Sampling Program Inputs with Mutation Analysis: Going Beyond Combinatorial Interaction Testing

Mike Papadakis, Christopher Henard, and Yves Le Traon
Interdisciplinary Centre for Security, Reliability and Trust (SnT)
University of Luxembourg, Luxembourg, Luxembourg
michail.papadakis@uni.lu, christopher.henard@uni.lu, yves.letraon@uni.lu

Abstract—Modern systems tend to be highly configurable. Testing such systems requires selecting test cases from a large input space. Thus, there is a need to systematically sample program inputs in order to reduce the testing effort. In such cases, testing the interactions between program parameters has been identified as an effective way to deal with this problem. In these lines, Combinatorial Interaction Testing (CIT) models the program input interactions and uses this model to select test cases. Going a step further, we apply mutation analysis on the CIT input model to select program test cases. Mutation operates by injecting defects to the program input model and measures the number of defects found by the selected test cases. Experiments performed on four real programs show that measuring the number of model-based defects gives a stronger correlation to code-level faults than measuring the number of the exercised interactions. Therefore, the proposed mutation analysis approach forms a valid and more effective alternative to CIT.

Keywords—Mutation Analysis, Combinatorial Interaction Testing, Fault Detection

I. INTRODUCTION

Modern software systems tend to be highly configurable and thus, they involve a vast number of parameters and configurations. This characteristic of the systems results in forming enormous input configuration spaces. We call these spaces as input or test spaces and we use the terms program inputs and test cases indifferently. Testing these systems requires exercising the programs behavior with different configurations. As a consequence, exhaustive testing is practically intractable. Since testing is a complex and costly process, a systematic and rigorous approach should be performed. To this end, the test selection should aim at choosing the tests that are potentially the best ones, i.e., revealing the highest number of defects. The question that it is raised is how to perform such a selection?

Testing the interactions between program parameters has been proposed as a possible method to answer this question [1], [2]. This proposition is based on the observation that most of the faults are triggered by the interactions between a small number of variables [3]. For instance, Kuhn et al. [3] have shown that interactions between two variables are able to disclose 80% of the bugs. Thus, program inputs can be systematically sampled from the test space with the aim of minimizing the number of the selected tests that exercise a specific number of interactions. In other words, this method eliminates the redundancy between the selected tests. The redundancy is measured in terms of interactions.

Testing the combinations of interactions between the input parameters has been shown to be a powerful technique [4]. This technique is called Combinatorial Interaction Testing (CIT) or t-wise testing and aims at selecting those tests that exercise all the interactions between any t parameters. Different values of t yield different selection criteria. Generally, a higher t value implies a stronger criterion. Also, selecting tests according to t-wise interactions subsumes, i.e., satisfies the requirements of t − 1. Applying CIT requires a model that represents the test space. This model is composed of the input parameters associated with the possible values that they can take and some constraints. The constraints, named here as input constraints represent requirements on the use of variables or on the use of variable values. These constraints are very important in order to effectively perform the testing process [4], [5]. Thus, they must be fulfilled by the test cases. Test cases violating these constraints are simple false positives [4].

Generally, t-wise testing serves as a yardstick towards assessing the ability of the selected test cases to reveal faults. In other words, t-wise forms a measurement of the test suite effectiveness. It is calculated by counting the number of valid combinations that are exercised by the test suite. To this end, the ratio of the exercised t-wise combinations to the total number of the existing ones is referred to as the interaction coverage [4]. Therefore, given the input model and two sets of test cases, interaction coverage identifies which set is the most effective one, i.e., the one with the higher coverage. Using this information, testers can prioritize their test suites by using first the tests that cover the most interactions [4], [6]., they can reduce the size of a test suite [1], [2] by removing test cases that do not cover additional interactions, and they can guide the generation or selection of new test cases based on the uncovered interactions [6], [7].

In the same lines, the present paper applies mutation analysis to the CIT input model. Thus, it introduces an alternative but more representative measure of the effectiveness of a test suite. Traditionally, mutation analysis is applied at the program code and aims at evaluating the quality of a test suite [8]. It operates by introducing artificial defects, called mutants, in the code of the tested program. Thus, multiple versions of the program under test are produced. Each version contains a defect that is introduced by making a slight modification. The test suites are then evaluated based on their ability to distinguish the introduced problems [8]. Contrary to the
Thus, the boolean model $M_b$ equivalent to $M$ is defined as:

$$M_b := a_{p1} \in \{true, false\}, b_{p1} \in \{true, false\}, c_{p2} \in \{true, false\}, d_{p2} \in \{true, false\}, e_{p2} \in \{true, false\}, f_{p3} \in \{true, false\}, (a_{p1} \Rightarrow \neg b_{p1}), (b_{p1} \Rightarrow \neg a_{p1}), (c_{p2} \Rightarrow \neg d_{p2}), (c_{p2} \Rightarrow \neg e_{p2}), (d_{p2} \Rightarrow \neg e_{p2}), (d_{p2} \Rightarrow \neg c_{p2}), (e_{p2} \Rightarrow \neg c_{p2}), (e_{p2} \Rightarrow \neg d_{p2}).$$

II. Example

This section introduces an illustrative example to explain how mutation analysis can be applied. The approach is based on a model of the program inputs, as those typically used by CIT approaches, e.g., $M$. Such a model encompasses the different parameters and the constraints between these parameters.

A. An Input Model

Consider the following model $M$ involving three parameters $p_1, p_2$ and $p_3$. Each parameter is a variable of the model. The parameter $p_1$ can take the two values $a$ and $b$, $p_2$ can take the three values $c, d, e$ and the last parameter $p_3$ can only take the $f$ value. Thus, the model $M$ is defined as follows:

$$M := p_1 \in \{a, b\}, p_2 \in \{c, d, e\} \text{ and } p_3 \in \{f\}.$$
With respect to CIT, \( T_1 \) covers \( \frac{15}{66} \) 2-wise interactions while \( T_2 \) covers \( \frac{24}{66} \). As a result, for CIT, the second test suite is more effective as it covers more interactions than the first one.

### C. Case 2: Presence of Input Constraints

When there are input constraints in the model \( M \), we simply transform them to boolean ones and we add them to the constraints of \( M_b \). For instance, suppose \( M \) contains the input constraint \((p_1 = a) \Rightarrow (p_2 = c)\). It is transformed into \((a_1 \Rightarrow c_2)\). Thus, in this case, the boolean model of \( M \) is: \( M_b := \{a_1, b_1, c_2\} \). We evaluate each input towards the mutated model to find any defects found by a test suite. We can say that \( I_1 \) did not cover any of the two mutants and that \( I_2 \) violated half of the mutants. Thus, with respect to our approach, \( I_2 \) is more effective.

With respect to CIT, for instance 2-wise interactions, \( I_1 \) covers 15 interactions. An example of such an interaction is \((a_1 \Rightarrow b_1 \Rightarrow f_3)\). The total number of 2-wise interactions of the model \( M_b \) is 66. Thus, the CIT measure for \( I_1 \) is \( \frac{15}{66} \). Consider now the following two test suites:

- \( T_1 = \{I_1\} \)
- \( T_2 = \{I_1, I_2\} \)

With respect to our approach, both the test suites are similarly effective since they both violate one of the two mutants. Thus, we measure \( \frac{1}{2} \) for both \( T_1 \) and \( T_2 \), which is the number of violated mutants to the total ones.

Fig. 1. The mutation analysis approach. The technique operates on a model of the program inputs by creating defective (i.e., mutated) model versions. The application of the approach is equivalent to CIT since it uses the same input model. However, instead of measuring the interactions covered by the test cases, as done by CIT, it measures the ability of the test cases to distinguish the defective versions.
as \((p_i = v) \Rightarrow (p_j \neq w)\). Thus, a program input composed of both \(p_i = v\) and \(p_j = w\) does not satisfy the constraint since \(p_j\) should not be set to \(w\) when \(p_i\) is set to \(v\). We say that a test case satisfies the model of the program inputs if all the constraints of this model are satisfied at the same time.

Each constraint involves several variables that correspond to the parameters of the program. A variable has a domain which corresponds to the values that the parameter can take. For instance, the variable \(p_1\) may take two values \(v\) and \(w\). In that case, \(p_1\) has a domain involving two values, \(v\) and \(w\).

In practice, our approach requires a flattening of the model to make it boolean. This is a typical process undertaken by most of the CIT tools, e.g., [7]. It also gives the opportunity to the mutation approach to produce mutants in the case that no input constraints exist. Instead of having a model with variables corresponding to parameters, the flattened model contain variables that represent all the possible values for all the parameters. Each one of these variables can be true or false, depending on whether this variable is assigned an input value. For instance, if a parameter can take three different values, the flattening transforms the parameter variable, which has a domain of three values into three different variables. Doing so transforms the model to a boolean one as required by the proposed approach.

### B. The CIT Approach

The CIT approach works by counting the unique number of interactions (or combinations) between any \(t\) parameters’ values exercised by the test suite. Such interactions are called \(t\)-wise interactions, as they involve combinations between the \(t\) parameters. Thus, given the input model, we evaluate all the possible \(t\) combinations of the input parameters. Then, based on the input constraints, we eliminate the invalid ones, i.e., the combinations that are prohibited by the constraints. Finally, the effectiveness measure of the test suite is calculated based on the \(t\)-wise coverage which is the number of \(t\) combinations that are covered by the test cases.

### C. The Mutation Approach

The mutation approach operates by altering the boolean model of the program inputs. It actually produces defective models by altering the model constraints. Thus, it creates several versions of the model, called mutants. Each mutant contains only one altered constraint. The constraints are altered based on a set of syntactic rules called mutation operators. Thus, by applying the mutation operators on all the model constraints, we end up with the sought set of mutants.

### Table I

<table>
<thead>
<tr>
<th>Input</th>
<th>Applies on</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>A constraint (C = v_1 \lor \ldots \lor v_i \lor \ldots \lor v_k)</td>
<td>A variable (v_i) of the constraint</td>
<td>One constraint (C' = v_1 \lor \ldots \lor \neg v_i \lor \ldots \lor v_k)</td>
</tr>
<tr>
<td>A constraint (C = v_1 \lor \ldots \lor v_i \lor \ldots \lor v_k)</td>
<td>A disjunction operator of the constraint</td>
<td>Two constraints: (C' = v_1 \lor \ldots \lor v_i) and (C'' = v_{i+1} \lor \ldots \lor v_k)</td>
</tr>
</tbody>
</table>

The constraints of the flattened model are boolean and they are represented as a disjunction between variables. Thus, each constraint \(C\) between \(k\) variables has the general form \(C = \bigvee_{i=1}^{k} v_k\), where \(v_k\) is a variable (corresponding to one parameter’s value) either set to true or false.

In this work, we employ two mutation operators adapted from [9]. These operators are presented in Table I. The first operator alters a constraint by taking one of its variables and negating it. In other words, if the selected variables is true, it becomes false, and conversely. The second operator aims at creating two constraints from the original one. To this end, one of the disjunction operator in the constraint is replaced by a conjunction operator. Thus, this second operator splits a constraint into two, increasing by one the total number of constraints of the model.

A test suite is evaluated based on its ability to distinguish the defective models from the original one. We refer to the mutants that are distinguished by the test cases as killed and to those that can not be distinguished as live. However, how can we check whether a mutant is killed or not? To answer this question, we need to consider that our models are composed of boolean constraints. Therefore, the evaluation is straightforward. It is examined whether a test case satisfies the constraints of the mutant model with a satisfiability (SAT) solver. The mutant is killed when its constraints are violated by the test case. In the opposite case it is live. It is noted that only the constraints of the mutant models can be violated. The constraints of the original model must always be satisfied in order to have valid program inputs. By determining the number of mutants that are killed by all the test cases, the overall effectiveness measure of the test suite is calculated.

### IV. Experimental Methodology

The aims of the conducted experiment are summarized in the following research questions:

- **RQ1**: How well does the mutation-based approach evaluate the quality of the selected test suite?
- **RQ2**: How well does the CIT approach evaluate the quality of the selected test suite?
- **RQ3**: How does the mutation approach compares with CIT?

Answering the first question is important in order to substantiate the practical use of mutation. Answering the second question is important in reinforcing the empirical evidence in favor of CIT.

Now, suppose that mutation and CIT are capable of predicting accurately the fault detection ability of test suites. In
this case, practitioners will be able to evaluate the quality of their test suites. Going a step further, they will be able to prioritize their tests (by pointing first the test cases that cover the majority of the interactions or kill most of the mutants), reduce the suites’ size (by removing redundant tests) and guide the test generation process (by generating test cases that cover new interactions or kill additional mutants). However, in this case, which one should be used? This is our third research question which aims at identifying which of the two examined approaches should be used in practice.

A. Definition of the Experiment

To answer the stated research questions, we analyze the ability of test cases to cover t-wise interactions, to kill our mutants, and to expose code-based faults. We employ four subjects that are accompanied with test cases and faulty versions (Section IV-B). We then sample at random 30 test suites from the initial test suite of each program. Thus, we sample suites of random size from \(4 \leq n \leq N\), where \(N\) is the size of the initial test suite. The minimum size of 4 test cases per suite has been chosen in order to have a sufficient sample for the correlation analysis. Then, we measure three metrics. a) the number of interactions covered, b) the number of mutants killed and c) the number of faults found (Section IV-C). These measures are recorded for every test case of the selected suites.

Then, these metrics are examined in order to identify possible correlations between them and to answer to RQs 1 and 2, Section IV-D. To this end, we perform a statistical analysis to quantify these correlations. In other words, we try to measure the extent to which the relationship of covering interactions and killing mutants relates with fault detection. Thus, for each subject program, 30 Kendall rank coefficients are obtained by evaluating the correlation between the killed mutants and the faults found by the test cases, case denoted as MF; and 30 coefficients are obtained from the correlation between the t-wise coverage and the faults found. This latter comparison is denoted as tWF. In this work, we consider t-wise coverage for \(t = 2, ..., 4\).

Finally, we compare the two methods based on the level of correlations, thus, answering to RQ3. An overview of the followed process is depicted by Figure 2.

B. Subjects

We use the following programs: flex, gzip, make and sed. These subjects are taken from the Software-artefact Infrastructure Repository (SIR) [10] and their details are presented in Table I. The examined versions of these program were randomly selected. Hence, for each subject, Table II records its size in uncommented lines of code \([\text{LOC}]\) the number of faults per version taken from the faults matrix provided by the SIR, the number of variables and constraints of its respective model, the number of mutants and killable (i.e., there is at least one test configuration that cannot satisfy the faulty model) mutants obtained by applying the mutation operators, the number of the test cases contained in the initial test suite and the number of t-wise interactions for \(t = 1, ..., 4\).

The input models are taken from the study of Petke et al. [4]. This study concerns the test case prioritization according to CIT and thus, their models are well suited for the present study. These models were built based on the descriptions of the Test Suite Specification Language (TSL) that are proposed with the utilized programs [4]. Thus, the parameters and values of these models represent the program input space. As described in Section III, we transform these models to boolean ones in order to evaluate the various test cases.

C. Evaluating Test Suites

1) T-wise: Given a test suite, we evaluate its t-wise coverage based on the following process. All the t-wise interactions between the parameters’ values covered by the first test case of the suite are recorded. Then, we consider the second test case and we add all the interactions that are not exercised by the first one. This process is repeated for all the test cases of the suite and it gives the cumulative number of unique t-wise interactions covered by each one of suite’ test cases. Thus, given a test suite of n test cases \(tc_1, ..., tc_n\), we obtain the t-wise coverage represented by the tuples \((tc_i, c_i)\).

2) Fault Detection: Given a test suite, we evaluate its fault detection ability based on the following process. We first take the faults found by the first test case of the suite by using the fault matrix provided by the SIR. This matrix contains the faults found by each test case. These faults are not considered while evaluating the next test cases. Then, the number of faults found by the second test case is recorded. This process is

\[\text{http://cloc.sourceforge.net/}\]
TABLE II
THE FOUR SUBJECTS PROGRAMS USED IN THE EXPERIMENTS.

<table>
<thead>
<tr>
<th>Subject</th>
<th>flex</th>
<th>gzip</th>
<th>make</th>
<th>sed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncommented lines of code</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v1</td>
<td>9,581</td>
<td>4,604</td>
<td>14,459</td>
<td>-</td>
</tr>
<tr>
<td>v2</td>
<td>-</td>
<td>5,092</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>v3</td>
<td>-</td>
<td>-</td>
<td>7,161</td>
<td>-</td>
</tr>
<tr>
<td>v4</td>
<td>11,470</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>v7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>14,177</td>
</tr>
<tr>
<td>Faults</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v1</td>
<td>19</td>
<td>16</td>
<td>19</td>
<td>-</td>
</tr>
<tr>
<td>v2</td>
<td>-</td>
<td>7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>v3</td>
<td>-</td>
<td>-</td>
<td>6</td>
<td>-</td>
</tr>
<tr>
<td>v4</td>
<td>16</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>v7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Model variables</td>
<td>23</td>
<td>29</td>
<td>20</td>
<td>34</td>
</tr>
<tr>
<td>Model constraints</td>
<td>43</td>
<td>91</td>
<td>21</td>
<td>143</td>
</tr>
<tr>
<td>Model mutants</td>
<td>139</td>
<td>295</td>
<td>63</td>
<td>527</td>
</tr>
<tr>
<td>Killable model mutants</td>
<td>139</td>
<td>292</td>
<td>63</td>
<td>373</td>
</tr>
<tr>
<td>Test cases</td>
<td>500</td>
<td>159</td>
<td>768</td>
<td>144</td>
</tr>
<tr>
<td>All 2-wise interactions</td>
<td>1,035</td>
<td>1,653</td>
<td>780</td>
<td>2,278</td>
</tr>
<tr>
<td>All 3-wise interactions</td>
<td>15,180</td>
<td>30,856</td>
<td>9,880</td>
<td>50,116</td>
</tr>
<tr>
<td>All 4-wise interactions</td>
<td>163,185</td>
<td>424,270</td>
<td>91,390</td>
<td>814,385</td>
</tr>
<tr>
<td>Valid 2-wise interactions</td>
<td>939</td>
<td>1,388</td>
<td>736</td>
<td>1,421</td>
</tr>
<tr>
<td>Valid 3-wise interactions</td>
<td>11,478</td>
<td>19,980</td>
<td>8,268</td>
<td>23,075</td>
</tr>
<tr>
<td>Valid 4-wise interactions</td>
<td>95,176</td>
<td>194,974</td>
<td>63,475</td>
<td>265,698</td>
</tr>
</tbody>
</table>

repeated for all the test cases of the test suite and provides the cumulative number of unique faults found by each test case. Thus, given a test suite of \( n \) test cases \( tc_1, \ldots, tc_n \) we obtain the number of faults found after executing each test case. It is represented by the tuples \( (tc_i, f_i) \).

3) Mutation: We evaluate a test suite according to mutation based on the following process. Initially, the number of mutants killed by the first test case is determined. These mutants are removed and the second test case is evaluated according to all the remaining mutants. This process is repeated for the whole suite. Thus, the process gives the cumulative number of the unique killed mutants after executing each test case. Hence, given a test suite of \( n \) test cases \( tc_1, \ldots, tc_n \), we obtain the number of mutants killed after executing each test case. It is represented by the tuples \( (tc_i, m_i) \).

D. Rank Correlation Analysis

Evaluating a test suite according to \( t \)-wise, fault detection and mutation, as described in the previous section gives the following information after executing \( k \leq n \) tests of the test suite:

1) The current \( t \)-wise coverage achieved after considering the \( i \)-th test case of the test suite,

2) The current number of faults found after executing the \( i \)-th test case of the test suite,

3) The current number of mutated models that cannot be satisfied after considering the \( i \)-th test case of the test suite.

Given these three measures, we evaluate whether 1) correlates with 2), whether 3) correlates with 2), and which of these two correlations is better. In order to evaluate these correlations, we compute the Kendall \( \tau \) rank correlation coefficient. This coefficient is considered as the most robust and usefully interpreted statistical measure for this question [11, 12, 13]. Thus, the coefficient is one hand calculated given the correlation between the tuples \( (tc_i, c_i) \), \( (tc_i, f_i) \) and on the other hand given the tuples \( (tc_i, m_i) \) and \( (tc_i, f_i) \).

The Kendall \( \tau \) is in the range \(-1 \leq \tau \leq 1\). A coefficient of 1 indicates that the correlation between the two considered ranking is perfect. A coefficient around 0 denotes that the two observed sample are independent. Finally, a \( \tau \) equals to \(-1\) represents the complete absence of correlation between the two considered rankings.

V. EXPERIMENTAL RESULTS

This section reports results regarding the ability of the CIT and the mutation approaches to reveal actual faults. Then, it compares the two approaches.

A. Correlation Analysis (RQ1 and RQ2)

Figure 3 presents the distribution the Kendall coefficients for all the subject programs. MF denotes the coefficients resulting from the correlation analysis between the mutants killed and the faults found by the test cases. The correlation coefficient between the \( t \)-wise coverage of the test cases and the faults found on the program are denoted as \( tWF \), with \( t = 2, \ldots, 4 \).

From these results we can infer three interesting conclusions. First, we observe that both CIT and mutation strongly correlate with fault detection for all the employed subjects. This is due to the fact that most of the coefficients are greater than 0.5 (median values) and almost all are at least 0.3, thus indicating very good correlations. Second, we observe big differences on the correlations between the subject programs and between the different versions. For example, sed v3 has coefficients close to 0.5 while gzip v2 is close to 0.2. Similarly, gzip has differences between v1 and v2 where the coefficients are close to 0.5 and 0.2 respectively. Third, we observe that higher \( t \)-values results in weaker correlations than lower \( t \) values in all the examined cases. The next section compares \( t \)-wise and CIT based on these results.

B. Comparing CIT and Mutation (RQ3)

In order to compare the CIT approach with the mutation one, consider the MF and the \( tWF \) coefficients. From Figure 3 it is evident that MF coefficients tend to be closer to 1 than the \( tWF \) ones. Even when they are not very good, such as the case of gzip v2, they are greater than the \( tWF \) coefficients. The difference is big for three out of the 7 cases and always greater to all the \( tWF \) coefficients. This denotes a higher correlation
Results for `flex` v1 and v4. FX denotes `flex` vX, MF denotes the correlation mutants killed and faults found, and tWF denotes the correlation t-wise coverage and faults found.

Results for `gzip` v1 and v2. GX denotes `gzip` vX, MF denotes the correlation mutants killed and faults found, and tWF denotes the correlation t-wise coverage and faults found.

Results for `make` v1. MX denotes `make` vX, MF denotes the correlation mutants killed and faults found, and tWF denotes the correlation t-wise coverage and faults found.

Results for `sed` v1 and v7. FX denotes `sed` vX, MF denotes the correlation mutants killed and faults found, and tWF denotes the correlation t-wise coverage and faults found.

Fig. 3. Distribution of the Kendall $\tau$ rank coefficients on different versions of the subject programs. Each boxplot represents the distribution of the 30 $\tau$ coefficients of correlation between either the mutants killed and the faults found or between the t-wise ($t = 2, ..., 4$) coverage and the faults found.

between the mutants and the faults than between the t-wise coverage and the faults.

For instance, consider the results for `flex` v1, depicted by Figure 3a. The maximum coefficient $\tau$ is very close to 1. The median $\tau$ among the 30 coefficients is above 0.87. It thus denotes a strong correlation for the MF case. Regarding the t-WF results, median coefficients are below 0.5, which denote moderate correlations between t-wise and the faults found by the test cases. Conclusively, it can be argued that the comparison is in favor of the MF since it gives correlations greater than the t-WF.

The results presented so far consider the correlation of CIT and mutation with the code-based faults. However, this correlation may be misled by the difficulty of finding the faults. In other words, if the faults are very easy to find or very difficult to find, this will have a direct effect on the measured correlations. In the same lines, we can compare the CIT and the mutation approaches.

Figure 4 presents the easiness of finding faults, killing mutants, covering 2-wise and covering 3-wise interactions...
per subject. The easiness of finding a fault represents the percentage of test cases that find this fault. Similarly, the easiness of killing a mutant represents the percentage of test cases that kill this mutant, and the easiness of covering a t-wise interaction is the percentage of test cases that cover this interaction. FX, GX, MX, and SX respectively represent the flex, gzip, make and sed subjects, with the corresponding X version. For each subject, the fault box represents the easiness of the faults, the mutants box represent the easiness of the mutants and the t-wise box the easiness of the t-wise interactions.

From these results, we can explain why both CIT and mutation have low correlation values for gzip v2. This is due to the fact that the faults of this version are very easy to detect, contrary to mutants and interactions. It is the same for make. However, in this case, both mutants and interactions are easy to detect, resulting in satisfactory correlations.

Conclusively, from the presented results, it becomes evident that killing mutants is more or less as difficult as covering 2-wise and 3-wise interactions. Indeed, the easiness median values for mutants are comparable with the easiness medians of 2-wise and 3-wise interactions for sed, greater for gzip and lower for flex and make. Considering the easiness of mutants and faults, we observe that mutants tend to behave similarly as faults.

Finally, covering 3-wise interactions is harder than covering the 2-wise ones. This is expected since covering all 3-wise interactions results in covering all the 2-wise ones. As a result, higher interactions strengths (t ≥ 4) are more difficult to cover.

VI. DISCUSSION

The findings of the conducted experiment suggest that the mutation approach can form an alternative method to CIT. Based on the results, we can infer that mutation is probably more effective in predicting the actual fault detection of a test suite. In view of this, some additional considerations regarding the comparison with CIT and the application cost of the approach are needed. This section discusses these considerations as long as threats to the validity of the conducted experiment.

A. Additional Consideration about Mutation and CIT

By considering the CIT and the mutation approaches, we can observe that they more or less both use the same information, which is the input model. However, they provide much different results. Why this difference? In other words, why mutants can be more powerful than the input combinations?

Generally, it is very hard to fully answer this question. A full answer requires extensive and independent studies. However, we believe that the power of mutation lies on the fact that it considers the input constraints of the tested systems. Recall that our benchmark programs are real word programs. Thus, they have input constraints. These constraints play an important role in the testing process of the system. Not only they define the valid program inputs but they also reflect a logic of the underlying system. Therefore, they provide some useful information to the testers which is actually missed by CIT. Mutation takes advantage of this information by mutating these constraints. Mutating the input constraints forces tests to exercise limit cases that trigger faults. From the testing perspective, input constraints provide information similar to the one provided by boundary conditions of a system. They also have a role which is similar to the role of the preconditions of a system. Testing preconditions and boundary conditions of a system has been identified as an important step of the testing process [14]. Hence, testing them is necessary for establishing a rigorous testing approach. Here, it should be noted that CIT uses the input constraints only for computing the valid combinations of program inputs. Thus, it completely ignores both their importance and the information they provide.

Generally, mutation analysis relies on the power of the utilized mutants. The present studies uses two boolean operators mainly chosen based on the authors’ experience from the feature modeling of software product lines [9, 15, 6]. Feature models are boolean models like the flattened ones used by the present study and hence, the employed operators form a good choice. In future, we plan to investigate the use of other operators, e.g., [16] and higher order ones [17, 16]. Nevertheless, the performed analysis shows that the utilized mutants simulate very well the behavior of the actual faults.
Here, it should be mentioned that the use of mutant selection strategies can increase the difficulty of finding them. However, the main question is whether doing so results in accurate estimations of the actual fault detection. Particularly, we do not want to underestimate or overestimate our measures. This matter falls outside the scope of this paper since it is a general research challenge of the whole mutation testing area.

B. Cost of the Approach

One of the main issues of mutation analysis is its computational cost. This is due to the need to introduce and execute a vast number of mutants. However, in our context the computational cost of mutation analysis is not very important due to the following three reasons. First, we only check whether a mutant violates the boolean constraints of the mutated models. This is a simple SAT verification process which is actually quite fast. Second, the number of mutants is small. Actually much smaller than the number of interactions, see Table II. For 2-wise, the number of mutants is 5 times lower than the number of interactions. As a result, mutation requires less operations than CIT. Third, we do not execute the system. Test selection and evaluation based on the input model is much faster than the actual program execution. Furthermore, in practice, testers will have to verify the program behavior, i.e., resolving the oracle problem, which is a typical manual activity. Hence, human time dominates the computational expenses of the approach.

Finally, it should be mentioned that equivalent mutants, i.e., mutants that can not be killed by any test case do not introduce a big overhead. Actually, the number of such mutants is much less than the number of invalid pairs (Table II). Therefore, the computational cost of mutation is lower than the cost of CIT.

C. Threats to Validity

There are several influencing factors that can threaten the validity of the conducted experiment. Regarding the generalization of the findings, i.e., external validity, it is possible that the selected programs are not representative. This may also be the case for the utilized test suites and the faulty versions. Thus, on larger or other types of programs, mutants and input interactions might not be the most appropriate choice. Similarly, the examined approaches might not being effective in revealing other faults. However, the chosen subjects are real world programs widely used in the literature e.g. [4], [20], [21]. Additionally, both the test suites and faults were developed by researchers independently of the present study. The problem we are facing is the absence of programs with high quality test cases, well defined input models or specifications, and faulty versions. Clearly, more studies are in need to answer this concern with confidence.

With respect to the confidence on the reported results, i.e., internal validity, issues on the utilized input models, the employed test sets and the correctness of the used tools can be identified. It is possible that errors on the input models and the used tools may have influenced the reported results. To reduce this threat we performed several manual checks on both the implementation and the employed input models. Here, it must be noted that the input models were independently developed by other researchers [4]. They were also checked by us in order to give confidence about their correctness. Additionally, as already mentioned, the employed suites are widely used in software engineering experiments e.g. [4], [20], [21].

Finally, some threats regarding the evaluation metrics used, i.e., construct validity, can be identified. It is likely that the number of faults found by the approaches do not express the real fault detection ability of the test suites. Additionally, it is possible that the faults number and difficulty can influence the significance of the performed statistical analysis. To reduce this threat, we employed the Kendall τ coefficient which is a non-parametric hypothesis test, i.e., it does not require a very big sample size, and it measures the similarity of the data order when ranked by the studied effectiveness measures, i.e., the mutants found or the interactions covered. We also measured the correlations after executing every test case of various test suite sizes to eliminate the effects of the test suite size.

VII. RELATED WORK

Mutation analysis is a powerful technique with multiple applications. Generally, code-based mutants have been used to guide the test generation process to assist the debugging activities and to evaluate the fault detection ability of a test suite. The technique has also been applied to test specification models and to capture semantic errors of the programs. For instance, Motto et al. used mutation to test model transformations. Other applications of this technique include Petri nets and Feature Models. Contrary to these work, the present paper applies mutation analysis to the model of the program inputs and measures the correlation between model-based mutants and code-based faults.

CIT is a well researched technique with multiple criteria and combination strategies. However, very few work consider the fault detection ability of CIT e.g. [4], [3], [5], [29], [30], [31]. In the most recent one, Petke et al. [4] show that higher t strengths result in finding more faults than lower strengths. Our work differs from this one in three ways. First, we use mutation while they only consider t-wise. Second, we consider the correlation between faults and t-wise interactions. They only investigate whether covering higher interaction strengths results in higher fault detection. Third, we use randomly selected test suites while they use test suites selected with a covering array tool. Similarly, Arcuri and Briand showed that random testing can perform similarly to CIT on large-scale models. However, their results hold only in the case where there are no input constraints.

Considering boolean specifications, mutation faults have been used to select minimum test suites. Similarly, Kaminski et al. use a logic mutation approach to measure test data quality. Their approach relies on the notion of higher order mutants and aim at improving logic-based testing. In another work, Kaminski et al. target at augmenting logic-based criteria inspired by the mutation approach. Contrary
to these approaches, the present paper applies mutation on the program input model. Additionally, we measure the fault detection ability of this approach. Andrews et al. showed that generated mutants can be used to predict the detection effectiveness of real faults [18]. They investigate the relative cost and effectiveness of different testing coverage criteria. Here, we do not focus on whether or not the generated mutants of the model are representative of real defects.

Finally, the Kendall coefficient has been used in several work to measure the the correlation between two measured quantities. For instance, Gilgoric er al. [11] performed a correlation analysis using this coefficient in order to evaluate the relationship between coverages and mutation score. In this work, we perform a correlation analysis by measuring the Kendall \( \tau \) between a) the number of input parameter interactions covered by a given test suite and b) the number of the introduced mutants distinguished by the test suite with its actual fault detection.

VIII. CONCLUSION

In this paper, we proposed a mutation analysis approach as an alternative technique to CIT. We conducted a correlation analysis between a) the CIT and b) the mutation approach with their actual fault detection. Our results suggest that our mutants have a stronger correlation with code-level faults than the input interactions of the CIT approach. Therefore, mutation forms a valid measure of the test suites’ quality.

Our future work includes experimentations with additional mutant operators, like replacing one parameter with another one. Additionally, possible ways of combining the examined techniques are also under investigation. Finally, we intend to broaden our study with additional programs and faults.

IX. ACKNOWLEDGMENT

The authors would like to thank Justyna Petke and Shin Yoo for for their help in understanding and setting the input models used by the present study.

REFERENCES