This article has been accepted for publication in IEEE Transactions on Emerging Topics in Computing. 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/TETC.2023.3322563

Using Diversities to Model the Reliability of Two-version Machine Learning Systems

Fumio Machida, *Senior Member, IEEE*

***Abstract*— The N-version machine learning system (MLS) is an architectural approach to reduce error outputs from a system by redundant configuration using multiple machine learning (ML) modules. Improved system reliability achieved by N-version MLSs inherently depends on how diverse ML models are employed and how diverse input data sets are given. However, neither error input spaces of individual ML models nor input data distributions are obtainable in practice, which is a fundamental barrier to understanding the reliability improvement by N-version architectures. In this paper, we introduce two diversity measures quantifying the similarities of ML models' capabilities and the interdependence of input data sets causing errors, respectively. The defined measures are used to formulate the reliability of an elemental N-version MLS called dependent double-modules double-inputs MLS. The system is assumed to fail when two ML modules output errors simultaneously for the same classification task. The reliabilities of different architecture options for this MLS are comprehensively analyzed through a compact matrix representation form of the proposed reliability model. The theoretical analysis and numerical results show that the architecture exploiting two diversities achieves preferable reliability under reasonable assumptions. Intuitive relations between diversity parameters and architecture reliabilities are also demonstrated through numerical examples.**

***Index Terms*— Diversity, Machine learning system, Redundant architecture, Reliability, Software fault-tolerance**

I. Introduction

QUALITY assurance of machine learning systems (MLSs) is becoming a major concern of system providers who adopt advanced machine learning (ML) functions in their products or services. In the development of MLSs, desirable input-output relations are not fully specified in advance since the behavior of ML functions is determined by the samples used in the ML training process [21]. Correct outputs are not always guaranteed in real user environments, even if the trained ML functions achieve high accuracy for the testing data set. Therefore, understanding the reliability consequence of error outputs from an ML function is crucial in MLS design, particularly for safety-critical domains such as autonomous vehicles [1].

The N-version machine learning system is an approach to improving the reliability of system outputs by introducing redundant architecture [2]. The idea is analogous to the traditional software fault-tolerant technique referred to as N-version programming that employs $N \geq 2$ functionally-equivalent independent programs from the same initial specification [3][4]. An MLS may employ N different ML models that are trained independently for the same task. Since individual ML models output errors differently, the probability of simultaneous errors from the N-version MLS can be reduced. To make an N-version MLS effective, it is essential to include diverse versions of ML models. Compared to software programs, multiple versions of ML models can be generated easily with much smaller costs by using different algorithms, training data, and hyperparameters [5]. Moreover, outputs from a single ML model can also be diversified by perturbating input data for inference [6][18]. We can obtain diverse outputs not only from different ML models but also from a single ML model by changing input data for inference. Both the diversity among ML models and the diversity among input data sets significantly affect the reliability of the system output. Nevertheless, modeling the reliability of N-version MLS with two diversities is still underexplored [2].

In this paper, we present an analytical model for characterizing the reliability of output from a basic N-version MLS composed of two ML modules for classification tasks and two sensors generating different input data. An ML module installs an ML model to classify the input data and is connected to either one of the sensors. The MLS outputs classification results when two ML modules agree with the output. We assume that the errors of the two ML models are largely similar, and the probability distributions of the two input data are not independent of each other. A representative example of such a system is a perception system in an autonomous vehicle equipped with two image sensors and two image classifiers. We call this type of system *dependent double-modules double-inputs MLS*, whose applications are not limited to autonomous driving. The challenge in the reliability modeling for dependent double-modules double-inputs MLSs resides in the statistical dependence between ML inference errors. The reliability formulation is simple if we can assume two outputs from two modules are statistically independent. However, such an assumption does not hold in practice since the capabilities of ML models have a certain similarity, and the input data from sensors also have similarities. Therefore, we introduce two diversity measures, *conjunction of errors* and *intersection of errors*, representing the similarity of input data causing errors and the similarity of error input spaces for two ML models, respectively.

We use the diversity measures to formulate the reliability of a dependent double-modules double-inputs MLS. Depending on the choice of input data and ML model for individual ML

This work was supported in part by a grant of JSPS KAKENHI Grant Number JP19K24337 and JP22K17871.

F. Machida is with the Department of Computer Science, University of Tsukuba, Tsukuba, Japan, (e-mail: machida@cs.tsukuba.ac.jp).

modules, there can be six architecture options (shown in Figure 2 in Section II-C). The presented reliability model is subsequently used to show some properties of the reliabilities of different architectures under specific assumptions on the relation between input data distribution and error input spaces. Finally, we conduct numerical experiments to evaluate the reliability of dependent double-modules double-inputs MLSs in a hypothetical setting. Our numerical results show the advantage of two-input architectures compared with a conventional modular redundancy scheme. The result also gives an intuitive view of the properties derived from the proposed model.

The contributions are summarized as follows.

1. We propose a new reliability model for a basic N-version MLS by defining two diversity measures that characterize the dependencies of errors by two input data and the two error input spaces attributed by corresponding ML models. Our formulation with a compact matrix representation gives a useful tool to investigate the properties of architectures' reliability in relation to the diversity measures.
2. By restricting the type of input data distributions on error input spaces, we show some important properties that can guide the choice of the preferable architecture in terms of system output reliability. Except for limiting cases, our analysis implies that the architecture exploiting two diversities tends to achieve preferable reliability to the architectures relying on a single diversity.
3. We provide the results of numerical experiments that help understand the dependencies between two input data and two ML models for computing the reliability of a dependent double-modules double-inputs MLS. The experimental results confirm the properties derived from the proposed model.

The rest of the paper is organized as follows. Section II introduces a motivating example and specifies the problem scope; the reliability analysis of a dependent double-modules double-inputs MLS. To compare six possible architecture options, in Section III, we present the diversity measures and use them to formulate the reliabilities of MLS outputs. Section IV shows some important properties we can derive from the constructed reliability models. Section V shows the results of numerical experiments to show the difference in reliabilities achieved by our redundant configuration scheme. Section VI describes a potential application scenario and limitations of the work. Section VII discusses related work, and finally, Section VIII gives conclusions and potential future studies.

II. PRELIMINARIES

In this section, we first show a motivating example and present the reliability design issue. Then we define our problem scope and introduce the notations used throughout the paper.

A. A motivating example

As an example of MLS applications, we consider an autonomous vehicle equipped with image sensors and ML modules to classify input images. For safe autonomous driving on the road, the system needs to recognize the traffic signs, signals, other vehicles, and other obstacles using the equipped sensors. Relying on a single sensor and a single ML function is not encouraged since the output of an ML module is highly error-prone to the samples from the real world [7][8]. Thus, N-version MLSs can be adopted to improve the reliability of system output. Figure 1 shows a scenario in which an autonomous vehicle needs to recognize the traffic signal in front of the car using two cameras and two ML modules.

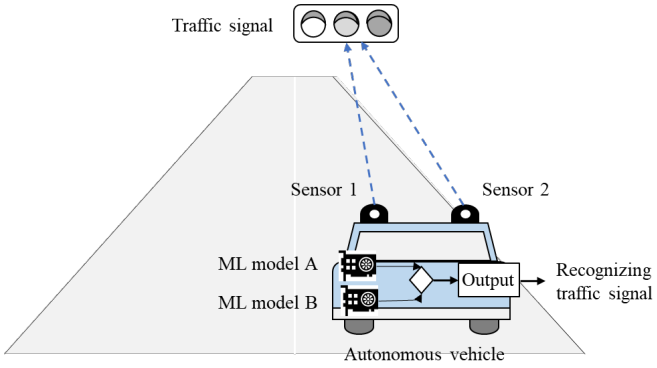

Figure 1. An example scenario for an autonomous vehicle on the road.

The cameras can generate different images from slightly different angles. The ML models can behave differently on the same input image by using different ML algorithms or training data sets. When the outputs of these ML modules disagree, the system does not output any inference results to avoid errors. In other words, the system outputs an error only when both modules agree on the inference result but the answer is incorrect.

The design issue we encounter in configuring such a dependent double-modules double-inputs MLS is the combination of input data and ML models for individual modules. A wrong combination of input data and ML model even may decrease the system output reliability. In order to discuss this design issue and clarify our problem scope, we introduce formal notations in the following section.

B. Reliability of MLS

We define the reliability of MLS as the probability that an MLS output agrees with the ground truth in the real world (e.g., red signal). Unlike the accuracy measures of classification tasks, we do not distinguish false positives from false negatives. Errors can also be caused by the implementation bugs of ML programs [9][15]. Any outputs that do not match the ground truth are considered errors causing unreliable system outputs.

To model the reliability of MLS, we introduce the following notations. Let $x_i,\ i = \{1,2,\dots\}$ be an input data from a sensor i and let $m_j, j = \{a, b, \dots\}$ be a ML model. An ML module is a unit of the MLS, which installs one ML model m_j and selects one input data source i. The ML module outputs errors when the installed ML model does not classify the input data correctly. Let S be the total set of possible input data, and let $E_j \subset S$ be the subset of S that makes ML model m_j outputs errors. The probability that the module outputs an error can be represented by $P[x_i \in E_j]$. Therefore, the reliability of MLS using this ML module solely is given by $1 - P[x_i \in E_j]$. Throughout this

This article has been accepted for publication in IEEE Transactions on Emerging Topics in Computing. 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/TETC.2023.3322563

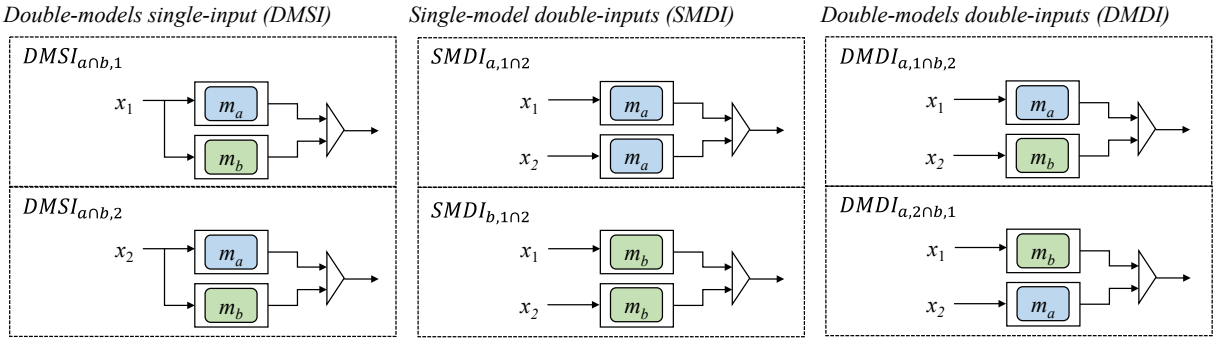

Figure 2. Architecture options for dependent double-modules double-inputs MLS.

paper, we assume $P[x_i \in E_j] \in (0,1)$ unless otherwise stated. The possibility that the input x_i occurs in a real environment can also depend on the sensor's capability. To consider the randomness of sensor input, denote X_i as the random variable representing the input data x_i and define $\mu_{X_i}(x_i)$ as the corresponding distribution function. When we define error function f_j as

$$f_j(x_i) = \begin{cases} 1, & x_i \in E_j \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

the error probability can be expressed as

$$P[x_i \in E_j] = \int f_j(x_i) d\mu_{X_i}(x_i). \tag{2}$$

Note that $P[x_i \in E_j]$ is the expectation of the value of f_j for given input data distribution, and the value should be in $(0,1)$. f_j is also called score function and is used for representing the reliability of multi-version software systems [37][38]. We use these notations in the following reliability models.

C. Double-modules double-inputs MLS

With the reliability function defined above, next, we formulate the reliability of MLS with redundant configuration. Given two different sensor inputs x_1 and x_2 and two different ML models m_a and m_b, there could be six different architecture choices which are categorized into three types, as shown in Figure 2. Note that no matter which architecture is adopted, the final output of the MLS is determined by voting on two modules' outputs. Voting strategies are commonly adopted in multi-version machine learning systems [5][26][36]. In our study, we assume that the MLS outputs the prediction results only when two modules output an identical result. Therefore, we consider the reliability of system output by the probability that both two modules' outputs are correct. If any one of the modules outputs an error, the voter cannot judge which output is correct and hence discard all the outputs. The system may issue an alert to higher-level modules or users when encountering such conditions several times. We do not consider this case as a system failure in this paper. The reliabilities of the six architectures are characterized by the input data distributions and error functions as formulated below.

1. *Double-models single-input (DMSI) architecture*

In this architecture, two ML modules employ different ML models while choosing the same input data. Depending on which input data (x_1 or x_2) is chosen, there are two architecture options. The reliabilities of these architectures can be represented by

$$\begin{aligned} R_{DMSI_{a\cap b,1}} &= 1 - P[x_1 \in E_a \cap E_b], \\ R_{DMSI_{a\cap b,2}} &= 1 - P[x_2 \in E_a \cap E_b]. \end{aligned} \tag{3}$$

Given the input data distributions $\mu_{X_1}(x_1)$ and $\mu_{X_2}(x_2)$, the reliabilities can also be expressed as

$$\begin{aligned} R_{DMSI_{a\cap b,1}} &= 1 - \int f_a(x_1) f_b(x_1) d\mu_{X_1}(x_1), \\ R_{DMSI_{a\cap b,2}} &= 1 - \int f_a(x_2) f_b(x_2) d\mu_{X_2}(x_2). \end{aligned} \tag{4}$$

2. *Single-model double-inputs (SMDI) architecture*

In this architecture, both ML modules employ the same ML model but choose different sensor input data resulting in two potentially different outputs. Depending on which ML model (m_a or m_b) is employed, there are two architecture options. The reliabilities of these architectures can be represented by

$$\begin{aligned} R_{SMDI_{a,1\cap 2}} &= 1 - P[x_1 \in E_a, x_2 \in E_a], \\ R_{SMDI_{b,1\cap 2}} &= 1 - P[x_1 \in E_b, x_2 \in E_b]. \end{aligned} \tag{5}$$

Let $\mu_{X_1,X_2}(x_1, x_2)$ be the joint distribution function of two input data. The reliabilities can also be expressed as

$$\begin{aligned} R_{SMDI_{a,1\cap 2}} &= 1 - \int f_a(x_1) f_a(x_2) d\mu_{X_1,X_2}(x_1, x_2), \\ R_{SMDI_{b,1\cap 2}} &= 1 - \int f_b(x_1) f_b(x_2) d\mu_{X_1,X_2}(x_1, x_2). \end{aligned} \tag{6}$$

3. *Double-models double-inputs (DMDI) architecture*

This architecture uses two different inputs and two different ML models. Two ML modules deploy different ML models and choose different sensor inputs. Depending on the combination of input data and ML model, there are two architecture options in this case as well. The reliabilities of these architectures can be represented by

$$\begin{aligned} R_{DMDI_{a,1\cap b,2}} &= 1 - P[x_1 \in E_a, x_2 \in E_b], \\ R_{DMDI_{a,2\cap b,1}} &= 1 - P[x_2 \in E_a, x_1 \in E_b]. \end{aligned} \tag{7}$$

With the joint distribution function of two input data, the reliabilities can also be expressed as

$$\begin{aligned} R_{DMDI_{a,1\cap b,2}} &= 1 - \int f_a(x_1) f_b(x_2) d\mu_{X_1,X_2}(x_1, x_2), \\ R_{DMDI_{a,2\cap b,1}} &= 1 - \int f_b(x_1) f_a(x_2) d\mu_{X_1,X_2}(x_1, x_2). \end{aligned} \tag{8}$$

The architecture reliability comparison is easy if we can

assume the two modules' outputs are independent. When we know the reliability of individual modules' outputs, the reliability of redundant configuration can be simply computed by combinatorial reliability models such as reliability block diagram [11] or fault tree [12][13]. However, the independence assumptions on the input data distributions and the errors of different ML models are unlikely to hold in practice. Moreover, we hardly obtain complete knowledge about the degrees of dependence between two ML models and two input data sets [6]. The architecture reliability comparison under such conditions is not a trivial issue.

D. Problem

With the above notations of MLS architectures' reliability, the general problem we try to address in this paper can be described as follows.

MLS redundant configuration problem. *Given two different sensor inputs x_1 and x_2 with two different ML models m_a and m_b for the same classification task, determine the best or preferable architecture options in terms of system reliability without knowing the complete information about the input data distributions and error input sets of the ML models.*

Answers to this problem must give guides to choosing a suitable configuration of redundant MLS architecture. Since we do not have complete knowledge about the input data distributions or the errors of ML models, the best option may not be determined due to the lack of information. However, any additional information may help screen inappropriate options regarding system reliability. Finding preferable architecture options from any available information is practically meaningful in the design of reliable MLS.

III. Diversity measures and reliability model

Instead of directly approaching input data distributions and error spaces of ML models, we attempt to characterize the architecture reliability through the diversities among the modules. In this section, we define two diversity measures and formulate the architecture reliabilities in a compact matrix representation form.

A. Diversity measures

Input data from different sensors must be very similar since both sensors observe the same target. On the other hand, the error tendencies of different ML models must be similar since both the models are trained for the same task and possibly trained from the same data sets. The degree of these similarities must influence the system reliability and may bring useful information to determine the preferable architecture option. To incorporate the factor of dependence quantitatively in the reliability model, we introduce two measures of diversity.

Intersection of errors *(model similarity). Let E_a and E_b be the subsets of input space S that make machine learning models m_a and m_b output errors, respectively. For an input data x_i sampled from distribution $\mu_{X_i}(x_i)$, the intersection of errors $\alpha_{b|a,i}$ is defined by the conditional probability*

$$\alpha_{b|a,i} = P[x_i \in E_b | x_i \in E_a] = \frac{P[x_i \in E_a \cap E_b]}{P[x_i \in E_a]}, \quad \text{where } P[x_i \in E_a] > 0. \tag{9}$$

The intersection of errors represents the degree of overlap between the sets E_a and E_b. The larger the elements of E_b overlaps the elements of E_a, the value of $\alpha_{b|a,i}$ becomes larger. Since E_a and E_b are attributed to the capabilities of different ML models, their intersection indicates how two models resemble in terms of error input space. The smaller intersection decreases the probability of common errors of m_a and m_b, which corresponds to the smaller model similarity.

Conjunction of errors *(input similarity). Let x_1 and x_2 be the input data that follow distributions $\mu_{X_1}(x_1)$ and $\mu_{X_2}(x_2)$, respectively. For a machine learning model m_j whose error space is given by E_j, the conjunction of errors $\beta_{j,2|1}$ is defined by the conditional probability*

$$\beta_{j,2|1} = Pr[x_2 \in E_j | x_1 \in E_j] = \frac{P[x_1 \in E_j, x_2 \in E_j]}{P[x_1 \in E_j]}, \quad \text{where } P[x_1 \in E_j] > 0. \tag{10}$$

The conjunction of errors represents the possibility of both inputs x_1 and x_2 fall into error outputs by the process of the same ML model m_j. The larger similarity there is between distributions $\mu_{X_1}(x_1)$ and $\mu_{X_2}(x_2)$ in error space E_j, the value of $\beta_{j,2|1}$ becomes the larger. The difference in distributions can be regarded as the diversity of the sensors' capabilities. The smaller conjunction decreases the probability of double errors due to the smaller input similarity.

With the defined diversity measures, the reliabilities of DMSI and SMDI architectures can be expressed as

$$\begin{aligned} R_{DMSI_{a\cap b,1}} &= 1 - \alpha_{b|a,1} \cdot P[x_1 \in E_a] \\ &= 1 - \alpha_{a|b,1} \cdot P[x_1 \in E_b], \\ R_{SMDI_{a,1\cap 2}} &= 1 - \beta_{a,2|1} \cdot P[x_1 \in E_a] \\ &= 1 - \beta_{a,1|2} \cdot P[x_2 \in E_a]. \end{aligned} \tag{11}$$

From the expressions, we can derive the bounds of $R_{DMSI_{a\cap b,1}}$ and $R_{SMDI_{a,1\cap 2}}$ as follows.

$$\begin{aligned} 1 - \min(P[x_1 \in E_a], P[x_1 \in E_b]) \le R_{DMSI_{a\cap b,1}} \le 1, \\ 1 - \min(P[x_1 \in E_a], P[x_2 \in E_a]) \le R_{SMDI_{a,1\cap 2}} \le 1. \end{aligned} \tag{12}$$

The upper bound is given when the value of diversity is equal to zero, which means there is no intersection or conjunction of errors between the two modules. On the other hand, the lower bound is given when the value of diversity is equal to one, which means two models are identical or inputs are identical.

B. Reliability model for DMDI

Since DMDI architecture uses two sensor input data and two distinct ML models, both model diversity and input diversity can influence the architecture reliability. Let us first consider the reliability of $DMDI_{a,1\cap b,2}$. The probability that the two modules in $DMDI_{a,1\cap b,2}$ output errors simultaneously is given by $P[x_1 \in E_a, x_2 \in E_b]$. Assume that input x_1 causes an error of m_a, the error probability of MLS can be given by conditioning whether x_2 also causes an error of m_a.

$$P[x_1 \in E_a, x_2 \in E_b] = P[x_2 \in E_b | x_2 \in E_a, x_1 \in E_a] \cdot P[x_2 \in E_a | x_1 \in E_a] \cdot P[x_1 \in E_a] + P[x_2 \in E_b | x_2 \in \overline{E_a}, x_1 \in E_a] \cdot P[x_2 \in \overline{E_a} | x_1 \in E_a] \cdot P[x_1 \in E_a], \quad (13)$$

where $\bar{E} = S \setminus E$ represents the complementary set of E and $P[x_2 \in E_a | x_1 \in E_a] \in (0,1)$. The first term of the above expression corresponds to the probability that x_2 occurs in $E_a \cap E_b$, while the second term corresponds to the probability that x_2 occurs in $\overline{E_a} \cap E_b$. Define the parameters

$$\alpha_{b,2|a,1\cap 2} = P[x_2 \in E_b | x_2 \in E_a, x_1 \in E_a], \quad \alpha_{b,2|a,1\cap \bar{2}} = P[x_2 \in E_b | x_2 \in \overline{E_a}, x_1 \in E_a]. \quad (14)$$

Since $P[x_2 \in \overline{E_a} | x_1 \in E_a]$ complements $P[x_2 \in E_a | x_1 \in E_a]$, using the input diversity $\beta_{a,2|1}$ the reliability is expressed as

$$R_{DMDI_{a,1\cap b,2}} = 1 - \left[\alpha_{b,2|a,1\cap 2} \cdot \beta_{a,2|1} + \alpha_{b,2|a,1\cap \bar{2}} \cdot (1 - \beta_{a,2|1})\right] \cdot P[x_1 \in E_a]. \quad (15)$$

If E_a and E_b are identical, $\alpha_{b,2|a,1\cap 2} = 1$ and $\alpha_{b,2|a,1\cap \bar{2}} = 0$, which results in $R_{DMDI_{a,1\cap b,2}} = R_{SMDI_{a,1\cap 2}}$. Expression (15) characterizes the reliability of $DMDI_{a,1\cap b,2}$ by the combination of model similarity and input similarity. We discuss the relationship between different architecture options through the set of diversity-associated parameters in the following section.

C. Matrix representation

In the derivation of the reliability of $DMDI_{a,1\cap b,2}$ above, we condition the error probability by the occurrence of error conjunction with $x_1 \in E_a$. Applying the same conditioning to the expression for the reliability of $DMSI_{a\cap b,1}$, we have

$$P[x_1 \in E_a, x_1 \in E_b] = P[x_1 \in E_b | x_2 \in E_a, x_1 \in E_a] \cdot \mathrm{P}[x_2 \in E_a | x_1 \in E_a] \cdot P[x_1 \in E_a] + P[x_1 \in E_b | x_2 \in \overline{E_a}, x_1 \in E_a] \cdot \mathrm{P}[x_2 \in \overline{E_a} | x_1 \in E_a] \cdot P[x_1 \in E_a], \quad (16)$$

where $P[x_2 \in E_a | x_1 \in E_a] \in (0,1)$. When we define

$$\alpha_{b,1|a,1\cap 2} = P[x_1 \in E_b | x_2 \in E_a, x_1 \in E_a], \quad \alpha_{b,1|a,1\cap \bar{2}} = P[x_1 \in E_b | x_2 \in \overline{E_a}, x_1 \in E_a], \quad (17)$$

the reliability of $DMSI_{a\cap b,1}$ can be expressed as

$$R_{DMSI_{a\cap b,1}} = 1 - \left[\alpha_{b,1|a,1\cap 2} \cdot \beta_{a,2|1} + \alpha_{b,1|a,1\cap \bar{2}} \cdot (1 - \beta_{a,2|1})\right] \cdot P[x_1 \in E_a]. \quad (18)$$

In a similar manner, the reliabilities of $DMDI_{a,2\cap b,1}$ and $DMSI_{a\cap b,2}$ are derived by conditioning whether error conjunction occurs at $x_2 \in E_a$.

$$R_{DMDI_{a,2\cap b,1}} = 1 - \left[\alpha_{b,1|a,1\cap 2} \cdot \beta_{a,1|2} + \alpha_{b,1|a,\bar{1}\cap 2} \cdot (1 - \beta_{a,1|2})\right] \cdot P[x_2 \in E_a], \quad (19)$$

$$R_{DMSI_{a\cap b,2}} = 1 - \left[\alpha_{b,2|a,1\cap 2} \cdot \beta_{a,1|2} + \alpha_{b,2|a,\bar{1}\cap 2} \cdot (1 - \beta_{a,1|2})\right] \cdot P[x_2 \in E_a], \quad (20)$$

where

$$\alpha_{b,1|a,\bar{1}\cap 2} = P[x_1 \in E_b | x_1 \in \overline{E_a}, x_2 \in E_a], \quad \alpha_{b,2|a,\bar{1}\cap 2} = P[x_2 \in E_b | x_1 \in \overline{E_a}, x_2 \in E_a]. \quad (21)$$

The expressions (15)(18)(19)(20) can be merged into the following matrix representation.

$$\boldsymbol{R}_{b|a} = \boldsymbol{J}_2 - \boldsymbol{A}_{b|a} \cdot \boldsymbol{B}_a^{\top} \cdot \boldsymbol{P}_a, \quad (22)$$

where

$$\boldsymbol{R}_{b|a} = \begin{bmatrix} R_{DMSI_{a\cap b,1}} & R_{DMDI_{a,2\cap b,1}} \\ R_{DMDI_{a,1\cap b,2}} & R_{DMSI_{a\cap b,2}} \end{bmatrix}, \quad \boldsymbol{J}_2 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \boldsymbol{A}_{b|a} = \begin{bmatrix} \alpha_{b,1|a,1\cap 2} & \alpha_{b,1|a,1\cap \bar{2}} & \alpha_{b,1|a,\bar{1}\cap 2} \\ \alpha_{b,2|a,1\cap 2} & \alpha_{b,2|a,1\cap \bar{2}} & \alpha_{b,2|a,\bar{1}\cap 2} \end{bmatrix},$$

$$\boldsymbol{B}_a = \begin{bmatrix} \beta_{a,2|1} & 1 - \beta_{a,2|1} & 0 \\ \beta_{a,1|2} & 0 & 1 - \beta_{a,1|2} \end{bmatrix}, \quad \boldsymbol{P}_a = \begin{bmatrix} P[x_1 \in E_a] & 0 \\ 0 & P[x_2 \in E_a] \end{bmatrix}.$$

The matrix $\boldsymbol{R}_{b|a}$ represents the reliabilities of four different architectures. $\boldsymbol{J}_2$ is a two-dimensional all-ones matrix. The elements of $\boldsymbol{A}_{b|a}$ are associated with the intersection of errors in E_b over E_a, while the elements of $\boldsymbol{B}_a$ represent the conjunction or non-conjunction of errors in E_a, and $\boldsymbol{P}_a$ represents the error probabilities on E_a by inputs x_1 and x_2.

The matrix term $\boldsymbol{B}_a^{\top} \cdot \boldsymbol{P}_a$ in (22) can be represented by the functions of $R_{SMDI_{a,1\cap 2}}$ as follows.

$$\boldsymbol{B}_a^{\top} \cdot \boldsymbol{P}_a = \begin{bmatrix} \beta_{a,2|1} \cdot P[x_1 \in E_a] & \beta_{a,1|2} \cdot P[x_2 \in E_a] \\ (1 - \beta_{a,2|1}) \cdot P[x_1 \in E_a] & 0 \\ 0 & (1 - \beta_{a,1|2}) \cdot P[x_2 \in E_a] \end{bmatrix} = \begin{bmatrix} 1 - R_{SMDI_{a,1\cap 2}} & 1 - R_{SMDI_{a,1\cap 2}} \\ P[x_1 \in E_a] + R_{SMDI_{a,1\cap 2}} - 1 & 0 \\ 0 & P[x_2 \in E_a] + R_{SMDI_{a,1\cap 2}} - 1 \end{bmatrix}. \quad (23)$$

Therefore, the expression (22) characterizes the relationships among the reliabilities of five different architectures out of six options (i.e., only $R_{SMDI_{b,1\cap 2}}$ is not associated).

A similar derivation can be carried out by conditioning the error conjunction or non-conjunction in E_b. As a result, the dual of $\boldsymbol{R}_{b|a}$ can be obtained by exchanging E_a and E_b in (22).

$$\boldsymbol{R}_{a|b} = \boldsymbol{J}_2 - \boldsymbol{A}_{a|b} \cdot \boldsymbol{B}_b^{\top} \cdot \boldsymbol{P}_b,$$

where

$$\boldsymbol{R}_{a|b} = \begin{bmatrix} R_{DMSI_{a\cap b,1}} & R_{DMDI_{a,1\cap b,2}} \\ R_{DMDI_{a,2\cap b,1}} & R_{DMSI_{a\cap b,2}} \end{bmatrix}, \quad \boldsymbol{A}_{a|b} = \begin{bmatrix} \alpha_{a,1|b,1\cap 2} & \alpha_{a,1|b,1\cap \bar{2}} & \alpha_{a,1|b,\bar{1}\cap 2} \\ \alpha_{a,2|b,1\cap 2} & \alpha_{a,2|b,1\cap \bar{2}} & \alpha_{a,2|b,\bar{1}\cap 2} \end{bmatrix}, \quad (24)$$

$$\boldsymbol{B}_b = \begin{bmatrix} \beta_{b,2|1} & 1 - \beta_{b,2|1} & 0 \\ \beta_{b,1|2} & 0 & 1 - \beta_{b,1|2} \end{bmatrix}, \quad \boldsymbol{P}_b = \begin{bmatrix} P[x_1 \in E_b] & 0 \\ 0 & P[x_2 \in E_b] \end{bmatrix}.$$

Since expression (24) includes the term

$$\boldsymbol{B}_b^{\top} \cdot \boldsymbol{P}_b = \begin{bmatrix} \beta_{b,2|1} \cdot P[x_1 \in E_b] & \beta_{b,1|2} \cdot P[x_2 \in E_b] \\ (1 - \beta_{b,2|1}) \cdot P[x_1 \in E_b] & 0 \\ 0 & (1 - \beta_{b,1|2}) \cdot P[x_2 \in E_b] \end{bmatrix} = \begin{bmatrix} 1 - R_{SMDI_{b,1\cap 2}} & 1 - R_{SMDI_{b,1\cap 2}} \\ P[x_1 \in E_b] + R_{SMDI_{b,1\cap 2}} - 1 & 0 \\ 0 & P[x_2 \in E_b] + R_{SMDI_{b,1\cap 2}} - 1 \end{bmatrix}. \quad (25)$$

the relationships among the reliabilities of five different

architectures out of six options are represented by expression (24) (i.e., only $R_{SMDI_{a,1\cap 2}}$ is not associated). By definition, $\boldsymbol{R}_{b|a}$ and $\boldsymbol{R}_{a|b}$ satisfy the following relation.

$$\boldsymbol{R}_{a|b} = \boldsymbol{R}_{b|a}^{\top} \tag{26}$$

Consequently, the relationships between the reliabilities of six different architectures can be characterized by diversity-property matrixes $\boldsymbol{A}_{b|a}, \boldsymbol{A}_{a|b}, \boldsymbol{B}_a$ and $\boldsymbol{B}_b$ under the relation (26).

In the above derivation, for a given event $x_1 \in E_a$, first, we assume error conjunctions in E_a and then consider the intersection to E_b. One can also derive a similar relationship in inverse order; i.e., first assume that the error occurs in the intersection of E_a and E_b and then consider the conjunction of errors by another input. The reliability matrices can be represented as $\boldsymbol{R}_{2|1}$ and $\boldsymbol{R}_{1|2}$ which satisfy the relation $\boldsymbol{R}_{1|2} = \boldsymbol{R}_{2|1}^{\top}$. Since the derivation is mostly the same as described above, we omit to show this case for brevity.

Both the equalities $\boldsymbol{R}_{a|b} = \boldsymbol{R}_{b|a}^{\top}$ and $\boldsymbol{R}_{1|2} = \boldsymbol{R}_{2|1}^{\top}$ can fully represent the relations among six architectures' reliabilities and diversity-associated parameters.

IV. Architecture Property Analysis

Using the reliability model shown in expressions (22)(24)(26), we discuss the properties of these architectures that may help the decision of relevant architecture choice in terms of MLS reliability. The exact values of diversity parameters such as $\boldsymbol{A}_{b|a}$ and $\boldsymbol{B}_a$ are not obtainable in practice. However, we can argue a preferable architecture through the analysis of the relation of these diversity values. In the following discussion, we first show the general property derived from the model and then present some special properties that can be shown under a specific type of joint distribution for the input data set.

A. General properties

An interesting question about the presented architecture model is which architecture achieves better reliability than another architecture under specific conditions on the diversity metrics. In other words, it is interesting to know if the value of the diversity metric can determine the preference of the architecture in terms of reliability. To investigate this, first, consider the reliabilities difference between the elements of $\boldsymbol{R}_{b|a}$ and $R_{SMDI_{a,1\cap 2}}$. Define the reliability difference matrix by

$$\boldsymbol{H}_{b|a} = \boldsymbol{R}_{b|a} - R_{SMDI_{a,1\cap 2}} \cdot \boldsymbol{J}_2. \tag{27}$$

By definition, $R_{SMDI_{a,1\cap 2}}$ can be expressed either by $\beta_{a,2|1} \cdot P[x_1 \in E_a]$ or by $\beta_{a,1|2} \cdot P[x_2 \in E_a]$. We can write

$$R_{SMDI_{a,1\cap 2}} \cdot \boldsymbol{J}_2 = \boldsymbol{J}_2 - \boldsymbol{B}_a^{+\top} \cdot \boldsymbol{P}_a, \tag{28}$$

where

$$\boldsymbol{B}_a^{+} = [\beta_{a,2|1} \quad \beta_{a,1|2}]^{\top} \cdot [1 \quad 1] = \begin{bmatrix} \beta_{a,2|1} & \beta_{a,2|1} \\ \beta_{a,1|2} & \beta_{a,1|2} \end{bmatrix}. \tag{29}$$

Applying (22) and (28) to (27),

$$\begin{aligned} \boldsymbol{H}_{b|a} &= \boldsymbol{J}_2 - \boldsymbol{A}_{b|a} \cdot \boldsymbol{B}_a^{\top} \cdot \boldsymbol{P}_a - (\boldsymbol{J}_2 - \boldsymbol{B}_a^{+\top} \cdot \boldsymbol{P}_a) \\ &= \left(\boldsymbol{B}_a^{+\top} - \boldsymbol{A}_{b|a} \cdot \boldsymbol{B}_a^{\top}\right) \cdot \boldsymbol{P}_a. \end{aligned} \tag{30}$$

The expression (30) shows that the conditions where DMDI and DMSI architectures achieve higher reliability than $R_{SMDI_{a,1\cap 2}}$ can be provided by the sign of $\boldsymbol{B}_a^{+\top} - \boldsymbol{A}_{b|a} \cdot \boldsymbol{B}_a^{\top}$. Thus, we have the general property which characterizes the reliability difference of the architectures as described below.

Lemma 1. *Given a parameter matrix $\boldsymbol{A}_{b|a}$, the reliabilities of $DMSI_{a\cap b,1}$ and $DMDI_{a,1\cap b,2}$ monotonically increase against the reliability of $SMDI_{a,1\cap 2}$ as $\beta_{a,2|1}$ increases, while the reliabilities of $DMDI_{a,2\cap b,1}$ and $DMSI_{a\cap b,2}$ monotonically increase against the reliability of $SMDI_{a,1\cap 2}$ as $\beta_{a,1|2}$ increases.*

The proof of the lemma is presented in Appendix. The lemma implies that the architectures employing double models (i.e., DMDI and DMSI) tend to be preferable compared with the single model architecture (i.e., SMDI) when input x_1 has higher conjunction with x_2 in E_a. Note that the condition where the double-models architecture becomes preferable to single-model architecture is given by $\boldsymbol{H}_{b|a} > 0$. In order to investigate the condition, we need to understand the lower and upper bounds of $\boldsymbol{H}_{b|a}$ with respect to $\beta_{a,2|1}$ and $\beta_{a,1|2}$, which are given in the following lemma.

Lemma 2. *Given a parameter matrix $\boldsymbol{A}_{b|a}$, the lower and upper bounds of $\boldsymbol{H}_{b|a}$ as the functions of $\beta_{a,2|1}, \beta_{a,1|2} \in (0,1)$ are given by*

$$\inf \boldsymbol{H}_{b|a} = \begin{bmatrix} -\alpha_{b,1|a,1\cap\bar{2}} & -\alpha_{b,1|a,\bar{1}\cap 2} \\ -\alpha_{b,2|a,1\cap\bar{2}} & -\alpha_{b,2|a,\bar{1}\cap 2} \end{bmatrix} \cdot \boldsymbol{P}_a,$$

$$\sup \boldsymbol{H}_{b|a} = \begin{cases} \boldsymbol{H}_{b|a}^{\sup 1} \cdot \boldsymbol{P}_a, & P[x_2 \in E_a] < P[x_1 \in E_a] \\ \boldsymbol{H}_{b|a}^{\sup 2} \cdot \boldsymbol{P}_a, & P[x_2 \in E_a] \geq P[x_1 \in E_a], \end{cases}$$

$$\boldsymbol{H}_{b|a}^{\sup 1} = \begin{bmatrix} \breve{\alpha}_1 \cdot \dfrac{P[x_2 \in E_a]}{P[x_1 \in E_a]} - \alpha_{b,1|a,1\cap\bar{2}} \cdot \dfrac{P[x_2 \in \overline{E_a}]}{P[x_1 \in E_a]} & \breve{\alpha}_1 \\ \breve{\alpha}_2 \cdot \dfrac{P[x_2 \in E_a]}{P[x_1 \in E_a]} - \alpha_{b,2|a,1\cap\bar{2}} \cdot \dfrac{P[x_2 \in \overline{E_a}]}{P[x_1 \in E_a]} & \breve{\alpha}_2 \end{bmatrix},$$

$$\boldsymbol{H}_{b|a}^{\sup 2} = \begin{bmatrix} \breve{\alpha}_1 & \breve{\alpha}_1 \cdot \dfrac{P[x_1 \in E_a]}{P[x_2 \in E_a]} - \alpha_{b,1|a,\bar{1}\cap 2} \cdot \dfrac{P[x_1 \in \overline{E_a}]}{P[x_2 \in E_a]} \\ \breve{\alpha}_2 & \breve{\alpha}_2 \cdot \dfrac{P[x_1 \in E_a]}{P[x_2 \in E_a]} - \alpha_{b,2|a,\bar{1}\cap 2} \cdot \dfrac{P[x_1 \in \overline{E_a}]}{P[x_2 \in E_a]} \end{bmatrix},$$

where $\breve{\alpha}_1 = 1 - \alpha_{b,1|a,1\cap 2}$ and $\breve{\alpha}_2 = 1 - \alpha_{b,2|a,1\cap 2}$. (31)

The proof is presented in Appendix. Since all the elements of the lower bound $\inf \boldsymbol{H}_{b|a}$ are negative, double-model architectures achieve lower reliabilities than the single-model architecture at the lower bound. By Lemma 1, the values of $\boldsymbol{H}_{b|a}$ increases monotonically in terms of $\beta_{a,2|1}$ and $\beta_{a,1|2}$, the sign of $\sup \boldsymbol{H}_{b|a}$ determines changes in the preference. From Lemma 1 and Lemma 2, we have the following proposition.

Proposition 1. *Given a parameter matrix $\boldsymbol{A}_{b|a}$, when (i,j)-element of $\sup \boldsymbol{H}_{b|a}$ is positive, there exists a unique changing point in the increasing value of $\beta_{a,2|1}$ (when j=1) or $\beta_{a,1|2}$ (when j=2) at which the reliability of $SMDI_{a,1\cap 2}$ becomes lower than the reliability of $DMSI_{a\cap b,1}$ (i,j=1), $DMDI_{a,1\cap b,2}$ (i=2, j=1), $DMDI_{a,2\cap b,1}$ (i=1, j=2), or $DMSI_{a\cap b,2}$ (i,j=2).*

The values of $\beta_{a,2|1}$ and $\beta_{a,1|2}$ that give the changing points

satisfy $\boldsymbol{H}_{b|a} = 0$. From (10), the condition can also be derived from the equation $\boldsymbol{B}_a^{+\top} = \boldsymbol{A}_{b|a} \cdot \boldsymbol{B}_a^{\top}$.

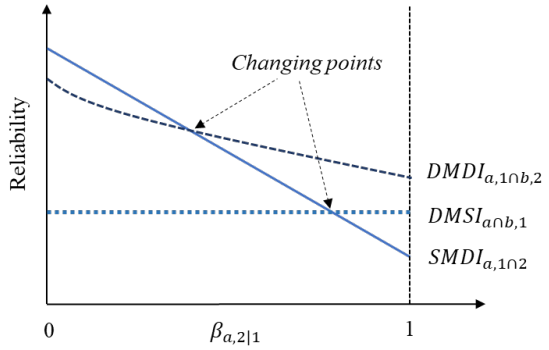

Figure 3. The example shows that the reliability of $SMDI_{a,1\cap 2}$ decreases as the value of $\beta_{a,2|1}$ increase.

Figure 3 shows an illustrative example of Proposition 1. The reliability of $SMDI_{a,1\cap 2}$ decreases by the increased $\beta_{a,2|1}$, representing the higher conjunctions of errors. While $SMDI_{a,1\cap 2}$ achieves the highest reliability when $\beta_{a,2|1}$ is low, it drops to the least option when $\beta_{a,2|1}$ is close to one. The implication is that $SMDI_{a,1\cap 2}$ is a preferable architecture when we can obtain any confidence in the small value of $\beta_{a,2|1}$ (i.e., error conjunction of inputs x_2 with inputs x_1 seldom occurs).

A similar property can be derived from the elements of $\boldsymbol{R}_{a|b}$ in comparison with $R_{SMDI_{b,1\cap 2}}$. Define the reliability difference matrix by

$$\boldsymbol{H}_{a|b} = \boldsymbol{R}_{a|b} - R_{SMDI_{b,1\cap 2}} \cdot \boldsymbol{J}_2. \tag{32}$$

The matrix can be expressed as

$$\boldsymbol{H}_{a|b} = \left(\boldsymbol{B}_b^{+\top} - \boldsymbol{A}_{a|b} \cdot \boldsymbol{B}_b^{\top}\right) \cdot \boldsymbol{P}_b,$$
$$\boldsymbol{B}_b^{+} = [\beta_{b,2|1} \quad \beta_{b,1|2}]^{\top} \cdot [1 \quad 1] = \begin{bmatrix} \beta_{b,2|1} & \beta_{b,2|1} \\ \beta_{b,1|2} & \beta_{b,1|2} \end{bmatrix}. \tag{33}$$

Note that (33) represents the same relation with (30), only by substituting a with b. Consequently, the duals of Lemma 1, Lemma 2, and Proposition 1 can be shown as well.

Furthermore, the difference in the elements of $\boldsymbol{R}_{2|1}$ and $DMSI_{a\cap b,1}$ can be characterized by the reliability difference matrix $\boldsymbol{H}_{2|1} = \boldsymbol{R}_{2|1} - DMSI_{a\cap b,1} \cdot \boldsymbol{J}_2$, while the difference in the elements of $\boldsymbol{R}_{1|2}$ and $DMSI_{a\cap b,2}$ can be characterized by the reliability difference matrix $\boldsymbol{H}_{1|2} = \boldsymbol{R}_{1|2} - DMSI_{a\cap b,2} \cdot \boldsymbol{J}_2$. For $\boldsymbol{H}_{2|1}$ and $\boldsymbol{H}_{1|2}$, by similar derivation steps, we can show the monotonicity, the bounds, and the changing points in terms of $\alpha_{b|a,1}$, $\alpha_{a|b,1}$, $\alpha_{b|a,2}$ and $\alpha_{a|b,2}$.

B. *Properties under restricted distributions*

The previous section shows the general property of the difference in architectures' reliabilities regarding diversity metrics without restricting the type of input data distributions. However, Proposition 1 is still insufficient to judge the preferable architecture for given conditions on the diversity measures. For example, the preference between $DMDI_{a,1\cap b,2}$ and $DMDI_{a,2\cap b,1}$ cannot be characterized through the property of either $\boldsymbol{H}_{b|a}$ or $\boldsymbol{H}_{a|b}$. To derive any useful information to select the architecture, we may need more information about the input data distributions. In the following, we consider three special cases where additional assumptions restricting the type of input data distributions help architecture selection.

1) *Case 1: Input superiority is known*

First, we assume that input data X_2 is more error-prone than X_1 for any error space E_j, and clarify the conditions on diversity measures to decide the preferable architecture in this case. The assumption is likely to hold in practice when there is a non-trivial difference in sensors' capabilities (e.g., an old sensor always induces more errors than the new sensor). The assumption can formally be described as, for any subset $E^* \subseteq E_a \cup E_b$, the joint distribution $\mu_{X_1,X_2}(x_1, x_2)$ satisfies $P[x_1 \in E^*] \leq P[x_2 \in E^*]$. Under this assumption, we have relations

$$P[x_1 \in E_a] \leq P[x_2 \in E_a],$$
$$P[x_1 \in E_b] \leq P[x_2 \in E_b]. \tag{34}$$

The above two inequalities also yield relations

$$P[x_1 \in E_a \cap E_b] \leq P[x_1 \in E_a, x_2 \in E_b] \leq P[x_2 \in E_a \cap E_b],$$
$$P[x_1 \in E_a \cap E_b] \leq P[x_1 \in E_b, x_2 \in E_a] \leq P[x_2 \in E_a \cap E_b]. \tag{35}$$

Under this assumption, the architecture reliabilities satisfy

$$R_{DMSI_{a\cap b,1}} \geq \left(R_{DMDI_{a,1\cap b,2}}, R_{DMDI_{a,2\cap b,1}}\right) \geq R_{DMSI_{a\cap b,2}}. \tag{36}$$

The inequality indicates that $DMDI_{a,1\cap b,2}$, $DMDI_{a,2\cap b,1}$ and $DMSI_{a\cap b,2}$ are not considered the best architecture option regardless of the values of diversity measures. Then, the question is which architecture would be the best among $DMSI_{a\cap b,1}$, $SMDI_{a,1\cap 2}$ and $SMDI_{b,1\cap 2}$. The answer to this question is provided in the following proposition.

Proposition 2. *Assume that the joint distribution $\mu_{X_1,X_2}(x_1, x_2)$ satisfies $P[x_1 \in E^*] \leq P[x_2 \in E^*]$ for any subset $E^* \subseteq E_a \cup E_b$. The most reliable architecture is given by either*

$$\begin{cases} SMDI_{a,1\cap 2}, & \text{if } \beta_{a,2|1} \leq \dfrac{\alpha_{b,1|a,1\cap\bar{2}}}{1 - \alpha_{b,1|a,1\cap 2} + \alpha_{b,1|a,1\cap\bar{2}}} \text{ and } \\ & \qquad \beta_{a,2|1} \leq \beta_{b,2|1} \cdot \dfrac{P[x_1 \in E_b]}{P[x_1 \in E_a]}, \\ SMDI_{b,1\cap 2}, & \text{if } \beta_{b,2|1} \leq \dfrac{\alpha_{a,1|b,1\cap\bar{2}}}{1 - \alpha_{a,1|b,1\cap 2} + \alpha_{a,1|b,1\cap\bar{2}}} \text{ and } \\ & \qquad \beta_{a,2|1} \geq \beta_{b,2|1} \cdot \dfrac{P[x_1 \in E_b]}{P[x_1 \in E_a]}, \\ DMSI_{a\cap b,1}, & \text{otherwise.} \end{cases} \tag{37}$$

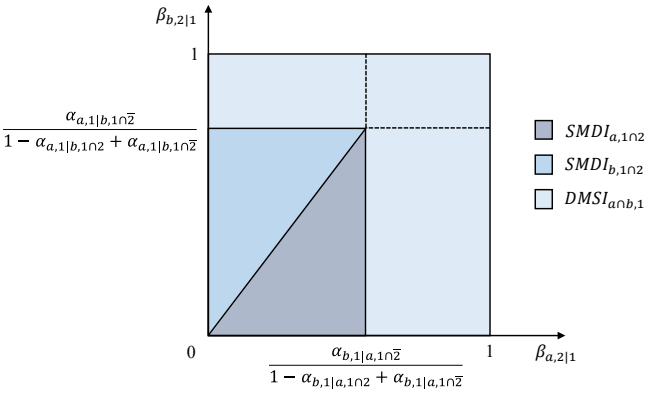

Figure 4. Boundary conditions of the most reliable architecture.

The proof of the lemma is presented in Appendix. Figure 4 visualizes the above conditions on $(\beta_{a,2|1}, \beta_{b,2|1})$ coordinates. As shown in Figure 4, intuitively $DMSI_{a\cap b,1}$ becomes the best architecture choice when we observe higher conjunction of errors (i.e., larger $\beta_{a,2|1}$ and $\beta_{b,2|1}$).

2) Case 2: Model superiority is known

Next, we assume that the errors in E_b occurs more frequently than the errors in E_a by any input data set and derive the conditions for deciding the preferable architecture in this case. The assumption restricts the shape of input data distributions, in particular over error spaces E_a and E_b, while it seems relatively acceptable in practice as the error frequencies are largely dominated by the ML models' capabilities. The assumption can be described formally as; for any input data x_i follow the input data distribution $\mu_{X_1}(x_1)$ or $\mu_{X_2}(x_2)$, it satisfies $P[x_i \in E_a] \leq P[x_i \in E_b]$. Under this assumption, we have the following relations.

$$\begin{aligned} P[x_1 \in E_a] &\leq P[x_1 \in E_b], \\ P[x_2 \in E_a] &\leq P[x_2 \in E_b]. \end{aligned} \tag{38}$$

The two inequalities also yield the relations

$$\begin{aligned} P[x_1 \in E_a, x_2 \in E_a] &\leq P[x_1 \in E_a, x_2 \in E_b] \\ &\leq P[x_1 \in E_b, x_2 \in E_b], \\ P[x_1 \in E_a, x_2 \in E_a] &\leq P[x_1 \in E_b, x_2 \in E_a] \\ &\leq P[x_1 \in E_b, x_2 \in E_b]. \end{aligned} \tag{39}$$

Under this assumption, the architecture reliabilities satisfy

$$R_{SMDI_{a,1\cap 2}} \geq \left(R_{DMDI_{a,1\cap b,2}}, R_{DMDI_{a,2\cap b,1}}\right) \geq R_{SMDI_{b,1\cap 2}}. \tag{40}$$

In this case, either $SMDI_{a,1\cap 2}$, $DMSI_{a\cap b,1}$ or $DMSI_{a\cap b,2}$ becomes the most reliable architecture, which can be determined by the following proposition. We omit the proof, but it can be shown using the matrix $\boldsymbol{H}_{2|1}$ and $\boldsymbol{H}_{1|2}$ in a similar derivation step as shown in the proof of Proposition 2.

Proposition 3. *Assume that $P[x_i \in E_a] \leq P[x_i \in E_b]$ satisfies for any input data x_i follow the input data distribution $\mu_{X_1}(x_1)$ or $\mu_{X_2}(x_2)$. The most reliable architecture is given by either*

$$\begin{cases} DMSI_{a\cap b,1}, & \text{if } \alpha_{b|a,1} \leq \dfrac{\beta_{a,2|a\cap\bar{b},1}}{1-\beta_{a,2|a\cap b,1}+\beta_{a,2|a\cap\bar{b},1}} \text{ and} \\ & \quad \alpha_{b|a,1} \leq \alpha_{b|a,2} \cdot \dfrac{P[x_2 \in E_a]}{P[x_1 \in E_a]}, \\ DMSI_{a\cap b,2}, & \text{if } \alpha_{b|a,2} \leq \dfrac{\beta_{a,1|a\cap\bar{b},2}}{1-\beta_{a,1|a\cap b,2}+\beta_{a,1|a\cap\bar{b},2}} \text{ and} \\ & \quad \alpha_{b|a,1} \geq \alpha_{b|a,2} \cdot \dfrac{P[x_2 \in E_a]}{P[x_1 \in E_a]}, \\ SMDI_{a,1\cap 2}, & \text{otherwise}, \end{cases}$$

where

$$\begin{aligned} \beta_{a,2|a\cap b,1} &= P[x_2 \in E_a | x_1 \in E_a, x_1 \in E_b], \\ \beta_{a,2|a\cap\bar{b},1} &= P[x_2 \in E_a | x_1 \in E_a, x_1 \in \overline{E_b}], \\ \beta_{a,1|a\cap b,2} &= P[x_1 \in E_a | x_2 \in E_a, x_2 \in E_b], \\ \beta_{a,1|a\cap\bar{b},2} &= P[x_1 \in E_a | x_2 \in E_a, x_2 \in \overline{E_b}]. \end{aligned} \tag{41}$$

Proposition 3 indicates that $SMDI_{a,1\cap 2}$ becomes the best architecture choice when we observe a higher intersection of errors (i.e., larger $\alpha_{b|a,1}$ and $\alpha_{b|a,2}$).

3) Case 3: Complementary models and inputs are given

In the above two possible scenarios with reasonable assumptions, we find that DMDI architectures are neither the best architecture nor the worst architecture choice. An interesting question is whether DMDI architectures can become the best option in terms of reliability under a reasonable assumption. The answer could be yes when the two input data distributions have a complementary relationship about the error-proneness of the different ML models. The assumption can be formally defined such that for any subsets $E_a^* \subseteq E_a$ and $E_b^* \subseteq E_b$, the joint distribution $\mu_{X_1,X_2}(x_1, x_2)$ satisfies $P[x_1 \in E_a^*] \leq P[x_2 \in E_a^*]$ and $P[x_2 \in E_b^*] \leq P[x_1 \in E_b^*]$. Under this assumption, we have

$$\begin{aligned} P[x_1 \in E_a] &\leq P[x_2 \in E_a], \\ P[x_2 \in E_b] &\leq P[x_1 \in E_b]. \end{aligned} \tag{42}$$

The two inequalities also yield the relations

$$\begin{aligned} P[x_1 \in E_a, x_2 \in E_b] &\leq P[x_1 \in E_a \cap E_b] \\ &\leq P[x_1 \in E_b, x_2 \in E_a], \\ P[x_1 \in E_a, x_2 \in E_b] &\leq P[x_2 \in E_a \cap E_b] \\ &\leq P[x_1 \in E_b, x_2 \in E_a]. \end{aligned} \tag{43}$$

Under this assumption, the architecture reliabilities satisfy

$$R_{DMDI_{a,1\cap b,2}} \geq \left(R_{DMSI_{a\cap b,1}}, R_{DMSI_{a\cap b,2}}\right) \geq R_{DMDI_{a,2\cap b,1}}. \tag{44}$$

Using $\boldsymbol{H}_{b|a}$ and $\boldsymbol{H}_{a|b}$, we can show the next proposition that clarifies the conditions where $DMDI_{a,1\cap b,2}$ can achieve the best reliability among other architectures.

Proposition 4. *Assume that the joint distribution $\mu_{X_1,X_2}(x_1, x_2)$ satisfies $P[x_1 \in E_a^*] \leq P[x_2 \in E_a^*]$ and $P[x_2 \in E_b^*] \leq P[x_1 \in E_b^*]$ for any subsets $E_a^* \subseteq E_a$ and $E_b^* \subseteq E_b$. The most reliable architecture is given by either one of the following.*

$$\begin{cases} SMDI_{a,1\cap 2}, & \text{if } \beta_{a,2|1} \leq \dfrac{\alpha_{b,2|a,1\cap\bar{2}}}{1-\alpha_{b,2|a,1\cap 2}+\alpha_{b,2|a,1\cap\bar{2}}} \text{ and} \\ & \quad \beta_{a,2|1} \leq \beta_{b,1|2} \cdot \dfrac{P[x_2 \in E_b]}{P[x_1 \in E_a]}, \\ SMDI_{b,1\cap 2}, & \text{if } \beta_{b,1|2} \leq \dfrac{\alpha_{a,1|b,\bar{1}\cap 2}}{1-\alpha_{a,1|b,1\cap 2}+\alpha_{a,1|b,\bar{1}\cap 2}} \text{ and} \\ & \quad \beta_{a,2|1} \geq \beta_{b,1|2} \cdot \dfrac{P[x_2 \in E_b]}{P[x_1 \in E_a]}, \\ DMDI_{a,1\cap b,2}, & \text{otherwise}. \end{cases} \tag{45}$$

Figure 5 visualizes the above conditions on $(\beta_{a,2|1}, \beta_{b,1|2})$ coordinates.

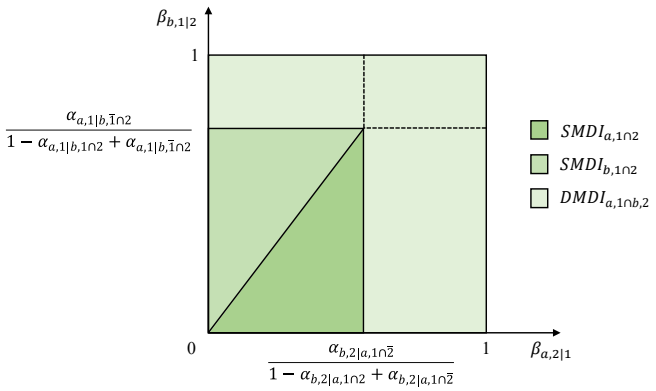

Figure 5. Conditions where DMDI architectures achieve the highest reliability.

By symmetric assumption, we can also derive the conditions where $DMDI_{a,2\cap b,1}$ can achieve the best reliability. Alternatively, when we assume the input data distributions $\mu_{X_1}(x_1)$ and $\mu_{X_2}(x_2)$ satisfy $P[x_1 \in E_a] \leq P[x_1 \in E_b]$ and $P[x_2 \in E_b] \leq P[x_2 \in E_a]$ for any x_1 and x_2, we can show the conditions where $DMDI_{a,1\cap b,2}$ can achieve higher reliability than $DMSI_{a\cap b,1}$ and $DMSI_{a\cap b,2}$.

In a summary of the above discussions, we can have guides to decide the appropriate architecture in terms of reliability by restricting the shape of input data distributions. For instance, if we observe more errors from the sensor input x_2 than the input x_1, from Proposition 2, it is not encouraged to choose $DMSI_{a\cap b,2}$ considering its reliability. In practice, the values of diversity measures are not very close to either zero or one. From (36) and (40), DMDI could be considered a neutral and conservative choice when there is not much information about the preference of input data and ML models. However, it is not a trivial decision to select $DMDI_{a,1\cap b,2}$ or $DMDI_{a,2\cap b,1}$ because when either one choice achieves the best reliability, the other one becomes the least reliable architecture, as seen in (44).

C. Property under the conditional independence

In the proposed reliability model, we used the conditional diversity parameters such as $\alpha_{b,2|a,1\cap 2}$ and $\alpha_{b,2|a,1\cap\bar{2}}$. Although they have monotonic relation to $\alpha_{b|a,2} = P[x_2 \in E_b | x_2 \in E_a]$, their exact values are hard to be inferred from observations. Therefore, even if the input data distributions are restricted, as discussed in the previous section, it is still difficult to estimate how much the selected architecture achieves higher reliability than the others. In contrast, when we assume the independence of individual ML modules' outputs, the difference in architectures' reliabilities can be easily computed. In this section, we attempt to narrow this gap by introducing the assumption of conditional independence [35] between the different diversity measures. With the conditional independence assumption, the intersection of error spaces can be considered independently of the effects of error conjunctions. This means $\alpha_{b,2|a,1\cap 2} = P[x_2 \in E_b | x_2 \in E_a, x_1 \in E_a]$ becomes equal to $\alpha_{b|a,2} = P[x_2 \in E_b | x_2 \in E_a]$, since the occurrence of the event $x_1 \in E_a$ does not affect the conditional event $P[x_2 \in E_b | x_2 \in E_a]$. Assuming that $P[x_2 \in E_a] < 1$, we can also have

$$\begin{aligned}\alpha_{b,2|a,1\cap\bar{2}} &= P[x_2 \in E_b | x_2 \in \overline{E_a}, x_1 \in E_a] \\ &= P[x_2 \in E_b | x_2 \in \overline{E_a}] \\ &= \frac{P[x_2 \in E_b] - \alpha_{b|a,2} \cdot P[x_2 \in E_a]}{1 - P[x_2 \in E_a]}.\end{aligned} \tag{46}$$

The reliability of $DMDI_{a,1\cap b,2}$ expressed in (1) can be rewritten by two diversity-related parameters $\alpha_{b|a,2}$ and $\beta_{a,2|1}$ as

$$\begin{aligned}&R_{DMDI_{a,1\cap b,2}} = 1 - P[x_1 \in E_a] \cdot \\ &\left[\alpha_{b|a,2} \cdot \beta_{a,2|1} + \frac{P[x_2 \in E_b] - \alpha_{b|a,2} P[x_2 \in E_a]}{1 - P[x_2 \in E_a]}\left(1 - \beta_{a,2|1}\right)\right] \\ &= 1 - \frac{P[x_1 \in E_a]}{1 - P[x_2 \in E_a]} \cdot \\ &\left[\alpha_{b|a,2} \cdot \left(\beta_{a,2|1} - P[x_2 \in E_a]\right) + P[x_2 \in E_b] \cdot \left(1 - \beta_{a,2|1}\right)\right].\end{aligned} \tag{47}$$

Since the reliabilities of $DMSI_{a\cap b,2}$ and $SMDI_{a,1\cap 2}$ can also be expressed using these parameters

$$\begin{aligned}R_{DMSI_{a\cap b,2}} &= 1 - \alpha_{b|a,2} \cdot P[x_2 \in E_a], \\ R_{SMDI_{a,1\cap 2}} &= 1 - \beta_{a,2|1} \cdot P[x_1 \in E_a],\end{aligned} \tag{48}$$

the reliabilities of these architectures can be directly compared under given individual modules' reliabilities with two additional variables $\alpha_{b|a,2}$ and $\beta_{a,2|1}$. This model corresponds to the previous model presented in [2]. From (47) and (48), we obtain the next proposition.

Proposition 5. *Assume that the intersection of errors is conditionally independent of the conjunction of errors, i.e.,* $\alpha_{b,2|a,1\cap 2} = \alpha_{b|a,2}$ *and* $\alpha_{b,2|a,1\cap\bar{2}} = P[x_2 \in E_b | x_2 \in \overline{E_a}]$. *Given the modules' reliabilities* $P[x_i \in E_j] \in (0,1), i = \{1,2\}, j = \{a,b\}$, *the most reliable architectures among* $DMDI_{a,1\cap b,2}$, $DMSI_{a\cap b,2}$ *and* $SMDI_{a,1\cap 2}$ *can be determined by the following conditions on the values of* $\alpha_{b|a,2}$ *and* $\beta_{a,2|1}$.

$$\begin{cases} DMSI_{a\cap b,2}, & \text{if } \omega(\alpha_{b|a,2}, \beta_{a,2|1}) - \alpha_{b|a,2} \cdot P[x_2 \in E_a] \geq 0 \text{ and} \\ & \beta_{a,2|1} \geq \alpha_{b|a,2} \cdot \dfrac{P[x_2 \in E_b]}{P[x_1 \in E_a]}, \\ SMDI_{b,1\cap 2}, & \text{if } \omega(\alpha_{b|a,2}, \beta_{a,2|1}) - \beta_{a,2|1} \cdot P[x_1 \in E_a] \geq 0 \text{ and} \\ & \beta_{a,2|1} \leq \alpha_{b|a,2} \cdot \dfrac{P[x_2 \in E_b]}{P[x_1 \in E_a]}, \\ DMDI_{a,1\cap b,2}, & \text{otherwise.} \end{cases}$$

where

$$\begin{aligned}&\omega(\alpha_{b|a,2}, \beta_{a,2|1}) = \frac{P[x_1 \in E_a]}{1 - P[x_2 \in E_a]} \cdot \\ &\left[\alpha_{b|a,2} \cdot \left(\beta_{a,2|1} - P[x_2 \in E_a]\right) + P[x_2 \in E_b] \cdot \left(1 - \beta_{a,2|1}\right)\right].\end{aligned} \tag{49}$$

Figure 6 visualizes the reliability differences of $DMDI_{a,1\cap b,2}$, $DMSI_{a\cap b,2}$ and $SMDI_{a,1\cap 2}$ by varying the parameter values $\alpha_{b|a,2}$ and $\beta_{a,2|1}$ under given modules' probabilities $P[x_1 \in E_a] = 0.1$, $P[x_2 \in E_a] = 0.2$ and $P[x_2 \in E_b] = 0.3$. The difference of architecture's reliabilities is quantitatively comparable when we obtain the estimates of diversity measures as well as modules' error probabilities.

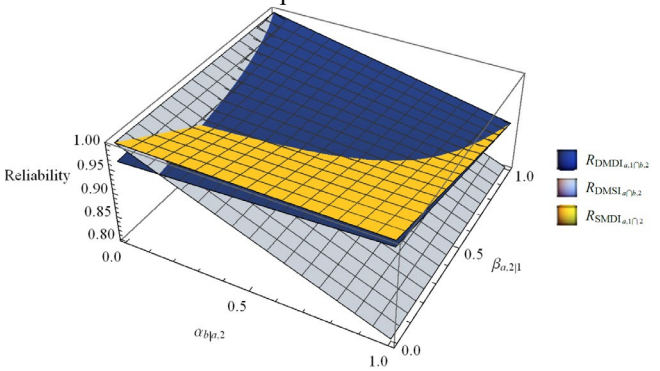

Figure 6. Reliability differences of $DMDI_{a,1\cap b,2}$, $DMSI_{a\cap b,2}$ and $SMDI_{a,1\cap 2}$.

In a similar way, the reliabilities of $DMDI_{a,2\cap b,1}$, $DMSI_{a\cap b,1}$ and $SMDI_{b,1\cap 2}$ can be directly compared by using the two variables $\alpha_{b|a,1}$ and $\beta_{b,2|1}$. However, to compare all six architectures, we need at least four diversity-associated variables, such as $(\alpha_{b|a,1}, \alpha_{b|a,2}, \beta_{a,2|1}, \beta_{b,2|1})$.

V. NUMERICAL EXPERIMENTS

In this section, we present the results of numerical experiments to show the reliabilities of dependent double-modules double-inputs MLS under hypothetical distribution and error functions. We examine that the properties derived from our reliability model presented in the previous section give the guides to choose the relevant architecture options without having complete knowledge of input data distribution and error functions.

A. Hypothetical setting

For the purpose of the experiments, we extend the domain of the joint distribution $\mu_{X_1,X_2}(x_1,x_2)$ to $\mathbb{R}^2$. The two continuous random variables X_1 and X_2 are assumed to follow $\mu_{X_1,X_2}(x_1,x_2)$. We adopt a bivariate normal distribution

$$\mu_{X_1,X_2}(x_1,x_2)=\frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{1-\rho^2}}\cdot\exp\left\{-\frac{z}{2(1-\rho^2)}\right\},$$

where

$$z=\frac{\left(x_1-\lambda_{X_1}\right)^2}{\sigma_{X_1}^2}+\frac{\left(x_2-\lambda_{X_2}\right)^2}{\sigma_{X_2}^2}-\frac{2\rho\left(x_1-\lambda_{X_1}\right)\left(x_2-\lambda_{X_2}\right)}{\sigma_{X_1}\sigma_{X_2}}, \tag{50}$$

λ_{X_1} and λ_{X_2} are the means, σ_{X_1} and σ_{X_2} are the variances, and ρ is the correlation coefficient of X_1 and X_2. The dependence of two input data x_1 and x_2 can be characterized by the value of ρ. The marginal distributions are given by $X_1\sim\mathcal{N}\left(\lambda_{X_1},\sigma_{X_1}^2\right)$ and $X_2\sim\mathcal{N}\left(\lambda_{X_2},\sigma_{X_2}^2\right)$, where $\mathcal{N}(\lambda,\sigma^2)$ represents the normal distribution with mean λ and variance σ^2. For the sake of visualization and reliability computation, we define error input spaces for individual ML models m_a and m_b by closed intervals on $\mathbb{R}$ as $E_a=[e_a^{\min},e_a^{\max}]$ and $E_b=\left[e_b^{\min},e_b^{\max}\right]$. An example relation between two different error spaces with a joint distribution of two input data can be visualized as shown in Figure 7, where the horizontal and vertical axes represent the values of x_1 and x_2, respectively.

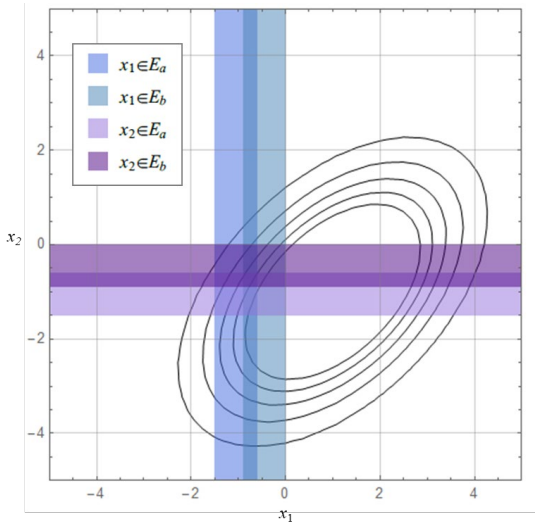

Figure 7. Error input spaces over binormal distribution with $\rho=0.5$.

The colored regions show the error spaces for $x_1\in[-1.5,0.6], x_1\in[-0.9,0], x_2\in[-1.5,0.6]$ and $x_2\in[-0.9,0]$, while the circles represent the counter lines shaped by the bivariate normal distribution with the parameters $\left(\lambda_{X_1},\lambda_{X_2},\sigma_{X_1}^2,\sigma_{X_2}^2,\rho\right)=(1,-1,1.5,1.5,0.5)$. Note that the above hypothetical distribution and the error function are introduced as instances of the general distribution and the error function used in the reliability model. The visualization of this configuration gives an intuitive view of the interrelation between dependent input data and dependent error spaces. For example, consider the point $(x_1,x_2)=(-0.1,2)$, ML model m_b outputs error for x_1, but neither model outputs error for x_2. The probability density of such a case can be given by $\mu_{X_1,X_2}(-0.1,2)$. The probability of error output by an ML module can be computed by integrating the probability density of the corresponding input data and the error space. For example, the probability of $x_1\in E_a$ is given by

$$P[x_1\in E_a]=\int f_a(x_1)d\mu_{X_1,X_2}(x_1,x_2)=\int f_a(x_1)d\mu_{X_1}(x_1)$$
$$=\int_{e_a^{\min}}^{e_a^{\max}}\mu_{X_1}(x_1)dx_1. \tag{51}$$

B. Baseline comparison

The benefit of the proposed dependent double-modules double-inputs MLS can be measured by the improved reliability compared with the MLS relying on a single module. The reliability can also be improved by adopting a simple redundancy scheme without using different sensor inputs, which corresponds to DMSI architectures in our study. Table I shows the reliabilities achieved by the different architectures, when we set $\left(\lambda_{X_1},\lambda_{X_2},\sigma_{X_1}^2,\sigma_{X_2}^2,\rho\right)=(1,-1,1.5,1.5,0)$, $E_a=[-0.4,0.1]$, $E_b=[0,0.5]$.

TABLE I. RELIABILITY IMPROVEMENT BY DIFFERENT ARCHITECTURES

Category	Architecture	Reliability
Single model single input	$SMSI_{a,1}$	0.901071
	$SMSI_{a,2}$	0.887099
	$SMSI_{b,1}$	0.883051
	$SMSI_{b,2}$	0.906163
Single model double inputs	$SMDI_{a,1\cap2}$	0.988831
	$SMDI_{b,1\cap2}$	0.989026
Double models single input	$DMSI_{a\cap b,1}$	0.978239
	$DMSI_{a\cap b,2}$	0.979185
Double models double inputs	$DMDI_{a,1\cap b,2}$	0.990717
	$DMDI_{a,2\cap b,1}$	0.986796

We observe that $DMDI_{a,1\cap b,2}$ achieves the highest reliability among other architecture options. In Table I, $R_{DMDI_{a,1\cap b,2}}$ is 0.990717 that is higher than $R_{DMSI_{a\cap b,1}}$ and $R_{DMSI_{a\cap b,2}}$. The results clearly show that the reliability of conventional DMSI systems can be further improved by exploiting two input data. In the following, we further investigate the impact of diversities.

C. Architecture comparison by varying input similarity

To analyze the impact of input similarity, next, we vary the correlation coefficient ρ in the range of [-0.9,0.9] while keeping other parameters the same as the previous example. The dependency of two input data distributions resulting from $\mu_{X_1,X_2}(x_1,x_2)$ can be characterized by the correlation coefficient ρ. When $|\rho|$ is close to 1, X_1 has a strong positive or negative correlation with X_2.

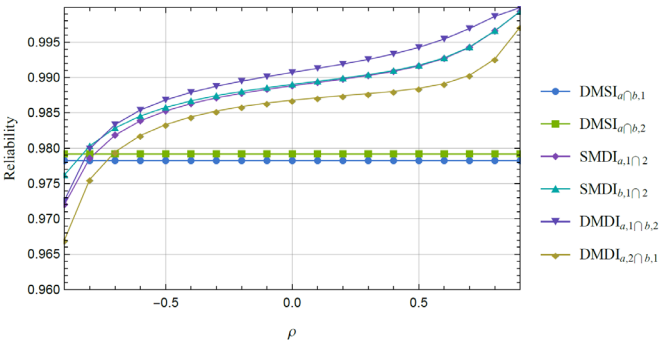

Figure 8. Reliabilities of six architectures by varying ρ.

The computed reliabilities of the six architectures are shown in Figure 8. As can be seen, the reliabilities of DMDI and SMDI architectures are affected by changing the values of ρ, while the reliabilities of DMSI architectures are not affected. The results make sense because DMSI architecture does not use multiple input data. Hence the correlation of different input data does not influence their reliabilities. Interestingly, DMSI architectures achieve the highest reliabilities when ρ is close to -1 among the six architectures while they become the least reliabilities when $\rho > -0.7$. In most of the range, $DMDI_{a,1\cap b,2}$ is considered as the best architecture in terms of reliability.

In practice, we do not know the exact distribution $\mu_{X_1,X_2}(x_1, x_2)$ or error input spaces E_a and E_b. The comparative results of architecture reliabilities shown in Figure 8 are not obtainable from the real observation. However, the preferable architectures can be inferred partly from Proposition 4, since, under the given distribution and the error spaces, we can expect $P[x_1 \in E_a] \leq P[x_2 \in E_a]$ and $P[x_2 \in E_b] \leq P[x_1 \in E_b]$ are likely hold. When we observe that the above relations are generally held in the target application, by Proposition 4, either $DMDI_{a,1\cap b,2}$, $SMDI_{a,1\cap 2}$ or $SMDI_{b,1\cap 2}$ can be selected as the preferable architectures.

We investigate the influence of the conjunction of errors $\beta_{a,2|1}$ and $\beta_{b,2|1}$. The value of ρ is associated with the values of $\beta_{a,2|1}$ and $\beta_{b,2|1}$. When error spaces and input data distribution are given, the diversity measure can be computed by

$$\beta_{j,2|1} = \Pr[x_2 \in E_j | x_1 \in E_j] = \frac{\int_{e_j^{\min}}^{e_j^{\max}} \int_{e_j^{\min}}^{e_j^{\max}} \mu_{X_1,X_2}(x_1, x_2) dx_2\, dx_1}{\int_{e_j^{\min}}^{e_j^{\max}} \mu_{X_1}(x_1) dx_1}, j = a, b. \quad (52)$$

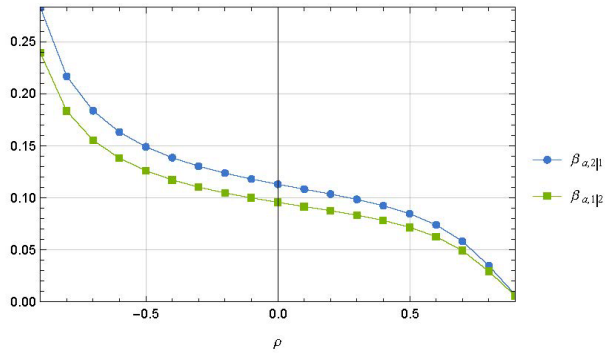

Figure 9. Values of input diversities by varying ρ.

By varying the value of ρ, the values of $\beta_{a,2|1}$ and $\beta_{b,2|1}$ change as plotted in Figure 9. As can be seen, there are negative correlations between them. The higher ρ leads to the smaller $\beta_{a,2|1}$ and $\beta_{a,1|2}$, resulting in the higher reliabilities of DMDI and SMDI architectures. From Proposition 1, the conditions where the reliability of $SMDI_{a,1\cap 2}$ becomes lower than the reliabilities of DMSI and DMDI architectures are given by $\boldsymbol{B}_a^{+\mathsf{T}} > \boldsymbol{A}_{b|a} \cdot \boldsymbol{B}_a^{\mathsf{T}}$. For $DMSI_{a\cap b,1}$ and $DMDI_{a,1\cap b,2}$, the conditions on $\beta_{a,2|1}$ are given by;

$$\begin{cases} R_{SMDI_{a,1\cap 2}} < R_{DMSI_{a\cap b,1}} \Leftrightarrow \\ \quad \beta_{a,2|1} > \dfrac{\alpha_{b,1|a,1\cap\bar{2}}}{1 - \alpha_{b,1|a,1\cap 2} + \alpha_{b,1|a,1\cap\bar{2}}} \left(= DMSI_{a\cap b,1} \text{ bound}\right), \\ R_{SMDI_{a,1\cap 2}} < R_{DMDI_{a,1\cap b,2}} \Leftrightarrow \\ \quad \beta_{a,2|1} > \dfrac{\alpha_{b,2|a,1\cap\bar{2}}}{1 - \alpha_{b,2|a,1\cap 2} + \alpha_{b,2|a,1\cap\bar{2}}} \left(= DMDI_{a,1\cap b,2} \text{ bound}\right). \end{cases} \quad (53)$$

Figure 10 plots the values of the above bounds together with $\beta_{a,2|1}$ by varying ρ.

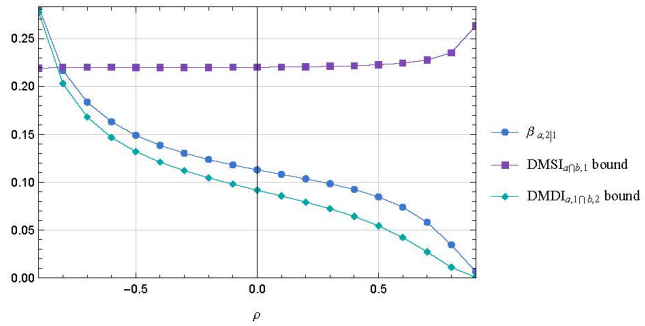

Figure 10. Bounds of $\beta_{a,2|1}$ for DMSI and DMDI architectures.

$\beta_{a,2|1}$ is constantly larger than $DMDI_{a,1\cap b,2}$ bound regardless of ρ, which means $DMDI_{a,1\cap b,2}$ is always a better choice than $SMDI_{a,1\cap 2}$. On the other hand, $\beta_{a,2|1}$ is smaller than $DMSI_{a\cap b,1}$ bound in most of the range of ρ (≥ -0.8), which indicates that $SMDI_{a,1\cap 2}$ is the better choice than $DMSI_{a\cap b,1}$ unless $\beta_{a,2|1}$ tends to be large.

D. Architecture comparison by varying model similarity

Next, we investigate the effects of dependency between two ML models that have different error spaces. In this experiment, we fix the parameters of bivariate normal distribution $(\lambda_{X_1}, \lambda_{X_2}, \sigma_{X_1}^2, \sigma_{X_2}^2, \rho) = (1, -1, 1.5, 1.5, 0.5)$ and the error space $E_b = [0, 0.5]$. The other error space is given by $E_a = [e_a^{\min}, e_a^{\min} + 0.5]$ where $e_a^{\min}$ is varying in the range of [-0.5,0]. Figure 11 shows E_a and E_b in the interval [-0.5, 0.5].

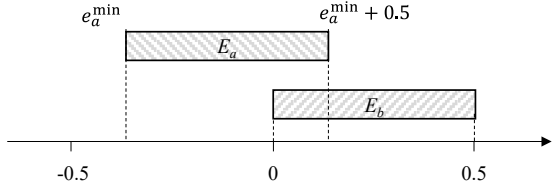

Figure 11. The intersection of errors between $\boldsymbol{E_a}$ and $\boldsymbol{E_b}$.

When $e_a^{\min} = -0.5$, there is no intersection between E_a and E_b, and thus the value of model similarity is equal to zero. On the other hand, when $e_a^{\min} = 0$, E_a becomes identical to E_b, and hence the value of model similarity becomes one. Figure 12 shows the reliabilities of the six different architectures by varying the value of $e_a^{\min}$.

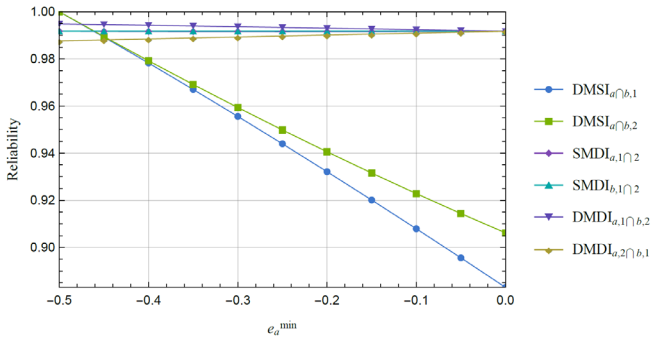

Figure 12. Reliabilities of six architectures by varying e_a^{min}.

As can be seen, the reliabilities of both DMSI architectures decrease by increasing the values of $e_a^{\min}$, while the reliabilities of SMDI and DMDI architectures are not much affected. Since the intersection of error spaces between E_a and E_b increases by the increase in $e_a^{\min}$, the error probabilities of DMSI architectures also increase accordingly. When $e_a^{\min}$ is close to -0.5, the intersection over E_b becomes extremely small, and then DMSI architectures achieve the highest reliabilities. In most of the range, however, $DMDI_{a,1\cap b,2}$ achieves the highest reliability among the six architectures. Regardless of the position of E_a, the condition of Proposition 4 is likely to hold. Thus, by Proposition 4, we can infer that $DMDI_{a,1\cap b,2}$ or SMDI architectures are preferable, even if we do not know the exact distribution of input data or the error input spaces.

We investigate the influence of the intersection of errors $\alpha_{b|a,1}$ and $\alpha_{b|a,2}$ in this experimental setting. For the given input data distributions and the error spaces $E_b = [0,0.5]$ and $E_a = [e_a^{\min}, e_a^{\min} + 0.5]$ where $e_a^{\min} = [-0.5,0]$, the diversity measures can be computed by

$$\alpha_{b|a,i} = \Pr[x_i \in E_b | x_i \in E_a] = \frac{\int_{e_b^{\min}}^{e_a^{\min}+0.5} \mu_{X_i}(x_i)dx_i}{\int_{e_a^{\min}}^{e_a^{\min}+0.5} \mu_{X_i}(x_i)dx_i}, \quad i = 1,2. \tag{54}$$

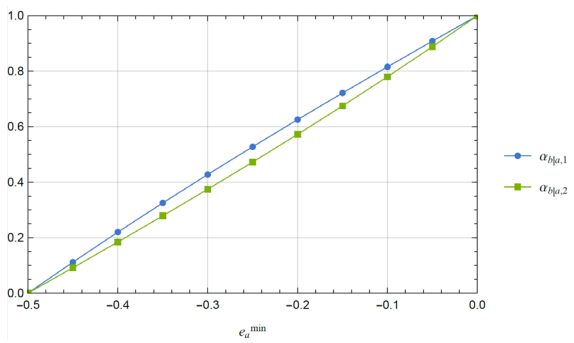

Figure 13. Values of model diversities by varying e_a^{min}.

By varying the value of $e_a^{\min}$, the values of $\alpha_{b|a,1}$ and $\alpha_{b|a,2}$ change as shown in Figure 13. As expected, we can observe both $\alpha_{b|a,1}$ and $\alpha_{b|a,2}$ have positive correlations with $e_a^{\min}$, resulting in the lower reliabilities of DMSI architectures. Therefore, when we observe large values of $\alpha_{b|a,1}$ or $\alpha_{b|a,2}$, by the variant of Proposition 1, DMSI architectures are unlikely the appropriate architecture options in terms of reliability.

E. Special distribution case

For an instance of the restricted type of distribution discussed in Section IV-B, we consider the case where errors in E_b occur more frequently than errors in E_a, and examine Proposition 3. Consider the bivariate normal distribution with the parameters $(\lambda_{X_1}, \lambda_{X_2}, \sigma_{X_1}^2, \sigma_{X_2}^2) = (0,0,1.5,1.5)$. Regardless of the value of correlation coefficient ρ, the marginal distribution is given by $\mathcal{N}(0,1.5)$. When we fix the error spaces $E_a = [-1.4, -0.9]$ and $E_b = [-1.0, -0.5]$, it satisfies $P[x_i \in E_a] \leq P[x_i \in E_b]$ for most of input data x_i that follows the input data distribution $\mathcal{N}(0,1.5)$. The intuition of this relation can be given in Figure 14, as we can see that the region of E_b is always closer to the peak of probability density at (0.0) than the region of E_a.

From Proposition 3, the most reliable architecture is either $SMDI_{a,1\cap 2}$, $DMSI_{a\cap b,1}$ or $DMSI_{a\cap b,2}$ depending on the values of $\alpha_{b|a,1}$ and $\alpha_{b|a,2}$. Since the given bivariate normal distribution has symmetry in marginal distributions, the condition can be simplified as

$$\begin{cases} DMSI_{a\cap b,1}, & \alpha_{b|a,1} \leq \dfrac{\beta_{a,2|a\cap\bar{b},1}}{1-\beta_{a,2|a\cap b,1}+\beta_{a,2|a\cap\bar{b},1}}, \\ SMDI_{a,1\cap 2}, & \text{otherwise.} \end{cases} \tag{55}$$

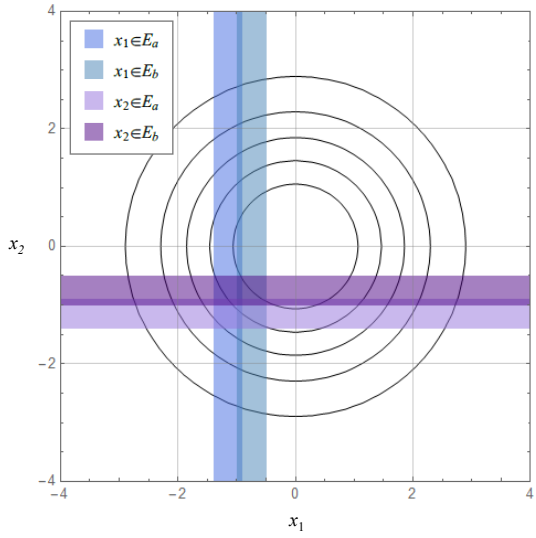

Figure 14. An example of error spaces satisfying $P[x_i \in E_a] \leq P[x_i \in E_b]$.

Figure 15 shows the architecture reliabilities directly computed from the given distribution by varying ρ in the rage of $[-0.9,0.9]$. Due to the symmetry marginal distributions, we can observe $R_{DMSI_{a\cap b,1}} = R_{DMSI_{a\cap b,2}}$ and $R_{DMDI_{a,1\cap b,2}} = R_{DMDI_{a,2\cap b,1}}$. In the most of range of ρ, $SMDI_{a,1\cap 2}$ achieves the highest reliability.

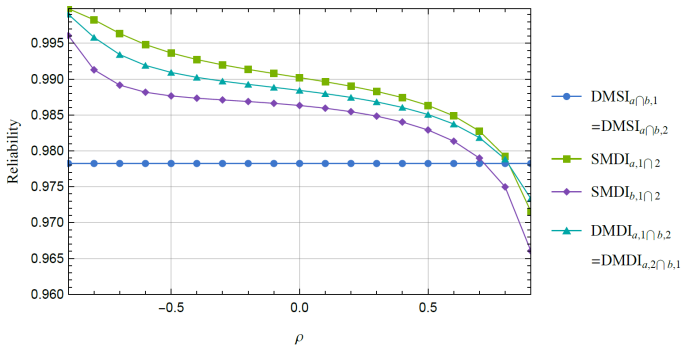

Figure 15. Reliabilities of different architectures by varying ρ.

Figure 16 plots the $\alpha_{b|a,1}$ and the boundary condition given by Proposition 3.

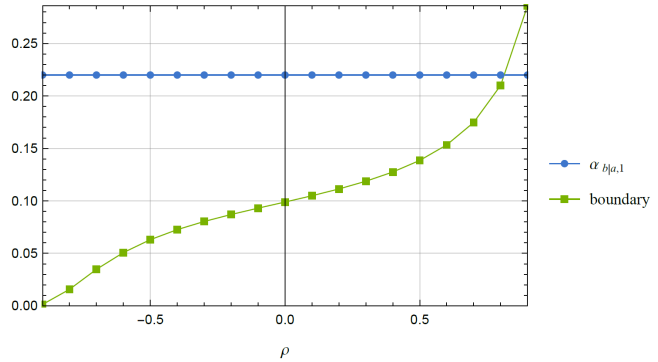

Figure 16. Bounds of $\alpha_{b|a,1}$ for DMSI architectures.

The value of $\alpha_{b|a,1}$ becomes lower than the boundary when $\rho = 0.9$, at which $DMSI_{a\cap b,1}$ achieves higher reliability than $SMDI_{a,1\cap 2}$ as observed in Figure 15. The result confirms the condition of Proposition 3. Unless we are confident that $\alpha_{b|a,1}$ is very small value from empirical observation, $SMDI_{a,1\cap 2}$ can be chosen as the preferable architecture option.

Remark. The bivariate normal distribution and the error spaces used in the above example help to visualize the dependence of error input space over the joint distributions of dependent two input data as shown in Figure 7. Note that the properties examined in this section are not limited to the above hypothetical setting since we derived them theoretically in Section IV.

VI. Application and Limitation

As presented in the motivating example, the potential application of the proposed dependent double-modules double-inputs MLS is a perception system in a self-driving vehicle. For example, traffic signal recognition is one of the important perception tasks in autonomous vehicle architecture [39]. A common approach to exploiting redundant architecture for improved reliability relies on modular redundancy, corresponding to DMSI architectures in this paper. Our study analytically clarified how reliability could be further improved by the architectures using diverse inputs. In the self-driving vehicle scenario, the system may be equipped with multiple cameras, and hence the system can adopt SMDI and DMDI architectures.

The experiment results show that our reliability model and the derived properties provide guides to choosing the preferable architecture options based on the diversity measures. However, to determine the best architecture option, finding the values of diversity measures is necessary. In practice, we may need to rely on empirical estimates of the diversity measures from the limited samples that are subject to estimation errors. Although finding the best architecture option is not always easy, our reliability model can be useful for selecting the preferable architectures by screening undesirable architecture options. For instance, consider the traffic signal recognition scenario in Section II-A. We can train multiple ML models independently and test them with real-world test samples. ML models can be ranked by the test accuracy that provides information about the model's superiority. In this case, SMDI architecture relying on a lower-ranked model is not a recommendable option, and thus they can be removed from the candidates. From Proposition 3, we can narrow down the candidates to SMDI with the best-ranked model or DMSI using the top two-ranked models. Suppose errors from the top two-ranked models are mostly overlapped, indicating a higher intersection of errors between the two models. In that case, SMDI with the best-ranked model is presumed to be the best option by Proposition 3.

The presented models are limited to two-version architectures. The reliability can be further improved by increasing the redundancy (e.g., using three or more versions). The reliability models with three or more versions of ML systems need to be considered in that case. However, real-world application systems such as autonomous vehicles may have stringent cost and latency requirements that may not allow extra redundancy. When the redundancy level is limited to two due to cost or performance constraints, the guidelines derived from our analysis are useful.

VII. Related Work

MLS quality assurance has recently become a hot issue in software and system engineering research. MLS is fundamentally built as a software system, and therefore developers of MLS apply and improve the traditional software engineering methodologies to assure the quality of ML application systems [14]. In this regard, machine learning testing is one of the important challenges actively discussed today [10]. Compared to traditional software testing, MLS testing needs to deal with an oracle problem [19]. The correct output for arbitrary input is not given when testing an MLS. The software systems without reliable test oracle are known as non-testable programs [20]. To address the oracle issue of MLS, Murphy and Kaiser [21] presented a metamorphic testing approach in which a pseudo-oracle [22] for new test cases is created by modifying the inputs used in the initial test cases. Metamorphic testing has been recently applied for testing image classifiers [15], a deep learning-based forecaster [16], and graph convolutional neural networks [17]. Further recent advances in ML testing techniques are reviewed in a comprehensive survey report [23]. Our study is not an MLS testing method, but the presented architecture reliability analysis can complement developing highly reliable MLSs.

Another important technical challenge in MLS quality management is the runtime validation of MLS execution. Even if an MLS passes reasonably designed test scenarios, the outputs of MLS are significantly influenced by real input data given in users' environments. The discrepancy between the assumed input data in the development phase and the real input data observed in the operation phase likely causes undesirable outputs at runtime. Wu et al. leveraged the data validation technique to detect real-world corner cases for DNN-based systems [24]. Considering a cyber-physical system employs ML components, Dreossi et al. proposed a method to analyze the input space for ML classifiers that can lead to undesirable consequences in the cyber-physical system [25]. While we also consider the operational phase of MLS, our focus is not on validation but evaluation of reliability by N-version MLS.

Redundant configurations of ML components can improve the reliability of MLS. N-version machine learning has been presented to improve the reliability of MLS's outputs [2]. The relationship between the reliability of N-version MLS and the diversities in ML models and input data has been analyzed in our previous study [2]. The experimental studies also showed that the coverage of errors could be improved by using diverse ML models and different input data sets [6]. Similar experiments have been done by Xu et al. in which N-version deep neural networks are used to complement individual error cases and make a fault-tolerant system [5]. A Multi-version approach has also been used for ML applications using deep neural networks to make more robust steering control in an autonomous vehicle [26]. PolygraphMR also introduced modular redundancy to reduce wrong answers with high confidence [18]. In this paper, extending the diversity measures introduced in [2], we present more comprehensive and general reliability models. To the best of our knowledge, our study is the first work to show the reliability properties of dependent double-modules double-inputs MLS via diversity measures.

Using multiple learners to compose a better ML model is a commonly adopted ML technique known as the ensemble method [27]. Ensemble methods are also shown to be effective in improving the ML model's reliability against faults in the training data sets [42]. To obtain a better ensemble, individual learners should have differences among them. It is studied that the minimum margin of the ensemble is associated with the diversity among the individual learners [28]. Several diversity measures were proposed to analyze the association with the accuracy of ensemble methods [29][30][31][32]. Although diversity is believed to play a fundamental role in ensemble accuracy, the right formulation and measures for diversity are unsolved issues [33]. Empirical studies show that diverse ML models can be constructed by using different data sets, sampling methods, training algorithms, and neural network architectures [5][6][34]. While the similarity metrics such as the Bhattacharyya coefficient [40] and the Fréchet distance [41] can be used for measuring the distance between two input data distributions, the reliability cannot be characterized without the error function and the joint distribution of two input. In this paper, we focus on the reliability of MLS outputs that are directly related to the accuracies of individual ML models used in the system. While highly accurate ML models likely improve the reliability of MLS, it is valid only when real input data distributions follow the assumption or expectation. Instead of the accuracy of the ML model, we discuss the reliability of MLS attributed to redundant configurations that consist of different input data and ML models. Since both input data sets and ML models are not independent, we cannot formulate the problem by a simple combinatorial logic and similarity metrics for distributions, which is the challenge addressed in this paper.

VIII. CONCLUSION

This paper proposed the reliability model for an N-version MLS that employs two ML modules for classification tasks and determines the system output on a consensus basis. The system has six architecture options depending on the choice of input data and ML models used in the modules. We defined two diversity measures to quantify the similarities of ML models' capabilities and the interdependence of errors by input data sets. The diversity measures are used to formulate the reliabilities of six different architectures. With the proposed matrix representation of the reliability model, we showed some properties that can guide the choice of preferable architecture when given individual ML models' reliabilities and diversity measures. Among possible architectures, DMDI architecture, which exploits both model diversity and input diversity, can be regarded as a neutral and a conservative option since the values of diversity measures are unlikely to be close to zero or one in practice. We also presented the numerical results to give an intuitive understanding of our proposed models and properties.

A future study could explore extending the reliability model to multi-modules multi-inputs MLS. This paper focuses on a dependent double-modules double-inputs MLS that can be a building block of a larger N-version system. Another important future research direction is the validation of the models with real-world datasets. Our previous work also presented some experimental results that show two potential ways to diversify the outputs of redundant MLS [6]. Because the input data distributions and error spaces are hard to obtain empirically, future studies are required to narrow the gap between the limited empirical results and the theoretically derived properties.

REFERENCES

[1] R. McAllister, Y. Gal, A. Kendall, M. van der Wilk, A. Shah, R. Cipolla, A. Weller, Concrete problems for autonomous vehicle safety: advantages of Bayesian deep learning, In Proc. of the 26th International Joint Conference on Artificial Intelligence (IJCAI), pp. 4745-4753, 2017.

[2] F. Machida, N-version machine learning models for safety critical systems, In Proc. of the DSN Workshop on Dependable and Secure Machine Learning, pp. 48-51, 2019.

[3] A. Avizienis and L. Chen, On the implementation of N-version programming for software fault tolerance during execution. In Proc. of IEEE International Computer, Software and Application Conference (COMPSAC), pp. 149–155, 1977.

[4] A. Avizienis, The methodology of n-version programming, Software fault tolerance, Vol. 3, pp. 23-46, John Wiley & Sons, New York, 1995.

[5] H. Xu, Z. Chen, W. Wu, Z. Jin, S. Kuo, M. R. Lyu, NV-DNN: towards fault-tolerant DNN systems with N-version programming, In Proc. of the

Workshop on Dependable and Secure Machine Learning, pp. 44-47, 2019.
[6] F. Machida, On the diversity of machine learning models for system reliability, In Proc. of IEEE Pacific Rim Int'l Symp. on Dependable Computing (PRDC), pp. 276-285, 2019.
[7] K. Pei, Y. Cao, J. Yang, and S. Jana, DeepXplore: Automated whitebox testing of deep learning systems, In Proc. of the 26th Symposium on Operating Systems Principles (SOSP), pp. 1-18, 2017.
[8] Y. Tian, K. Pei, S. Jana, and B. Ray, Deeptest: Automated testing of deep-neural-network-driven autonomous cars, In Proc. of the 40th Int'l Conf. on Software Engineering, pp. 303-314, 2018.
[9] Y. Zhang, Y. Chen, S. C. Cheung, Y. Xiong, and L. Zhang, An empirical study on tensorflow program bugs, In Proc. of the 27th ACM SIGSOFT Int'l Symp. on Software Testing and Analysis, pp. 129-140, 2018.
[10] H. B. Braiek and F. Khomh, On testing machine learning programs, Journal of Systems and Software, vol. 164, p. 110542, 2020.
[11] IEC61078, Reliability Block Diagram Method, IEC Standard No. 61078, 1991.
[12] IEC61025, Fault Tree Analysis. IEC Standard No. 61025, 2nd edn., 2006.
[13] E. Ruijters and M. Stoelinga, Fault tree analysis: A survey of the state-of-the-art in modeling, analysis and tools, Computer science review, vol. 15, pp. 29-62, 2015.
[14] D. Sculley, et al., Hidden technical debt in machine learning systems, In Advances in neural information processing systems, pp. 2503-2511, 2015.
[15] A. Dwarakanath, M. Ahuja, S. Sikand, R. M. Rao, R. J. C. Bose, N. Dubash, and S. Podder, Identifying implementation bugs in machine learning based image classifiers using metamorphic testing, in Proc. of the 27th ACM SIGSOFT Int'l Symp. on Software Testing and Analysis, pp. 118-128 2018.
[16] A. Dwarakanath, M. Ahuja, S. Podder, S. Vinu, A. Naskar, and M. Koushik, Metamorphic testing of a deep learning based forecaster, In Proc. of 2019 IEEE/ACM 4th Int'l Workshop on Metamorphic Testing (MET), pp. 40-47, 2019.
[17] Y. Wang, W. Wang, Y. Cai, B. Hooi, B. C. Ooi, Detecting implementation bugs in graph convolutional network based node classifiers, In Proc. of the 31st Int'l Symp. on Software Reliability Engineering (ISSRE 2020), 2020.
[18] S. Latifi, B. Zamirai, and S. Mahlke, Polygraphmr: Enhancing the reliability and dependability of cnns, In Proc. of 50th IEEE/IFIP Int'l Conf. on Dependable Systems and Networks (DSN), pp. 99-112, 2020.
[19] E. Barr, M. Harman, P. McMinn, M. Shahbaz, and S. Yoo, The oracle problem in software testing: A survey, IEEE transactions on software engineering, vol. 41, no. 5, pp. 507-525, 2015.
[20] E. J. Weyuker, On testing non-testable programs, Computer Journal, vol.25, no.4, pp.465-470, 1982.
[21] C. Murphy, G. E. Kaiser, and M. Arias, An approach to software testing of machine learning applications. In SEKE, vol. 167, 2007.
[22] M. D. Davis and E. J. Weyuker, Pseudo-oracles for non-testable programs, In Proc. of the ACM'81 Conference, pp. 254-257, 1981.
[23] J. M. Zhang, M. Harman, L. Ma, and Y. Liu, Machine learning testing: Survey, landscapes and horizons, IEEE Transactions on Software Engineering, 2020.
[24] W. Wu, H. Xu, S. Zhong, M. Lyu, and I. King. Deep validation: Toward detecting real-world corner cases for deep neural networks. In Proc. of the 49th IEEE/IFIP Int'l Conf. on Dependable Systems and Networks (DSN), pp. 125-137, 2019.
[25] T. Dreossi, A. Donzé, and S. A. Seshia, Compositional falsification of cyber-physical systems with machine learning components, In NASA Formal Methods Symposium, pp. 357-372, 2017.
[26] A. Wu, A. H. M. Rubaiyat, C. Anton, and H. Alemzadeh, Model Fusion: weighted N-version programming for resilient autonomous vehicle steering control, In Proc. of IEEE Int'l Symp. on Software Reliability Engineering Workshops, pp. 144-145, 2018.
[27] T. Dietterich, Ensemble methods in machine learning, In Proc. of international workshop on multiple classifier systems, pp. 1-15, 2000.
[28] R. E. Schapire, Y. Freund, P. Bartlett, W. S. Lee, Boosting the margin: A new explanation for the effectiveness of voting methods, The annals of statistics, vol. 26, no. 5, pp. 1651-1686, 1998.
[29] G. Brown, J. Wyatt, R. Harris, and X. Yao, Diversity creation methods: a survey and categorisation, Information Fusion, vol. 6, no. 1, pp.5-20, 2015.
[30] D. Partridge, W. Krzanowski, Software diversity: practical statistics for its measurement and exploitation, Information and software technology, vol. 39, no. 10, pp.707-717, 1997.
[31] G. Brown, An information theoretic perspective on multiple classifier systems, In Int'l Workshop on Multiple Classifier Systems, pp. 344-353, 2009.
[32] E. K. Tang, P. N. Suganthan, and X. Yao, An analysis of diversity measures, Machine learning, vol. 65, no.1, pp. 247-271, 2006.
[33] Z.-H. Zhou, Ensemble methods: foundations and algorithms, CRC press, 2012.
[34] Z. Gong, P. Zhong, and W. Hu, Diversity in machine learning, IEEE Access, vol. 7, pp. 64323-64350, 2019.
[35] A. P. Dawid, Conditional independence in statistical theory, Journal of the Royal Statistical Society. Series B (Methodological), vol. 41, no. 1, pp. 1-31, 1979.
[36] W. Xu, D. Evans, and Y. Qi, Feature Squeezing: Detecting adversarial examples in deep neural networks, 25th Annual Network and Distributed System Security Symposium, 2018.
[37] D. E. Eckhardt, L. D. Lee, A theoretical basis for the analysis of multiversion software subject to coincident errors, IEEE Trans. Software Eng., Vol. SE-11, No. 12, pp. 1511-1517, 1985.
[38] B. Littlewood, D.R. Miller, Conceptual modeling of coincident failures in multiversion software, IEEE Trans. on Software Eng., Vol. 15, No. 12, pp.1596-1614, 1989.
[39] Apollo 5.0 Perception module, https://github.com/ApolloAuto/apollo/blob/master/docs/06_Perception/perception_apollo_5.0.md
[40] A. Bhattacharyya, On a measure of divergence between two statistical populations defined by their probability distributions, Bulletin of the Calcutta Mathematical Society, vol. 35, pp. 99-109, 1943.
[41] D. C. Dowson, and B.V. Landau, The Frechet distance between multivariate normal distributions, Journal of Multivariate Analysis, vol. 12, no. 3, pp. 450-455, 1982.
[42] A. Chan, A. Gujarati, K. Pattabiraman, and S. Gopalakrishnan, The Fault in Our Data Stars: Studying Mitigation Techniques against Faulty Training Data in ML Applications, In Proc. of 52nd IEEE/IFIP Int'l Conf. on Dependable Systems and Networks (DSN), pp. 163-171, 2020.

Fumio Machida is an associate professor in the Computer Science Department at University of Tsukuba. He received his PhD degree from Tokyo Institute of Technology in 2018. He was a principal researcher at NEC Corporation until 2019. He was a visiting scholar in the Department of Electrical and Computer Engineering at Duke University in 2010. He was a recipient of the young scientists' prize of Japan in 2014. His research interests include modeling and analysis of system dependability, software aging and rejuvenation, and cloud and edge computing systems. He is a senior member of the IEEE and the member of ACM.