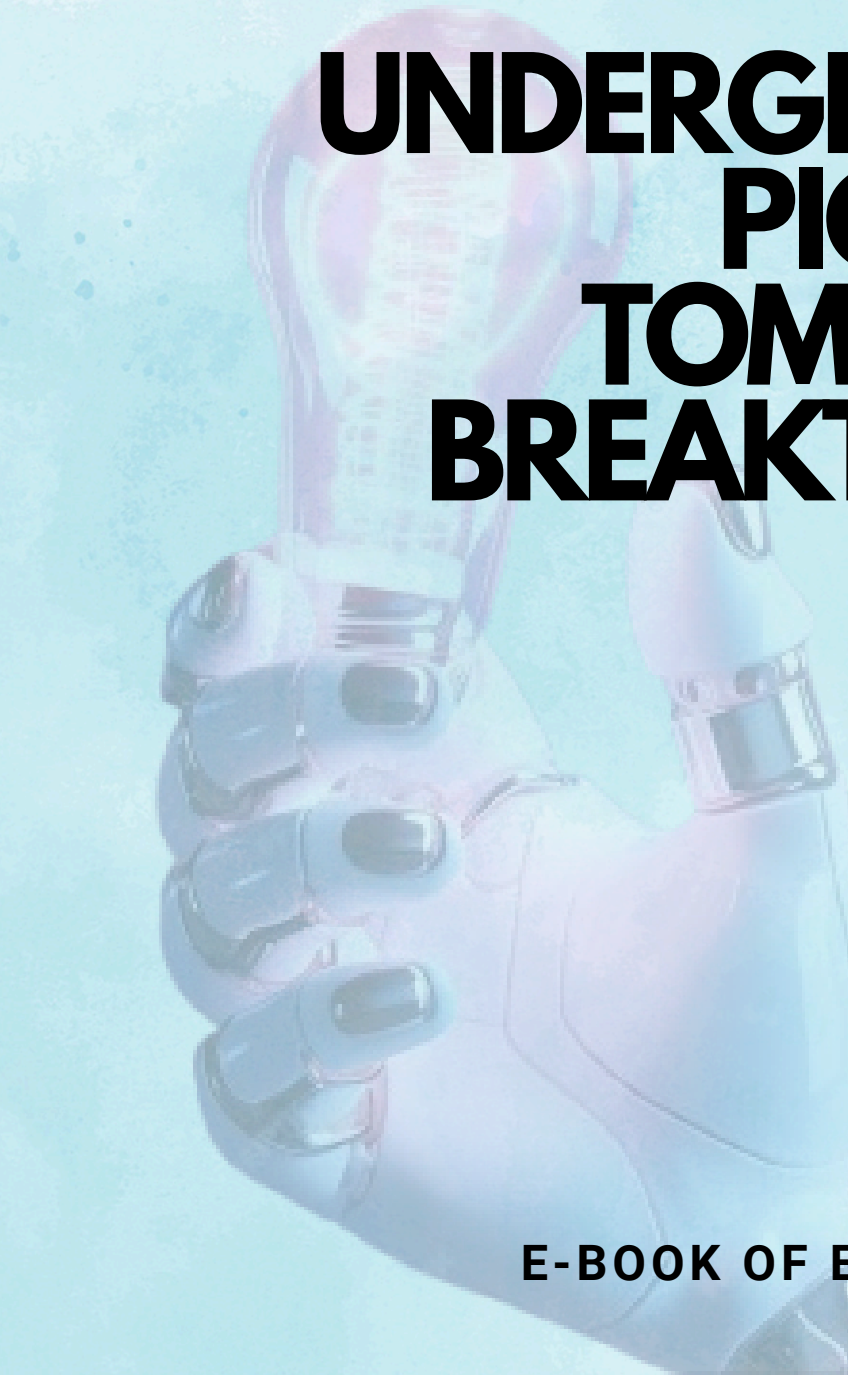
UNIVERSITI TEKNOLOGI MARA

I-URID 2025

INTERNATIONAL UNDERGRADUATE RESEARCH INNOVATION, INVENTION AND DESIGN

UNDERGRADUATES PIONEERING TOMORROW'S BREAKTHROUGH



E-BOOK OF EXTENDED ABSTRACTS

CONFIDENTLY WRONG: AUDITING CLASSIFIERS USING REVERSIBLE GENERATOR MLP (RevGEN-MLP)

Wan Muhammad Isyraf Wan Hasmar Azim, and R.U. Gobithaasan*

School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM Penang, Malaysia

*gobithaasan@usm.my

ABSTRACT

Machine learning classifiers are known to be overconfident when facing unseen or anomalous inputs. These present a dangerous vulnerability that reduces the trust in the systems that rely on artificial intelligence as it becomes increasingly widespread. We present a novel method of auditing these machine learning models by generating confidently classified anomalies using Reversible Generator Multilayer Perceptron (RevGEN-MLP). These generated anomalies act as a tool for testing the robustness of different models, identifying which models are more vulnerable to overconfidence on a specific dataset. Our method can be adapted to datasets with continuous numerical features and can be tailored for domain-specific robustness testing. We demonstrate the application of this technique on two different datasets and show that the generated anomalies not only deceive the original model but can also transfer to other classifiers. Results reveal that a classifier's vulnerability to such anomalies can vary by dataset, with Random Forest observed to be more robust towards these generated inputs compared to other neural networks and K-Nearest Neighbour classifiers.

Keywords: Artificial Intelligence (AI), Model Robustness, Overconfidence, Reversible Neural Networks

INTRODUCTION

Artificial intelligence (AI) has become a fundamental part of modern systems and has led to significant improvements in predictive analytics and decision-making. As we continue integrating AI systems into our everyday lives, it is critical to ensure the trustworthiness of these systems. Ideally, these models would only be confident when working with data that is similar to that on which it was trained. In practice, many systems, especially those designed for classifications, are overconfident on anomalous and unrealistic data. These issues reflect poor model calibration and are common in real-world systems. This work defines any anomalous or unrealistic input that is confidently classified as a sample that deceives or tricks the model.

For neural networks that are built for classification, softmax probabilities are often used as a confidence score to indicate how confident a model is in its classification. However, softmax probabilities are widely recognised as an unreliable confidence score, as it is prone to being overconfident on inputs it has never seen (Lee et al., 2018). Neural network classifiers are normally not bijective, hence there are multiple distinct inputs that can produce outputs with similar softmax probabilities. This leads to overconfidence as the neural network can assign high confidence to anomalies whose features overlap with those of the training data. Overconfidence can also be caused by limitations of the dataset, as there are many regions in feature space where models extrapolate with high confidence (Pearce et al., 2021). By exploiting these weaknesses, we can produce new inputs to evaluate and audit the robustness of other classification models as well. Thus, we present a new model called Reversible Generator MLP (RevGEN-MLP) that generates confidently classified anomalies.

We focus our analysis on the Iris and White Wine Quality datasets, which are well-known tabular benchmark datasets for classification. The Iris dataset consists of 150 samples with four features each and is assigned to three balanced classes of Iris species (Fisher, 1936). The White Wine Quality dataset consists of 4898 samples, each with 11 features that represent wine characteristics, and comes with 7 quality ratings from 3 to 9, where 3 is bad and 9 is excellent. To simplify training, we combine the ratings into two classes,

high-quality (7-9) and low-quality (3-6) wine. However, the resulting classes are imbalanced with a low-to-high quality ratio of 3838:1060 (Cortez et al., 2009).

OBJECTIVES

i. To construct a reversible neural network that acts as a generator for confidently classified anomalies.
ii. To generate new inputs that trick other commonly used classifiers and find shared vulnerabilities.
iii. To analyse robustness of different models against overconfidence on anomalous inputs.

ARCHITECTURE

A neural network can be thought of as a composite function where each layer, denoted L_i is a component function. Each layer has the following form,

$$L_i(x) = \phi\,(Wx + b), \tag{1}$$

where ϕ is the activation function, $W \in R^{Out \times In}$ is the matrix containing all the weights of the layer, $x \in R^{In}$ is the input, $b \in R^{Out}$ is the bias, Out and In are the output and input dimensions. The design of RevGEN-MLP for classification and generating new samples requires the following constraints:

1. Dimension Consistency: The dimensions of each layer must be the same as the input
2. Invertible Weight Matrix: The weight matrix, $\boldsymbol{W}$ for each layer must be invertible
3. Invertible Activation Function: The activation functions for each layer, ϕ must be invertible

The activation function of the hidden layers in the model is Leaky ReLU as it is invertible and for classification, we use the softmax function. Although the softmax function is not invertible, we can obtain a pseudo-inverse by storing the denominator during the forward pass and using it to find the original input. This limits the model to being invertible only for outputs with stored softmax denominators. It has the following form,

$$x_i = ln\ ln\ p_i + ln\ ln\ C\,, \tag{2}$$

where p_i is the softmax probability of class i and C is the stored denominator from the forward pass. The inverse of a layer will then have the following form,

$$L_i^{-1}(x) = W^{-1}\,(\,\phi^{-1}(x) - b). \tag{3}$$

The network outputs class probabilities through a softmax layer with k neurons corresponding to the number of classes ($k = 3$ for Iris and $k = 2$ for White Wine Quality). Since the network is invertible, the dimensions of the output must match the input (n features). Therefore, the remaining $n - k$ outputs are latent variables copied from the previous hidden layer, as shown in Figure 1. Training will only be done with the output values containing the k softmax class probabilities using the cross-entropy loss function. The design also lets us keep the class probabilities constant and modify the latent variables before inverting, guaranteeing any generated sample has the exact same probabilities as the original, which is different from other generative models without this guarantee.

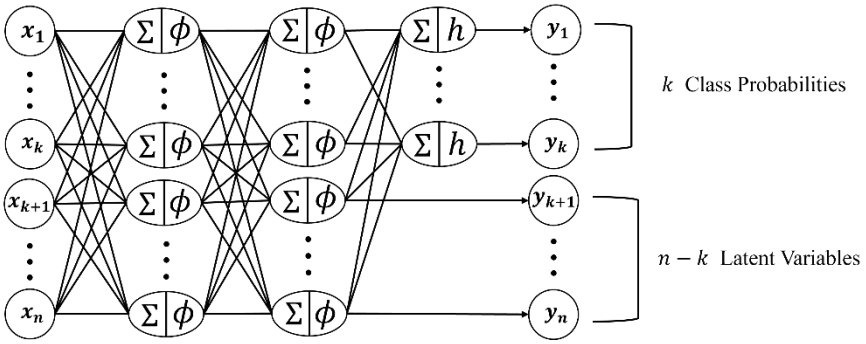


Figure 1: General Architecture Diagram for RevGEN-MLP (ϕ: Leaky ReLU, h: Softmax)

GENERATING NEW SAMPLES

The training and testing set for all models is the same and follows the 80:20 train-test split. As shown in Figure 2, we generate new samples by perturbing the latent variables in the outputs of confidently and correctly classified test data with varying shift parameters ϵ, until flagged as anomalies by an Isolation Forest at a chosen anomaly threshold, λ. The generated samples will remain confidently classified by RevGEN-MLP and may be more likely to fool other well-trained models. This is because RevGEN-MLP can potentially share similar high-confidence regions with other well-trained models, and the generated samples could transfer, although this is not guaranteed.

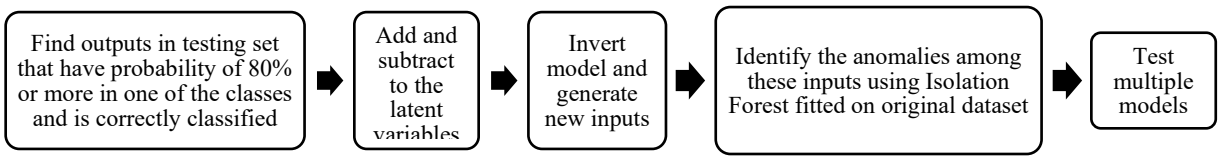


Figure 2: Flowchart of Method used for Generating Confidently Classified Anomalies

RESULTS AND DISCUSSION

We trained the RevGEN-MLP model on the Iris and White Wine Quality datasets and used the generated anomalies to test a non-invertible neural network, a Random Forest, and a K-Nearest Neighbor (K-NN) classifier (Wan & Gobithaasan, 2025). Anomalies were identified using an Isolation Forest, with more negative anomaly scores indicating it as more anomalous. The number of epochs was chosen after experimentation to find the best set of hyperparameters that leads to good performance and avoids overfitting. For Iris, RevGEN-MLP had one hidden layer and achieved a test accuracy of 91.1%, precision of 91.6%, and a recall of 91.1% after 160 epochs. The neural network used for comparison had one hidden layer with 20 neurons using Leaky ReLU, trained for 250 epochs. For White Wine Quality, RevGEN-MLP had 3 hidden layers and achieved a test accuracy of 83.37%, precision of 75.82%, and a recall of 72.14% after 101 epochs. The neural network used for comparison had 4 hidden layers with ReLU activation (number of neurons: 64,32,16,8) trained for 440 epochs. The Random Forest models used an unscaled version of the data for both datasets, had 200 estimators, whereas the K-NN classifier had K = 5 and used the scaled versions of the data.

Anomaly scores from Isolation Forests denoted as λ are relative to the dataset, hence to evaluate robustness fairly, we generate new inputs using two thresholds. The first threshold, λ_1 is chosen to be the 5th percentile of anomaly scores among the original dataset, this represents moderately anomalous samples and is directly comparable between the two datasets (Iris: $\lambda_1 = -0.09136$, Wine Quality: $\lambda_1 = -0.00921$). On the other hand, the second threshold, $\lambda_2 = -0.15$ is a fixed threshold for both datasets and it is due to a higher negative value than the minimum anomaly score observed in the samples of both dataset (Iris: -0.1457, Wine Quality: -0.1425). This ensures that samples generated with λ_2 are likely to be extreme anomalies or unseen data.

Table 1: Experimental Results

Model	Test Metrics		Anomalies Confidently Classified (Fooling Rate)	
	Iris	Wine Quality	Iris	Wine Quality
Neural Network	Accuracy : 93.3% Precision : 93.5% Recall : 93.3%	Accuracy: 85.4% Precision: 78.7% Recall : 77.2%	λ_1: 56/66 (84.85%) λ_2: 54/68 (79.41%)	λ_1: 558/580 (96.21%) λ_2: 1050/1063 (98.78%)
Random Forest	Accuracy : 91.1% Precision : 91.6% Recall : 91.1%	Accuracy : 89.1% Precision : 86.6% Recall : 79.4%	λ_1: 55/66 (83.33%) λ_2: 30/68 (44.12%)	λ_1: 181/580 (31.2%) λ_2: 100/1063 (9.4%)
K-NN	Accuracy : 91.1% Precision : 93.0% Recall : 91.1%	Accuracy : 84.2% Precision : 77.1% Recall : 73.9%	λ_1: 57/66 (86.36%) λ_2: 42/68 (61.76%)	λ_1: 504/580 (86.9%) λ_2: 898/1063 (84.48%)

Table 1 shows that even models with good test performance can still be tricked by samples generated by RevGEN-MLP. As the anomaly score decreases (more anomalous), the Random Forest model is tricked by fewer samples than the other models. This indicates it is more robust than neural networks and the K-NN classifier for these two examples. However, it should be noted that we are still able to find anomalies that trick the Random Forest, and thus, it is still not foolproof. On the other hand, we have been able to generate anomalies that consistently trick neural networks of different architectures in both datasets. This confirms the well-known vulnerabilities of softmax confidence and neural networks on unseen data.

For the Iris dataset, 62.12% (41/66) of generated anomalies under λ_1 were confidently classified by all tested models and decreased to 8.82% (6/68) under λ_2. For the White Wine Quality dataset, this was observed as 30.17% (175/580) under λ_1 and 8.37% (89/1063) under λ_2. These anomalies represent a shared weakness, which we believe is likely the result of the dataset-driven vulnerabilities and is not model specific. Thus, these generated samples can be used to test other models in deployment. Notably, this approach is not just a result of exploiting known weaknesses in softmax confidence, rather it has also deceived tree-based ensemble models like Random Forest and distance-based ones such as K-NN, indicating transferability and general vulnerability. This makes the method suitable for auditing different models, as it reveals regions of shared overconfidence that are likely learned by multiple classifiers trained on the same dataset. Table 2 details generated data with shown values rounded.

Table 2: Example of a Generated Anomalous White Wine Quality Sample (values rounded for readability only)

Fixed Acidity (g/dm^3)	Volatile Acidity (g/dm^3)	Citric Acid (g/dm^3)	Residual Sugar (g/dm^3)	Chlorides (g/dm^3)	Free Sulfur Dioxide (mg/dm^3)	Total Sulfur Dioxide (mg/dm^3)	Density (g/cm^3)	pH	Sulphates (g/dm^3)	Alcohol (%)
6.2906	0.42731	**- 0.26**	**- 6.053**	0.021421	**- 10.484**	61.0321	**0.9869**	3.356	0.4454	8.996

The sample in Table 2 is confidently classified (with a probability of at least 80%) as a low-quality wine by all models. However, its feature values are unrealistic as the residual sugar, citric acid, and free sulfur dioxide are negative, while the density is lower than the minimum (minimum density = 0.9871). This example highlights a confidently classified input that is unrealistic but has consistently fooled several classifiers.

COMMERCIAL POTENTIAL AND IMPACT

Our proposed method has been submitted for copyright, and it has a significant potential impact on the reliability and robustness of AI-based systems. It has high commercial value as an auditing tool used by auditing or consulting firms such as EY, Deloitte, KPMG, PwC, and Accenture in their technology divisions. It addresses the critical issues of overconfident machine learning models on unseen anomalous

inputs and provides a novel way to audit model robustness across different domains such as finance, healthcare, and autonomous systems. This approach works for testing different models, making it suitable to be implemented in existing toolboxes such as the IBM Adversarial Robustness Toolbox (ART), which includes other types of robustness testing distinct from ours. Safety will be increasingly important as AI is gradually integrated into various industries. As we contribute to robust AI systems, this aligns with the sustainable practices as outlined by the UN Sustainable Development Goals (SDGs), particularly SDG 9 (Industry, Innovation and Infrastructure). This tool aids in the development of resilient AI infrastructure, which results in safer and more trustworthy AI. Other related work in robustness testing includes Fast Gradient Sign Method (FGSM) and AdvGAN. Both methods focus on finding perturbations that alter input data to cause misclassification rather than focusing on overconfidence. In contrast, Expected Calibration Error (ECE) is a metric that measures how closely the confidence scores reflect reality, but it requires true labels and is usually applied to original data instead of using generated high-confidence anomalies for testing.

CONCLUSION

We have introduced a new model called RevGEN-MLP that is capable of auditing classification models by generating confidently classified anomalies to test the robustness of models on specific datasets. This approach exposes a hidden weakness shared by multiple models trained on the same dataset. These vulnerabilities are not captured by traditional metrics such as accuracy, as it deals with unseen anomalies without labels. By testing these generated samples on models with different confidence scores, we are able to analyse anomalies that are dangerous to real-life systems. We have also shown that Random Forest is more robust against the confidently classified anomalies we generated when compared to neural networks and K-NN classifiers. These generated anomalies can be integrated into real-world auditing frameworks and guide model selection beyond just classification performance. This approach lays the foundation for safer and more reliable AI systems, contributing to AI governance and sustainable innovation. In the future, this work can be extended to a family of models called RevGEN-Net with advanced invertible architectures, such as using normalizing flows to construct RevGEN-Flow.

ACKNOWLEDGEMENT

This work was partly supported by the Ministry of Higher Education (MOHE) under the Fundamental Research Grant Scheme (FRGS/1/2024/STG06/USM/02/11). The authors thank the School of Mathematical Sciences, Universiti Sains Malaysia, for funding the copyright fees and the I-URIID 2025 registration fees.

REFERENCES

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553. https://doi.org/10.1016/j.dss.2009.05.016

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188.

Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 7167–7177.

Pearce, T., Brintrup, A., & Zhu, J. (2021). Understanding Softmax Confidence and Uncertainty. *CoRR*, *abs/2106.04972*. https://arxiv.org/abs/2106.04972

Wan Isyraf & R.U. Gobithaasan (2025). Construction of Invertible Neural Network. Copyright submitted on 24 July 2025.

Wan Isyraf & R.U. Gobithaasan (2025). Implementation of RevGEN-MLP. https://github.com/Wmisyraf02/RevGEN-MLP Accessed 6th Sept 2025