CHALMERS
UNIVERSITY OF TECHNOLOGY

Learning How to Search: Generating Exception-Triggering Tests Through Adaptive Fitness Function Selection

Citation for the original published paper (version of record):

Almulla, H., Gay, G. (2020). Learning How to Search: Generating Exception-Triggering Tests Through Adaptive Fitness Function
Selection. Proceedings - 2020 IEEE 13th International Conference on Software Testing, Verification and Validation, ICST 2020: 63-73. http://dx.doi.org/10.1109/ICST46399.2020.00017

Learning How to Search: Generating Exception-Triggering Tests Through Adaptive Fitness Function Selection

Hussein Almulla
Department of Computer Science & Engineering
University of South Carolina
Columbia, SC, USA
halmulla@email.sc.edu

Gregory Gay
Department of Computer Science & Engineering
Chalmers and the University of Gothenburg
Gothenburg, Sweden
greg@greggay.com

***Abstract*—Search-based test generation is guided by feedback from one or more fitness functions—scoring functions that judge solution optimality. Choosing informative fitness functions is crucial to meeting the goals of a tester. Unfortunately, many goals—such as forcing the class-under-test to throw exceptions—*do not* have a known fitness function formulation. We propose that meeting such goals requires treating fitness function identification as a secondary optimization step. An *adaptive* algorithm that can vary the selection of fitness functions could adjust its selection throughout the generation process to maximize goal attainment, based on the current population of test suites. To test this hypothesis, we have implemented two reinforcement learning algorithms in the EvoSuite framework, and used these algorithms to dynamically set the fitness functions used during generation.**

We have evaluated our framework, EvoSuiteFIT, on a set of 386 real faults. EvoSuiteFIT discovers and retains more exception-triggering input and produces suites that detect a variety of faults missed by the other techniques. The ability to adjust fitness functions allows EvoSuiteFIT to make strategic choices that efficiently produce more effective test suites.

***Index Terms*—Automated Test Generation, Search-Based Software Engineering, Reinforcement Learning**

I. Introduction

Test creation is an expensive, effort-intensive task. If test creation could be even partially automated, the benefit to developers in terms of effort and cost would be immense. Naturally, a large body of research has amassed around automated test input generation [3]. One area that has shown great promise is *search-based test generation* [3], [23].

Input selection can naturally be seen as a search problem [15]. Testers approach input selection with a **goal** in mind—perhaps they would like to *cause the program to crash*, *maximize code coverage*, *detect a set of known faults*, or any number of other potential goals. Of the near-infinite number of possible input that could be provided to a program, the tester seeks input that achieves the chosen goal. This search can be automated. Given a goal, an optimization **algorithm** can systematically sample the space of possible test input in search of a solution to that goal, guided by feedback from one or more **fitness functions**—numeric scoring functions that judge the optimality of the chosen input [11]. In other words: $algorithm + fitness\ functions \implies goal$.

Supported by National Science Foundation grant CCF-1657299.

Effective search-based generation relies on the selection of the right feedback mechanism—the right fitness functions. The best fitness functions offer the information needed to rapidly increase attainment of the goal. For example, a common goal in test generation is maximum Branch Coverage. Branch Coverage is a measurement of *how much* of the code has been executed. For each program statement that can cause the execution path to diverge—such as `if` and `case` statements—test input should ensure that at all potential outcomes are covered at least once [25]. The most effective fitness functions for attaining Branch Coverage take each subgoal we wish to cover and judge *how close* the chosen test input was to achieving covering those goals. This concept, the *branch distance* [4], offers the algorithm the feedback needed to inch closer and closer to covering each outcome.

From this example, we can see that the selection of fitness functions is crucially important to maximizing attainment of our goal. For the goal of attaining Branch Coverage, we have effective fitness functions that lead to rapid improvement in coverage. Unfortunately, many goals *do not* have a known, effective fitness function formulation. In fact, many goals do not inherently lend themselves to such a formulation. Consider a common goal—"cause the program to crash", often measured by counting the number of *exceptions*—program-interrupting error messages—thrown during test execution [27]. Exceptions indicate the faults and abnormal operating conditions in programs. Thus, tests that trigger exceptions are valuable.

However, as we cannot know ahead of time how many or what exceptions are possible to throw, "throw more exceptions" is not a goal that translates into an informative fitness representation. Prior work has proposed the use of a simple count of thrown exceptions as a fitness function [28]. Unfortunately, this count yields poor results in terms of both goal attainment and fault detection, as it offers the algorithm no guidance for improving its guesses [11], [12].

This does *not* mean that there is no way to effectively achieve this goal. Rather, we simply do not yet know what

fitness functions will be effective. There are many fitness functions available for use in search-based test generation, devised for attainment of other goals. Careful selection of one or more of those functions could yield high attainment of our goal, exception throwing, as well. In fact, we may even attain higher goal attainment by reevaluating our choice of fitness functions at regular intervals throughout the generation process, adapting based on the evolving population of test suites. We hypothesize that an *adaptive* algorithm—one that can vary the selection of fitness functions—could make strategic selections that maximize attainment of our goal.

To evaluate this hypothesis, we propose a *hyperheuristic search* that optimizes the test generation process [17]. Through the use of reinforcement learning [26], this approach is able to select the most appropriate set of fitness functions for the class-under-test (CUT) and testing goal, and adjust that set as needed during generation. Throughout the test generation process, this algorithm retains test cases that cause exceptions to be thrown, yielding effective test suites. We have implemented two reinforcement learning algorithms—Upper Confidence Bound (UCB) and Differential Semi-Gradient Sarsa (DSG-Sarsa) [26]—in the EvoSuite generation framework [29].

We have evaluated the modified framework, EvoSuiteFIT, on 386 Java case examples in terms of the ability of generated test suites to discover and retain exceptions and to detect faults. We compare to two baselines: a count of exceptions thrown—used in prior work to encourage exception-throwing test suites—and EvoSuite's default configuration—a combination of eight fitness functions that serves as a "best guess" at what will produce effective test suites. We observe that:

- EvoSuiteFIT discovers and retains more exception-triggering inputs than the baseline techniques, yielding up to a 203.03% improvement in the number of exceptions thrown by the final test suite. DSG-Sarsa yields more consistent performance than UCB.
- Both EvoSuiteFIT techniques produce suites that detect faults missed by the other techniques. UCB detects 11.92 and 249.57% more faults than the baseline techniques, and 4.3% more than DSG-Sarsa.
- The ability to avoid the calculation of unhelpful fitness functions mitigates the additional computational overhead imposed by reinforcement learning.
- The ability to adjust the set of fitness functions at regular intervals in the generation process allows EvoSuiteFIT to make strategic choices that refine the test suite. This is not possible when a static set of fitness functions is used throughout the entire generation process.

The use of reinforcement learning algorithms allows EvoSuiteFIT to identify combinations of fitness functions effective at triggering exceptions in a CUT, and strategically vary that set of functions throughout the ongoing generation process. We hypothesize that other goals without known effective fitness function representations could also be maximized in a similar manner. We make EvoSuiteFIT available to others for use in test generation research or practice.

II. Background

A. Search-Based Test Generation

Test case creation can naturally be seen as a search problem [15]. Of the thousands of test cases that could be generated for any CUT, we want to select—systematically and at a reasonable cost—those that meet our goals [23], [1]. Given a well-defined testing goal, and fitness functions denoting *closeness to the attainment of that goal*, optimization algorithms can sample from a large and complex set of options as guided by a chosen strategy (the *metaheuristic*) [5].

Metaheuristics are often inspired by natural phenomena, such as swarm behavior [8] or evolution [16]. While the particular details vary between algorithms, the general process employed by a metaheuristic is as follows: (1) One or more solutions are generated, (2), The solutions are scored according to the fitness functions, and (3), the feedback from the fitness functions is used to reformulate the solutions for the next round of evolution. This process continues over multiple generations, ultimately returning the best-seen solutions. By determining how solutions are evolved and selected over time, the choice of metaheuristic and fitness functions impact the effectiveness and efficiency of the search process [9].

B. Reinforcement Learning

The n-armed bandit problem [21] describes a situation where you are repeatedly faced with a choice of n different options. After each choice, you receive a reward chosen from a probability distribution dependent on the action selected. Reinforcement learning algorithms are designed to learn the optimal choice of action to maximize the reward earned [26].

Each action has an expected reward when it is selected. Over time, the reinforcement learning algorithm will try different actions and refine its estimations of their value. During each round, the reinforcement learning algorithm will choose an action based on the expected reward of applying it in the current problem state. After applying the action, the algorithm will receive a reward value. The algorithm will update the expected reward for the chosen set using the new reward.

At any time, there will be a portfolio with the greatest estimated value. If the algorithm selects that portfolio, it *exploits* its current knowledge to gain immediate reward. If, instead, it selects a portfolio with an unknown or potentially lower reward, it is *exploring* the option space to improve its estimate of a portfolio's value. Reinforcement learning algorithms are designed to effectively balance exploration and exploitation for different problem spaces [26], [18], [17].

III. Approach

Search-based test generation requires the selection of one or more fitness functions to guide the search process. Careful selection is crucial, as the fitness functions act as strategies to shape the resulting test suite. Fitness functions should be selected to maximize attainment of the tester's overall goal. In practice, however, many goals do not translate cleanly to an effective fitness function representation—one that offers feedback to the search process to attain rapid attainment.

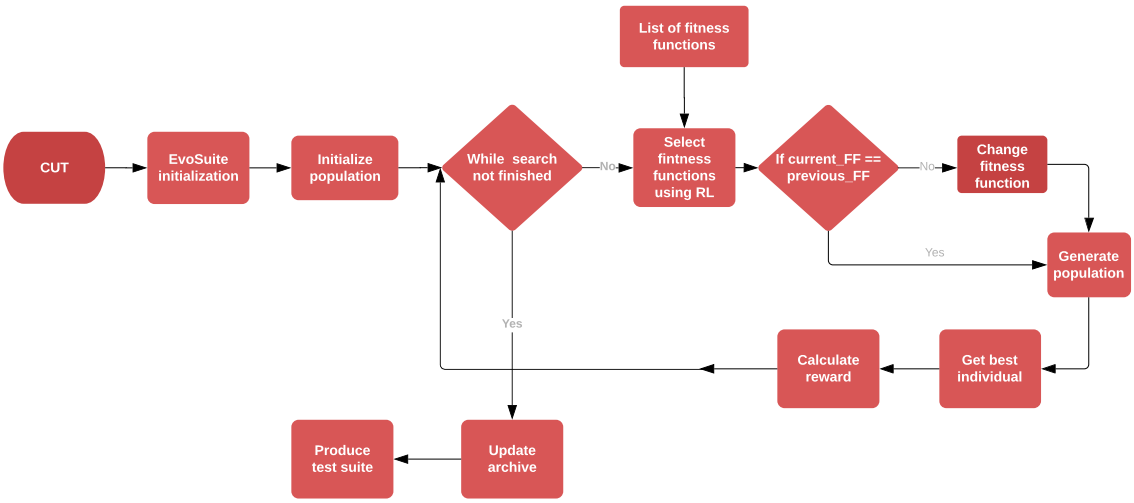


Fig. 1: An overview of how reinforcement learning fits into the test generation process.

In this work, we focus on the goal of causing the class-under-test (CUT) to throw exceptions. Exceptions—program-interrupting error messages—are thrown when the software is placed outside of normal operating conditions, and indicate situations the program is not able to handle gracefully [27]. Therefore, it is desirable to force the program into such states in order to identify areas of improvement or faults to fix. It is impossible to know all possible exceptions that can be thrown by a CUT, so there exists no feedback-based fitness function to encourage the discovery of more exceptions. A count of the number of exceptions thrown has been used in previous work [28], [11], [12]. However, this has been shown to be a poor choice for both finding faults [11] or triggering exceptions [12], with general functions such as Branch Coverage able to yield better results in both areas.

We hypothesize, however, that careful selection—at different points in the generation process—of fitness functions could result in test suites that trigger more exceptions than simple baselines or naive combinations of functions. If this is true, identifying this set of fitness functions becomes a secondary search problem—one that could be tackled as an additional step within the normal test generation process.

We propose the use of reinforcement learning techniques to adapt the set of fitness functions for a chosen CUT and the goal of throwing exceptions. Adjusting the set of fitness functions could be considered as an instance of the n-armed bandit problem [21]. Given a measurable goal, each action—each choice of one or more fitness functions—has an expected reward when it is selected. *If we use this function combination, we will cause additional exceptions to be thrown*. Specifically, we measure reward as the sum of the exceptions discovered during generation and the exceptions thrown by the current solution, encouraging both discovery and retention of exceptions. Because test generation is a stateful process—the population of test suites at round N depends on the population from round $N-1$—reinforcement learning affords not just an opportunity to identify effective fitness functions, but to strategically adjust the functions based on the changing population of test suites.

We have implemented two reinforcement learning algorithms—Upper Confidence Bound (UCB) and Differential Semi-Gradient Sarsa (DSG-Sarsa)—as part of the EvoSuite test generation framework [29]. The EvoSuite framework uses a genetic algorithm to evolve test suites over a series of generations, forming a new population by retaining, mutating, and combining the strongest solutions. It is actively maintained and has been successfully applied to a variety of projects [31]. We call our approach EvoSuiteFIT.

The modified process is illustrated in Figure 1. At a user-defined interval, the reinforcement learning algorithm will alter the current set of fitness functions and refine its estimation of their ability to increase the count of exceptions thrown. Throughout generation, we will also retain an *archive* of tests that cause exceptions to be thrown to ensure that the final test suite is effective. Note that the process we propose is generic, with only the reward function reflecting the goal of exception discovery. Any measurable goal can be used as a reward function, enabling the use of the same process to optimize other hard-to-satisfy goals.

EvoSuiteFIT is available from
https://github.com/hukh/evosuite/tree/evosuitefit

A. Upper Confidence Bound (UCB) Algorithm

The UCB algorithm is well-suited to address n-armed bandit problems [26]. Each time a choice is made, UCB selects an action that has a higher expected reward than the other possible actions. Each action returns a numerical value that is considered as the reward of taking that action.

For a selected action A at time step t (represented as A_t), the reward R_t represents the corresponding reward of taking action A_t. Using this notation, the expected reward of action

a is $q_* \doteq E[R_t|A_t = a]$. We apply UCB to select the action, as defined by Sutton [26]:

$$A_t \doteq max[Q_t(a) + c\sqrt{\frac{ln(t)}{N_t(a)}}] \quad (1)$$

where A_t represents the index of the combination that gives the highest expected reward. The c term represents the confidence level, determining the balance between exploration—refinement of reward expectation at a potential cost to reward—and exploitation—taking advantage of the action currently thought to be best—in the algorithm. The value of c needs to be larger than 0, otherwise the algorithm will behave in a purely greedy manner. $ln(t)$ represent the natural log of the t value. $Q_t(a)$ denotes the estimated worth of choosing a fitness functions (a):

$$Q_t(a) = \frac{1}{N_t}\sum_{i=1}^{t-1} R_i(a) \quad (2)$$

This equation represents the total reward of a combination a divided by the number of times that combination had been selected until the time t. In this project, t denotes the number of generations that the search has progressed through.

B. Differential Semi-Gradient Sarsa (DSG-Sarsa) Algorithm

A special class of reinforcement learning algorithms are known as *approximate solution methods* [26]. Many reinforcement learning approaches associate rewards with particular states, and work best in a constrained state space. Approximate methods are appropriate for problems with a large or unconstrained state space—i.e., test generation—where finding exact solutions is not feasible with limited time [6]. Approximate methods generalize from previously encountered states. As test case generation has a very large state space, we have explored the use of an approximate solution method, DSG-Sarsa [26].

DSG-Sarsa is semi-gradient, enabling continual and online learning. Relevant to our application domain, the algorithm is well-suited to problems in which there is no termination state. Each round, the action—choice of fitness functions—is applied, and the test suite evolves to a new state S', with observed reward R. We then use this information to choose a new action A', using the formula:

$$\hat{q}(S, A, W) \doteq W^\top \cdot X(S, A) = \sum w_i x_i(S, A) \quad (3)$$

This action-value function is calculated by the inner product of weights and feature vectors. X(S,A) is the feature vector: $X(s, A) = (x_1(S, A), x_2(S, A), \ldots x_d(S, A))$. The feature vector describes the current state of a test suite using a set of attributes. In this case, the state is represented using the current set of fitness functions, the current fitness value for that set of functions, the suite size, the number of exceptions thrown, and the number of exceptions covered. These features are utilized because that can describe a particular test suite.

W represents a weight vector, used to bias action selection [26]. A weight is provided for each feature, and represents the importance of its contribution to the action value. The weight for an action is updated each round using the semi-gradient with delta, controlled by the learning rate:

$$W_{t+1} \doteq W_t + \alpha\delta\nabla\hat{q}(S_t, A_t, W_t) \quad (4)$$

Where δ is an error function representing the difference between the immediate reward R and the average reward $\bar{R}_t$ and the difference between the value of a target $\hat{q}(S_{t+1}, A_{t+1}, W_t)$ and the value of the old estimate $\hat{q}(S_t, A_t, W_t)$.

$$\delta_t = R_{t+1} - \bar{R}_t + \hat{q}(S_{t+1}, A_{t+1}, W_t) - \hat{q}(S_t, A_t, W_t) \quad (5)$$

$\bar{R}_t$ is the is estimated average reward at time t. $\bar{R}_t$ is calculated using the following equation:

$$\bar{R}_{t+1} = R_t + \beta\delta \quad (6)$$

α and β are algorithm parameters that represent the step size of updating the weight and the average reward. The notation t represent the the time step (the number of generations, in our case).

By using the average reward, we consider the immediate reward as important as a delayed one. This means that we treat all fitness function combinations impartially without bias toward the combinations that were selected first. This mean there is no priority for the chosen combinations.

C. Implementation in EvoSuite

We have implemented both reinforcement learning algorithms in EvoSuite, and integrate their use into the standard Genetic Algorithm (GA)—adding an additional fitness function selection stage. At a user-defined interval, the RL algorithm will choose a new set of one or more fitness functions. The modified process is illustrated in Figure 1.

We use combinations of the following fitness functions:

- **Exception Count:** A simple count of the number of exceptions thrown by a test suite.
- **Branch Coverage:** See Section I.
- **Direct Branch Coverage:** Branch coverage may be attained by calling a method *directly*, or *indirectly*—calling a method from another method. Direct branch coverage requires coverage through a direct call.
- **Line Coverage:** A test suite satisfies line coverage if it executes each non-comment source code line at least once. To cover each line, EvoSuite tries to ensure that each basic code block is reached. For each conditional statement that is a control dependency for some other line in the code, the branch of the statement leading to the dependent code must be executed.
- **Method Coverage:** Method Coverage simply requires that all methods in the CUT are executed at least once. The fitness function for method coverage is discrete, as a method is either called or not called.
- **Method Coverage (Top-Level, No Exception):** Test suites sometimes achieve high levels of method coverage by calling methods in an invalid state or with invalid parameters. This variant requires that methods be called directly and terminate without throwing an exception.

- **Output Coverage:** Output coverage rewards diversity in output by mapping return types to a list of abstract values [2]. For numeric data types, distance functions offer feedback using the difference between the chosen value and target abstract values.
- **Weak Mutation Coverage:** Test effectiveness is often judged using synthetic faults, called mutants [19]. Weak mutation coverage is satisfied if, for each mutated statement, at least one test detects the mutation. The search is guided by the *infection distance*, a variant of branch distance tuned towards reaching mutated statements [10].

Rojas et al. provide more details on each of these fitness functions [28]. The RL algorithm chooses a combination of one to four of these fitness functions each time it makes a selection. Initial experimentation revealed that the most effective combinations of functions all included the exception count. Therefore, we filtered the set of choices down to all combinations of one to four fitness functions that include the exception count as one of the choices. This means that the RL algorithm can choose from 64 actions (different sets of fitness functions).

In the beginning, EvoSuiteFIT will make sure that all the actions have been tried once before it starts using the standard UCB or DSG-Sarsa selection mechanisms. This allows seeding of reward estimations. Before the initial selection occurs, the list of actions is randomized to avoid an ordering bias. This is important, as the population of test suites is shaped by the action used each generation. After this stage, every time the RL algorithm makes a selection, the set of chosen fitness functions will change unless the currently-selected combination is exploited.

After changing the fitness functions, EvoSuiteFIT will proceed through the normal population evolution mechanisms, judging solutions using the new set of fitness functions. We use the reformulated population to calculate the reward. Then, we used this reward to update the expectations of the RL algorithm. For UCB, we store the accumulated reward of each combination alongside the number of times each is selected N_t, so we can calculate the average reward. Over time, the combination that gains the highest reward will be more likely to be selected again until reaching convergence. For DSG-Sarsa, after getting the reward, the new combination is selected using the policy. Based on the new and current combination, the new and current state, and the reward, the average reward and the weight of the state is updated. Then the current fitness function combination will change to the new one.

After experimentation, we found that changing the set of fitness functions every three generations allows enough time to adequately adjust reward expectations. Fewer generations does not allow sufficient time for the chosen fitness function combination to reshape the test suite. This means that the GA will have three generations to reshape the population before the reward is evaluated.

In EvoSuiteFIT, during the search and optimization process, test cases that cover a set of chosen goals can be retained in a test archive to prevent loss in coverage as the test suites are reshaped. In traditional EvoSuite, this archive is based on the chosen fitness functions. However, as we use RL to change the fitness function, we have altered how the test archive is used. We store test cases known to trigger discovered exceptions. After the search process completes, the archive is used to produce the final test suite. This prevents the loss of test cases that trigger exceptions due to changes in the fitness functions.

IV. Study

To better understand the effectiveness, use, and applicability of our hyperheuristic approach, we have assessed EvoSuiteFIT against 386 case examples from the Defects4J dataset [20]. In doing so, we wish to address the following research questions:

1) Is either EvoSuiteFIT approach more effective, in terms of the number of discovered exceptions and the number of exceptions thrown by the final test suite, than test generation using static fitness function choices?
2) Is either EvoSuiteFIT approach more effective, in terms of fault detection effectiveness than test generation using static fitness function choices?
3) What impact does the computational overhead from reinforcement learning have on the test generation process?
4) Are there trends that can be discerned in the behavior of EvoSuiteFIT based on class or project features?

In order to investigate these questions, we have performed the following experiment:

1) **Collected Case Examples:** We have used 386 case examples, from six Java projects, as test generation targets (Section IV-A).
2) **Generated Test Suites:** For each class, we generate 10 suites per approach. Approaches include the two reinforcement learning algorithms—UCB and DSG-Sarsa—and two baselines—generation guided by exception count and a combination of all eight fitness functions. A search budget of 10 minutes is used per suite (Section IV-B).
3) **Removed Non-Compiling and Flaky Tests:** Any tests that do not compile, or that return inconsistent results, are removed (Section IV-B).
4) **Assessed Effectiveness:** We measure the number of exceptions thrown and detected by each test suite, the number of faults detected by each approach, and the number of generations of evolution that occur during the generation process (Section IV-C).

Tools and experiment scripting are available from https://github.com/Greg4cr/evosuitefit-exp

A. Case Examples

Defects4J is an extensible database of real faults extracted from Java projects [20][1]. The version we employed, Defects4J 1.4, consists of 395 faults from six projects: Chart (26 faults),

[1]Available from http://defects4j.org

Closure (133 faults), Lang (65 faults), Math (106 faults), Mockito (38 faults), and Time (27 faults). Nine of the case examples were excluded from our analysis—Closure faults 38, 44, 47, and 51, Math faults 13, 31, and 59, Mockito fault 6, and Time fault 21—as no technique caused exceptions to be thrown. For each fault, Defects4J provides access to the faulty and fixed versions of the code, developer-written test cases that expose the fault, and a list of classes and lines of code modified by the patch that fixes the fault.

Each fault is required to meet three properties. First, a pair of code versions must exist that differ only by the minimum changes required to address the fault. The "fixed" version must be explicitly labeled as a fix to an issue, and changes imposed by the fix must be to source code, not to other project artifacts such as the build system. Second, the fault must be reproducible—at least one test must pass on the fixed version and fail on the faulty version. Third, the fix must be isolated from unrelated code changes such as refactoring.

B. Test Suite Generation

For each class from each case example from Defects4J, we have generated test suites using each reinforcement learning approach—UCB and DSGSarsa. In addition, we generate tests for two baseline approaches representing current practice:

- **Exception Count**: A common fitness function representation of the goal of throwing exceptions is to simply count the number of exceptions thrown by a test suite. This would be the likely starting point for a tester interested in inducing exceptions.
- **Combination of all Eight Functions ("Default")**: The default configuration of EvoSuite is a combination of all eight criteria also used in the RL process. This configuration is used because it attains reasonable fulfillment of each individual function, and in theory will produce multifaceted test suites effective at fault-finding [28]. This configuration represents a "best guess" at what would produce effective test suites, and would be considered a reasonable approach in the absence of a known, informative fitness function.

Test suites are generated that target the classes reported as relevant to the fault by Defects4J. Tests are generated using the fixed version of the CUT and applied to the faulty version because EvoSuite generates its own assertions for use as oracles. In practice, this translates to a regression testing scenario, where tests are generated using a version of the system understood to be "correct" in order to guard against future issues [31]. Tests that fail on the faulty version, then, detect behavioral differences between the two versions[2].

To perform a fair comparison between approaches, each is allocated a ten minute search budget for test generation. In past work, 10 minutes was used as the maximum generation time and represented a point of "diminishing returns" for detection of the faults in Defects4J [30].

[2]Note that this is identical practice to other studies using EvoSuite with Defects4J, i.e. [31], [30]

To control experiment cost, we deactivated assertion filtering—all possible regression assertions are included. All other settings were kept at their default values. As results may vary, we performed 10 trials for each fault and search budget. This resulted in the generation of 15,800 test suites (ten trials, four approaches, 395 faults), representing over 2,633 hours of computation time. We performed experiments on Amazon EC2 infrastructure.

Generation tools may generate flaky (unstable) tests [31]. For example, a test case that makes assertions about the system time will only pass during generation. We automatically remove flaky tests. First, all non-compiling test suites are removed. Then, each remaining test suite is executed on the fixed version five times. If the test results are inconsistent, the test case is removed. This process is repeated until all tests pass five times in a row. On average, less than one percent of tests tends to be removed from each suite.

C. Data Collection

In order to address our research questions, we collect the following data for each test suite:

- **Number of Unique Exceptions Discovered During Generation**
- **Number of Unique Exceptions Thrown by the Final Test Suite**: Tests that trigger an exception can be lost during the generation process. We calculate this number by monitoring test suite execution.
- **Number of Faults Detected**
- **Number of Generations of Evolution**: The amount of time that it takes to complete one generation of evolution is not static, and each approach may complete a different number of generations during the test generation process based on the time needed to calculate each employed fitness function. Reinforcement learning will add additional overhead to this process, further decreasing the number of completed generations. We collect the number of generations to assess the impact of fitness function choice and RL overhead.
- **Decisions Made by EvoSuiteFIT**: The reinforcement learning algorithms reformulate the fitness function combination in use at regular intervals. Each time a combination is selected, we log the decision made. This can assist in understanding how the reinforcement learning algorithms function, and how they make decisions in service of goal attainment.

V. Results & Discussion

We are interested in understanding the effectiveness of EvoSuiteFIT in terms of discovering and exposing exceptions and faults. We are also interested in the impact of the overhead of reinforcement learning on the generation process and trends in how the RL algorithm makes selections. The following subsections detail our observations.

TABLE I: Median count of exceptions thrown and exceptions discovered for each technique, along with the median ratio of thrown to discovered. Counts are normalized between 0-1 for each fault to allow comparison across case examples.

	DSG-Sarsa		UCB		Exception Count		Default	
System	**Thrown**	**Discovered**	**Thrown**	**Discovered**	**Thrown**	**Discovered**	**Thrown**	**Discovered**
Chart	1.00	1.00	1.00	1.00	0.40	0.40	0.76	0.76
Closure	1.00	1.00	1.00	1.00	0.33	0.33	0.67	0.67
Lang	1.00	1.00	1.00	1.00	0.68	0.68	0.92	0.95
Math	1.00	1.00	1.00	1.00	0.60	0.60	0.75	0.75
Mockito	1.00	1.00	1.00	1.00	0.67	0.67	0.89	0.90
Time	1.00	1.00	1.00	1.00	0.42	0.43	0.73	0.73
Overall	1.00	1.00	1.00	1.00	0.50	0.50	0.75	0.75

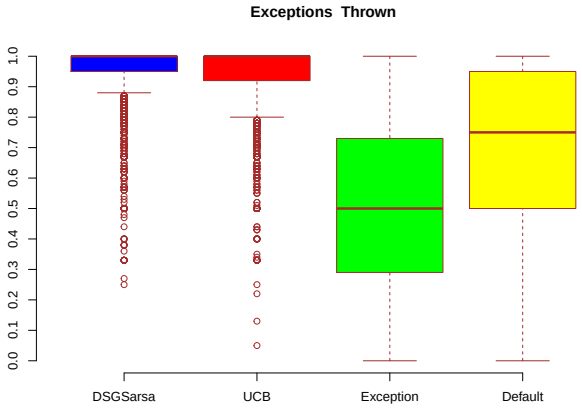


Fig. 2: Unique exceptions thrown by each technique. Counts are normalized between 0-1 for each case example.

A. Ability to Discover and Retain Exceptions

Our first question asks whether reinforcement learning can be used to more effectively meet our goal of throwing more exceptions than baseline approaches. We assess this along two dimensions. First, we look at the number of unique exceptions **discovered** by each approach during test generation. Second, we look at the number of unique exceptions **thrown** by the final test suite produced by each method. Are exception-inducing tests cases both produced *and* retained by the generation framework? An effective approach must excel at both.

We do not know a priori how many exceptions can be thrown by a class. However, we do know that the number of possible exceptions varies from class to class. Therefore, it is not reasonable to compare raw counts of exceptions between each case example. If we discover thirty exceptions when testing one class, and five when testing another, we should not compare five to thirty. Instead, we *normalize* exception counts between 0-1 for each case example, using the formula $(\frac{\text{Number of Exceptions Discovered/Thrown}}{\text{Maximum Number of Observed Exceptions for the Current Class}})$. Scaling all counts in this manner allows fair comparison.

The median count of exceptions thrown and discovered for each technique is listed in Table I for each project and overall. Boxplots of the exceptions thrown by each technique are shown in Figure 2. Results for exceptions discovered are similar, so are omitted. The results show that both reinforcement learning techniques have a higher median performance in both measurements than the two baselines. This is true across all systems, with EvoSuiteFIT attaining up to a 203.03% improvement in median exceptions thrown over the basic exception count and a 49.25% improvement over the eight-way default configuration. EvoSuiteFIT also tends to retain all discovered exceptions, while the default configuration may discard a small number of exception-triggering tests if offered improvements in the other fitness functions.

TABLE II: P-Values for Mann-Whitney rank-sum test for exceptions discovered. Exceptions thrown are the same.

	DSG-Sarsa	UCB	Exception	Default
DSG-Sarsa	-	$<$ **0.01**	$<$ **0.01**	$<$ **0.01**
UCB	0.99	-	$<$ **0.01**	$<$ **0.01**
Exception	1.00	1.00	-	1.00
Default	1.00	1.00	$<$ **0.01**	-

TABLE III: Results of Vargha-Delaney A Measure for exceptions thrown/discovered. Large positive effect sizes are bolded.

	DSG-Sarsa	UCB	Exception	Default
DSG-Sarsa	-	0.53, 0.53	**0.90**, **0.90**	**0.90**, **0.80**
UCB	0.47, 0.47	-	**0.89**, **0.89**	**0.80**, 0.77
Exception	0.10, 0.10	0.11, 0.11	-	0.31, 0.31
Default	0.19, 0.21	0.21, 0.32	0.69, 0.69	-

Figure 2 further shows the ability of reinforcement learning techniques to trigger unique exceptions. Both RL techniques not only offer a higher median than the competing techniques, but also have a narrower interquartile spread, showing relatively consistent performance. DSGSarsa and UCB attain the same median performance. However, DSG-Sarsa yields more consistent performance, as shown by the decreased spread of results. Both reinforcement learning techniques demonstrate superior ability to discover exception-triggering input over traditional baselines. Additionally, due to the use of a test archive to retain exception-triggering input, the two techniques do a better job of retaining exceptions.

We can perform statistical analysis to assess our observations. For each pair of techniques and baselines, we formulate hypotheses and null hypotheses:

- H_1: Test suites generated using technique *A* will have a different distribution of exception discovery results than suites generated using technique *B*.
- H_2: Test suites generated using technique *A* will have a different distribution of exception retention results than suites generated using technique *B*.
- $H0_1$: Observations of exception discovery for both techniques are drawn from the same distribution.
- $H0_2$: Observations of exception retention for both techniques are drawn from the same distribution.

Our observations are drawn from an unknown distribution; To evaluate the null hypotheses without any assumptions

TABLE IV: Number of faults detected by each approach.

System	DSG-Sarsa	UCB	Exception Count	Default
Chart	21	22	10	17
Closure	9	8	5	20
Lang	38	42	11	34
Math	73	73	13	61
Mockito	5	6	3	5
Time	16	18	5	14
Overall	162	169	47	151

on distribution, we use a one-sided (strictly greater) Mann-Whitney-Wilcoxon rank-sum test [33], a non-parametric test for determining if one set of observations is drawn from a different distribution that another set. We apply the test for each pairing of techniques and baselines with $\alpha = 0.05$.

The resulting p-values are listed in Table II. P-values are the same for both discovered and thrown. The results confirm our informal observations. For DSG-Sarsa, we can reject both null hypotheses for UCB and two baselines. For UCB, we can reject the null hypotheses for the two baselines. For the default baseline, we can reject the null hypotheses for the exception count baseline, but not for either EvoSuiteFIT technique.

We have also used the Vargha-Delaney A measure to assess effect size [32]. The results for both exception discovery and retention are listed in Table III, with large effect sizes in bold (≥ 0.80). This test further confirms our observations. In terms of both exceptions retained and discovered, DSG-Sarsa outperforms the two baselines with a large effect size. UCB outperforms the exception count baseline with a large effect size in both exception discovery and retention, and outperforms the default baseline with a large effect size for exception retention and a medium effect size in discovery. DSG-Sarsa outperforms UCB, but with a negligible effect size.

> Both EvoSuiteFIT techniques discover and retain more exception-triggering input than the baseline techniques, with DSG-Sarsa yielding more consistent results.

B. Fault Detection Effectiveness

In theory, forcing the class-under-test to throw exceptions will help developers discover faults in the system. Therefore, our second research question revolves around the ability of the generated test suites to trigger and detect failures. Both DGS-Sarsa and UCB trigger more exceptions than the baseline fitness functions. Does this translate into greater fault detection?

Table IV lists the number of faults detected by each technique. We can immediately see that both EvoSuiteFIT techniques generate suites that are able to detect faults that are missed by suites generated using the baselines. UCB, in particular, detects the most faults—identifying seven more faults than DSG-Sarsa (4.32%), 18 more than default (11.92% more), and 122 more than the exception count (259.57%).

The Default combination outperforms EvoSuiteFIT in one system—Closure—discovering 11 more faults than DSG-Sarsa. In these cases, it is likely that triggering these fault requires deep exploration of the code structure, and these faults may not be detected by triggering exceptions, requiring the production of incorrect output instead. The default configuration is outperformed across all other systems.

TABLE V: Median time per generation (in seconds). EX+(1,2,3) = exception count + 1-3 fitness functions. The comparison data [30] lacked EX+1 and EX+3 data for Mockito.

	DSG-Sarsa	UCB	EX+1	EX+2	EX+3	Default
Chart	0.24	0.26	0.36	0.49	0.75	3.29
Closure	0.32	0.49	1.40	1.61	2.70	5.71
Lang	0.30	0.44	0.23	0.34	1.03	4.38
Math	0.14	0.22	0.18	0.25	0.42	3.03
Mockito	0.03	0.03	-	0.03	-	0.08
Time	0.33	0.43	0.32	0.50	0.97	3.72
Overall	0.22	0.31	0.72	0.64	1.23	3.84

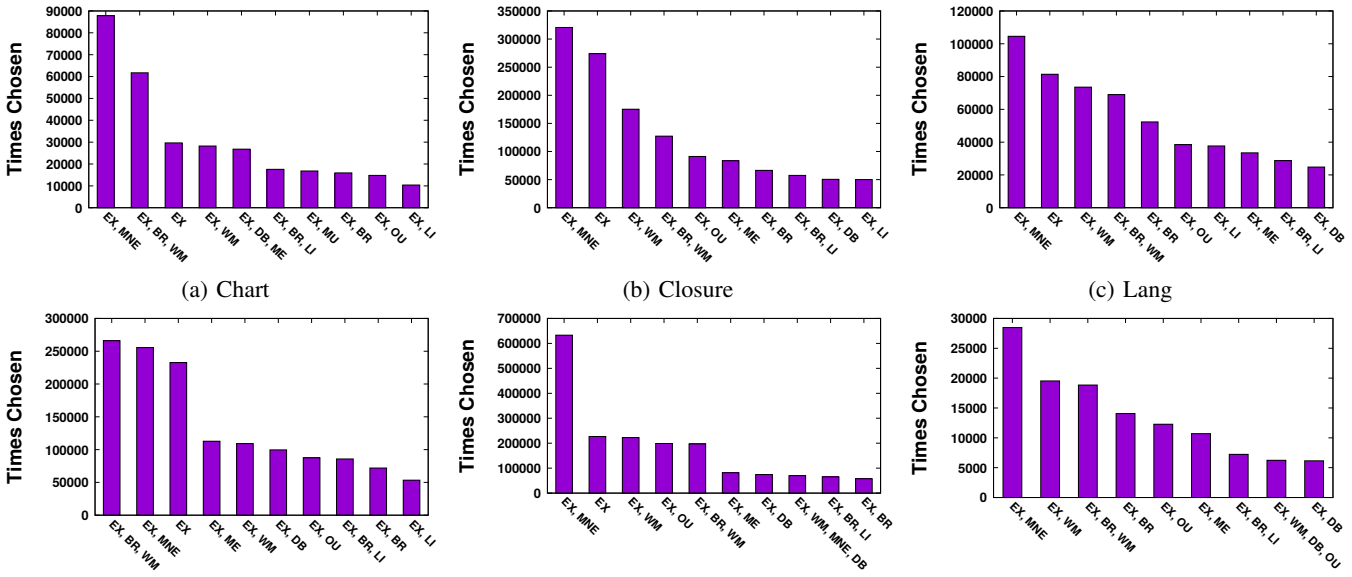
We previously found that DSG-Sarsa yielded slightly better and more consistent performance. This does not necessarily translate into higher fault detection, as UCB detected more faults. The difference between the two may come down to differing strategies in how fitness functions are chosen. In search-based test generation, fitness functions bias the selection of input. The reinforcement learning strategy, by impacting how and which fitness functions are selected, will further impact the selection process. Differences in how UCB and DSG-Sarsa make this selection will influence the likelihood of fault detection. Further analysis is required to understand the full impact that reinforcement learning strategy can have on fault detection capability. Still, the broad hypothesis that triggering exceptions can aid fault discovery appears to have some merit.

> Both EvoSuiteFIT techniques produce suites that detect faults missed by the other techniques. UCB detects 11.92 and 249.57% more faults than the baseline techniques, and 4.3% more than DSG-Sarsa.

C. Impact of Reinforcement Learning Overhead

Search-based test generation approaches are generally benchmarked using a fixed time budget [31]. During this period, the amount of work completed by each algorithm may not be equal. The number of generations of evolution completed will largely depend on the choice of fitness functions, with the cost being determined largely by the total cost to calculate fitness. The addition of reinforcement learning will further impact this cost. We are interested in understanding whether the cost of reinforcement learning has more of an effect than the cost of fitness calculation, and the further impact of being able to change that set of fitness functions.

Table V lists the median time per generation for DSG-Sarsa, UCB, and the Default combination. An issue in the version of EvoSuite deployed prevented us from collecting accurate generation times for the exception count alone, but—as it is an extremely simple count that does not require sophisticated instrumentation—it can be assumed that the exception count is far less expensive than any other option. Using data from past experiments [30], [12], we can also compare the time per generation with a sample of 33,759 test suites generated using combinations of the exception count and



(a) Chart (b) Closure (c) Lang

(d) Math (e) Mockito (f) Time

Fig. 3: Top ten fitness function combinations chosen by DSG-Sarsa for each system. EX = Exception Count, BR = Branch Coverage, DB = Direct Branch Coverage, LI = Line Coverage, OU = Output Coverage, ME = Method Coverage, MNE = Method (No Exception), WM = Weak Mutation Coverage

1-3 additional fitness functions—options that can be chosen by reinforcement learning. This sample lacked EX+1 or EX+3 cases for Mockito, but can still be used to understand the impact of reinforcement learning.

From Table V, we can see that the median cost per generation tends to increase as additional fitness functions are added to the calculation, with the cost per generation for the Default combination often being many times higher than that of either EvoSuiteFIT approach. While reinforcement learning may add to the cost of generation, its overhead is far less than that required to use a larger number of fitness functions.

Most of the potential users of a test generation framework would, rightfully, not be interested in tinkering with fitness functions until they stumbled on the right approach. In the absence of perfect knowledge, using the "default"—all eight fitness functions—is a reasonable idea. However, it is also an expensive option. Reinforcement learning can reduce the time required to get effective test cases.

Both reinforcement learning approaches perform similarly to static EX+1 combinations, if not faster. Figure 3 helps explain why. In Figure 3, we show the ten actions chosen most often by DSG-Sarsa for each system. While the EvoSuiteFIT techniques can choose combinations of up to four fitness functions, it rarely does so in practice. Often, EX+1 combinations are chosen, or even the simple exception count. Because EvoSuiteFIT can strategically change its fitness function selection, any overhead added by reinforcement learning is mitigated by speed gains from the ability to avoid the use of unhelpful fitness functions.

> The ability to avoid calculation of unhelpful fitness functions mitigates reinforcement learning overhead.

D. Trends in Reinforcement Learning Behavior

EvoSuiteFIT is able to freely alternate between 64 combinations of fitness functions. To understand why reinforcement learning is effective, it is important to understand how RL techniques choose between these actions. We list the most commonly-chosen actions taken by DSG-Sarsa in Figure 3.

From Figure 3, we can see that the systems differ in terms of the frequency choices are made, and to a small extent, in terms of the combinations chosen. For example, DSG-Sarsa frequently used a combination of exception count, Direct Branch Coverage, Weak Mutation Coverage, and Output Coverage for the Time system. However, although the ordering differs, there are also a lot of commonalities between the choices made.

For the most part, the favored combinations are simple—pairing the exception count with, at most, one additional fitness function. It is reasonable that simple combinations would be used frequently. Larger combinations introduce a risk of conflicting goals, and are harder to maximize. Simple combinations offer a reasonable feedback mechanism to increase the exception count. Complex combinations may offer less feedback by adding noise to the search.

Many of the fitness function combinations chosen by DSG-Sarsa would yield poor results when used on their own, as static fitness functions for suite generation. For example, DSG-Sarsa often uses the pure exception count, when this yields poor results when used as the sole fitness function. Similarly,

we know from past unpublished experiments that the EX-MNE combination produces poor results.

However, it is important to remember that test generation is a stateful process. Each round of the generation process builds on the results of previous rounds. There are times where the choices that DSG-Sarsa makes are relevant given the state of generation, even if those choices yield poor results when used in a static context. For example, if a suite *already* has achieved a high level of code coverage, it might make sense to switch to pure use of the exception count to further tune the population. Similarly, the EX-MNE combination makes sense as a strategic choice because it adds a light feedback mechanism to the exception count. Method (No Exception) Coverage requires that methods execute without exception. This does not mean that the test itself cannot throw an exception. Rather, it means that test generation will be encouraged to execute some code *before* the exception is thrown. This combination may be ineffective in a static context, as it does not offer enough feedback to fully explore the code space. However, it can be very effective if chosen at the right stage of the generation process, as part of an adaptive process.

The ability to adjust the set of fitness functions at regular intervals in the generation process allows EvoSuiteFIT to make strategic choices that refine the test suite. This is not possible when a static set of fitness functions is used throughout the generation process.

VI. Related Work

Hyperheuristic search—often based on reinforcement learning—has been employed in addressing a number of search-based software engineering problems. Jia et al. used reinforcement learning to select the metaheuristic algorithm for Combinatorial Interaction Testing, improving performance by learning the best algorithm for test generation for targeted problems [17], [18]. Similarly, Zamli et al. used hyperheuristic search to learn the selection and acceptance mechanisms used by the metaheuristic in Combinatorial Interaction Testing [34]. Guizzo et al. have used a reinforcement learning-based hyperheuristic search to tune the algorithm for optimizing the integration and test ordering problem [13], [14]. In addition, Kumari and Srinivas have used hyperheuristic search to tune software design—with the algorithm learning how to cluster classes for maximum cohesion and minimum coupling [22].

In all of these cases, the hyperheuristic is used to tune the algorithm itself, and not the fitness functions. Fitness function selection has been performed by hyperheuristic search in other domains, such as production scheduling [7], [24]. However, our approach is the first automated technique for optimizing the set of fitness functions used during test generation.

VII. Threats to Validity

External Validity: Our study has focused on six systems—a relatively small number. Nevertheless, we believe that such systems are representative of, at minimum, other small to medium-sized Java systems. We believe that Defects4J offers enough fault examples that our results are generalizable to other, sufficiently similar, projects. As Defects4J is used across multiple research fields, the use of this dataset also allows comparisons of our approach with other research, and allows others to replicate our experiments.

We have implemented our reinforcement learning techniques in a single test generation framework. There are many search-based methods of generating tests and these methods may yield different results. Unfortunately, no other generation framework offers the same number and variety of fitness functions. Therefore, a more thorough comparison of tool performance cannot be made at this time. By using the same framework to generate all test suites, we can compare our approach to the baselines on an equivalent basis.

To control experiment cost, we have only generated ten test suites for each combination of fault, budget, and configuration. It is possible that larger sample sizes may yield different results. However, given the consistency of our experiment results, we believe that this is a sufficient number of repetitions to draw stable conclusions.

Conclusion Validity: When using statistical analyses, we have attempted to ensure the base assumptions behind these analyses are met. We have favored non-parametric methods, as distribution characteristics are not generally known a priori, and normality cannot be assumed.

VIII. Conclusions

Choosing informative fitness functions is crucial to meeting the goals of a tester. Unfortunately, many testing goals—such as forcing the class-under-test to throw exceptions—*do not* have a known, effective fitness function formulation. We propose that the key to meeting such goals is to treat fitness function identification as a learning problem. An *adaptive* algorithm—one that can vary the selection of fitness functions—could adjust fitness functions throughout the generation process to maximize attainment of the chosen goal. To test this hypothesis, we have implemented two reinforcement learning algorithms in the EvoSuite framework.

Both EvoSuiteFIT techniques discover and retain more exception-triggering input than two baseline techniques. Both techniques also produce suites that detect a variety of faults missed by the other techniques. The ability to adjust the set of fitness functions at regular intervals in the generation process allows EvoSuiteFIT to make strategic choices that refine the test suite. Further, the ability to avoid the calculation of unhelpful fitness functions mitigates the additional computational overhead imposed by reinforcement learning. This is not possible when a static set of fitness functions is used throughout the generation process.

We make EvoSuiteFIT available to others for use in test generation research or practice. We hypothesize that other goals without known effective fitness function representations could also be maximized in a similar manner, such as triggering more generic crashes or Strong Mutation coverage. In future work, we will examine apply EvoSuiteFIT to such goals.

REFERENCES

[1] S. Ali, L. C. Briand, H. Hemmati, and R. K. Panesar-Walawege. A systematic review of the application and empirical investigation of search-based test case generation. Software Engineering, IEEE Transactions on, 36(6):742–762, 2010.

[2] N. Alshahwan and M. Harman. Coverage and fault detection of the output-uniqueness test selection criteria. In Proceedings of the 2014 International Symposium on Software Testing and Analysis, ISSTA 2014, pages 181–192, New York, NY, USA, 2014. ACM.

[3] S. Anand, E. K. Burke, T. Y. Chen, J. Clark, M. B. Cohen, W. Grieskamp, M. Harman, M. J. Harrold, and P. McMinn. An orchestrated survey of methodologies for automated software test case generation. Journal of Systems and Software, 86(8):1978–2001, 2013.

[4] A. Arcuri. It really does matter how you normalize the branch distance in search-based software testing. Software Testing, Verification and Reliability, 23(2):119–147, 2013.

[5] L. Bianchi, M. Dorigo, L. Gambardella, and W. Gutjahr. A survey on metaheuristics for stochastic combinatorial optimization. Natural Computing, 8(2):239–287, 2009.

[6] L. Buoniu, D. Ernst, B. De Schutter, and R. Babuka. Approximate reinforcement learning: An overview. In 2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), pages 1–8, April 2011.

[7] B. Crawford, R. Soto, E. Monfroy, W. Palma, C. Castro, and F. Paredes. Parameter tuning of a choice-function based hyperheuristic using particle swarm optimization. Expert Systems with Applications, 40(5):1690 – 1695, 2013.

[8] M. Dorigo and L. M. Gambardella. Ant colony system: a cooperative learning approach to the traveling salesman problem. Evolutionary Computation, IEEE Transactions on, 1(1):53–66, 1997.

[9] R. Feldt and S. Poulding. Broadening the search in search-based software testing: It need not be evolutionary. In Search-Based Software Testing (SBST), 2015 IEEE/ACM 8th International Workshop on, pages 1–7, May 2015.

[10] G. Fraser and A. Arcuri. Achieving scalable mutation-based generation of whole test suites. Empirical Software Engineering, 20(3):783–812, 2014.

[11] G. Gay. The fitness function for the job: Search-based generation of test suites that detect real faults. In Proceedings of the International Conference on Software Testing, ICST 2017. IEEE, 2017.

[12] G. Gay. Generating effective test suites by combining coverage criteria. In Proceedings of the Symposium on Search-Based Software Engineering, SSBSE 2017. Springer Verlag, 2017.

[13] G. Guizzo, G. M. Fritsche, S. R. Vergilio, and A. T. R. Pozo. A hyper-heuristic for the multi-objective integration and test order problem. In Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation, GECCO '15, pages 1343–1350, New York, NY, USA, 2015. ACM.

[14] G. Guizzo, S. R. Vergilio, and A. T. R. Pozo. Evaluating a multi-objective hyper-heuristic for the integration and test order problem. In 2015 Brazilian Conference on Intelligent Systems (BRACIS), pages 1–6, Nov 2015.

[15] M. Harman and B. Jones. Search-based software engineering. Journal of Information and Software Technology, 43:833–839, December 2001.

[16] J. H. Holland. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press, 1992.

[17] Y. Jia. Hyperheuristic search for sbst. In Proceedings of the Eighth International Workshop on Search-Based Software Testing, SBST '15, pages 15–16, Piscataway, NJ, USA, 2015. IEEE Press.

[18] Y. Jia, M. B. Cohen, M. Harman, and J. Petke. Learning combinatorial interaction test generation strategies using hyperheuristic search. In Proceedings of the 37th International Conference on Software Engineering, ICSE '15, pages 540–550, Piscataway, NJ, USA, 2015. IEEE Press.

[19] R. Just. The major mutation framework: Efficient and scalable mutation analysis for java. In Proceedings of the 2014 International Symposium on Software Testing and Analysis, ISSTA 2014, pages 433–436, New York, NY, USA, 2014. ACM.

[20] R. Just, D. Jalali, and M. D. Ernst. Defects4J: A database of existing faults to enable controlled testing studies for Java programs. In Proceedings of the 2014 International Symposium on Software Testing and Analysis, ISSTA 2014, pages 437–440, New York, NY, USA, 2014. ACM.

[21] M. N. Katehakis and A. F. Veinott Jr. The multi-armed bandit problem: decomposition and computation. Mathematics of Operations Research, 12(2):262–268, 1987.

[22] A. C. Kumari and K. Srinivas. Hyper-heuristic approach for multi-objective software module clustering. Journal of Systems and Software, 117:384 – 401, 2016.

[23] P. McMinn. Search-based software test data generation: A survey. Software Testing, Verification and Reliability, 14:105–156, 2004.

[24] G. Ochoa, J. A. Vazquez-Rodriguez, S. Petrovic, and E. Burke. Dispatching rules for production scheduling: A hyper-heuristic landscape analysis. In 2009 IEEE Congress on Evolutionary Computation, pages 1873–1880, May 2009.

[25] M. Pezze and M. Young. Software Test and Analysis: Process, Principles, and Techniques. John Wiley and Sons, October 2006.

[26] Richard S. Sutton and Andrew G. Barto. Reinforcement Learning, Second Edition An Introduction. 2018.

[27] M. P. Robillard and G. C. Murphy. Designing robust java programs with exceptions. In Proceedings of the 8th ACM SIGSOFT International Symposium on Foundations of Software Engineering: Twenty-first Century Applications, SIGSOFT '00/FSE-8, pages 2–10, New York, NY, USA, 2000. ACM.

[28] J. M. Rojas, J. Campos, M. Vivanti, G. Fraser, and A. Arcuri. Combining multiple coverage criteria in search-based unit test generation. In M. Barros and Y. Labiche, editors, Search-Based Software Engineering, volume 9275 of Lecture Notes in Computer Science, pages 93–108. Springer International Publishing, 2015.

[29] J. M. Rojas, M. Vivanti, A. Arcuri, and G. Fraser. A detailed investigation of the effectiveness of whole test suite generation. Empirical Software Engineering, 22(2):852–893, Apr 2017.

[30] A. Salahirad, H. Almulla, and G. Gay. Choosing the fitness function for the job: Automated generation of test suites that detect real faults. Software Testing, Verification and Reliability, 29(4-5):e1701, 2019. e1701 stvr.1701.

[31] S. Shamshiri, R. Just, J. M. Rojas, G. Fraser, P. McMinn, and A. Arcuri. Do automatically generated unit tests find real faults? an empirical study of effectiveness and challenges. In Proceedings of the 30th IEEE/ACM International Conference on Automated Software Engineering (ASE), ASE 2015, New York, NY, USA, 2015. ACM.

[32] A. Vargha and H. D. Delaney. A critique and improvement of the cl common language effect size statistics of mcgraw and wong. Journal of Educational and Behavioral Statistics, 25(2):101–132, 2000.

[33] F. Wilcoxon. Individual comparisons by ranking methods. Biometrics Bulletin, 1(6):pp. 80–83, 1945.

[34] K. Z. Zamli, F. Din, G. Kendall, and B. S. Ahmed. An experimental study of hyper-heuristic selection and acceptance mechanism for combinatorial t-way test suite generation. Information Sciences, 399:121 – 153, 2017.