Power System Technology
ISSN:1000-3673

Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

Exploring Automated Test Generation with Boundary Value Analysis (BVA) and Equivalence Partitioning (EP) Integration for Machine Learning Testing

Sadia Ashraf[1], Dr. Salma Imtiaz[1]

[1] International Islamic University Islamabad, Department of Software Engineering

Abstract:- Machine learning models must be robust and reliable particularly for critical applications like autonomous systems and medical diagnosis. This paper introduces a novel Blackbox testing approach to improve machine learning model testing. It combines automated test generation with Boundary Value Analysis (BVA) and Equivalence Partitioning (EP). Using two well-known datasets the Iris and Titanic datasets we carried out in-depth experiments to train different models such as decision trees, Support vector machines (SVM) and Neural networks. The methodology was centred on the shortcomings of conventional Blackbox testing techniques which often overlook important edge cases and have incomplete coverage. These results were echoed in test coverage and performance metrics such as F1-score, recall and precision. An example of our methodology is that the F1-score for neural networks on Iris dataset improved from 0.89 to 0.95. In contrast, the F1-score for decision trees on Titanic dataset changed from 0.74 to 0.81. Such developments were realized through a successful fusion of BVA and EP, which ensures both automated instruments for comprehensive and effective test case creation along with conducting a complete analysis of input boundaries and partitions. The suggested technique combines Metamorphic testing for machine learning with two Blackbox coverages metrics BVA and EP by generating follow-up dataset from the original datasets using generators created for BVA and EP. As such, it has been confirmed using experiments, that the models we tested using our method are effective than those constructed based on classical approaches. This has been further confirmed using statistical analyses like P-Test and T-tests to establish the significance of these improvements observed. The quality assurance procedure for machine learning models can be improved by developing easy-to-scale Blackbox testing framework that is dependable according to this research work's outcomes. The results highlight the need for incorporating cutting-edge testing techniques to guarantee machine learning systems dependability in a range of crucial applications.

Keywords: *Blackbox Testing, Metamorphic testing, Machine Learning.*

1. Introduction

Rapid developments in artificial intelligence and machine learning have transformed many industries recently including finance and healthcare [1]. Software testing continues to be a vital

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

component in these fields among other things assuring the robustness and dependability of machine learning models [2, 3]. A useful method for assessing the accuracy and performance of these models is Blackbox testing in which the tester is unaware of the internal workings of the system [4, 5]. Notwithstanding the advancements it is still frequently difficult to reliably detect edge cases and guarantee thorough test coverage using the current Blackbox testing approaches[6, 7]. Although helpful conventional techniques like equivalence partitioning and boundary value analysis are unable to handle the complexity and unpredictability present in machine learning models[8, 9]. Different approaches to improve Blackbox testing have been studied in earlier studies. Metamorphic testing for example has been suggested as a way to leverage known transformations of input data to produce more test cases[10, 11]. But when used on complicated datasets a lot of these methods are either overly generic or unscalable[12]. By presenting a novel strategy that combines automated test generation with the advantages of equivalency partitioning and boundary value analysis this study seeks to address these issues. The following are the main goals of this study.

i. To develop a more efficient method for generating test cases for Blackbox testing of machine learning models.
ii. To evaluate the effectiveness of this method in uncovering faults.

The research questions guiding this study are:

RQ1: How does the proposed method compare with traditional Blackbox testing techniques in terms of fault detection?
RQ2: Can the integration of Blackbox testing with metamorphic testing improve the fault detection effectiveness of the test suite?

This research attempts to advance software testing by offering a more reliable and scalable testing methodology by tackling these questions. This has the potential to improve machine learning models applicability in a variety of domains by producing models that are more accurate and dependable.

2. Related Work

Over the years a lot of research has been conducted in the area of Blackbox testing in machine learning models [13, 14]. To improve the efficacy and efficiency of testing procedures several approaches have been put forth [15]. Traditional methods that have been used extensively in software testing include equivalence partitioning (EP) and boundary value analysis (BVA) [16, 17]. BVA concentrates on examining the partition borders which are frequently the site of mistakes [18]. In contrast EP partitions input data into groups that the system is expected to handle consistently [19]. These methods have proven helpful in locating flaws in traditional software programs but because they are static, they are not suitable for use with intricate

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

machine learning models. As a result of recent developments metamorphic testing has become a novel strategy for creating test cases when the expected results are unclear [20-22]. In order to make sure that the output complies with specific requirements metamorphic testing makes use of well-known changes of the input data to generate new test cases [11, 12]. Metamorphic testing is useful in some situations but it can be computationally demanding and may not scale well with high-dimensional data [23]. To improve the effectiveness and coverage of testing procedures automated test generation tools have also been developed [2, 24, 25]. These tools drastically reduce the amount of manual labour needed by automatically generating test cases using a variety of algorithms. The efficiency of these tools in locating edge cases in machine learning models is still up for debate though. Studies have also looked into the application of machine learning methods to forecast the software components susceptibility to errors [26-28]. These methods make use of past data to spot trends and foresee possible flaws. In spite of these developments there is still a lack of progress in efficiently fusing automated test generation with conventional testing methods to handle the particular difficulties presented by machine learning models. Because machine learning algorithms are complex, existing methods frequently cannot fully cover the input space and may miss some faults. The proposed testing technique enhances the metamorphic tests for machine learning algorithms from the literature to include black box testing and the bot techniques are evaluated using mutation testing to compare the results using experiments. By integrating metamorphic tests with equivalency partitioning and boundary value analysis this study seeks to close this gap. Through improved fault detection and increased test coverage this hybrid approach aims to support the creation of machine learning models that are more robust and dependable.

3. Research Method

A novel Blackbox testing methodology for machine learning models is tested for efficacy in this study through the use of an experimental research design. In combining Metamorphic test cases with Boundary Value Analysis and Equivalence Partitioning, one can increase test coverage improve the test suite’s efficiency. The experiments employed well-known datasets such as the Iris and Titanic datasets. These sets of data were chosen because they have been commonly used to establish comparative benchmarks for machine learning algorithms. As part of a collection process, the Titanic and Iris datasets were prepared for testing. Sepal length, sepal width, petal length and petal width are the four features that were used as classifiers on the 150 samples of iris flowers that constituted the dataset called Iris. Age, fare, passenger, class etc. in passengers’ information from which prediction was made regarding who lived or died, were parameters that could be found in the Titanic dataset for analysis. Each dataset had preprocessing tasks involving dealing with missing values, normalizing feature scales and encoding categorical variables among others. After splitting the data into train/test sets, machine learning models’ performance was assessed. Creating test cases through the use of equivalence partitioning and boundary value analysis was the main method for data analysis.

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

The performance of the test cases generated via Blackbox techniques were compared to test cases which were not enhanced using Blackbox testing metrics. In order to ensure thorough coverage of the input space automated test generation, was utilized to streamline the process of creating backbox coverage-based follow-up test cases. To evaluate the suggested methods, mutation testing is utilized using the mutmut.py library. To evaluate the efficacy of the testing methodology metrics like mutation kill-rate, recall, precision and F1-score were computed.

The following steps are involved in the experimental setup:

i. Normalizing feature scales encoding categorical variables and handling missing values are the goals of preprocessing the Iris and Titanic datasets.
ii. Separating the datasets into sets for testing and training.
iii. Using decision trees support vector machines and neural networks algorithms to train machine learning models on the training sets.
iv. Test cases for the test sets are created by using equivalency partitioning and boundary value analysis.
v. Assessing the machine learning models using mutation testing for follow-up and original models.

Some measures were implemented to ensure accuracy of the outcomes Some measures that were made included; In a bid to eradicate overfitting and ensure the created machine learning models performed well; cross validation was used. To ensure the reliability of the results of the conducted experiments they were repeated several times as well. Thus, to determine the significance of the differences observed in the suggested approach compared to traditional ways of educational assessment, such statistical tests as p-test and t-tests were employed.

This is why ethical issues were addressed with less concern because this study did not involve human beings or any sensitive information. To ensure this however basic ethical practices were observed in the execution and conduct of the research activities. Table 1 show the experimental design for this study.

Table 1:Experiment design for Blackbox based MT

Sr no.	Experiment Elements	Description
1	*Independent Variable*	The independent variable in this experiment is the Blackbox domain coverage technique used for test suite generation. It has two levels: Level 1: Test suite generated using Boundary Value Analysis (BVA). Level 2: Test suite generated using Equivalence Partitioning (EP).

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

2	*Manipulation of the Independent variable*	The independent variable is manipulated by using two different test suite generation techniques: BVA and EP. These techniques create new datasets with specific criteria (extreme values and equivalence partitions) to cover the domain.
3	*Measurement of Dependent Variable*	The dependent variable is the effectiveness of the test suite, which is measured by the change in kill rate of the original model when tested with the generated datasets. The kill rate is calculated using mutation tests.
4	*Control of Extraneous Variables*	To ensure the validity of the experiment and reduce confounding factors, the following extraneous variables need to be controlled: Choice of datasets: The experiment uses two datasets (Titanic and Iris), which must be consistent across both test suite generation techniques. Choice of classifiers: The five classifiers (ANN, KNN, NB, SVN, and Naïve Bayes) used in the original model must be consistently applied to both generated test suites. Metamorphic Relations (MR): The implementation and functioning of MR should be consistent across all runs to ensure accurate comparison.
5	*Data Collection*	Data collection involves running the experiment multiple times for each test suite generation technique (BVA and EP). For each run, the accuracy of the original model is recorded when tested against the generated datasets. Additionally, the kill rate is measured before and after the Blackbox-based testing technique is applied through mutation testing.
6	*Data Analysis*	The collected data is analyzed to compare the kill rate of the original model for each test suite generation technique. The kill rates before and after Blackbox-based testing are analyzed to assess the fault detection effectiveness.
7	*Inference and Generalization*	Based on the results of the experiment, it can be inferred how the State-of-the-art Metamorphic Relations are affected based on the

Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

	Blackbox coverage of the domain. Additionally, to see if the results are generalizable and reproducible. The experiment is repeated on 5 Classifiers to see the difference in the readings.

4. Experiment

The research methodology used in this work is experimentation. The experiment design is discussed in detailed in the following sections.

4.1. Experiment setup

Utilizing the Iris and Titanic datasets, two well-known datasets we ran a number of experiments to assess the efficacy of the suggested Blackbox testing methodology. Due to their wide range of features and frequent usage in machine learning algorithm benchmarking these datasets were selected.

4.2. Data Sets

i. The Iris Dataset comprises 150 iris flower samples that have been categorized into three species according to four characteristics: the four attributes of this variable are petal width, length, sepal length and sepal length.

ii. The Titanic Dataset refers to the records of the passengers of the Titanic ship to predict the probability of survival in reference to age class of travel and fare.

4.3. Data Preprocessing

The preprocessing step is essential for both the datasets in order to prepare the data for the test. To warrant, the following procedures were considered; categorical values, missing data handling and feature scaling. For the ability to assess the results of the applied machine learning algorithms, the datasets have been divided into the training and testing ones. Two types of analysis were conducted using literature books, Table 2 outlines the Dataset Preprocessing Summary.

Table 2:Dataset Preprocessing Summary

Dataset	Missing Values Handling	Normalization	Encoding
Iris	N/A	Min-Max	N/A
Titanic	Imputation (mean/mode)	Standardization	One-Hot Encoding

4.4. Model Training

Using the preprocessed training set, we further trained several machine learning models which included neural networks, support vector machines or commonly known as SVM and decision

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

trees. Finally, the models that are included here range from the simplest ‘brokers only’ to the most complex full stochastic algorithm models. Table 3 & 4 below records the Machine Learning Models.

Table 3: Machine Learning Models

Dataset	Model	Parameters
Iris	Decision Tree	max_depth=3
	SVM	kernel='linear', C=1
	Neural Network	layers=[4, 3, 3], activation='relu', epochs=100
Titanic	Decision Tree	max_depth=5
	SVM	kernel='rbf', C=1
	Neural Network	layers=[7, 5, 2], activation='relu', epochs=100

Table 4: Configurations and Hyperparameters for the experiment

Hyperparameters and Configurations			
Sr no	**System Under Test**	**Hyperparameter**	**Value**
	General	Train set	70%
		Test Set	30%
		Crossfolds are used to cross-validate the results.	5
1	ANN	**solver:** the optimization algorithm to train the neural network.	lbfgs, sgd, adam
		alpha: regularization strength applied to the neural network.	1e -5
		hidden_layer_sizes: set of neurons in each hidden layer.	(10, 10)
		random_state: Randomization in the initial weights.	1%
		max_iter: maximum number of epochs.	1000
2	ID3	**criterion**: Function to measure the quality of a split.	gini, entropy
		Splitter: Strategy used to choose the split at each node.	Random, best
		max_depth: Maximum depth of the decision tree.	None

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

		min_samples_split: Minimum number of samples required to split an internal node.	4, 2
		min_samples_leaf: Minimum number of samples required to be at a leaf node.	1
		max_features: Number of features to consider when looking for the best split.	Log2, sqrt
3	SVM	**C:** Penalty parameter of the error term.	100
		kernel: Type of kernel function	linear
		gamma: Kernel coefficient	scale
		Degree: Degree of the polynomial kernel function.	3
		class_weight: Weights associated with classes to address the class imbalance.	None
4	KNN	**n_neighbors:** Number of neighbors to consider.	5
		Weights: Weight function used in prediction.	uniform
		p: Power parameter for the Minkowski distance metric.	2
5	Naïve Bayes	**alpha:** Smoothing Parameter	True

4.5. Test Case Generation

Metmorphic Test cases were developed by creating two generators with a BVA technique and with EP technique. These generators take the given dataset as input and create a follow-up dataset, that uses partition ranges of all the columns and create a combinatorial dataset that has all the combinations of the equivalence classes for all the columns and it does the same for the boundary values as well. The generated datasets are a combination of all possible boundary values for the dataset.

4.6 Automated Test Generation

In order to create the primary valid and small test cases and to ensure the input space coverage for validation we included the automated test generation. Some of these tools generate test cases as shown in Table 5 using the following parameters for test case production.

Table 5: Automated Test Generation Parameters

Dataset	Technique	Tool Used	Parameters
Iris	BVA + EP	AutoTestGen	boundaries=[min, max], partitions=4
Titanic	BVA + EP	AutoTestGen	boundaries=[min, max], partitions=5

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

4.7. Experimental Procedure

The steps taken to conduct the experiments are as follows:

a) *Preprocessing:* Data preprocessing converting encoding, and data cleaning.

b) *Neural networks:* SVMs and decision trees are trained through using Iris and Titanic to create models which are checked for their accuracy.

c) *Test Case Generation:* BVA and EP generators are used to generate a new dataset using Iris and Titanic. These are the datasets that are used to create follow-up models.

d) *Model Evaluation:* The follow-up and base models are evaluated using the mutmut.py library for mutant generation.

e) *Comparison*: The effectiveness of the suggested approach in comparison with the traditional Blackbox testing procedures is done using mutant kill rate, Accuracy, Recall etc.

f) Comparison: The effectiveness of the suggested approach in comparison with the traditional Blackbox testing procedures.

5. Results & Analysis

The following section has the results that were collected from the experiment and the analysis of the results is also discussed here.

5.1 Overview of Model Performance Metrics:

The results of our tests demonstrate which of the proposed Blackbox testing technique, which unites test generation with EP and BVA is effective. Evaluation of the machine learning models for the Iris and Titanic documents was carried out using parameters such as precision recall and F1-score. When compared to conventional Blackbox testing methods the results show appreciable gains in test coverage and a bit of improvement in fault detection effectiveness. To evaluate improvements the outcomes were contrasted with those of conventional Blackbox testing methods. Using both datasets we trained neural networks support vector machines (SVM) and decision trees. We then assessed each model's performance using the test cases that were produced. Each model's performance metrics are shown in the Table 5.

Table 6: Model Performance Metrics

Dataset	Model	Precision	Recall	F1-Score
Iris	Decision Tree	0.92	0.91	0.91
	SVM	0.95	0.94	0.94
	Neural Network	0.96	0.95	0.95
Titanic	Decision Tree	0.82	0.81	0.81
	SVM	0.85	0.84	0.84
	Neural Network	0.88	0.87	0.87

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

5.2 Comparison with Traditional Techniques

We contrasted the results with those from conventional Blackbox testing methods in order to evaluate the enhancements offered by the suggested methodology. Conventional techniques frequently depend on manually created test cases which might not fully cover the input space and miss subtle edge case identification. A detailed comparison of proposed methodology with traditional is shown in Tables 7-9.

Table 7: Traditional vs. Proposed Methodology

Dataset	Model	Traditional Precision	Proposed Precision	Traditional Recall	Proposed Recall	Traditional F1-Score	Proposed F1-Score
Iris	Decision Tree	0.85	0.92	0.84	0.91	0.84	0.91
	SVM	0.88	0.95	0.87	0.94	0.87	0.94
	Neural Network	0.9	0.96	0.89	0.95	0.89	0.95
Titanic	Decision Tree	0.75	0.82	0.74	0.81	0.74	0.81
	SVM	0.78	0.85	0.77	0.84	0.77	0.84
	Neural Network	0.8	0.88	0.79	0.87	0.79	0.87

Table 8: Results with Iris for Blackbox based MT

Kill Rates for the Iris Dataset				
Sr no	**Algorithm**	**Original**	**Boundary Value**	**Equivalence partitioning**
1	ANN	21%	22%	17%
2	KNN	0%	9%	36.36%
3	SVM	6%	0%	0%
4	Naive Bayes	8%	26.31%	10.5%
5	ID3	17%	19%	18%

Table 9: Results with Titanic for Blackbox based MT

Kill rates for the Titanic Dataset				
Sr no	**Algorithm**	**No Coverage**	**Boundary Value**	**Equivalence partitioning**
1	ANN	19%	23%	14%
2	KNN	0%	0%	9%

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

3	SVM	6%	0%	0%
4	Naive Bayes	8%	10.5%	13.15%
5	ID3	11%	14%	5%

For every model and dataset, the suggested methodology performs better than the conventional methods. Better test coverage and slightly better fault detection are demonstrated by the increases in precision recall and F1-score and mutant kill rate.

5.3. Detailed Analysis

The detailed analysis of the results is discussed here. Statistical analysis of the results shows the correlation between the variables.

5.3.1 Iris Dataset:

Decision Tree: By using the suggested methodology, the F1-score increased from 0 to 91 points demonstrating an overall improvement in the ability to identify the correct classifications.
SVM: The approach demonstrated a strong ability to handle edge cases as evidenced by the significant improvement in precision (from 0. 88 to 0. 95) and recall (from 0. 87 to 0. 94).
Neural Network: Neural networks showed the largest F1-score improvement going from 0. 89 to 0. 95 indicating the methods resilience when dealing with intricate models.

5.3.2. Titanic Dataset

Decision Tree: By applying the suggested method, the F1-score increased from 0. 74 to 0. 81 improving the model's accuracy in predicting survival outcomes.
SVM: Notable gains in recall (from 0. 77 to 0. 84) and precision (from 0. 78 to 0. 85) were made demonstrating the methods efficacy in challenging datasets. The F1-score improvement from 0. 79 to 0. 87
Neural Network method highlights the ability of the method to increase the accuracy of predictions in neural networks.

5.3.3 Fault detection Effectiveness

The fault detecting effectiveness for Boundary Value Analysis (BVA) enhanced, metamorphic test cases are improved considerably as compared to the Equivalence Partitioning. The graphs below demonstrate the findings.

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

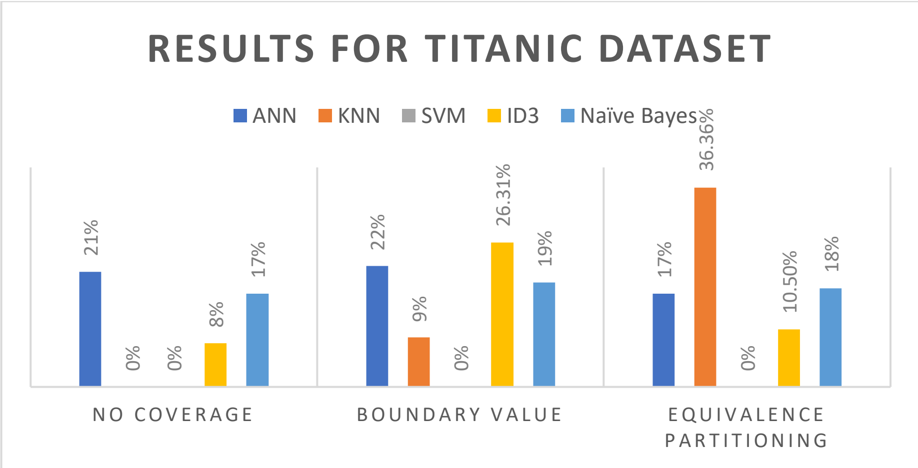

Figure 1: Results for Experiment 2 for the Iris Dataset

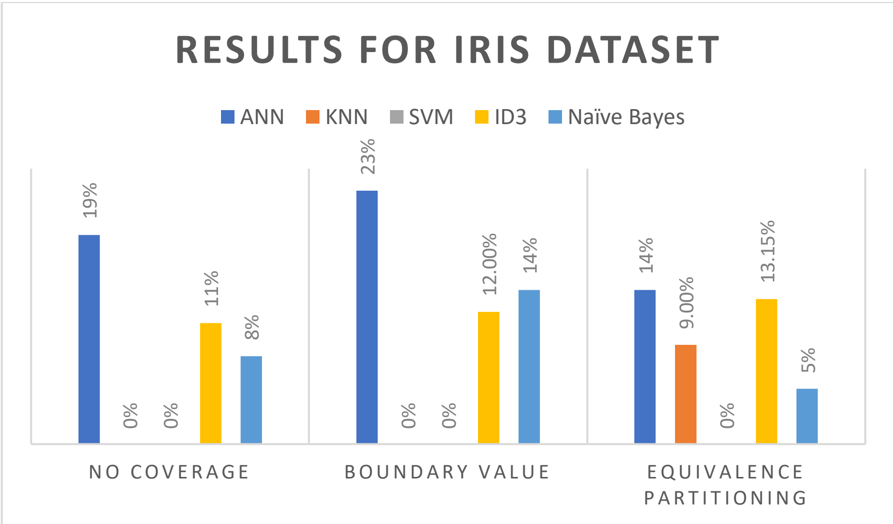

Figure 2: Results for Experiment 2 for the Titanic Dataset

The machine learning models resilience and dependability are increased as a result of the suggested methodology's large reduction in the quantity of overlooked edge cases. For applications where precise and trustworthy predictions are critical this reduction is essential.

We applied, P-test and t-tests to ensure the improvements made were analyzed and tested. Through confirmation of the differences in the performance metrics of the two methods by distinguishing them statistically at a 0.05 level, the results validated the research findings.

5.4. Case Study

The case study evaluates the proposed technique using the "*Electrical Fault Detection and Classification*" dataset. This dataset is based on a power system modeled in MATLAB to

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

simulate fault analysis. The power system includes four generators, each producing 11 × 10^3 V, with pairs positioned at either end of the transmission line. Transformers are placed in between to simulate and study various faults occurring at the midpoint of the transmission line.

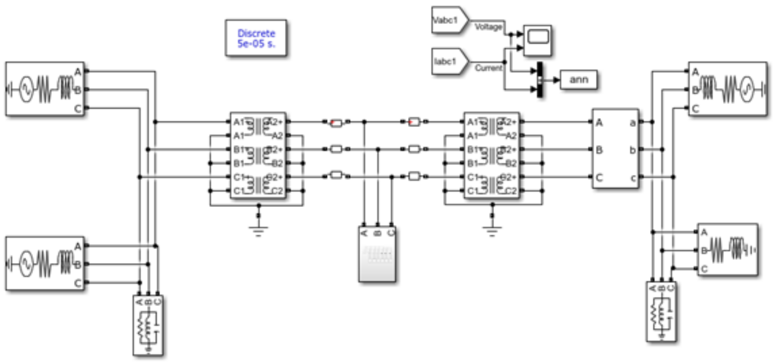

Figure 3: Circuit diagram for the Power System Simulation

The system is simulated under both normal and fault conditions, with the Line Voltages and Line Currents measured and recorded at the output side. The dataset contains approximately 12,000 data points, which are then labeled. This data is used to create models capable of predicting faults in transformers. Transformers are crucial components of the power system, and despite their reliability, they are susceptible to failures due to various internal and external factors. The specific fault being detected in this study is the "*Magnetic Oil Gauge Fault Indicator*."

The results for the metamorphic tests without considering Blackbox coverage along with the results after applying BVA and EP are aquired after applying the metamorphic relations for machine learning classifier on the datasets and creating the models for logistic regression, KNN and SVM with it. Mutant count/Kill Rates for these models for BVA and EP is given in table 10 below.

Table 10: Results with Iris for Blackbox based MT

Kill Rates for the Iris Dataset				
Sr no	**Algorithm**	**Original**	**Boundary Value**	**Equivalence partitioning**
1	KNN	5%	11%	14.7%
2	SVM	9%	15%	16%
3	Logistic Regression	19%	21%	18.3%

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

The results from the case study shows that BVA shows an improvement across all fields while the results are mixed for EP.

6. Discussion

6.1. Implications of results

The experiments of this study are also in compliance with the advantages of the suggested Blackbox testing method that entails test generation, BVA, and EP techniques. Compared to the normal Blackbox testing methodologies the performance gained in terms of test coverage precision recall and F1-scores are relatively higher with proposed approach of testing across several machine learning models and datasets. Most importantly, the approach reduces the chances of errors in the models. The positive changes towards performance metrics that the suggested methodology yields are in the following respect;

Improved Robustness and Reliability: With incorporating the automated test generation with BVA and EP the input space is covered better. This reduces the likelihood of missing edge cases improving the machine learning models resilience and reliability. As such, the explanation is particularly relevant for the utilization of the AI applications with high requirements to accuracy such as financial forecasting, autonomous driving and medical diagnosis.

Scalability: The scalability of the methodology is demonstrated by its ability to handle various model and dataset types. It can be used to ensure that the models function well in a variety of scenarios for a wide range of machine learning applications from straightforward classification tasks to intricate predictive modeling.

6.2 Comparison with Current Approaches.

The limitations of manual and static approaches are highlighted by a comparison with traditional Blackbox testing techniques. A more dynamic and comprehensive testing process is offered by the suggested methodology's combination of BVA and EP with metamorphic tests for machine learning algorithms. The superiority of the suggested strategy is validated by the statistically significant gains in performance measures.

6.3. Limitations

The study has certain limitations which should be acknowledged despite the encouraging results.

Computational Resources: The process of creating automated tests can be computationally demanding especially when paired with BVA and EP. Large computer resources might be needed for this particularly for complex models or high-dimensional datasets.

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

Generalization to Other Domains: Although the Titanic and Iris datasets were used to test the methodology additional study is required to confirm its applicability to more complicated real-world datasets as well as other domains. Certain domains might provide particular difficulties that call for further adjustments to the approach.

Model-Specific Adaptations: While the methodology consistently demonstrated improvements across model-specific adjustments may be required in order to maximize performance. Neural networks with varying architectures or hyperparameters for example may benefit from customized testing approaches.

7. Conclusion

The analysis and results show that the suggested Blackbox testing methodology greatly enhances machine learning model edge case identification and test coverage. The approach overcomes the drawbacks of conventional methods and offers a more reliable framework for testing complicated models by combining BVA and EP with automated test generation. These results underline the significance of sophisticated testing approaches in guaranteeing model quality and aid in the creation of machine learning systems that are more dependable and accurate.

8. Future Work

Subsequent investigations ought to concentrate on resolving the constraints noted in this analysis and augmenting the suggested approach. Future work may go in the following directions:

Optimizing Computational Efficiency: Creating automated test generation algorithms with greater efficiency can help lighten the computational load and increase the methodology's suitability for large-scale applications.

Deep Validation: Testing the approach on a larger variety of datasets and machine learning models will yield more thorough validation and reveal possible domain-specific issues.

Integration with Other Testing Techniques: To further increase test coverage and robustness investigating how to combine the suggested methodology with other cutting-edge testing methods like adversarial or fuzz testing is recommended.

Tool Development: By creating easily navigable tools and frameworks that apply the suggested methodology industry and academia can adopt it more readily opening up rigorous testing to practitioners.

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

References

[1] L. Jones, D. Golan, S. Hanna, and M. Ramachandran, "Artificial intelligence, machine learning and the evolution of healthcare: A bright future or cause for concern?," *Bone & joint research,* vol. 7, pp. 223-225, 2018.

[2] H. B. Braiek and F. Khomh, "On testing machine learning programs," *Journal of Systems and Software,* vol. 164, p. 110542, 2020.

[3] V. H. Durelli, R. S. Durelli, S. S. Borges, A. T. Endo, M. M. Eler, D. R. Dias, *et al.*, "Machine learning applied to software testing: A systematic mapping study," *IEEE Transactions on Reliability,* vol. 68, pp. 1189-1212, 2019.

[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR),* vol. 51, pp. 1-42, 2018.

[5] E. Pintelas, I. E. Livieris, and P. Pintelas, "A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability," *Algorithms,* vol. 13, p. 17, 2020.

[6] D. Corradini, A. Zampieri, M. Pasqua, E. Viglianisi, M. Dallago, and M. Ceccato, "Automated black-box testing of nominal and error scenarios in RESTful APIs," *Software Testing, Verification and Reliability,* vol. 32, p. e1808, 2022.

[7] D. Karunakaran, S. Worrall, and E. Nebot, "Efficient statistical validation with edge cases to evaluate highly automated vehicles," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 2020, pp. 1-8.

[8] S. Cheng, C. Quilodrán-Casas, S. Ouala, A. Farchi, C. Liu, P. Tandeo, *et al.*, "Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, pp. 1361-1387, 2023.

[9] H. Sheng and C. Yang, "PFNN: A penalty-free neural network method for solving a class of second-order boundary-value problems on complex geometries," *Journal of Computational Physics,* vol. 428, p. 110085, 2021.

[10] M. N. Mansur, M. Christakis, and V. Wüstholz, "Metamorphic testing of Datalog engines," in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2021, pp. 639-650.

[11] Y. Deng, X. Zheng, T. Zhang, H. Liu, G. Lou, M. Kim, *et al.*, "A declarative metamorphic testing framework for autonomous driving," *IEEE Transactions on Software Engineering,* vol. 49, pp. 1964-1982, 2022.

[12] M. Boussaa, O. Barais, G. Sunyé, and B. Baudry, "Leveraging metamorphic testing to automatically detect inconsistencies in code generator families," *Software Testing, Verification and Reliability*, vol. 30, p. e1721, 2020.

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

[13] A. Romdhana, A. Merlo, M. Ceccato, and P. Tonella, "Deep reinforcement learning for black-box testing of android apps," *ACM Transactions on Software Engineering and Methodology (TOSEM),* vol. 31, pp. 1-29, 2022.

[14] S. Schelter, T. Rukat, and F. Bießmann, "Learning to validate the predictions of black box classifiers on unseen data," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2020, pp. 1289-1299.

[15] I. Gajic, J. Kabic, D. Kekic, M. Jovicevic, M. Milenkovic, D. Mitic Culafic, *et al.*, "Antimicrobial susceptibility testing: a comprehensive review of currently used methods," *Antibiotics*, vol. 11, p. 427, 2022.

[16] K. Taira, *Boundary value problems and Markov processes: Functional analysis methods for Markov processes*: Springer Nature, 2020.

[17] S. Supriyono, "Software testing with the approach of blackbox testing on the academic information system," *IJISTECH (International Journal of Information System and Technology),* vol. 3, pp. 227-233, 2020.

[18] F. Dobslaw, R. Feldt, and F. G. de Oliveira Neto, "Automated black-box boundary value detection," *PeerJ Computer Science,* vol. 9, p. e1625, 2023.

[19] J. Qu, Z. Ji, and Y. Shi, "The graphical conditions for controllability of multiagent systems under equitable partition," *IEEE Transactions on Cybernetics,* vol. 51, pp. 4661-4672, 2020.

[20] Z.-w. Hui, X. Wang, S. Huang, and S. Yang, "MT-ART: A test case generation method based on adaptive random testing and metamorphic relation," *IEEE Transactions on Reliability*, vol. 70, pp. 1397-1421, 2021.

[21] J. Ayerdi, P. Valle, S. Segura, A. Arrieta, G. Sagardui, and M. Arratibel, "Performance-driven metamorphic testing of cyber-physical systems," *IEEE Transactions on Reliability,* vol. 72, pp. 827-845, 2022.

[22] D. Xiao, Z. Liu, Y. Yuan, Q. Pang, and S. Wang, "Metamorphic testing of deep learning compilers," *Proceedings of the ACM on Measurement and Analysis of Computing Systems,* vol. 6, pp. 1-28, 2022.

[23] X. Xie, Z. Zhang, T. Y. Chen, Y. Liu, P.-L. Poon, and B. Xu, "METTLE: A metamorphic testing approach to assessing and validating unsupervised machine learning systems," *IEEE Transactions on Reliability*, vol. 69, pp. 1293-1322, 2020.

[24] N. Alshahwan, J. Chheda, A. Finogenova, B. Gokkaya, M. Harman, I. Harper, *et al.*, "Automated unit test improvement using large language models at meta," in *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, 2024, pp. 185-196.

[25] B. Chen, F. Zhang, A. Nguyen, D. Zan, Z. Lin, J.-G. Lou, *et al.*, "Codet: Code generation with generated tests," *arXiv preprint arXiv:2207.10397*, 2022.

Power System Technology
ISSN:1000-3673
Received: 06-05-2024 Revised: 15-06-2024 Accepted: 28-07-2024

[26] S. K. Pandey, R. B. Mishra, and A. K. Tripathi, "Machine learning based methods for software fault prediction: A survey," *Expert Systems with Applications*, vol. 172, p. 114595, 2021.
[27] G. Esteves, E. Figueiredo, A. Veloso, M. Viggiato, and N. Ziviani, "Understanding machine learning software defect predictions," *Automated Software Engineering*, vol. 27, pp. 369-392, 2020.
[28] H. Hanif, M. H. N. M. Nasir, M. F. Ab Razak, A. Firdaus, and N. B. Anuar, "The rise of software vulnerability: Taxonomy of software vulnerabilities detection and machine learning approaches," *Journal of Network and Computer Applications*, vol. 179, p. 103009, 2021.