Machine Learning-Based Approaches for Verification, Validation, and Uncertainty Quantification of Automatically Generated Models

1. Introduction:

Automatically Generated Models (AGMs) are increasingly prevalent across diverse fields, offering the advantage of reduced manual effort in their creation and updates. These models, often stemming from the Discrete Event Simulation (DES) discipline and termed Automated Simulation Model Generation (ASMG), can leverage data-driven techniques to adapt swiftly to changing conditions, rendering them particularly valuable in dynamic environments such as manufacturing. However, ensuring the reliability of these AGMs necessitates specialized Verification, Validation, and Uncertainty Quantification (VVUQ) processes. Given that ASMG models frequently operate as black boxes, traditional VVUQ methodologies may prove inadequate. The core challenge lies in achieving a comprehensive understanding of the model's internal logic and confirming its accurate representation of the physical system. The application of sophisticated data-driven methods introduces several key challenges that can impede successful VVUQ, including model opacity, data dependency, dynamic model adaptation, and the difficulty in quantifying uncertainty. Machine learning (ML) presents a promising paradigm with a wide array of techniques capable of addressing these specific hurdles in the VVUQ of AGMs. This section explores the application of ML-based approaches in detecting deviations in AGM behavior, managing the complexities of data and feature selection, examining prior work in the field, highlighting current state-of-the-art research, and addressing the inherent challenges and requirements associated with AGM VVUQ.

2. Machine Learning Techniques for VVUQ:

Machine learning offers a diverse range of techniques that can be applied to the Verification, Validation, and Uncertainty Quantification of Automatically Generated Models. These techniques can be broadly categorized into supervised, unsupervised, semi-supervised, and reinforcement learning, each offering unique capabilities for addressing different aspects of the VVUQ process. Supervised learning methods can be employed when labeled data regarding model deviations or performance is available. For instance, historical data on AGM behavior, classified as either valid or invalid, can be used to train a supervised learning model to predict the validity of new model outputs [1]. Unsupervised learning techniques, on the other hand, are valuable for identifying anomalies or unexpected patterns in AGM outputs or internal states without requiring prior knowledge of specific failure modes [3]. These methods can detect deviations from the normal operational behavior learned from unlabeled data. Semi-supervised learning approaches offer a middle ground, proving useful in scenarios where labeled data for model deviations is scarce. These techniques can leverage large volumes of unlabeled data to establish a baseline of normal model behavior and then use a smaller set of labeled data to refine the detection of deviations [2]. While less commonly applied in this context, reinforcement learning holds potential for developing adaptive validation strategies or for optimizing VVUQ processes over time based on feedback from the validation outcomes.

The increasing interest in and application of ML across various scientific and engineering domains, including surrogate modeling, anomaly detection, and uncertainty quantification, underscores its potential for AGM VVUQ [5]. The rapid advancements in deep learning and the growing availability of computational power are key enablers for exploring sophisticated ML models in complex validation tasks. This suggests a trend towards utilizing more advanced ML techniques to address the intricate validation needs of AGMs that themselves are becoming increasingly complex, often employing deep learning methodologies [5]. The field of ML model

VVUQ, particularly within engineering disciplines, is recognized as a crucial area, with ongoing research and discussions focused on developing robust methodologies [7]. The fact that experts are actively engaged in exploring the challenges and refining the methodology indicates that this is an evolving domain where established, universally accepted methods may not yet exist, highlighting the importance of investigating diverse ML approaches. Integrating domain-specific knowledge with advanced ML techniques, such as the use of physics-informed summary statistics with generative models for uncertainty assessment in imaging problems, demonstrates a promising direction for enhancing the effectiveness and interpretability of VVUQ, especially for AGMs modeling physical systems [8]. Furthermore, machine-learned classifiers can serve as correctness properties for the runtime verification of autonomous systems, showcasing the capability of ML to act as "oracles" for validating the behavior of other complex systems, including AGMs, while also providing measures of uncertainty associated with their verification decisions [9]. Recent advancements also include the use of Large Language Models (LLMs) for automated test case generation and the validation of these tests using token probabilities, as well as the ability of predictive uncertainties in ML models to reflect observed errors [10]. This indicates that advanced ML models can contribute to the verification aspect of VVUQ, and their internal mechanisms can be leveraged for validation purposes, while the uncertainty estimates they provide can serve as valuable indicators of their reliability in VVUQ tasks.

3. Classification Methods for Model Deviation Detection:

Model deviations in Automatically Generated Models can be effectively framed as anomalies within the model's output, behavior, or internal states. These anomalies can manifest in various forms, including isolated instances of unusual behavior (point anomalies), deviations that are only apparent within a specific context (contextual anomalies), or groups of data points that collectively deviate from the norm (collective anomalies) [4]. When historical data on AGM failures or deviations is available and appropriately labeled, standard supervised classification algorithms can be effectively applied to this task. Algorithms such as Support Vector Machines (SVM), Neural Networks (NN), Decision Trees, and Random Forests can learn from this labeled data to classify new model outputs or behaviors as either normal or indicative of a deviation [1]. However, a common challenge in anomaly detection scenarios is the issue of class imbalance, where instances of normal behavior far outnumber the occurrences of deviations. This imbalance needs to be carefully addressed during model training to prevent the classifier from being biased towards the majority class [2].

In situations where labeled data on model deviations is scarce or unavailable, unsupervised classification techniques, also known as anomaly detection methods, become particularly valuable. Distance-based methods, such as k-Nearest Neighbors (k-NN) and Local Outlier Factor (LOF), identify anomalies based on the principle that they are isolated from the majority of normal data points in the feature space [1]. These methods do not require prior knowledge of what constitutes a deviation, making them suitable for detecting unforeseen failure modes in AGMs. Density-based methods, including Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and other density estimation techniques, identify anomalies as data points residing in low-density regions of the data space [2]. These methods can effectively distinguish between clusters of normal AGM behavior and deviations that fall outside these typical operational regions. The Isolation Forest algorithm offers an efficient approach for high-dimensional data by isolating anomalies through random data partitioning, where anomalies, being rare and distinct, tend to be isolated more quickly than normal data points [2]. This makes it well-suited for complex AGMs with numerous parameters and output variables. One-Class

Support Vector Machine (OCSVM) learns a boundary that encloses the normal data, identifying any instances falling outside this boundary as anomalies [1]. This method is especially useful when only data representing normal AGM behavior is readily available, allowing the model to learn what constitutes "normal" operation and flag any future deviations. Clustering-based methods, such as k-Means, can identify anomalies as data points that do not belong to any major cluster or form very small, isolated clusters [1]. These methods can reveal inherent patterns in AGM behavior and highlight instances that do not conform to these established patterns. Neural network-based methods, particularly autoencoders, can be trained on normal AGM data to learn a compressed representation and then reconstruct the original data. Anomalies, which differ significantly from the normal data, typically exhibit higher reconstruction errors, making them detectable [2]. These methods are capable of learning complex, non-linear patterns in AGM behavior and are effective in identifying subtle deviations. Furthermore, semi-supervised techniques can enhance deviation detection by leveraging a small amount of labeled anomaly data in conjunction with a larger set of normal data, improving the accuracy of anomaly detection models [2]. The diverse array of classification and anomaly detection techniques available in machine learning provides a comprehensive toolkit for identifying various types of model deviations in Automatically Generated Models.

4. Challenges in Data Preparation for ML-Based VVUQ:

Effective training of machine learning models for Verification, Validation, and Uncertainty Quantification of Automatically Generated Models is heavily reliant on the quality of the data used [5]. Several data quality issues can significantly impact the performance and reliability of these ML models. Missing values in the dataset can hinder the training process and potentially introduce biases if not handled appropriately, requiring strategies such as imputation or removal of incomplete data points [14]. Outliers, which are data points that deviate significantly from the rest, pose a challenge as they might represent genuine deviations in the AGM's behavior or simply be erroneous entries, necessitating careful identification and treatment [14]. Noisy data, characterized by irrelevant or erroneous information, can obscure the underlying patterns that ML algorithms aim to learn, thereby reducing the accuracy and generalization ability of the VVUQ models [5]. Inconsistent data, arising from variations in formats, units, or representations of the same information, can lead to confusion for ML algorithms and require standardization and cleansing [15]. Duplicate data entries can skew the training process by overemphasizing certain observations, potentially leading to biased models [16]. Addressing these data quality issues through rigorous data cleaning and preprocessing is a fundamental step in building reliable ML-based VVUQ systems [1].

Beyond basic data quality, the representativeness and potential biases within the data are critical considerations [15]. The training data used for VVUQ models must accurately reflect the operational environment in which the AGM will be deployed to ensure that the validation results are meaningful and applicable to real-world conditions [15]. Biases present in the data, which can arise from various sources such as historical human decisions or data collection processes, can lead to VVUQ models that unfairly or inaccurately assess the AGM's performance for certain subsets of data or operational scenarios [15]. Identifying and mitigating these biases through techniques like data re-balancing or the use of unbiased data sources is essential for ensuring the fairness and accuracy of the VVUQ process [19]. Furthermore, for certain ML algorithms, the scaling and normalization of data to a common range is important to prevent features with larger magnitudes from dominating the learning process [14]. The appropriate selection of scaling and normalization methods can significantly impact the performance of the VVUQ models. Finally, the way data is split into training, testing, and validation sets is crucial for avoiding overfitting, where the model learns the training data too well and fails to generalize to unseen data, and for obtaining a reliable estimate of the model's performance on new data [16]. For time-series data or data with temporal dependencies, traditional random splitting methods can lead to data leakage, requiring the use of specialized splitting techniques that preserve the temporal order of the data [16]. The broader context of VVUQ in engineering and simulation, as highlighted by ASME and SmartUQ, underscores the fundamental importance of sound data acquisition and preparation practices for the overall validity of the VVUQ outcomes, regardless of whether traditional simulation or ML-based methods are employed [28]. Poor data quality can severely impede the performance of ML models used for VVUQ, potentially resulting in inaccurate deviation detection or unreliable uncertainty quantification. If the ML model used to validate an AGM is trained on flawed data, it might either fail to detect actual issues or falsely flag normal behavior as a deviation, thereby undermining the entire VVUQ process. Similarly, if the data used to train the VVUQ model does not accurately reflect the real-world conditions under which the AGM operates, the validation results might not be reliable. An AGM used in a dynamic environment like manufacturing needs to be validated against data that captures the full range of conditions it will encounter. If the validation data is limited or biased towards certain conditions, the assessment of the AGM's reliability might be incomplete or misleading.

5. Feature Selection for Effective VVUQ:

Selecting the most relevant features from the Automatically Generated Model's data, which can include inputs, outputs, and internal states, is a critical step in developing effective machine learning models for Verification, Validation, and Uncertainty Quantification [30]. By focusing on the features that are most indicative of the AGM's validity and reliability, feature selection can lead to improved performance, enhanced interpretability, and increased efficiency of the ML-based VVUQ processes. Furthermore, reducing the number of input features through feature selection can help mitigate the curse of dimensionality, a phenomenon that can negatively impact the performance of ML models, especially when dealing with complex AGMs that have a large number of parameters [32].

Various feature selection techniques can be employed. Filter methods utilize statistical measures, such as correlation, chi-square test, and mutual information, to rank the features based on their relevance to the target variable (e.g., an indicator of model deviation) [30]. These methods are computationally efficient and can provide valuable insights into the relationships between the features and the target variable. For an AGM with numerous potential indicators of deviation, filter methods can quickly identify the most statistically significant ones to prioritize for training the VVUQ model. Wrapper methods, including techniques like Forward Selection, Backward Elimination, and Recursive Feature Elimination (RFE), evaluate different subsets of features by training and testing a specific ML model on each subset [30]. These methods can account for potential interactions between features and often result in better model performance compared to filter methods, although they are generally more computationally expensive. Certain deviations in an AGM might only be detectable through the combined effect of multiple parameters. Wrapper methods can identify these important combinations of features for the VVUQ model. Embedded methods integrate feature selection directly into the model training process. Examples include LASSO and Ridge regression, which apply penalties to the coefficients of less important features, effectively shrinking them towards or to zero, and tree-based models like Random Forests and Gradient Boosting, which provide feature importance scores based on how much each feature contributes to reducing impurity in the trees [30]. These methods offer a balance between computational efficiency and model performance. When training a complex model like a Random Forest to detect AGM deviations, the model's inherent feature importance ranking can directly provide insights into which aspects of the AGM are most indicative of problems.

Despite the benefits, feature selection for AGM VVUQ also presents challenges. Identifying relevant features can be difficult when the underlying causes of model deviations are not well understood. Additionally, there is a risk of overfitting if feature selection is not performed carefully, particularly when combined with cross-validation techniques [36]. To avoid introducing bias and ensure the generalizability of the VVUQ model, feature selection should ideally be performed within each fold of the cross-validation process. Incorporating domain knowledge about the AGM and the physical system it represents can also be crucial in guiding the feature selection process and ensuring that the selected features have meaningful interpretations [27]. Effective VVUQ of AGMs using ML heavily relies on selecting the right features that are truly indicative of the model's validity and reliability, and this selection process needs to be robust and avoid introducing biases. If irrelevant features are used to train the VVUQ model, it might focus on noise in the AGM's data rather than actual indicators of a problem. Conversely, selecting the right features can lead to a more accurate and interpretable validation process.

6. Previous Applications of ML in AGM Verification and Validation:

Prior research has demonstrated the potential of machine learning techniques across various aspects of Verification, Validation, and Uncertainty Quantification for complex models, including those that are automatically generated. In the realm of verification, ML classifiers have been employed to learn correctness properties of systems, enabling runtime verification by detecting deviations from expected behavior [9]. This approach involves training an ML model to recognize patterns of correct operation, with any significant deviation triggering an alert. For validation, ML-based surrogate models have been used to approximate the behavior of AGMs, allowing for faster comparisons against real-world data or high-fidelity simulations [5]. These surrogate models can predict the AGM's output under various conditions, and discrepancies between these predictions and the actual observed data can indicate potential issues with the AGM. Anomaly detection techniques, as discussed earlier, have also been applied to validate AGMs by identifying unexpected or unusual patterns in their outputs or internal states that might signal a deviation from intended behavior [5].

Uncertainty quantification is another area where ML has shown promise in the context of complex models, including AGMs. Techniques such as quantifying the prediction uncertainties of Deep Neural Networks (DNNs) used as surrogate models have been explored [37]. By estimating the range of possible outputs and the likelihood of these outcomes, these methods provide a measure of confidence in the AGM's predictions. The broader field of machine learning model validation and testing, encompassing unit, integration, functional, regression, and robustness testing, provides a foundation of methodologies that can be adapted for AGMs [24]. While not exclusively focused on AGMs, the principles of verifying individual components, ensuring proper integration, validating functionality against requirements, checking for regressions after modifications, and assessing robustness to unexpected inputs are all pertinent to ensuring the reliability of automatically generated models. Challenges encountered in validating ML models in specific domains, such as finance, including issues related to data integrity, bias, and the need for explainability, also hold relevance for AGM VVUQ, highlighting common concerns when applying ML in critical applications [24]. The academic literature, as reflected in journals like the Journal of Machine Learning Research (JMLR) and Transactions on Machine Learning Research (TMLR), contains significant research on machine learning methodologies relevant to VVUQ, including techniques for constraint reasoning integrated with ML for structured prediction, which could be used to verify that AGM outputs adhere to predefined constraints, and methods for model selection and performance evaluation, which are essential for ensuring the reliability of both the AGMs and the ML models used for their validation [41]. Overall, previous research supports the feasibility and value of employing diverse ML techniques across the spectrum of VVUQ for complex models like AGMs. The fact that researchers have already applied ML for verification (checking correctness), validation (comparing with real-world data), and uncertainty quantification (assessing reliability of predictions) of other complex systems suggests that these approaches can be adapted and extended for automatically generated models.

7. State-of-the-Art Research:

Current state-of-the-art research in the application of machine learning to the Verification, Validation, and Uncertainty Quantification of Automatically Generated Models reflects several key trends. There is a growing emphasis on Explainable AI (XAI) techniques to address the inherent opacity of some ML models used in VVUQ [45]. These techniques aim to provide insights into the decision-making processes of black-box models, enhancing trust and enabling the identification of potential biases or errors in the validation process. For dynamic AGMs that continuously learn and adapt, continuous validation and monitoring using ML techniques are receiving significant attention [25]. These approaches involve the ongoing assessment of the AGM's performance over time, including the detection of drift in its behavior and the retraining or fine-tuning of validation models to maintain accuracy. Advancements in uncertainty quantification for complex ML models used in VVUQ are also a prominent area of research [5]. Techniques like Bayesian Neural Networks, Deep Ensembles, and Monte Carlo Dropout are being refined to provide more accurate and reliable estimates of the uncertainty associated with model predictions. Furthermore, there is a trend towards the development of specialized frameworks and toolkits that incorporate ML for VVUQ in specific domains, aiming to streamline the validation process and make advanced techniques more accessible [29].

Influential papers in this field highlight the recent progress. The work by Yaseen and Wu on quantifying DNN prediction uncertainties for VVUQ of machine learning models represents a significant contribution to applying specific uncertainty quantification techniques to deep learning models used as surrogates for complex systems [37]. The introduction of VALTEST, a framework for validating LLM-generated test cases using token probabilities, showcases a novel application of advanced ML in the verification of software, a concept that can be extended to validating the logic of AGMs [10]. The ongoing efforts in developing standards and organizing symposia by organizations like ASME indicate a concerted community effort to establish best practices and guidelines for VVUQ, including the integration of ML methodologies [29]. Frameworks like EasyVVUQ aim to make state-of-the-art VVUQ algorithms, potentially including ML-based analyses, more readily usable in high-performance computing environments [60]. The broader landscape of machine learning research, as evidenced by the vast number of publications in venues like Papers with Code, reveals rapid advancements across various ML tasks, many of which can be leveraged for enhancing VVUQ techniques for AGMs [63]. Current research is actively exploring and refining ML techniques for various aspects of VVUQ, with a particular emphasis on handling the complexities of deep learning models and ensuring their reliability through uncertainty quantification and continuous monitoring. The ongoing research in areas like XAI, continuous validation, and uncertainty quantification specifically for ML models suggests a recognition of the unique challenges posed by these models and a drive to develop more robust and trustworthy validation methodologies that can be applied to AGMs as well.

8. Addressing Key Challenges of AGM VVUQ with ML:

Machine learning offers a suite of techniques to specifically address the key challenges associated with the Verification, Validation, and Uncertainty Quantification of Automatically Generated Models. Model opacity, a significant hurdle due to the black-box nature of many sophisticated AGMs, can be tackled using Explainable AI (XAI) methods. Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), as well as the use of counterfactual examples, can provide insights into the decision-making processes of complex ML models used for VVUQ, thereby offering a degree of transparency [45]. Furthermore, employing inherently interpretable ML models like decision trees or linear regression for VVUQ can be advantageous when transparency is a primary concern. The concept of using surrogate models, where a black-box AGM is approximated by a more interpretable ML model for validation purposes, also helps in addressing opacity [45].

The challenge of data dependency in AGMs, which rely heavily on the quality and representativeness of the data they are trained on, can be mitigated by using ML techniques to assess and improve the data itself. ML-based data cleaning methods, anomaly detection algorithms applied to the data, and techniques for imputing missing values can ensure that the data used for both training the AGM and for its VVUQ is of high quality [14]. Data augmentation techniques can also be employed to enhance the robustness of VVUQ models against variations or limitations in the available data [5]. For AGMs that exhibit dynamic model adaptation, continuously learning and evolving over time, ML-based continuous validation techniques are essential [25]. Drift detection algorithms can identify changes in the AGM's behavior, and online learning methods can adapt the VVUQ models in real-time as new data is ingested. Retraining and fine-tuning of VVUQ models can also be performed to keep pace with the evolving characteristics of the AGM. Quantifying uncertainty, a critical aspect of VVUQ, can be achieved through various ML techniques such as Bayesian Neural Networks, Deep Ensembles, and Monte Carlo Dropout [5]. These methods provide estimates of the uncertainty associated with the predictions and behavior of AGMs, offering a measure of their reliability. Moreover, uncertainty quantification can also be applied to the VVUQ process itself, providing a level of confidence in the validation results.

9. Meeting VVUQ Requirements with ML:

Machine learning-based approaches are well-suited to meet the key requirements for the Verification, Validation, and Uncertainty Quantification of Automatically Generated Models. Model interpretability can be achieved through the use of inherently transparent ML models like decision trees and linear models for VVUQ. When more complex ML validators are employed, post-hoc explanation techniques such as SHAP, LIME, and surrogate models can be used to shed light on their behavior. Upholding data quality is facilitated by ML-based data cleaning, validation, and preprocessing techniques, ensuring the reliability of the data used for both AGM development and VVUQ. The requirement for validatable algorithms is met by utilizing well-established and theoretically sound ML algorithms for VVUQ. Furthermore, the robustness and security of these algorithms, including considerations for adversarial attacks, are important aspects in their selection and application. Continuous validation, crucial for dynamic AGMs, is enabled by ML-based continuous monitoring, drift detection, and retraining strategies that can validate AGMs in real-time as new data is ingested. The integration of VVUQ processes into existing model and production infrastructure can be streamlined through the adoption of MLOps practices and frameworks [53]. Finally, scalability, a key consideration for large and complex AGMs with evolving data, can be achieved through the use of distributed ML techniques and cloud computing platforms [28]. For example, ML can enhance constraint suggestions and detect anomalies in data quality time series in a scalable data quality validation platform [70].

10. Conclusion:

Machine learning offers a powerful and versatile set of tools for advancing the field of Verification, Validation, and Uncertainty Quantification for Automatically Generated Models. ML-based approaches can effectively address the unique challenges posed by AGMs, including their opacity, reliance on data quality, dynamic adaptation capabilities, and the need for quantifying uncertainty in their predictions. The careful selection and application of appropriate ML techniques, coupled with a strong focus on data quality and feature selection, are essential for developing robust and reliable VVUQ processes. Ongoing research and development in this area continue to enhance the capabilities of ML in ensuring the trustworthiness and accuracy of AGMs in a wide range of applications. By leveraging the power of machine learning, the reliability and trustworthiness of these increasingly important models can be significantly enhanced through sophisticated validation methodologies.