research direction that can provide practical improvements. Another line of research to be explored is to address the imbalance that is introduced in some dichotomies when a number of classes is not a power of 2. It will be interesting applying some techniques aiming at compensating this imbalance in those particular cases.

CRediT authorship contribution statement

Lorena Álvarez-Pérez: Writing – original draft, Validation, Supervision, Funding acquisition. **Álvaro Callejas-Ramos:** Validation, Software.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Lorena Alvarez-Perez reports financial support was provided by Spanish Ministry of Science, Innovation and Universities. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partially supported by Grant PID2021-125652OB-I00 (ML-SAM) funded by the Spanish MCIU/AEI/10.13039/501100011033 and by "ERDF A way of making Europe".

Data availability

Data will be made available on request.

References

Benítez-Buenache, A., Álvarez-Pérez, L., Figueiras-Vidal, Aníbal R., 2021. On the design of Bayesian principled algorithms for imbalanced classification. Knowl.-Based Syst. 221.

Benítez-Buenache, A., Álvarez-Pérez, L., John Mathews, V., Figueiras-Vidal, Aníbal R., 2019. Likelihood ratio equivalence and imbalanced binary classification. Expert Syst. Appl. 130, 84–96.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, USA.

Bouazizi, M., Ohtsuki, T., 2019. Multi-class sentiment analysis on twitter: Classification performance and challenges. Big Data Min. Anal. 2 (3), 181–194.

Bregman, L.M., 1967. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. 7 (3), 200–217.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Breukelen, M., Duin, R.P.W., Tax, D.M.J., den Hartog, J.E., 1998. Handwritten digit recognition by combined classifiers. Kybernetika 34 (4), 381–386.

Chaitanya, A., Mishra, A., Hoque, M.T., Tu, S., 2017. Multiclass patent document classification. Artif. Intell. Res. 7 (1), 1–14.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'16, Association for Computing Machinery, New York, NY, USA, pp. 785–794.

Cid-Sueiro, J., Arribas, J.I., Urban-Muñoz, S., Figueiras-Vidal, A.R., 1999. Cost functions to estimate a posteriori probabilities in multiclass problems. IEEE Trans. Neural Netw. 10 (3), 645–656.

Cid-Sueiro, J., Figueiras-Vidal, A.R., 2001. On the structure of strict sense Bayesian cost functions and its applications. IEEE Trans. Neural Netw. 12 (3), 445–455.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13 (1), 21–27.

Deterding, D.H., 1990. Speaker Normalisation for Automatic Speech Recognition. University of Cambridge.

Dietterich, T.G., Bakiri, G., 1995. Solving multiclass learning problems via error-correcting output codes. J. Artificial Intelligence Res. 2 (1), 263–286.

Duan, K., Keerthi, S.S., Chu, W., Shevade, S.K., Poo, A.N., 2003. Multi-category classification by soft-max combination of binary classifiers. In: Windeatt, T., Roli, F. (Eds.), Multiple Classifier Systems. MCS 2003. In: Lecture Notes in Computer Science, vol. 2709, Springer, Berlin, Heidelberg.

Duin, R., 1998. Multiple Features. UCI Machine Learning Repository, http://dx.doi.org/10.24432/C5HC70.

Frank, A., Asuncion, A., 2010. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, URL: http://archive.ics.uci.edu/ml.

Frey, P.W., Slate, D.J., 1991. Letter recognition using Holland-style adaptive classifiers. Mach. Learn. 6 (2), 161–182.

Funahashi, K.-I., 1989. On the approximate realization of continuous mappings by neural networks. Neural Netw. 2 (3), 183–192.

Goldbloom, A., Hammer, B., 2010. Kaggle: Your machine learning and data sciencie. Google LLC. URL: https://www.kaggle.com/.

Gupta, S., Amin, S., 2022. Scalable design of error-correcting output codes using discrete optimization with graph coloring. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS'22, Curran Associates Inc., Red Hook, NY, USA, pp. 10094–10107.

Krawczyk, B., 2016. Learning from imbalanced data: Open challenges and future directions. Prog. Artif. Intell. 5, 221–232.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1, 541–551, (Winter 1989).

Nguyen, H.D., LaValva, L.J., Hot, S.-S., Khan, M.S., Kaegi, N., 2021. Optimal N-ary ECOC matrices for ensemble classification. In: 2021 IEEE Symposium Series on Computational Intelligence. SSCI, Orlando, FL, USA, pp. 1–8. http://dx.doi.org/10.1109/SSCI50451.2021.9660146.

Nocentini, O., Kim, J., Bashir MZ, M.Z., Cavallo, F., 2022. Image classification using multiple convolutional neural networks on the fashion-MNIST dataset. Sens. (Basel) 22 (23).

Renals, S., Rohwer, R., 1989. Phoneme classification experiments using radial basis functions. In: International 1989 Joint Conference on Neural Networks. vol. 1, Washington, DC, USA, pp. 461–467.

Rifkin, R., Klautau, A., 2004. In defense of one-vs-all classification. J. Mach. Learn. Res. 5, 101–141.

Safavian, S.R., Landgrebe, D., 1991. A survey of decision tree classifier methodology. IEEE Trans. Syst. Man Cybern. 21 (3), 660–674.

Siuly, Li, Y., 2014. A novel statistical algorithm for multiclass eeg signal classification. Eng. Appl. Artif. Intell. 34 (C), 154–167, (September 2014).

Slate, D., 1991. Letter Recognition. UCI Machine Learning Repository, http://dx.doi.org/10.24432/C5ZP40.

Szűcs, G., 2023. Multiclass classification by min–max ECOC with hamming distance optimization. Vis. Comput. 39, 3949–3961. http://dx.doi.org/10.1007/s00371-022-02540-z.

Vanschoren, Joaquin, van Rijn, Jan N., Bischl, Bernd, Torgo, Luis, 2013. OpenML: networked science in machine learning. SIGKDD Explorations 15 (2), 49–60. http://dx.doi.org/10.1145/2641190.2641198.

Walsh, J.L., 1923. A closed set of normal orthogonal functions. Amer. J. Math. 45 (1), 5–24.

Xiao, H., Rasul, K., Vollgraf, R., 2017. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747.

Yoon, J., Kim, D.H., Kim, Y., Park, K., 2022. Prediction of knee prosthesis using patient gender and BMI with non-marked X-ray by deep learning. Front. Surg. 9, 857551. http://dx.doi.org/10.3389/fsurg.2022.857551.

Yu, A., Jing, S., Lyu, N., Wen, W., Yan, Z., 2024. Error correction output codes for robust neural networks against weight-errors: A neural tangent kernel point of view. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. NIPS'24, Curran Associates Inc., Red Hook, NY, USA, pp. 82493–82513.

Yuan, X., Xie, L., Abouelenien, M., 2018. A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data. Pattern Recogn. 77 (C), 160–172, May 2018.