9-18-2021

**Ensemble Learning using Error Correcting Output Codes: New Classification Error Bounds**

Hieu Nguyen  
*Rowan University, nguyen@rowan.edu*

Mohammed Sarosh Khan

Nicholas Kaegi

Shen-Shyang Ho  
*Rowan University, hos@rowan.edu*

Jonathan Moore

*See next page for additional authors*

Follow this and additional works at: https://rdw.rowan.edu/csm_facpub

Part of the Computer Sciences Commons

**Recommended Citation**


This Article is brought to you for free and open access by the College of Science & Mathematics at Rowan Digital Works. It has been accepted for inclusion in Faculty Scholarship for the College of Science & Mathematics by an authorized administrator of Rowan Digital Works.
Authors
Hieu Nguyen, Mohammed Sarosh Khan, Nicholas Kaegi, Shen-Shyang Ho, Jonathan Moore, Logan Borys, and Lucas Lavalva
Abstract—New bounds on classification error rates for the error-correcting output code (ECOC) approach in machine learning are presented. These bounds have exponential decay complexity with respect to codeword length and theoretically validate the effectiveness of the ECOC approach. Bounds are derived for two different models: the first under the assumption that all base classifiers are independent and the second under the assumption that all base classifiers are mutually correlated up to first-order. Moreover, we perform ECOC classification on six datasets and compare their error rates with our bounds to experimentally validate our work and show the effect of correlation on classification accuracy.

Index Terms—Error correcting output codes, ensemble learning, correlation

I. INTRODUCTION

Error correcting output codes (ECOC) is an ensemble classification technique in machine learning that is motivated by coding theory where transmitted or stored information is encoded by binary strings (codewords) with high Hamming distance which allows for unique decoding of bit errors [1]. There are many variants and extension of the ECOC techniques such as the use of ternary [2] and N-ary codes [3], optimizing individual classifier performance concurrently by exploiting their relationships [4], and optimizing the learning of the base classifiers together as a multi-task learning problem [5]. Some theoretical error bounds for ECOC can be found in [1] and [6]. Moreover, Passerini et al. [6] provided leave-one-out error bound for using kernel machines as base classifiers for ECOC classifier. More recently, the ECOC technique has been extended to handle the zero-shot learning problem [7], the life-long learning problem [8], and handling adversarial examples in neural network by integrating ECOC with increasing ensemble diversity [9].

In a conventional ECOC classifier, each class of a given dataset is assigned a codeword and a learned model \(L\) is trained through an ensemble of binary classifiers constructed from the columns of the corresponding ECOC matrix whose rows consists of the class codewords [1]. Each column defines a bipartition of the dataset by merging classes with the same bit value. Decoding (classification) is performed by matching the codeword predicted by \(L\) with the class codeword nearest in Hamming distance. In essence, ECOC is a generalization of one-vs-one and one-vs-all classification techniques, and as an ensemble technique, it is most effective when the binary classifiers make independent mistakes on a random sample.

In this paper we derive new bounds on ECOC classification error rates that improve on that obtained by [10] by first applying the Feller and Chernoff bounds that are well known in statistics to the case where all binary classifiers are mutually independent and then applying a more recent bound due to [11] where they are correlated. These new bounds theoretically establish the effectiveness of the ECOC approach in machine learning; in particular, we show under certain assumptions that ECOC classification error decays exponentially to zero with respect to codeword length. We also present experimental results to demonstrate the validity of these bounds by applying them to various datasets to show the effect of correlation on classification accuracy.

Let \(L = \{L_1, \ldots, L_n\}\) denote the aforementioned ensemble of \(n\) binary classifiers (or learners) for a data set \(S\) with \(C\) classes. Let \(e_i\) denote the error rate of \(L_i\). Since \(L_i\) is a binary classifier that only outputs 0 or 1 where \(L_i = 1\) indicating an error, we shall also call \(e_i\) the bit error rate since the outputs \(L_1, L_2, \ldots, L_n\) represent a binary string. The following result, due to [10] gives a crude bound on the accuracy of \(L\):

**Theorem 1 (GS Bound, [10]).** Let \(\bar{e} = (e_1 + \ldots + e_n)/n\) denote the average bit error rate. Then the ECOC classification error rate \(E\) of \(L\) is bounded by four times the average bit error rate, i.e.,

\[
E \leq 4 \frac{1}{n} (e_1 + \ldots + e_n) = 4\bar{e}
\]

We note that the GS bound makes no assumptions regarding whether or not the classifiers are independent or how much correlation exists between them. However, the GS bound is far from being sharp: assuming that \(\bar{e} = 0.1\), then \(E \leq 0.4\). Thus, the GS bound fails to answer whether it is theoretically possible for \(E < \bar{e}\), which would validate its effectiveness
as an ensemble technique. Moreover, the GS bound gives no explicit dependence of $E$ on $n$.

To the best of our knowledge and prior to this work, no error bound exists that rigorously demonstrates that $E < \bar{e}$ is theoretically possible in the ECOC setting. Progress so far has been limited to extending the GS bound to loss-based decoding schemes [12] and special distance measures [13]. In addition, theorems have been proven that bound the excess error rate of the ECOC classifier in terms of the excess error rates of the constituent binary classifiers [14], [15]. Here, “excess error rate” refers to the difference between the error rate and the ECOC classifier in terms of the excess error rates of the schemes [12] and special distance measures [13]. In addition, the GS bound gives no error bound exists that rigorously demonstrates that explicit dependence of $\lambda$ as an ensemble technique. Moreover, the GS bound gives no explicit dependence of $E$ on $n$.

Our main result establishes new bounds on $E$ by calling on results from statistical theory.

**Theorem 2** (Main Result). Let $M$ be the ECOC matrix corresponding to $L$ with row dimension $n$ and minimum row Hamming distance $2m$. Set $r = m/n$ and $\bar{e} = \frac{1}{n} \sum_{i=1}^{n} e_i$ with $\bar{e} \neq r$.

1) Chernoff Bound: If all binary classifiers are mutually independent, then

$$E \leq n^\lambda$$

where $\lambda = \frac{e^r - \bar{e}}{(r/\bar{e})^r}$. (2)

2) KZ Bound: If $e_i = \bar{e}$ with $\bar{e} \leq \frac{m-1}{n-1}$, and all binary classifiers are mutually correlated up to second-order only and specified by a uniform non-negative correlation coefficient $c$ that satisfies the Bahadur bound [36], then

$$E \leq n^\lambda + 0.5cn(n-1) \left( \frac{m-1}{n-1} - \bar{e} \right) \omega^n$$

(3)

where $\lambda$ is defined in part 1 and $\omega = \left( \frac{\bar{e}}{r} \right)^r \left( \frac{1 - \bar{e}}{1 - r} \right)^{1-r}$.

Assuming $r$ is fixed, these bounds imply that $E$ decays exponentially to zero with respect to $n$ (codeword length).

**II. INDEPENDENT BASE CLASSIFIERS**

In this section assume that all classifiers are mutually independent, but not necessarily identically distributed. This allows us to use the Poisson binomial distribution to describe the probability of error for our ensemble of classifiers and show that the corresponding ECOC error is bounded by the classical binomial distribution based on the maximum error rate of all the classifiers.

Although the assumption of independence rarely holds in practice for real-world data sets, it is still useful as a starting point for our theoretical analysis and for establishing baseline results. An important application where this assumption is considered involves the setting of multi-view learning within the context of co-training [16], where say two classifiers are trained separately on data representing two different views (or sets of attributes). In this setting one of the assumptions requires the classifiers to be conditionally independent given the class label. This assumption can be relaxed [17], [18]. We aim to do the same in the section where we take into account correlation between classifiers.

Denote by $S(n,k)$ the collection of all $k$-element subsets of $[n] = \{1, \ldots, n\}$. Given a subset $A$ of $[n]$, we define the outcome $L_A$ to be such that $L_i = 1$ if $i \in A$ and $L_i = 0$ if $i \notin A$, where $A$ denotes the complement of $A$ in $[n]$.

**Definition 3.** Let $E = \{e_1, \ldots, e_n\}$ be a set of error rates of $\{L_1, \ldots, L_n\}$, respectively. We define $p_E(n,k)$ to be the probability of the event where exactly $k$ out of the $n$ classifiers suffered bit errors, i.e., those outcomes $L_A$ where $\sum_{i=1}^{n} L_i = k$. Then $p_E(n,k)$ is given by (Poisson binomial distribution)

$$p_E(n,k) = \sum_{A \in S(n,k)} \left( \prod_{i \in A} e_i \right) \left( \prod_{j \notin A} (1 - e_j) \right)$$

(4)

If the classifiers are identically distributed so that $e_i = \bar{e}$ for all $i = 1, \ldots, n$, then we define this probability by (binomial distribution)

$$p(n,k,\bar{e}) = \binom{n}{k} \bar{e}^k (1 - \bar{e})^{n-k}$$

(5)

Recall that the minimum Hamming distance between any two rows or any two columns of an $n$-dimensional Hadamard matrix $H$ is $n/2$ (see [10]). In that case, when at least $n/4$ of the classifiers (corresponding to the columns of $H$) each makes an error, i.e., misclassifies a sample, then ECOC misclassification may occur. This is because the rows of a $H$ describes an error-correcting code that only guarantees correct decoding up to (but strictly less than) $n/4$ bit errors. Therefore, in order to bound $E$, we shall assume under a worst-case scenario that misclassification always occur when $k \geq n/4$, where $k$ is the number of classifiers that suffered bit errors.

The following theorem shows that $p_E(n,k)$ can be bounded by the maximum error rate of all the classifiers, assuming all are no larger than $k/n$.

**Theorem 4.** Let $n, k \in \mathbb{N}$ with $0 < k < n$. Let $E = \{e_1, \ldots, e_n\}$ be a set of error rates with $0 \leq e_i \leq k/n$ for all $i \in [n]$. Set $e_{\max} = \max(E)$. Then

$$p_E(n,k) \leq p(n,k,e_{\max})$$

(6)

The proof of this theorem requires the following lemmas, whose proofs are given in the appendix of this paper [19]. Before stating them, we first introduce notation: given $m \in [n]$, we define $A_m = [n] - \{m\}$ and $E_m = E - \{e_m\}$.

**Lemma 5.** We have

$$p_{E_m}(n-1,k-1) - p_{E_m}(n-1,k) > 0$$

(7)

for all $m \in [n]$ and $k = 2, \ldots, n$.

**Lemma 6.** $p_E(n,k)$ is strictly increasing with respect to $e_i$ over the interval $(0, k/n)$.

**Proof.** (of Theorem 4) Since $p_E(n,k)$ is monotone increasing in each variable $e_m$, it is maximal when each $e_m$ is replaced by $e_{\max}$. Thus,

$$p_E(n,k) \leq p(n,k,e_{\max})$$
as desired.

**Definition 7.** We define the maximum ECOC error rate $\varepsilon_E(n,m)$ as the probability of the event where at least $m$ out of $n$ independent binary classifiers produces an error and is given by the cumulative sum

$$\varepsilon_E(n,m) = \sum_{k=m}^{n} p_E(n,k)$$  \hspace{1cm} (8)

If the classifiers are identically distributed, then we define the probability of this event by

$$\varepsilon(n,m,\hat{e}) = \sum_{k=m}^{n} p(n,k,\hat{e})$$  \hspace{1cm} (9)

It is clear that $\mathcal{E} \leq \varepsilon_E(n,m)$. Moreover, we note that $\varepsilon_E(n,n/4)$ gives the maximum ECOC error rate for a Hadamard matrix $M$ of dimension $n = 4m$ with minimum row Hamming distance $n/2$. The following theorem, which follows immediately from Theorem 4, shows that $\varepsilon$ is bounded by the binomial distribution based on the largest bit error rate.

**Theorem 8.** Suppose $0 \leq e_i \leq m/n$ for all $i \in [n]$. Set $e_{\text{max}} = \max(E)$. Then

$$\varepsilon_E(n,m) \leq \varepsilon(n,m,e_{\text{max}})$$  \hspace{1cm} (10)

We now apply Feller’s result on $\varepsilon(n,m,\bar{e})$ to obtain the following simple rational bound:

**Lemma 9 (20).** For $m > n\bar{e}$, we have

$$\varepsilon(n,m,\bar{e}) \leq \frac{m(1 - \bar{e})}{(m - n\bar{e})^2}$$  \hspace{1cm} (11)

The following corollary shows that ECOC error rate tends to zero as the codeword length tends to infinity assuming the ratio $m/n$ stays fixed. This gives theoretical justification for the effectiveness of the ECOC approach for datasets with a large number of classes; of course, this assumes the existence of many relatively accurate independent binary classifiers.

**Corollary 10.** Suppose $r = m/n$ and $\bar{e}$ are fixed with $0 \leq e_i \leq \hat{e} < r$ for all $i \in [n]$. Then

$$\lim_{n \to \infty} \mathcal{E} = 0$$  \hspace{1cm} (12)

**Proof.** Set $e_{\text{max}} = \max(E)$. Then $e_{\text{max}} \leq \hat{e}$ and since $\hat{e} < r$, we have $m > n\hat{e}$. It follows from Theorem 8 and Lemma 9 that the chain of inequalities hold:

$$\mathcal{E} \leq \varepsilon_E(n,m) \leq \varepsilon(n,m,e_{\text{max}}) \leq \frac{m(1 - \hat{e})}{(m - n\hat{e})^2} \leq \frac{r(1 - \hat{e})}{n(\hat{r} - \hat{e})^2}$$

It is now clear that $\mathcal{E} \to 0$ as $n \to \infty$.

We emphasize that the improvement of the Chernoff bound (Corollary 12) over the GS bound (Theorem 1) is due to the assumption that all the binary classifiers are mutually independent. Figure 1 clearly demonstrates this for large $n$ and small $\bar{e}$.

We end this section by commenting that Corollary 12 is also valid for the non-binary ECOC setting where the coefficients of the ECOC matrix is chosen from a non-binary alphaset [3].

**III. CORRELATED BASE CLASSIFIERS**

In this section we assume dependence (correlation) between certain base classifiers to show how it affects ECOC accuracy. We first make the simple assumption that all binary classifiers $\{L_i\} \in \mathbb{B}_n$ are mutually independent except for a pair of dependent classifiers $L_{n-1}$ and $L_n$, which are allowed to depend on each other as follows. Recall that each $L_i$ takes on two possible values, namely $L_i = 0$ (correct prediction)
and $L_i = 1$ (incorrect prediction). As before, let $e_i$ denote the error rate of $L_i$, i.e., $P(L_i = 1) = e_i$. Since $L_{n-1}$ and $L_n$ are dependent on each other, we specify their correlation via the joint probability

$$P_{11} := P(L_{n-1} = 1 \text{ and } L_n = 1) = f \quad (16)$$

It follows that the remaining joint probabilities are given by

$$P_{10} := P(L_{n-1} = 0 \text{ and } L_n = 1) = e_{n-1} - f \quad (17)$$

$$P_{01} := P(L_{n-1} = 1 \text{ and } L_n = 0) = e_n - f \quad (18)$$

$$P_{00} := P(L_{n-1} = 1 \text{ and } L_n = 1) = 1 - e_{n-1} - e_n + f \quad (19)$$

We shall assume that $0 \leq e_{n-1} \leq 1/2$, $0 \leq e_n \leq 1/2$, and $0 \leq f \leq \min(e_{n-1}, e_n)$ so that all probabilities are non-negative. We then define the correlation between $L_{n-1}$ and $L_n$ as

$$c = \text{cor}(L_{n-1}, L_n) = \frac{f - e_{n-1}e_n}{\sqrt{e_{n-1}(1-e_{n-1})e_n(1-e_n)}} \quad (20)$$

In particular, if $L_{n-1}$ and $L_n$ are independent so that $f = e_{n-1}e_n$, then $c = 0$.

Given a subset $A \in S(n,k)$, we denote $\tilde{A} = [n] - A$ and define

$$P_n(A) := P(\{L_i = 1 : i \in A\} \text{ and } \{L_i = 0 : i \in \tilde{A}\}) \quad (21)$$

**Definition 13.** Let $E_n = \{e_1, \ldots, e_n\}$. We define the probability of the event where $k$ out of $n$ classifiers produces an error (with dependence between classifiers $L_{n-1}$ and $L_n$ as defined above) by

$$p_{E_n}(n, k, f) = \sum_{A \in S(n,k)} P(A). \quad (22)$$

If $e_i = \bar{e}$ for all $i \in [n]$, then we denote $p(n, k, \bar{e}, f) := p_{E_n}(n, k, f)$.

Define $\bar{e}_n = (e_{n-1} + e_n)/2$. The following lemma, whose proof is given in [19], shows the explicit dependence of $p_{E_n}(n, k, f)$ on $\bar{e}_n$ and $f$.

**Lemma 14.** We have

$$p_{E_n}(n, k, f) = \begin{cases} A, & \text{if } k \leq n - 2; \\ B, & \text{if } k = n - 1; \\ C, & \text{if } k = n. \end{cases} \quad (23)$$

where

$$A = fp_{E_n-2}(n-2, k-2) + 2(\bar{e}_n - f)p_{E_n-2}(n-2, k-1) + (1-2\bar{e}_n + f)p_{E_n-2}(n-2, k)$$

$$B = fp_{E_n-2}(n-2, n-3) + 2(\bar{e}_n - f)p_{E_n-2}(n-2, n-2)$$

$$C = fp_{E_n-2}(n-2, n-2)$$

Define $\bar{e}_{max} = \max(E_{n-2})$. We apply Theorem [4] to the above lemma to obtain the following bound.

**Corollary 15.** Suppose $\bar{e}_{max} \leq \frac{n-2}{n-2}$. Then

$$p_{E_n}(n, k, f) \leq \begin{cases} A, & \text{if } k \leq n - 2; \\ B, & \text{if } k = n - 1; \\ C, & \text{if } k = n. \end{cases} \quad (24)$$

$$A = fp(n-2, k-2, \bar{e}_{max}) + 2(\bar{e}_n - f)p(n-2, k-1, \bar{e}_{max}) + (1-2\bar{e}_n + f)p(n-2, k, \bar{e}_{max})$$

$$B = fp(n-2, n-3, \bar{e}_{max}) + 2(\bar{e}_n - f)p(n-2, n-2, \bar{e}_{max})$$

$$C = fp(n-2, n-2, \bar{e}_{max})$$

In the special case where all binary classifiers are identically distributed, i.e., $e_i = \bar{e}$ for all $i \in [n]$, then

$$A = fp(n-2, k-2, \bar{e}) + 2(\bar{e}_n - f)p(n-2, k-1, \bar{e}) + (1-2\bar{e} + f)p(n-2, k, \bar{e})$$

$$B = fp(n-2, n-3, \bar{e}) + 2(\bar{e}_n - f)p(n-2, n-2, \bar{e})$$

$$C = fp(n-2, n-2, \bar{e})$$

The next lemma, whose proof is given in [19], assumes all classifiers are identically distributed.

**Lemma 16.** $p(n, k, \bar{e}, f)$ is increasing with respect to $f$ for fixed $\bar{e} \in I$, where

$$I = \begin{cases} [0, k/n - \alpha(n,k)] & \text{if } k \leq n - 2; \\ [0, 1 - 2/n] & \text{if } k = n - 1; \\ [0, 1] & \text{if } k = n, \end{cases}$$

and $\alpha(n,k) = \frac{1}{n} \sqrt{\frac{k(n-k)}{n-1}}$.

**Definition 17.** We define the maximum ECOC error rate $\varepsilon_E(n, m, f)$ (assuming correlation given by $f$) as the probability of the event where at least $m$ out of $n$ binary classifiers produces an error and therefore is given by the cumulative sum

$$\varepsilon_E(n, m, f) = \sum_{k=m}^{n} p(n, k, \bar{e}, f) \quad (25)$$

If the classifiers are identically distributed, i.e., $e_i = \bar{e}$ for all $i = 1, \ldots, n$, then we define

$$\varepsilon(n, m, \bar{e}, f) := \varepsilon_E(n, m, f) = \sum_{k=m}^{n} p(n, k, \bar{e}, f) \quad (26)$$

The next two theorems describe the dependence of the maximum ECOC error rate on $f$. Their proofs can be found in [19].

**Theorem 18.** We have

$$\varepsilon(n, m, \bar{e}, f) = f\varepsilon(n-2, m-2, \bar{e}) + 2(\bar{e}_n - f)\varepsilon(n-2, m-1, \bar{e}) + (1-2\bar{e} + f)\varepsilon(n-2, m, \bar{e}) \quad (27)$$
Theorem 19. $\varepsilon(n, m, \bar{e}, f)$ is increasing with respect to $f$ for $0 \leq \bar{e} \leq \frac{m-1}{n-1}$ and decreasing with respect to $f$ for $\frac{m-1}{n-1} \leq \bar{e} \leq 1$.

The next theorem gives a simple bound for $\varepsilon(n, m, \bar{e}, f)$, which again implies that $\mathcal{E}$ decays exponentially to zero but assumes that $\bar{e}$ is fixed.

Theorem 20. Let $r = \frac{m-2}{n-2}$ and $\lambda = \frac{e}{(r/\bar{e})}$. Then

$$\varepsilon(n, m, \bar{e}, f) \leq \lambda \left( \frac{e}{r} \right)^{n-1} \left( 1 - \frac{e}{r} \right)^{m-1}$$

Proof. We apply Theorem (18) and Corollary [12]

$$\varepsilon(n, m, \bar{e}, f) \leq f \cdot \lambda^n + 2(\bar{e} - f) \cdot \lambda^{n-1}$$

$$+ (1 - 2\bar{e} + f) \cdot \lambda^{n-2}$$

since $\lambda_1 \geq \lambda_2 \geq \lambda_3$. Setting $\lambda = \lambda_1$ gives the desired result.

We also define the maximum ECOC error rate $\varepsilon(n, m, \bar{e}, f)$ as the probability of the event where at least $m$ out of $n$ binary classifiers produces an error and therefore is given by the cumulative sum

$$\varepsilon(n, m, \bar{e}, f) = \sum_{k=m}^{n} p(n, k, \bar{e}, f)$$

We next investigate the effect of having all classifiers mutually correlated up to second-order only (all higher-order correlations are zero). We define

$$Z_i = \frac{L_i - e_i}{\sqrt{e_i(1 - e_i)}}$$

$$f_{ij} = P(L_i = 1 \text{ and } L_j = 1)$$

$$c_{ij} = \text{corr}(Z_i, Z_j) = \frac{f_{ij} - e_i e_j}{\sqrt{e_i(1 - e_i)e_j(1 - e_j)}}$$

Let $A = \{a_1, \ldots, a_k\} \in S(n, k)$. Recall our definition of the outcome $L_A$ where $L_i = 1$ if $i \in A$ and $L_i = 0$ if $i \in \bar{A}$ where $k = \sum_{i=1}^{n} L_i$. Denote by $\pi_A$ the probability of the outcome $L_A$.

**Definition 21.** Suppose $e_i = \bar{e}$ and $c_{ij} = c$. We define the probability of the event where $k$ out of $n$ classifiers produces an error (with correlation given by (33)) by

$$p(n, k, \bar{e}, c) = \sum_{A \in S(n, k)} \pi_A.$$
Then using the bound
\[
\binom{n}{m} \leq \left( \frac{m}{n} \right)^m \left( 1 - \frac{m}{n} \right)^{n-m}^{-1},
\]
we have
\[
p(n - 1, m - 1, \bar{e}) \leq \left( \frac{r}{\bar{e}} \right) \frac{\bar{e}^m (1 - \bar{e})^{n-m}}{(m/n)^m (1 - m/n)^{n-m}} \leq \left( \frac{r}{\bar{e}} \right) \left( \frac{\bar{e}}{1 - \bar{e}} \right)^m \left( 1 - \bar{e} \right)^{n-m} \leq \left( \frac{r}{\bar{e}} \right) \left( \bar{e}^{r} \right) \left( 1 - \bar{e} \right)^{1-r} \right)
\]
Since \((\bar{e})^r \left( \frac{1 - \bar{e}}{1 - \bar{e}} \right)^{1-r} < 1\) for \(\bar{e} \neq r\), it follows that \(p(n - 1, m - 1, \bar{e}) \to 0\) exponentially as \(n \to \infty\). Thus, the same holds for \(\varepsilon(n, m, \bar{e}, c)\) as well.

IV. EXPERIMENTAL RESULTS

In this section we present experimental results to demonstrate the validity of our work by performing ECOC classification on various data sets and comparing the resulting classification error rates with those predicted by the Chernoff and KZ bounds established in the previous section.

In particular, we selected six public datasets to perform ECOC classification: Pendigits, Usps, Vowel, Letter Recognition (Letters), CIFAR-10, and Street View House Numbers (SVHN). Information regarding these datasets is given in Table I.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Samples</th>
<th># Features</th>
<th># Classes (C)</th>
<th>(r = m/n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pendigits</td>
<td>3,498</td>
<td>16</td>
<td>10</td>
<td>2/11</td>
</tr>
<tr>
<td>Usps</td>
<td>7,291</td>
<td>256</td>
<td>10</td>
<td>2/10</td>
</tr>
<tr>
<td>Vowel</td>
<td>990</td>
<td>10</td>
<td>11</td>
<td>2/11</td>
</tr>
<tr>
<td>Letters</td>
<td>20,000</td>
<td>16</td>
<td>26</td>
<td>6/26</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>60,000</td>
<td>Image</td>
<td>10</td>
<td>2/10</td>
</tr>
<tr>
<td>SVHN</td>
<td>99,289</td>
<td>Image</td>
<td>10</td>
<td>2/10</td>
</tr>
</tbody>
</table>

TABLE I: Datasets

1) ECOC Matrix: We employed a square ECOC matrix \(M\) for every dataset \((n = C)\) and constructed \(M\) from a \(\{0, 1\}\)-Hadamard matrix \(H\) of dimension \(2^k\), where \(k\) was chosen to be the smallest integer for which \(2^k \geq c\) and \(c\) denotes the number of classes. We then truncated an appropriate number of rows and columns from \(H\) (starting from the top left) to obtain our square matrix \(M\) of dimension \(n \times n\). The parameter \(r = m/n\) for each dataset is given in Table I.

2) Classification algorithms: For the datasets Pendigits, Usps, Vowel, and Letters, we employed two different models \(L\) for our base classifiers: decision tree (DT) and support-vector machines (SVM), using the Python modules (version 3.7) sklearn.tree.DecisionTreeClassifier and sklearn.svm.SVC with default settings, respectively, utilizing the scikit-learn machine learning library. Computations were performed on a standard laptop. For the image datasets CIFAR-10 and SVHN, we employed a pre-trained convolutional neural network, Resnet-18 (loaded from Pytorch), with an additional dense layer to produce binary output and using the Adam optimizer. Computations were performed for 10 epochs with a batch size of 128 and ran on the Open Science Grid [22, 23].

Thus, given a dataset, we performed 10th-fold cross-validation based on the experimental setup described above and recorded the ECOC error rate (experimental) for each fold, as well mean ECOC error rate and standard deviation for all ten folds. To compute the GS, Chernoff, and KZ bounds given by (1), (2), and (3), respectively, for each fold, we used the mean bit error rate \(\bar{e}\), obtained by averaging the bit error rates of all the binary classifiers. In addition, for the KZ bound we used the mean correlation for \(c\), obtained by averaging the coefficients of the correlation matrix of the binary classifiers. Full results, including values used for \(\bar{e}\) and \(c\), are given in [19] (Tables III-VIII).

A. Results and Discussion

Experimental results show that for all datasets the ECOC error rates \((\bar{e})\) are either below all three bounds (GS, Chernoff, and KZ) or clustered around the Chernoff and KZ bounds, where the latter occurs for Letters (DT and SVM) and Pendigits (SVM). This can be seen in the plots in Figures 2-7 for Pendigits, Letters, CIFAR-10, and SVHN (see [19] for plots of USPS and Vowels) where ECOC error rates are shown for each of the ten folds and in Table I where results are averaged over all folds (lowest value indicated in bold). These results demonstrate the validity of all three bounds. However, Figures 3-5 (Letters) and 7 (Pendigits) clearly show that the Chernoff and KZ bounds decay exponentially to zero with respect to \(n\) and thus are more effective for larger values of \(n\). Overall, we believe our experimental results demonstrate that the Chernoff and GS bounds are quite useful in practice.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we presented two new classification error bounds for the ECOC ensemble learning: the first under the assumption that all base classifiers are independent and the second under the assumption that all base classifiers are mutually correlated up to first-order. These bounds have exponential decay complexity with respect to codeword length and theoretically validate the effectiveness of the ECOC approach. Moreover, we perform ECOC classification on six datasets and compare their error rates with our bounds to experimentally validate our work and show the effect of correlation on classification accuracy. Future work include investigating the Chernoff bound for ECOC in settings with limited independence [24] and comparing the performance of binary vs \(N\)-ary ECOC with respect to the error bounds presented in this paper.
TABLE II: ECOC error rate (Mean and standard deviation of 10-fold cross-validation): Experimental vs GS, Chernoff, and KZ bounds.
REFERENCES


APPENDIX

In this appendix we present proofs of lemmas and theorems that were not included in our paper and additional experimental results.

A. Proofs of lemmas and theorems

Proof of Lemma 6. It suffices to prove (47) only for \( m = n \) because of symmetry. Observe that

\[
\frac{|S(n-1,k)|}{|S(n-1,k-1)|} = \left(\frac{n-1}{k-1}\right) = \frac{n-k}{k},
\]

or equivalently,

\[
k|S(n-1,k)| = (n-k)|S(n-1,k-1)|.
\]

Define \( S_m(n,k) \) to be the multi-set consisting of the union of \( m \) copies of \( S(n,k) \). Then \( |S_m(n-1,k)| = |S_{n-k}(n-1,k-1)| \) because of (42). We now differentiate and rearrange terms as follows:

\[
\frac{dp(n,k)}{de_n} = \sum_{A \in S(n-1,k-1)} \left( \prod_{i \in A} e_i \right) \left( \prod_{j \in \bar{A}} (1 - e_j) \right) \\
- \sum_{A \in S(n,k)} \left( \prod_{i \in A} e_i \right) \left( \prod_{j \in \bar{A}} (1 - e_j) \right) \\
= \sum_{A \in S_{n-k}(n-1,k-1)} \frac{1}{n-k} \left( \prod_{i \in A} e_i \right) \left( \prod_{j \in \bar{A}} (1 - e_j) \right) \\
- \sum_{A \in S(n-1,k-1)} \left[ \frac{1}{n-k} \left( \prod_{i \in A} e_i \right) \left( \prod_{j \in \bar{A}} (1 - e_j) \right) \right] \\
= \sum_{A \in S(n-1,k-1)} \left[ \sum_{l \in \bar{A}} \frac{1}{n-k} \left( \prod_{i \in A \setminus \{l\}} e_i \right) \left( \prod_{j \in A \setminus \{l\}} (1 - e_j) \right) \right] \\
> 0
\]

since

\[
\frac{1 - e_l}{n-k} > \frac{e_l}{k}, \quad i = 1, \ldots, n,
\]

which follows from the assumption that every \( e_l < k/n \).

Proof of Lemma 5. We first assume the case where \( k \leq n-2 \) and consider the following partition of \( S(n,k) \):

\[
S(n,k) = S(n-2,k) \cup S' \cup S'',
\]

where

\[
S' = \{A' = A \cup \{n-1\} : A \in S(n-2,k-1)\} \\
\cup \{A' = A \cup \{n\} : A \in S(n-2,k-1)\} \\
S'' = \{A'' = A \cup \{n-1,n\} : A \in S(n-2,k-2)\}
\]

for all \( m \in [n] \). By symmetry, it suffices to prove (43) for \( m = n \). We have

\[
\frac{dp(n,k)}{de_m} = \frac{dp_E(n,k)}{de_n} = \frac{d}{de_n} \left[ \sum_{A \in S(n,k)} \left( \prod_{i \in A} e_i \right) \left( \prod_{j \in \bar{A}} (1 - e_j) \right) \right] \\
= \sum_{A \in S(n,k)} \left( \prod_{i \in A} e_i \right) \left( \prod_{j \in \bar{A}} (1 - e_j) \right) \\
- \sum_{A \in S(n,k)} \left( \prod_{i \in A} e_i \right) \left( \prod_{j \in \bar{A}} (1 - e_j) \right) \\
= \sum_{A \in S(n,k)} \left( \prod_{i \in A} e_i \right) \left( \prod_{j \in \bar{A}} (1 - e_j) \right) \\
= p_{E_m}(n-1, k-1) - p_{E_m}(n-1, k).
\]

By the previous lemma, we have

\[
p_{E_m}(n-1, k-1) - p_{E_m}(n-1, k) > 0
\]

when \( 0 \leq e_i < k/n \) for every \( i \in A_m \), which proves that \( \frac{dp_E(n,k)}{de_m} > 0 \). Thus, \( p_E(n,k) \) is strictly increasing with respect to \( e_m \), and therefore strictly increasing with respect to each \( e_m \) due to symmetry.

Proof of Lemma 6. Observe that \( p_E(n,k) \) is linear in each \( e_m \). Assume \( k \geq 1 \). We claim that

\[
\frac{dp_E(n,k)}{de_m} = p_{E_m}(n-1, k-1) - p_{E_m}(n-1, k)
\]

(43)
Then
\[ p_{E_n}(n, k, f) = \sum_{A \in S(n, k)} P_n(A) \]
\[ = \sum_{A \in S(n-2, k)} P_n(A) + \sum_{A' \in S'} P_{n-1}(A') \]
\[ = (1 - e_{n-1} - e_n + f) \sum_{A \in S(n-2, k)} P_{n-2}(A) \]
\[ + (e_{n-1} - f) \sum_{A \in S(n-2, k-1)} P_{n-2}(A) \]
\[ + (e_n - f) \sum_{A \in S(n-2, k-1)} P_{n-2}(A) \]
\[ + f \sum_{A \in S(n-2, k-2)} P_{n-2}(A) \]
\[ = (1 - 2\varepsilon_n + f) p_{E_{n-2}}(n - 2, k) \]
\[ + 2(\bar{\varepsilon}_n - f)p_{E_{n-2}}(n - 2, k - 1) \]
\[ + f p_{E_{n-2}}(n - 2, k - 2) \]
This proves \([23]\) for this case. The other two cases, \(k = n - 1\) and \(k = n\), can be proven by a similar argument by considering appropriate partitions of \(S(n, n-1)\) and \(S(n, n)\), respectively. \(\square\)

**Proof of Lemma \([16]\)** We first assume the case where \(k \leq n - 2\). It suffices to prove that the derivative of \(p(n, k, e, f)\) with respect to \(f\) is non-negative. Since Lemma \([14]\) shows that \(p(n, k, e, f)\) is linear in \(f\), we have
\[ \frac{dp(n, k, e, f)}{df} = p(n - 2, k - 2, \bar{\varepsilon}) - 2p(n - 2, k - 1, \bar{\varepsilon}) \]
\[ + p(n - 2, k, \bar{\varepsilon}) \] (46)
\[ = \frac{(n - 2)}{k - 2} \varepsilon^{k-2}(1 - \varepsilon)^{n-k} \] (47)
\[ - 2 \frac{(n - 2)}{k - 1} \varepsilon^{k-1}(1 - \varepsilon)^{n-k-1} \] (48)
\[ + \frac{(n - 2)}{k} \varepsilon^{k}(1 - \varepsilon)^{n-k-2} \] (49)
\[ = \varepsilon^{k-2}(1 - \varepsilon)^{n-k-2} \left( \frac{n - 2}{k - 2} \right) \frac{R(\bar{\varepsilon})}{k(k - 1)}, \] (50)
where
\[ R(\bar{\varepsilon}) = (1 - \bar{\varepsilon})^2 k(k-1) - 2\bar{\varepsilon}(1 - \bar{\varepsilon})k(n-k) + \varepsilon^2(n-k)(n-k-1). \]
Solving \(R(\bar{\varepsilon}) = 0\) for \(\bar{\varepsilon}\) yields
\[ \bar{\varepsilon} = \frac{k}{n} \pm \frac{1}{n} \sqrt{\frac{k(n-k)}{n-1}} = \frac{k}{n} \pm \alpha(n, k). \]
It is straightforward to verify that \(R(\bar{\varepsilon}) \geq 0\) for \(0 \leq \bar{\varepsilon} \leq k/n - \alpha(n, k)\) and therefore, \(\frac{dp(n, k, e, f)}{df} \geq 0\) over the same domain. This proves that \(p(n, k, \bar{\varepsilon}, f)\) is increasing with respect to \(f\).

**Proof of Theorem \([18]\)** We compute the derivative of \(\varepsilon(n, m, \bar{\varepsilon}, f)\) with respect to \(f\) using identities \([46]\), \([51]\), and \([55]\), and the fact that the following sum telescopes:

\[ \frac{dp(n, n - 1, \bar{\varepsilon}, f)}{df} = p(n - 2, n - 3, \varepsilon) - 2p(n - 2, n - 2, \bar{\varepsilon}) \]
\[ = \frac{(n - 2)}{n - 3} \varepsilon^{n-3}(1 - \varepsilon) - 2 \frac{(n - 2)}{n - 2} \varepsilon^{n-2} \] (52)
\[ = \varepsilon^{n-3}[n - 2 - \frac{2}{n - 2}] \] (53)
\[ = \varepsilon^{n-3}[n - 1 - \frac{1}{n - 2}] \] (54)
It follows that \(\frac{dp(n, n - 1, \bar{\varepsilon}, f)}{df} \geq 0\) when \(0 \leq \bar{\varepsilon} \leq 1 - 2/n\) and proves that \(p(n, n - 1, \bar{\varepsilon}, f)\) is increasing with respect to \(f\).
Lastly, in the case where \(k = n\), we have As for the case where \(k = n - 1\), we have
\[ \frac{dp(n, n - 1, \bar{\varepsilon}, f)}{df} = p(n - 2, n - 2, \bar{\varepsilon}) \] (55)
which is clearly non-negative for all \(\varepsilon \in [0, 1]\) and proves that \(p(n, n, \bar{\varepsilon}, f)\) is increasing with respect to \(f\). \(\square\)
It follows that \( \frac{dz(n, m, \bar{e}, f)}{df} \geq 0 \) when \( p(n - 2, m - 2, \bar{e}) - p(n - 2, m - 1, \bar{e}) \geq 0 \), or equivalently,

\[
p(n - 2, m - 2, \bar{e}) \geq p(n - 2, m - 1, \bar{e})
\]

\[
\Rightarrow (n - 2) e^{n-2} (1 - \bar{e})^{n-1} \geq (n - 2) e^{n-1} (1 - \bar{e})^{n-2} - m - 1
\]

\[
\Rightarrow \bar{e} \leq \frac{m - 1}{n - 1}
\]

Thus, \( p(n, m, \bar{e}, f) \) is increasing with respect to \( f \) for \( 0 \leq \bar{e} \leq \frac{m - 1}{n - 1} \). By the same argument, \( p(n, m, \bar{e}, f) \) must be decreasing with respect to \( f \) for \( \frac{m - 1}{n - 1} \leq \bar{e} \leq 1 \).

**Proof of Theorem 2**. The probability \( \pi_A \) can be expressed in terms of \( P(A) = e^{k} (1 - \bar{e})^{n-k} \), the corresponding probability

\[
\pi_A = P(A) \left( 1 + \sum_{i<j} c_{ij} z_i z_j \right)
\]

assuming zero correlation, by
It follows that

$$\varepsilon(n, m, \bar{e}, c) = \sum_{k=m}^{n} \sum_{A \in S(n,k)} \pi_A \left( \varepsilon(n, m, \bar{e}) + \frac{c}{2\bar{e}(1-\bar{e})} \sum_{k=m}^{n} \binom{n}{k} \bar{e}^k (1-\bar{e})^{n-k} \right)$$

$$\left( k^2 - k + \bar{e}(n-1)(n\bar{e} - 2k) \right)$$

(84)

Fig. 8: USPS: Mean bit error vs ECOC error (DT using 10-fold cross-validation)

Fig. 9: USPS: Mean bit error vs ECOC error (SVM using 10-fold cross-validation)

Fig. 10: Vowels: Mean bit error vs ECOC error (DT using 10-fold cross-validation)

Fig. 11: Vowels: Mean bit error vs ECOC error (SVM using 10-fold cross-validation)

The rest of the proof follows exactly as that given in [11, Theorem 1].

B. Additional Experimental Results

In this section we provide experimental results that were not included in our paper. Figures 8-11 give plots of ECOC error rates for Usps and Vowels datasets. Tables 3-8 give full results for each dataset, namely mean bit error, mean correlation, and ECOC error for all ten folds.
<table>
<thead>
<tr>
<th>Fold</th>
<th>Model</th>
<th>Mean Bit Error</th>
<th>Mean Correlation</th>
<th>ECOC Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DT</td>
<td>0.0323 ± 0.00915</td>
<td>0.0154 ± 0.34584</td>
<td>0.0328</td>
</tr>
<tr>
<td>2</td>
<td>DT</td>
<td>0.0362 ± 0.00936</td>
<td>0.1535 ± 0.35876</td>
<td>0.0309</td>
</tr>
<tr>
<td>3</td>
<td>DT</td>
<td>0.0311 ± 0.00438</td>
<td>0.0956 ± 0.41445</td>
<td>0.0191</td>
</tr>
<tr>
<td>4</td>
<td>DT</td>
<td>0.0315 ± 0.00747</td>
<td>-0.0525 ± 0.3018</td>
<td>0.0258</td>
</tr>
<tr>
<td>5</td>
<td>DT</td>
<td>0.0342 ± 0.00878</td>
<td>-0.1070 ± 0.28860</td>
<td>0.0218</td>
</tr>
<tr>
<td>6</td>
<td>DT</td>
<td>0.0326 ± 0.00936</td>
<td>0.1356 ± 0.35876</td>
<td>0.0309</td>
</tr>
<tr>
<td>7</td>
<td>DT</td>
<td>0.0314 ± 0.00515</td>
<td>0.0350 ± 0.35522</td>
<td>0.0255</td>
</tr>
<tr>
<td>8</td>
<td>DT</td>
<td>0.0371 ± 0.01119</td>
<td>0.0637 ± 0.34278</td>
<td>0.0382</td>
</tr>
<tr>
<td>9</td>
<td>DT</td>
<td>0.0343 ± 0.00515</td>
<td>0.0350 ± 0.35522</td>
<td>0.0255</td>
</tr>
<tr>
<td>10</td>
<td>DT</td>
<td>0.0355 ± 0.00780</td>
<td>0.1476 ± 0.35864</td>
<td>0.0255</td>
</tr>
</tbody>
</table>

TABLE III: Pendigits: Mean bit error, mean correlation, ECOC error per fold

<table>
<thead>
<tr>
<th>Fold</th>
<th>Model</th>
<th>Mean Bit Error</th>
<th>Mean Correlation</th>
<th>ECOC Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DT</td>
<td>0.0742 ± 0.01642</td>
<td>0.0041 ± 0.33921</td>
<td>0.1029</td>
</tr>
<tr>
<td>2</td>
<td>DT</td>
<td>0.0710 ± 0.01685</td>
<td>0.0874 ± 0.35303</td>
<td>0.0850</td>
</tr>
<tr>
<td>3</td>
<td>DT</td>
<td>0.0811 ± 0.02434</td>
<td>0.1207 ± 0.34830</td>
<td>0.0288</td>
</tr>
<tr>
<td>4</td>
<td>DT</td>
<td>0.0630 ± 0.01976</td>
<td>0.1070 ± 0.43716</td>
<td>0.0741</td>
</tr>
<tr>
<td>5</td>
<td>DT</td>
<td>0.0704 ± 0.01766</td>
<td>0.1022 ± 0.34205</td>
<td>0.0919</td>
</tr>
<tr>
<td>6</td>
<td>DT</td>
<td>0.0798 ± 0.01976</td>
<td>0.0581 ± 0.41083</td>
<td>0.1001</td>
</tr>
<tr>
<td>7</td>
<td>DT</td>
<td>0.0754 ± 0.01674</td>
<td>0.1047 ± 0.31367</td>
<td>0.1001</td>
</tr>
<tr>
<td>8</td>
<td>DT</td>
<td>0.0663 ± 0.01578</td>
<td>0.0273 ± 0.36283</td>
<td>0.0919</td>
</tr>
<tr>
<td>9</td>
<td>DT</td>
<td>0.0689 ± 0.02031</td>
<td>0.0251 ± 0.38335</td>
<td>0.0741</td>
</tr>
<tr>
<td>10</td>
<td>DT</td>
<td>0.0721 ± 0.01807</td>
<td>0.0209 ± 0.36387</td>
<td>0.0808</td>
</tr>
</tbody>
</table>

TABLE IV: Usps: Mean bit error, mean correlation, ECOC error per fold

<table>
<thead>
<tr>
<th>Fold</th>
<th>Model</th>
<th>Mean Bit Error</th>
<th>Mean Correlation</th>
<th>ECOC Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DT</td>
<td>0.1111 ± 0.05779</td>
<td>0.0308 ± 0.30683</td>
<td>0.1515</td>
</tr>
<tr>
<td>2</td>
<td>DT</td>
<td>0.1185 ± 0.06600</td>
<td>0.0077 ± 0.33212</td>
<td>0.2222</td>
</tr>
<tr>
<td>3</td>
<td>DT</td>
<td>0.1001 ± 0.02739</td>
<td>0.0426 ± 0.31161</td>
<td>0.1212</td>
</tr>
<tr>
<td>4</td>
<td>DT</td>
<td>0.0970 ± 0.04514</td>
<td>-0.0298 ± 0.28943</td>
<td>0.1111</td>
</tr>
<tr>
<td>5</td>
<td>DT</td>
<td>0.1038 ± 0.03682</td>
<td>0.1067 ± 0.39916</td>
<td>0.1212</td>
</tr>
<tr>
<td>6</td>
<td>DT</td>
<td>0.0937 ± 0.06618</td>
<td>0.1193 ± 0.34373</td>
<td>0.1717</td>
</tr>
<tr>
<td>7</td>
<td>DT</td>
<td>0.1414 ± 0.05448</td>
<td>0.0819 ± 0.32819</td>
<td>0.2222</td>
</tr>
<tr>
<td>8</td>
<td>DT</td>
<td>0.1028 ± 0.03313</td>
<td>-0.0571 ± 0.38160</td>
<td>0.1919</td>
</tr>
<tr>
<td>9</td>
<td>DT</td>
<td>0.1056 ± 0.03885</td>
<td>0.1135 ± 0.32523</td>
<td>0.1414</td>
</tr>
<tr>
<td>10</td>
<td>DT</td>
<td>0.1166 ± 0.06360</td>
<td>0.0227 ± 0.28839</td>
<td>0.1414</td>
</tr>
</tbody>
</table>

TABLE V: Vowels: Mean bit error, mean correlation, ECOC error per fold
<table>
<thead>
<tr>
<th>Fold</th>
<th>Model</th>
<th>Mean Bit Error</th>
<th>Mean Correlation</th>
<th>ECOC Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DT</td>
<td>0.0706 ± 0.00791</td>
<td>0.0064 ± 0.19321</td>
<td>0.0695</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.0763 ± 0.00779</td>
<td>0.0053 ± 0.21969</td>
<td>0.1060</td>
</tr>
<tr>
<td>2</td>
<td>DT</td>
<td>0.0644 ± 0.00818</td>
<td>-0.0029 ± 0.20498</td>
<td>0.0520</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.0765 ± 0.00872</td>
<td>0.0073 ± 0.20761</td>
<td>0.1060</td>
</tr>
<tr>
<td>3</td>
<td>DT</td>
<td>0.0709 ± 0.00796</td>
<td>0.0089 ± 0.21072</td>
<td>0.0675</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.0785 ± 0.00981</td>
<td>0.0082 ± 0.21389</td>
<td>0.1150</td>
</tr>
<tr>
<td>4</td>
<td>DT</td>
<td>0.0688 ± 0.00761</td>
<td>0.0113 ± 0.19695</td>
<td>0.0635</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.0732 ± 0.00988</td>
<td>0.0235 ± 0.21649</td>
<td>0.1040</td>
</tr>
<tr>
<td>5</td>
<td>DT</td>
<td>0.0688 ± 0.00826</td>
<td>0.0129 ± 0.21209</td>
<td>0.0610</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.0768 ± 0.01037</td>
<td>0.0243 ± 0.22880</td>
<td>0.1055</td>
</tr>
<tr>
<td>6</td>
<td>DT</td>
<td>0.0652 ± 0.00815</td>
<td>0.0026 ± 0.19203</td>
<td>0.0555</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.0756 ± 0.00918</td>
<td>0.0115 ± 0.20290</td>
<td>0.0995</td>
</tr>
<tr>
<td>7</td>
<td>DT</td>
<td>0.0654 ± 0.00836</td>
<td>0.0175 ± 0.21199</td>
<td>0.0555</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.0722 ± 0.01040</td>
<td>0.0027 ± 0.20112</td>
<td>0.1065</td>
</tr>
<tr>
<td>8</td>
<td>DT</td>
<td>0.0716 ± 0.00674</td>
<td>-0.0011 ± 0.19719</td>
<td>0.0620</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.0761 ± 0.00909</td>
<td>0.0036 ± 0.19906</td>
<td>0.1045</td>
</tr>
<tr>
<td>9</td>
<td>DT</td>
<td>0.0731 ± 0.00803</td>
<td>-0.0034 ± 0.19219</td>
<td>0.0620</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.0789 ± 0.00974</td>
<td>0.0183 ± 0.21919</td>
<td>0.1125</td>
</tr>
<tr>
<td>10</td>
<td>DT</td>
<td>0.0668 ± 0.00802</td>
<td>0.0049 ± 0.22051</td>
<td>0.1045</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.0739 ± 0.00927</td>
<td>0.0112 ± 0.21925</td>
<td>0.1010</td>
</tr>
</tbody>
</table>

**TABLE VI:** Letters: Mean bit error, mean correlation, ECOC error per fold

<table>
<thead>
<tr>
<th>Fold</th>
<th>Mean Bit Error</th>
<th>Mean Correlation</th>
<th>ECOC Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0082 ± 0.00186</td>
<td>0.2153 ± 0.05684</td>
<td>0.0116</td>
</tr>
<tr>
<td>2</td>
<td>0.0089 ± 0.00223</td>
<td>0.1698 ± 0.04895</td>
<td>0.0109</td>
</tr>
<tr>
<td>3</td>
<td>0.0083 ± 0.00160</td>
<td>0.1922 ± 0.06399</td>
<td>0.0108</td>
</tr>
<tr>
<td>4</td>
<td>0.0081 ± 0.00178</td>
<td>0.1783 ± 0.05527</td>
<td>0.0107</td>
</tr>
<tr>
<td>5</td>
<td>0.0087 ± 0.00194</td>
<td>0.1644 ± 0.04570</td>
<td>0.0108</td>
</tr>
<tr>
<td>6</td>
<td>0.0082 ± 0.00156</td>
<td>0.1766 ± 0.05912</td>
<td>0.0092</td>
</tr>
<tr>
<td>7</td>
<td>0.0094 ± 0.00210</td>
<td>0.2134 ± 0.05468</td>
<td>0.0124</td>
</tr>
<tr>
<td>8</td>
<td>0.0088 ± 0.00162</td>
<td>0.2033 ± 0.04908</td>
<td>0.0125</td>
</tr>
<tr>
<td>9</td>
<td>0.0081 ± 0.00162</td>
<td>0.1723 ± 0.03762</td>
<td>0.0097</td>
</tr>
<tr>
<td>10</td>
<td>0.0091 ± 0.00222</td>
<td>0.1746 ± 0.05345</td>
<td>0.0121</td>
</tr>
</tbody>
</table>

**TABLE VII:** CIFAR-10: Mean bit error, mean correlation, ECOC error per fold

**TABLE VIII:** SVHN: Mean bit error, mean correlation, ECOC error per fold