Failure Analysis of a Complex Learning Framework Incorporating Multi-Modal and Semi-Supervised Learning

Laura L. Pullum, Christopher T. Symons
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
Oak Ridge, United States
PullumLL@ornl.gov, SymonsCT@ornl.gov

***Abstract*— Machine learning is used in many applications, from machine vision to speech recognition to decision support systems, and it is used to test applications. However, though much has been done to evaluate the performance of machine learning algorithms, little has been done to verify the algorithms or examine their failure modes. Moreover, complex learning frameworks often require stepping beyond black box evaluation to distinguish between errors based on natural limits on learning and errors that arise from mistakes in implementation. We present a conceptual architecture, failure model and taxonomy, and failure modes and effects analysis (FMEA) of a semi-supervised, multi-modal learning system, and provide specific examples from its use in a radiological analysis assistant system. The goal of the research described in this paper is to provide a foundation from which dependability analysis of systems using semi-supervised, multi-modal learning can be conducted. The methods presented provide a first step towards that overall goal.**

Keywords-machine learning; failure modes; failure analysis; dependability.

I. INTRODUCTION

The use of Machine Learning (ML) in real-world applications has steadily increased as the potential to support or replace human reasoning has grown. With that increased usage comes an increased need to ensure that learning applications operate according to design in both the learning phase and the application phase. As the problems addressed with learning algorithms increase in complexity, so do the algorithms themselves. This can be particularly problematic when it becomes necessary to verify and validate cutting edge algorithms, which are often treated as a black box. While it is often assumed that a black box approach is necessary, one cannot always ignore the intricacies of the specific algorithms that are employed. Some examples of this include learning that integrates data from multiple modalities (i.e., multi-modal learning) and learning from a combination of labeled and unlabeled data (i.e., semi-supervised learning). In the case of multi-modal learning each modality may offer an independent or interdependent view of the data. In the case of semi-supervised learning, examples may be leveraged very differently in training depending on whether or not a ground truth label is known for that data point.

Although ML algorithm performance evaluation measures such as accuracy and precision are well known, and ML's use in verification of other software has been proposed, there is a lack of research on the verification and failure analysis of ML algorithms themselves. This research is aimed at providing a foundation for verification and failure analysis of complex learning frameworks, such as those incorporating semi-supervised and multi-modal ML. By focusing on a particular, widely applicable framework, we present ways in which such software can fail, the potential effects of constituent component failure, and development of a fault tree for further evaluation.

To provide a reasonable framework for addressing verification and failure analysis of complex learning algorithms in real systems, we suggest a framework in the context of an analytical system that combines two modalities of data, with very different roles, and that seeks to involve both labeled and unlabeled data in the learning process. This particular analytical approach forces joint consideration of data with very different interacting roles. This type of learning is broadly applicable to a wide range of applications that attempt to utilize or learn from very different sources of information and that have limits imposed on the availability of ground truth data. This includes many decision support problems in the medical domain, where data describing a patient consists of a variety of information types, such as genetic information, blood tests, patient narratives, imaging, and so on. Moreover, access to ground truth concerning disease status is limited due to natural availability, cost, privacy issues, etc. Many other fields of application have this same "structure." For example, in broad terms many problems involving automatic object identification could be cast in this form; e.g., multiple sensors providing information about an object, multiple sources of intelligence data about the same incident, object, or entity (friend or foe, imminent danger or not, etc.).

Research was recently conducted to develop a multi-modal learning framework and tools for the analysis of radiology images and reports [1, 9]. We use this as a guide to analyzing a subset of multi-modal, semi-supervised learning applications.

Research sponsored in part by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

As a broad problem definition, we consider an application where two separate modes of information are available historically, while only one of these will be available at test time. Furthermore, we consider the data points with known ground truth labels to be a small subset of the data points that are in fact accessible for training. In specifying a concrete example of this type of application, we consider a decision support process designed to provide a check on or assistance to a radiologist's analysis of a mammogram. Thus, the first mode of information is an image, and it is on this form that the application will operate in the field. At training time however, a second modality of information is available in the form of a radiologist's report (natural language text) that describes each image. Since the system attempts to independently support a radiologist's assessment, this second modality will not be available during actual usage, while it is clearly a valuable piece of information in the right learning framework. Furthermore, due to privacy issues, difficulty of tying biopsy results back to early mammograms, etc., the data points known to be normal or cancerous are only a subset of the actual data available.

The semi-supervised machine-learning framework integrates text and image modalities by transforming both modalities into feature vectors, which are produced through text and image analysis and processing. These vectors are used to find a lower dimensional space for image analysis that is smooth with respect to the cancer-specific image similarities described in the radiological reports. When a classifier developed via the learning framework is given a set of mammography images as input, it would provide an automated ability to confirm a diagnosis, e.g., abnormal or normal, and a confidence measure for that diagnosis. We will use this system throughout the paper to illustrate the concepts presented.

II. Machine Learning Background

Machine learning [2] (ML) can be thought of as a software-component design process that builds a decision-making procedure that learns to map inputs to outputs based on examples it is given. This approach is particularly relevant when the software designer does not understand the process by which a decision is normally made, but can give examples of what is usually done. A common example is a child learning to distinguish sports cars from non-sporty cars. There is no concrete rule set used to make the distinction, and the definition is fuzzy and non-stationary. Therefore, writing a step-by-step procedure for making the decision is not appropriate. However, a learning algorithm could be given examples from each class, each of which is characterized by a set of descriptive features or attributes (color, shape, etc.), and learn to make decisions regarding previously unseen examples. ML is a field primarily defined by the goal of creating methods that generalize from examples that have been given in this way.

As a field, ML has dramatically advanced in recent years in ways that are encouraging more and more applications in real systems. Although much of the original theory and the classic learning algorithms are based on simplifying restrictions, real-world needs have driven the field to address problems such as handling very high-dimensional attribute spaces, dealing with noise and redundancy, limited or expensive ground truth, multiple modalities capturing overlapping information, non-stationarity, etc. Applicable verification and validation (V&V) techniques for ML components were already limited in the classic case, and many of the more advanced algorithms are even more difficult to handle with current software V&V constructs.

To provide focus, we address a specific learning framework that incorporates two of these newer learning approaches - semi-supervised learning and multi-modal learning. Semi-supervised learning [3] is a paradigm that learns from examples with ground truth labels (labeled examples) and at the same time incorporates information from examples with no known ground truth (unlabeled examples) into the learning process. Multi-modal learning [1, 4] attempts to incorporate information about a sample from different, overlapping views, such as an image and a textual description of it, genomic information with physical measurements, various sensor measurements centering around a single object, etc.

This paper provides part of a dependability framework for semi-supervised, multi-modal learning applications, specifically an architectural model, a fault model, fault taxonomy and failure modes and effects analysis (FMEA). These general (non-application specific) constructs are provided so that they can be applied to systems using semi-supervised and multi-modal learning. In addition, we provide real examples from a radiologist decision support system for mammographic data.

Fig. 1 illustrates the conceptual architecture for the semi-supervised, multi-modal learning system and its incorporation into an application. The system consists of components responsible for preprocessing and analyzing the data, and developing and applying the classifier. In our sample application, the training data consists of a set of mammograms, some of patients with cancer and some without, and the associated radiologists' reports. In some cases we know the outcome (have known ground truth), though not for a majority of the cases. The primary and secondary modalities are mammography images and radiologists' (text) reports, respectively. An example classification of an image is normal or abnormal, with a given confidence in the classification.

The failure analysis begins at the learning system architecture level, evaluating each component's functionality and data, and identifying the components with which it interacts (defined in Table 1). Similar information for the sample application is provided in Table 2. For example, in preprocessing the data, we want to isolate and retain only that data of interest. In the mammography example, the images are isolated to contain only the breast and pectoral muscle and unwanted text (e.g., header or footer text) is removed from the radiologist's report.

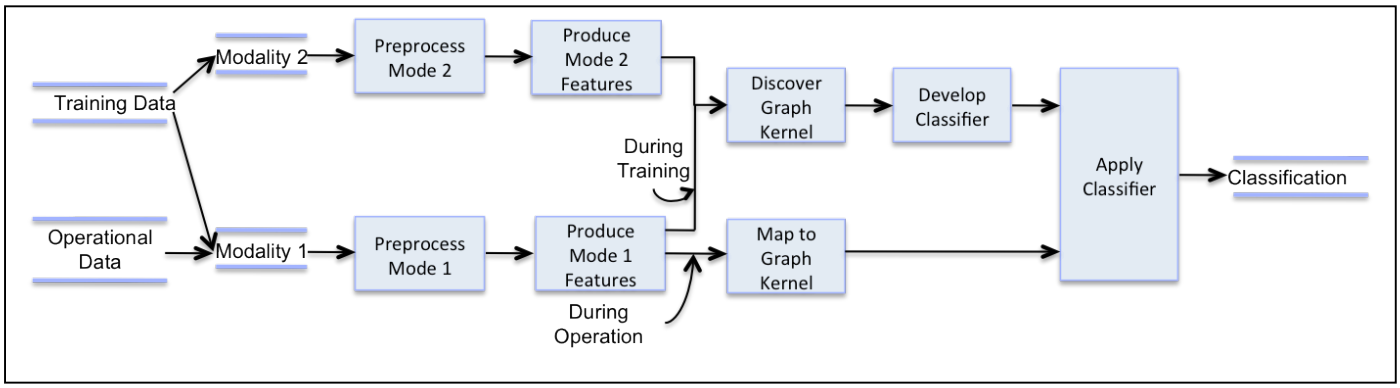

Figure 1. Conceptual architecture of semi-supervised, multi-modal learning application

TABLE I. MODULE DEFINITIONS

Module	Function	Input(s)	Output(s)
Preprocess [Primary | Secondary] Modality (PPM | PSM)	Isolate data to retain only data of interest	Data of [Primary | Secondary] Modality	Isolated data
Produce [Primary | Secondary] Modality Features (PMF | SMF)	Develop the modality's feature space	Isolated data	[Modality] feature vectors
Discover Graph Kernel (DGK)	Use both modalities' features to find a graph kernel for the primary modality classification	Feature vectors for both modalities; Links	Graph kernel
Develop Classifier (DC)	Develop classifier to classify primary modality	Graph kernel	Classifier
Apply Classifier (AC)	Classify data of primary modality; Provide confidence indicator	Classifier	Classification; Confidence level

III. FAILURE MODEL AND TAXONOMY

For the multi-modal, semi-supervised learning system, a failure model is comprised of combinations of false positive or false negative and confidence level (high, low) in the classification (see table 3). For the sample application, a *False Positive* occurs when the system classifies the image as abnormal when it is, in fact, normal. A *False Negative* occurs when the system classifies the image as normal when it is actually abnormal. A *Low* means the calculated confidence in the classification is low, providing a sense of uncertainty that partially mitigates the inaccuracy of the classification. A *High* indicates that the calculated confidence in the classification is high, despite its inaccuracy.

TABLE II. MODULE DEFINITIONS FOR SAMPLE APPLICATION

Module	Function	Input(s)	Output(s)
Preprocess Image	Isolate image to contain only breast and pectoral muscle	Mammography images	Isolated portions of images
Produce Image Features	Develop the image feature space	Isolated portions of images	Image feature vectors
Preprocess Text Reports	Clean up the text report by removing unwanted text	Radiologist text reports	Isolated text
Produce Text Features	Develop the text feature space	Isolated text	Text feature vectors
Discover Graph Kernel	Use the images and text features to find a graph kernel for image classification	Image feature vectors; Text feature vectors; Links	Graph kernel
Develop Classifier	Develop classifier to classify image, e.g., as normal or abnormal	Graph kernel	Classifier
Apply Classifier	Classify an image; Provide confidence indicator	Classifier	Classification; Confidence level

For the purpose of this study, we generalize the system failure mode to "Incorrect Classification or Incorrect Confidence". This is necessary because, at this level of analysis, one cannot determine, to any significant degree, the specific impact of a failure mode on the Classification or on the Confidence measure. The failure model defined in Table 3 will prove more important when module- and lower-level analyses are conducted.

TABLE III. SYSTEM FAILURE MODEL

Classification	Confidence
False Negative	High
	Low
False Positive	High
	Low

A next step in the research is to examine the system's failure modes and effects. A failure mode taxonomy defines the breadth and depth of failure modes to be considered in the research. Using a combination of failure mode taxonomies (e.g., [5, 6]) as a basis, we tailored a taxonomy for use in this analysis. The tailoring includes failure mode space reduction by considering only those failure modes that are possible given requirements-based constraints and the architectural level of the analysis. The resulting taxonomy of failure modes is provided in Fig. 2. This applies to both the baseline (general) and the sample classification system.

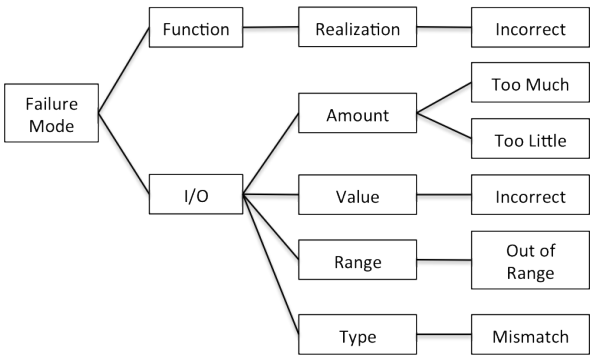

Figure 2. Tailored failure mode taxonomy

The taxonomy defines the failure modes as either function- or input/output (I/O)-related. The only functional failure mode considered in this analysis is the incorrect realization of the module's functionality. This can result from module implementation errors.

I/O failure modes refer to the inputs and outputs of a module. I/O.Amount refers to the number or quantity of input or output. For instance, in the example system, if the requirements state that the Preprocess Images module will input a set of 4 mammography images (i.e., top and side views of left and right breasts) and only one image is received, then this failure mode is referred to as I/O.Amount.Too_Little. An I/O.Value.Incorrect failure mode covers those cases when the input to or output from a module is incorrect. An I/O.Range.Out_of_Range failure mode occurs when the I/O value is outside its requirements-specified bounds/limits. An I/O.Type.Mismatch failure mode includes cases when the expected I/O type and the actual I/O type do not match. The taxonomy is used in defining the failure modes as illustrated in sections 4 and 5.

IV. FUNCTIONAL FMEA

The process of conducting software FMEA helps identify structural weaknesses in the design and identify missing or incorrect requirements. The primary purpose of FMEA is to identify possible failure modes of the system components and evaluate their impact on the system performance.

Software FMEA is conducted here on two levels – the system-level or functional FMEA (this section) and the more detailed level (section 5). The functional FMEA examines each module and for each functional failure mode, determines the local effect and the effect at the system level. The results of the functional FMEA for the multi-modal, semi-supervised learning system is provided in table 4.

For the mammography application, the functional FMEA can be applied mainly in the form and with the content of table 4. The Preprocess [Primary | Secondary] Modality component would be implemented as 2 modules, one each for text and image preprocessing, with the data processed being text and images, respectively. A similar structure is produced for the Produce [Primary | Secondary] Modality Features and the features (text or image) produced.

Additional analysis at this level can include the detectability and reversibility of each failure mode's effect, along with suggested fault mitigation techniques.

TABLE IV. FUNCTIONAL FMEA

Module	Failure Mode	Local Effect	System Effect
Preprocess [Primary | Secondary] Modality	Function. Incorrect_ Realization	Data are preprocessed incorrectly; Data features missing, incorrect, or superfluous	Incorrect Classification or Incorrect Confidence
Produce [Primary | Secondary] Modality Features	Function. Incorrect_ Realization	Incorrect features extracted; features missing, incorrect or superfluous	Incorrect Classification or Incorrect Confidence
Discover Graph Kernel	Function. Incorrect_ Realization	Worthless or skewed feature space	Incorrect Classification or Incorrect Confidence
		Module failure - crash	System failure – crash
Develop Classifier	Function. Incorrect_ Realization	Error in optimization routine that sets the weights; Effect - bad classifier	Incorrect Classification or Incorrect Confidence
Apply Classifier	Function. Incorrect_ Realization	Incorrect decision; Unwarranted or incorrect confidence in result	Incorrect Classification or Incorrect Confidence

V. DETAILED SOFTWARE FMEA

The detailed software FMEA (SW FMEA) examines each module for each I/O or data failure mode and describes the local effect and the effect at the system level. Tables 5-9 provide the failure modes and effects for each module in the learning system.

Recall for the mammography decision support system, the primary modality is the image data and the secondary modality is text. Some of the failure modes shown in figure 2 may not be applicable to all modules. For example, in the system under study, an out of range input does not apply to the Preprocess Images module. In table 5, we've added a column to indicate whether the failure mode is applicable to images only (I), text only (T) or both images and text (B) in the mammography

application. To conduct a detailed software FMEA on the sample application, one would take the following steps:

- Develop a SW FMEA data table for Preprocess Images based on table 5. Using the right-most column in table 5 we see that the rows included in the Preprocess Images SW FMEA are those with a B or an I in the cell.
- Similarly, develop a SW FMEA table for Preprocess Text and include those rows with a B or a T in the right-most column of table 5.
- Develop a SW FMEA for Produce Image Features based on table 6 and including those rows with a B or an I in the right hand column.
- Develop a SW FMEA for Produce Text Features based on table 6 and including those rows with a B or a T in the right hand column.
- Use table 7 for the Discover Graph Kernel SW FMEA and use text as the secondary modality.
- Use tables 8 and 9 as they are presented here for the Develop Classifier and Apply Classifier, respectively, SW FMEA.

TABLE V. SW FMEA FOR PREPROCESS [PRIMARY | SECONDARY] MODALITY MODULE

Failure Mode	Local Effect	System Effect	Modality
I/O.Amount. Too_Much	Missing information	Incorrect Classification or Incorrect Confidence	B
	Incorrect (wrong or ineffective) pre-processing	Incorrect Classification or Incorrect Confidence	T
	Module failure	System Failure (crash)	B
I/O.Amount. Too_Little	Missing information	Incorrect Classification or Incorrect Confidence	I
	Misinterpret data	Incorrect Classification or Incorrect Confidence	T
	Module failure	System Failure (crash)	B
I/O.Value. Incorrect_ Value	Bad or missing input/ information	Incorrect Classification or Incorrect Confidence	I
	Misinterpret data	Incorrect Classification or Incorrect Confidence	T
I/O.Range. Out_of_ Range	n.a.	n.a.	I
	Misinterpret text	Incorrect Classification or Incorrect Confidence	T
	Module failure	System Failure (crash)	B
I/O.Type. Data.Type_ Mismatch	Missing information	Incorrect Classification or Incorrect Confidence	I
	Input ignored; data not interpreted	Incorrect Classification or Incorrect Confidence	T
	Module failure	System Failure (crash)	B

TABLE VI. SW FMEA FOR PRODUCE [PRIMARY | SECONDARY] MODALITY FEATURES MODULE

Failure Mode	Local Effect	System Effect	Modality
I/O.Amount. Too_Much	Missing information	Incorrect Classification or Incorrect Confidence	B
	Module failure	System Failure (crash)	B
I/O.Amount. Too_Little	Worthless feature set	Incorrect Classification or Incorrect Confidence	I
	Misinterpret data	Incorrect Classification or Incorrect Confidence	T
	Module failure	System Failure (crash)	B
I/O.Value. Incorrect_ Value	Missing information	Incorrect Classification or Incorrect Confidence	I
	Misinterpret text	Incorrect Classification or Incorrect Confidence	T
I/O.Range. Out_of_ Range	n.a.	n.a.	I
	Misinterpret text	Incorrect Classification or Incorrect Confidence	T
I/O.Type. Data.Type_ Mismatch	Missing or incorrect images	Incorrect Classification or Incorrect Confidence	I
	Irrelevant feature set	Incorrect Classification or Incorrect Confidence	T

TABLE VII. SW FMEA FOR DISCOVER GRAPH KERNEL MODULE

Failure Mode	Local Effect	System Effect
I/O.Amount. Too_Much	Module failure	System Failure (crash)
	Out of Memory Crash	System Failure (crash)
I/O.Amount. Too_Little	Module failure	System Failure (crash)
I/O.Value. Incorrect_Value	No benefit from secondary modality	Incorrect Classification or Incorrect Confidence
	Bad graph kernel	Incorrect Classification or Incorrect Confidence
I/O.Range. Out_of_ Range	Bad similarity score, bad graph kernel	Incorrect Classification or Incorrect Confidence
I/O.Type. Data.Type_ Mismatch	Skewed similarity score, bad graph kernel	Incorrect Classification or Incorrect Confidence

TABLE VIII. SW FMEA FOR DEVELOP CLASSIFIER MODULE

Failure Mode	Local Effect	System Effect
I/O.Amount. Too_Much	Memory Crash	System Failure (crash)
I/O.Value. Incorrect_Value	Bad model	Incorrect Classification or Incorrect Confidence
	Bad classifier	Incorrect Classification or Incorrect Confidence
I/O.Range. Out_of_Range	Bad classifier	Incorrect Classification or Incorrect Confidence
	Module failure	System Failure (crash)

TABLE IX. SW FMEA FOR APPLY CLASSIFIER MODULE

Failure Mode	Local Effect	System Effect
I/O.Amount. Too_Little	Incorrect default to the first class	Incorrect Classification or Incorrect Confidence
	Module failure	System Failure (crash)
I/O.Value. Incorrect_Value	Misclassify	Incorrect Classification or Incorrect Confidence
	Module failure	System Failure (crash)
I/O.Range. Out_of_Range	0 weight in the classifier	Incorrect Classification or Incorrect Confidence
	Module failure	System Failure (crash)
I/O.Type. Data.Type_ Mismatch	Misclassify	Incorrect Classification or Incorrect Confidence
	Module failure	System Failure (crash)

VI. SUMMARY

In this paper, we provide insight into the black box typically assumed for analyzing complex learning frameworks by presenting a foundation for conducting failure analysis of a framework that incorporates multi-modal and semi-supervised learning. Specifically, we provide templates for a fault model, and FMEA for the learning framework, and describe its use in analyzing an implementation for a radiologist assistant system. In doing so, we have elucidated some of the issues associated with V&V of complex learning frameworks.

VII. FUTURE WORK

This research provides the basis for future research in complex learning framework dependability. In the FMEA discussion, we noted that several failures could be mitigated. A natural extension to the work presented is to provide fault mitigation suggestions, knowing that many of these will fall under the category of good design or programming practices for machine learning implementations.

Some of the more promising methods for V&V in machine learning, such as metamorphic testing [10], are not currently designed to handle major aspects of newer learning algorithms. For example, metamorphic relations would have to take into account different effects based on which modality is modified or whether labeled or unlabeled data is altered. If relations are confined to labeled data of the primary modality, many relations (such as re-prediction) still cannot be used, but more importantly, major portions of the learning process would be completely ignored. This is a particularly fertile area for future research, and we will consider methods, relations, etc. that can handle these new complexities.

ACKNOWLEDGMENT

Research sponsored in part by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

REFERENCES

[1] Symons, C.T., et al. 2009. A multimodal, semi-supervised learning system for building better decision support systems for the analysis of mammograms. In *Radiological Society of North America 95th Scientific Assembly and Annual Meeting Program* (Oak Brook, Ill. USA). RSNA 2009. SSA11-08.

[2] Mitchell, T.M. 1997. *Machine Learning*, McGraw-Hill.

[3] Chapelle, O., et al., eds. 2006. *Semi-Supervised Learning*, MIT Press.

[4] Nakamura, E.F., et al. 2007. Information fusion for wireless sensor networks: methods, models, and classifications. *ACM Computing Surveys*. 39(3).

[5] Li, B., Li, M., Ghose, S., and Smidts, C. 2003. Integrating Software into PRA. In *Proceedings of the 14th International Symposium on Software Reliability Engineering*. ISSRE'03.

[6] Avizienis, A., Laprie, J.-C., Randell, B., and Landwehr, C. 2004. Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions on Dependable and Secure Computing* 1 (Jan.-Mar., 2004), 11-33.

[7] Ye, F., and Kelly, T. 2004. Contract-based justification for COTS component within safety-critical applications. In *9th Australian Workshop on Safety Related Programmable Systems* (Brisbane, Australia, 2004). SCS'04. Australian Computer Society, Inc., 13-22.

[8] Pullum, L.L., and Dugan, J.B. 1996. Fault tree models for the analysis of complex computer-based systems. In *Proceedings of the Annual Reliability and Maintainability Symposium*. RAMS'96. 200-207.

[9] Pullum, L.L., et al. 2010. Architecture-level dependability analysis of a medical decision support system. In *Intl Conf on SW Eng (ICSE) Workshop on SW Eng in Health Care* (Cape Town, South Africa). SEHC'10.

[10] Xie, X., et al. 2009. Application of metamorphic testing to supervised classifiers. In *Proc. of the 9th International Conference on Quality Software*. QSIC, 135–144.