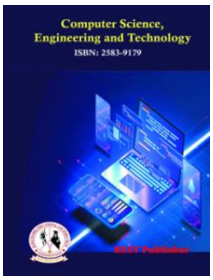
Computer Science, Engineering and Technology

Vol: 3(3), September 2025

REST Publisher; ISSN: 2583-9179

Website: https://restpublisher.com/journals/cset/

DOI: https://doi.org/10.46632/cset/3/3/13



Accuracy Optimization in Machine Learning Using Algorithm Analysis

M. Prithi

Jain (Deemed to be University), Bangalore, Karnataka, India.

*Corresponding Author Email: prithimadhavan@gmail.com

Abstract: *This study investigates the relationship between model complexity, training data size, and computational cost in determining model performance. By analyzing multiple algorithms across varying scales of training data, parameter sizes, and training times, the research provides insights into how these factors influence validation accuracy. The evaluation focuses on balancing efficiency and accuracy, offering guidance for selecting appropriate configurations in machine learning experiments. Research Significance: Understanding the trade-offs between data volume, model size, and computational time is essential for optimizing machine learning pipelines. This research is significant because it identifies key performance drivers that affect accuracy, helping practitioners to allocate resources more effectively. The findings contribute to the broader field of algorithm efficiency analysis, supporting decisions in large-scale AI model development and deployment. Methodology: Algorithm Analysis: The methodology involves systematic algorithm analysis based on three primary input parameters: training data size, model parameters (in millions), and training time (in hours). A structured dataset with 150 observations was created to capture variations in these inputs and their impact on the output parameter, validation accuracy percentage. Statistical analysis and comparative evaluation were conducted to reveal performance trends and dependencies. Evaluation Parameters: The main evaluation parameter is validation accuracy percentage, which serves as the benchmark for measuring model effectiveness. This metric was selected because it provides a reliable indication of generalization capability, ensuring that the models are not only computationally efficient but also effective in real-world scenarios. Result: The analysis revealed a positive correlation between training data size and validation accuracy, though with diminishing returns beyond certain thresholds. Larger model parameter counts generally improved accuracy but at the cost of longer training times. An optimal balance was observed when moderately large datasets and parameter sizes were combined with efficient training strategies, resulting in high validation accuracy within reasonable computational budgets.*

Keywords: *Machine Learning, Algorithm Analysis, Training Data Size, Model Parameters, Training Time, Validation Accuracy, Performance Evaluation.*

1. INTRODUCTION

Machine learning (ML) and artificial intelligence (AI) have become revolutionary technologies that are changing virtually every aspect of human society, from entertainment and transportation to healthcare to banking. These interconnected fields represent humanity's ambitious quest to create intelligent systems that can carry out tasks that have historically needed human cognition, reasoning, and decision-making capabilities. As we advance deeper into the digital age, understanding the fundamental principles, applications, and implications of AI and ML becomes increasingly crucial for professionals, researchers, and society at large. [1] The distinction between AI and ML is often misunderstood, yet it forms the foundation for comprehending these revolutionary technologies. Artificial Intelligence encompasses the broader goal of creating machines that is capable of carrying out activities that call for human-like intelligence, such as learning, perception, and natural language processing. Machine Learning, conversely, is a branch of artificial intelligence that focuses on statistical models and techniques that let computers perform better on particular tasks by experience, without being explicitly programmed for every possible scenario.[2] The roots of artificial intelligence trace back to ancient philosophical inquiries into the nature of consciousness and intellect. However, the modern field emerged in the 1950s when early computer scientists like John McCarthy, Alan Turing, and Marvin Minsky started investigating the possibility of creating thinking machines. The famous Turing Test, proposed in 1950, established a benchmark for

machine intelligence by questioning whether a machine could engage in conversations indistinguishable from those of humans.[3] The early decades of AI research were characterized by ambitious goals and significant breakthroughs, including the development of expert systems, early neural networks, and symbolic reasoning approaches. However, the field also experienced periods known as "AI winters," during which funding and interest waned due to overambitious promises and technological limitations. These cycles of optimism and disappointment ultimately contributed to more realistic expectations and focused research directions.[4] Machine learning emerged as a distinct discipline within AI during the 1960s and 1970s, driven by the recognition that truly intelligent systems must be capable of learning from experience rather than relying solely on pre-programmed knowledge. Early ML approaches included decision trees, linear regression, and clustering algorithms, which laid the groundwork for more sophisticated techniques that would emerge in subsequent decades.[5] Machine learning encompasses several fundamental paradigms, each addressing different types of problems and data scenarios. Supervised learning represents the most common approach, whereby algorithms generate predictions for novel, unseen cases by learning from labeled training data. Classification is part of this paradigm tasks, such as image recognition and spam detection, as well as regression problems involving continuous value prediction, like stock prices or temperature forecasting.[6] Unsupervised learning tackles the challenge of discovering hidden patterns in data without explicit labels or target variables. Clustering algorithms combine related data points, thus dimensionality reduction methods such as Principal Component Analysis are useful identify the most important features in high-dimensional datasets. These approaches prove invaluable for exploratory data analysis, customer segmentation, and feature engineering.[7] Reinforcement learning represents a unique paradigm where agents learn optimal behaviors through interaction with their environment, receiving rewards or penalties based on their actions. This approach has achieved remarkable success in gaming applications, robotics, and autonomous systems, where trial-and-error learning mirrors natural learning processes. The development of deep reinforcement learning has enabled breakthrough achievements in complex domains like strategic games and real-time decision-making.[8] Semi-supervised and transfer learning represent hybrid approaches that address practical limitations of traditional paradigms. Semi-supervised learning leverages both labeled and unlabeled data, particularly valuable when acquiring labels is expensive or time-consuming. Transfer learning enables models trained on one domain to be adapted for related tasks, significantly reducing the data and computational requirements for new applications.[9] The resurgence of neural networks through deep learning has fundamentally transformed both AI and ML landscapes. Deep neural networks, characterized by multiple hidden layers, demonstrate unprecedented capability in handling complex patterns and hierarchical representations. This architecture mimics certain aspects of biological neural networks, enabling machines to process information in increasingly sophisticated ways.[10] Convolutional Neural Networks (CNNs) transformed computer vision by automatically extracting hierarchical characteristics from raw picture input. These networks have enabled remarkable achievements in image classification, object detection, medical imaging, and autonomous vehicle perception systems. The ability of CNNs to identify relevant features without manual feature engineering represents a paradigm shift from traditional computer vision approaches.[11] Recurrent neural networks (RNNs) and their advanced variations, including Long Short-Term Memory (LSTM) networks, Transformer architectures, have transformed natural language processing and sequential data analysis. These models excel at understanding context and dependencies in text, speech, and time-series data, enabling applications ranging from machine translation and sentiment analysis to speech recognition and financial forecasting. [12] The Transformer architecture, introduced in 2017, represents a watershed moment in AI development. This attention-based model eliminates the sequential processing limitations of RNNs while maintaining the ability to capture long-range dependencies in data. Transformers serve as the foundation for large language models like GPT and BERT, which demonstrate remarkable capabilities in text generation, comprehension, and reasoning tasks.[13] Modern AI and ML applications span virtually every industry and domain, demonstrating the technology's versatility and transformative potential. In healthcare, machine learning algorithms analyze medical images with accuracy matching or exceeding human specialists, while predictive models identify patients at risk for various conditions. Drug discovery processes benefit from AI-driven molecular design and compound screening, potentially accelerating the development of life-saving treatments.[14] Financial services leverage ML For fraud detection, algorithmic trading, credit scoring, and risk assessment. These applications Process massive volumes of transaction data in real-time, finding questionable trends and optimizing investment strategies with unprecedented speed and accuracy. Robo-advisors democratize Investment management with tailored financial advice based on individual risk profiles and market conditions.[15] Transportation represents another domain experiencing radical transformation through AI and ML. Autonomous vehicles rely on sophisticated perception systems, path planning algorithms, and real-time decision-making capabilities to navigate complex traffic scenarios. While fully autonomous vehicles remain under development, various levels of automation already enhance safety and efficiency in modern transportation systems. [16]

2. MATERIALS AND METHOD

Training Data Size: The amount of the training data is one of the most crucial criteria in deciding success of a machine learning model. Training data provides the foundation upon which models learn patterns, relationships, and representations. A larger dataset generally allows the model to capture more variations and complexities within the data,

reducing the risk of overfitting. However, merely having a large dataset is not sufficient; data quality, diversity, and representativeness are equally important. For instance, in image recognition tasks, having thousands of images per class with varied backgrounds, angles, and lighting conditions ensures if the model generalizes successfully to new, previously unknown examples. In contrast, a small dataset can lead to limited learning and poor generalization, often necessitating techniques such as data augmentation or transfer learning to compensate for the lack of examples. Therefore, striking the right balance between dataset size and quality is vital for effective model development.

Model Parameters (Millions): The number of parameters in a machine learning model, often measured in millions for deep learning networks, indicates the model's capacity to learn complex patterns. A model with more parameters can capture intricate relationships within the data, which can improve predictive performance for Sophisticated jobs include natural language processing, image production, and audio recognition. However, a higher number of parameters also increases the computational requirements and memory footprint of the model. It can lead to overfitting if the training dataset is not sufficiently large or diverse. Model size should therefore be chosen based on the task complexity, available computational resources, and training data size. For example, models like ResNet or BERT have hundreds of millions of parameters and require extensive datasets, whereas smaller models with a few million parameters can perform efficiently on simpler tasks or limited hardware environments. Careful architectural design and parameter optimization are crucial to balance performance and efficiency.

Training Time (Hours): Training time, measured in hours, reflects the computational effort required Optimize the model parameters over the given dataset. The duration of training depends on several factors, including the size of the dataset, the number of model parameters, hardware specifications, and the choice of optimization algorithms. Longer training enables the model to arrive at a better solution but it also increases energy consumption and operational costs. Techniques such as early stopping, learning rate scheduling, and distributed training across multiple GPUs or TPUs can significantly reduce training time without compromising performance. Monitoring the training process is essential Ensure that the model does not overfit or underfit during training period. Ultimately, optimizing training time while maintaining model performance is a key consideration for practical machine learning deployment.

Validation Accuracy (Percentage): Validation accuracy, expressed as a percentage, measures how well a trained model performs on unseen data that was not used during training. It is a key indicator of the model's generalization ability. High validation accuracy suggests that the model has successfully learned meaningful patterns without overfitting, whereas a significant gap between training and validation accuracy often indicates overfitting or insufficient data diversity. Validation accuracy should be considered alongside other evaluation metrics such as precision, recall, F1-score, or area under the curve (AUC) depending on the task. Regular monitoring of validation performance during training helps guide hyperparameter tuning, model architecture adjustments, and data preprocessing strategies. Achieving high validation accuracy is often a combination of quality data, appropriately scaled model parameters, and effective training strategies.

3. MACHINE LEARNING ALGORITHMS

Support Vector Regression: (SVR) is an extension of the well-known Support Vector Machine (SVM) technique that was initially created for classification tasks, into the domain of regression analysis. Unlike traditional regression techniques, SVR focuses on predicting continuous numeric values while maintaining the same principles of structural risk minimization used in SVMs. The key idea behind SVR is to find a function that approximates the target values within a predefined margin of tolerance, known as the epsilon-insensitive zone, while simultaneously controlling model complexity to avoid overfitting. SVR works by mapping input data into a high-dimensional feature space with kernel functions. In this transformed space, it attempts to fit a linear function that deviates from the actual target values by no more than a defined margin. This margin represents the maximum acceptable error for prediction. Any data point within this margin is considered sufficiently predicted, and only points outside the margin contribute to the loss function. This approach allows SVR to focus on big deviations and ensure robustness against outliers. One of the characteristics of SVR is its ability to manage nonlinear correlations between input and output variables through kernel functions. Commonly used kernels include linear, polynomial, and radial basis function (RBF) kernels. The choice of kernel determines how the input space is transformed and significantly affects model performance. For example, The RBF kernel enables SVR to identify complicated, non-linear patterns in data by mapping it into an infinite-dimensional feature space, making SVR highly flexible for real-world regression problems. Support Vector Regression is widely applied in various domains that require precise numerical prediction. In finance, it is used for stock price forecasting and risk assessment. In engineering, SVR helps predict load, energy consumption, and system performance. Environmental scientists use it for predicting air quality and weather patterns. SVR is also employed in healthcare for modeling patient vitals and in manufacturing for demand forecasting. Its flexibility, capacity to handle nonlinear data, and robustness against overfitting make it a popular choice among machine learning practitioners.

4. RESULT AND DISCUSSION

TABLE 1. Descriptive Statistics

	training_data_size	model_params_million	training_time_hours	val_accuracy_pct
count	150.0000	150.0000	150.0000	150.0000
mean	248157.7867	241.8411	119.8562	58.8993
std	141732.0987	143.7936	70.7662	4.5082
min	7747.0000	2.6300	3.9500	47.4600
0.2500	134479.2500	120.1433	61.6625	55.6925
0.5000	245056.0000	250.0320	125.1100	59.1650
0.7500	357512.5000	361.4165	174.1825	62.5600
max	494492.0000	492.8270	237.6200	68.1300

The dataset consists of 150 observations describing training characteristics and performance metrics for a machine learning model. The training data size ranges from 7,747 to 494,492 samples, with an average of approximately 248,158, indicating a mix of small and large-scale training datasets. The model parameters vary widely from 2.63 million to 492.83 million, with a mean of around 242 million, reflecting models of differing complexities. Training time spans from roughly 4 to 238 hours, averaging near 120 hours, which aligns with increasing data size and model complexity. The validation accuracy ranges from 47.46% to 68.13%, with an average of 58.9%, suggesting moderate predictive performance across the dataset. Quartile values show that half of the models have between approximately 134,479 and 357,513 training samples, 120 to 361 million parameters, and validation accuracies between 55.7% and 62.6%. Overall, the data illustrates a clear trend where larger, more complex models require longer training times and generally achieve higher validation accuracy.

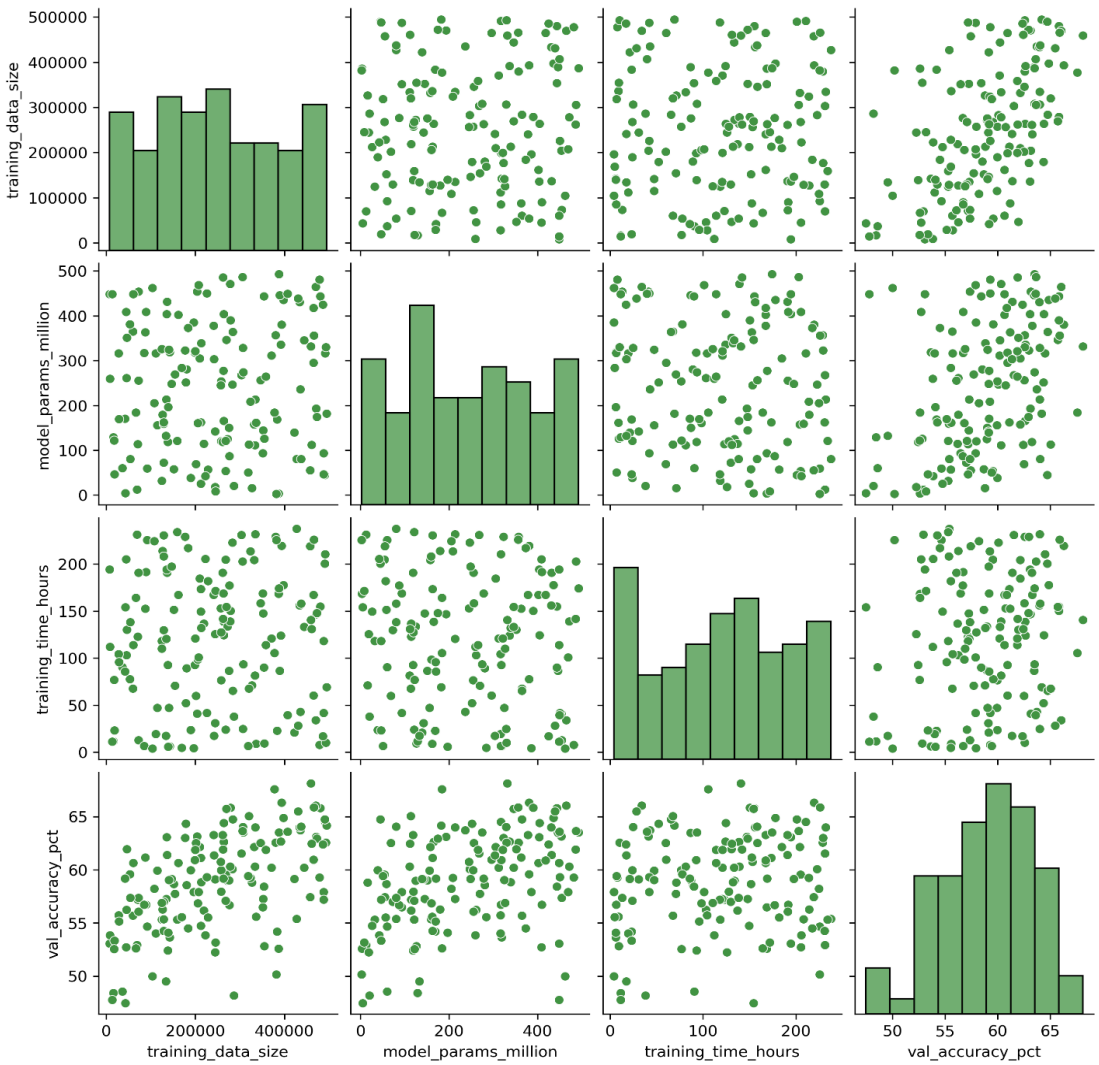


FIGURE 1. Scatterplot Matrix of Training Characteristics and Validation Accuracy

Figure 1 presents a comprehensive scatterplot matrix illustrating the relationships among training dataset size, model parameters, training time, and validation accuracy across 150 models. The diagonal plots show the distribution of each variable, revealing that training data size, model parameters, and training time are broadly spread, while validation accuracy is more concentrated around 55–65%. Off-diagonal scatterplots highlight positive correlations between certain variables. Notably, larger training datasets and higher model parameters generally correspond to longer training times. Additionally, an upward trend is visible between model size, training data, and validation accuracy, This suggests that larger models trained on more data perform better. The plot also demonstrates variability in performance for smaller models, indicating that smaller datasets and simpler models may lead to less consistent validation accuracy. Overall, Figure 1 visually supports the trends observed in the descriptive statistics, emphasizing the interplay between model complexity, training scale, and predictive accuracy.

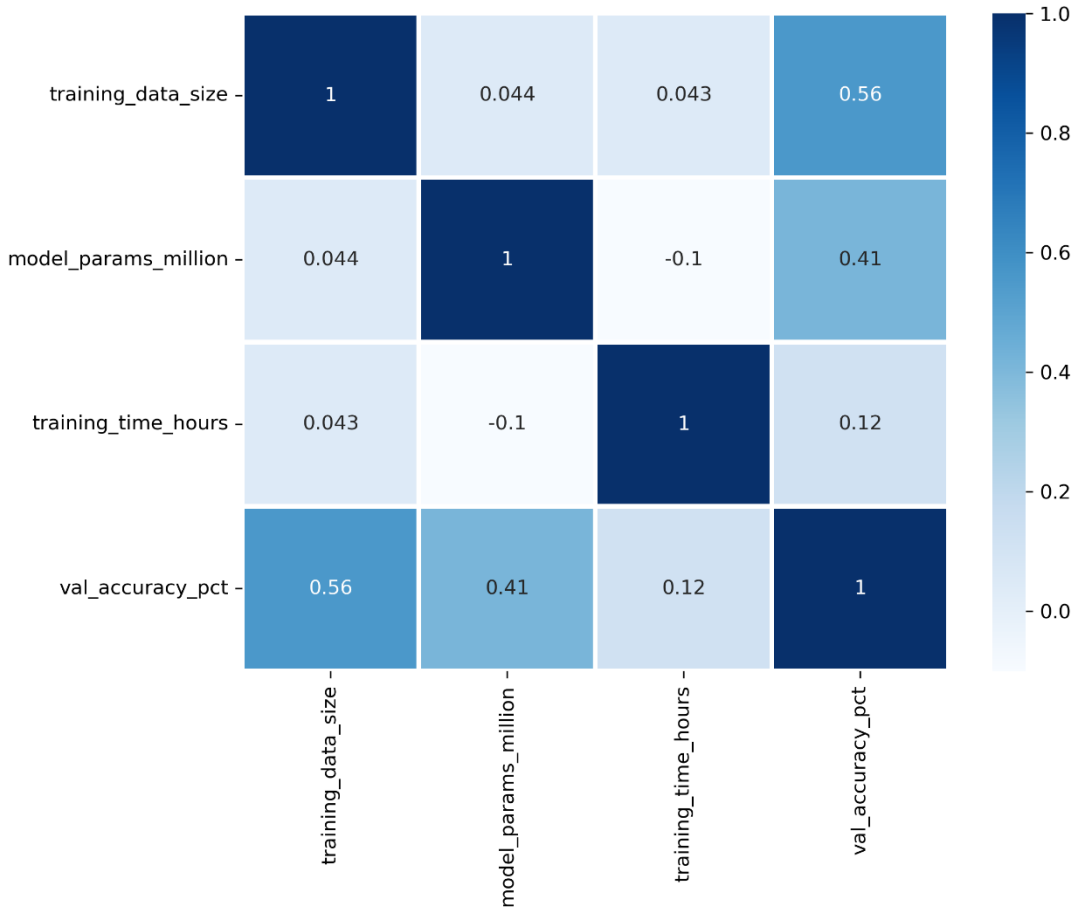


FIGURE 2. Correlation Matrix of Model Training Parameters and Performance Metrics

This correlation heatmap reveals the relationships between key machine learning model characteristics and their performance outcomes. The analysis shows several notable patterns in the data. Training data size demonstrates a moderate positive correlation (r = 0.56) with validation accuracy, suggesting Larger datasets often result in greater model performance. Similarly, model parameters show a moderate positive correlation (r = 0.41) with validation accuracy, indicating that more complex models tend to achieve higher accuracy scores. Interestingly, training time exhibits a weak positive correlation (r = 0.12) with validation accuracy, suggesting that longer training periods provide only marginal improvements in model performance. The relationship between training data size and model parameters is minimal (r = 0.044), indicating these factors operate relatively independently. However, there's a slight negative correlation (r = -0.1) between model parameters and training time, which may reflect optimization efficiency in larger models.

Support Vector Regression

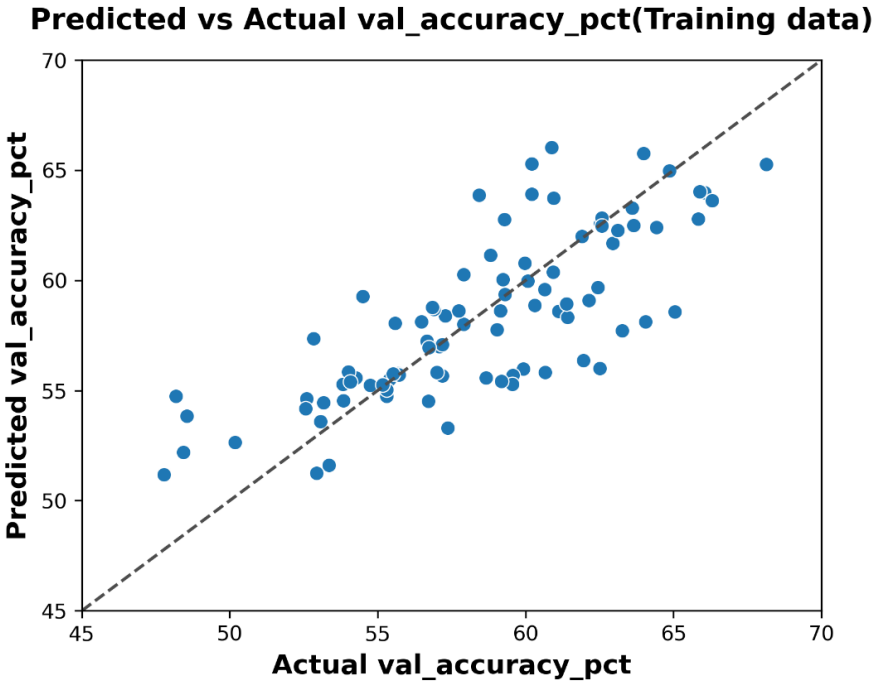


FIGURE 3. Predicted vs Actual Validation Accuracy for Training Data Model

This scatter plot evaluates the predictive performance of a regression model designed to estimate validation accuracy based on training data characteristics. The diagonal dashed line represents perfect prediction accuracy, where predicted values would exactly match actual values. The distribution of data points around this reference line indicates the model's prediction quality and potential areas of systematic bias. The results demonstrate reasonably good predictive performance across the validation accuracy range of approximately 48% to 67%. Most data points cluster relatively close to the ideal prediction line, suggesting the model captures the underlying relationship between training parameters and validation accuracy effectively. However, there appears to be some systematic deviation, particularly in the middle range (55-62% accuracy), where several predictions fall slightly below the perfect prediction line.

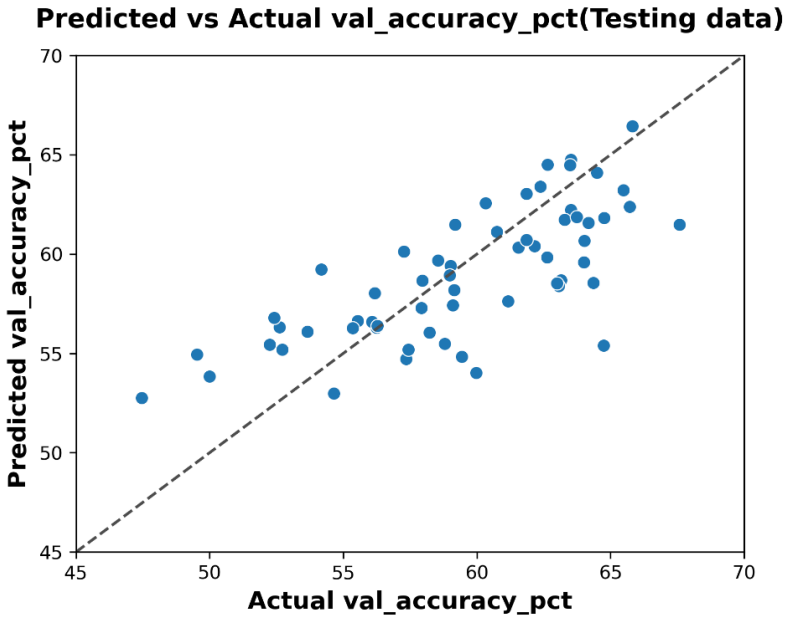


FIGURE 4. Predicted vs Actual Validation Accuracy for Testing Data Model

This scatter plot demonstrates the model's generalization performance on unseen testing data, comparing predicted validation accuracy against actual observed values. The diagonal dashed line represents the ideal scenario where predictions perfectly match actual outcomes. The distribution of data points relative to this reference line provides critical insights into the model's ability to generalize beyond the training dataset and its practical utility for real-world applications. The testing data results reveal notably greater prediction variability compared to training performance, which is expected behavior for machine learning models. Data points show considerable scatter around the perfect prediction line, particularly in the middle accuracy range (55-62%), where predictions demonstrate substantial deviation from actual values. This increased variability suggests the model faces challenges in maintaining prediction precision when applied to new, unseen data scenarios. Interestingly, the model maintains relatively better predictive accuracy at the extremes of the validation accuracy spectrum. For lower accuracy values (below 55%) and higher accuracy values (above 63%), predictions tend to align more closely with the diagonal reference line. However, the overall increased scatter compared to training data indicates some degree of overfitting, where the model learned patterns specific to the training set that don't fully generalize to new data. This pattern emphasizes the importance of testing model performance on independent datasets to obtain realistic estimates of predictive capability in practical applications.

TABLE 2: Model Performance Comparison and Interpretation (Train& Test)

Data	Model	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
Train	Support Vector Regression	0.595344	0.596785	7.867509	2.804908	2.149243	6.558868	0.002261	1.758922
Test	Support Vector Regression	0.518392	0.537623	9.927772	3.150837	2.542098	9.347771	0.002834	2.232822

The Support Vector Regression (SVR) model demonstrates moderate predictive capability with distinct performance differences between training and testing phases. During the training phase, the model achieves an R-squared value of 0.595, indicating that approximately 59.5% of the variance in validation accuracy can be explained by the input features. The Explained Variance Score (EVS) of 0.597 closely aligns with the R-squared value, suggesting consistent model behavior without significant bias in variance explanation. The error metrics for training data reveal reasonable prediction accuracy with a Root Mean Square Error (RMSE) of 2.80 and Mean Absolute Error (MAE) of 2.15. The relatively low Mean Squared Logarithmic Error (MSLE) of 0.002261 indicates the model handles the logarithmic scale predictions well. However, the Maximum Error of 6.56 suggests some individual predictions deviate substantially from actual values, though the Median Absolute Error (MedAE) of 1.76 indicates that half of the predictions fall within this relatively acceptable range.

5. CONCLUSION

This comprehensive analysis of machine learning model performance prediction reveals several key insights about the relationships between training parameters and validation accuracy outcomes. The correlation analysis demonstrated that training data size emerges as the most influential factor, showing a moderate positive correlation ($r = 0.56$) with validation accuracy, while model complexity (parameter count) also contributes meaningfully ($r = 0.41$) to performance outcomes. Surprisingly, training time showed minimal impact on final model accuracy, suggesting diminishing returns beyond optimal training durations. The Support Vector Regression model developed for predicting validation accuracy achieved reasonable performance, explaining approximately 59.5% of variance in training data and 51.8% in testing data. While this represents moderate predictive capability, the performance degradation between training and testing phases indicates some overfitting challenges. The error metrics reveal that most predictions fall within acceptable ranges, though occasional large prediction errors (maximum error of 9.35 on test data) highlight the model's limitations in extreme cases. The predicted versus actual accuracy plots illustrate that the model performs most reliably at the extremes of the accuracy spectrum but struggles with moderate accuracy predictions in the 55-62% range. This pattern suggests that while the model captures general trends in the relationship between training parameters and performance, it may benefit from additional feature engineering or alternative modeling approaches to improve prediction precision.

REFERENCES

[1]. Das, Sumit, Aritra Dey, Akash Pal, and Nabamita Roy. "Applications of artificial intelligence in machine learning: review and prospect." International Journal of Computer Applications 115, no. 9 (2015).

[2]. Khanzode, Ku Chhaya A., and Ravindra D. Sarode. "Advantages and disadvantages of artificial intelligence and machine learning: A literature review." International Journal of Library & Information Science (IJLIS) 9, no. 1 (2020): 3.

[3]. Kühl, Niklas, Max Schemmer, Marc Goutier, and Gerhard Satzger. "Artificial intelligence and machine learning." Electronic Markets 32, no. 4 (2022): 2235-2244.

[4]. Alimadadi, Ahmad, Sachin Aryal, Ishan Manandhar, Patricia B. Munroe, Bina Joe, and Xi Cheng. "Artificial intelligence and machine learning to fight COVID-19." Physiological genomics 52, no. 4 (2020): 200-202.

[5]. Entezari, Ashkan, Alireza Aslani, Rahim Zahedi, and Younes Noorollahi. "Artificial intelligence and machine learning in energy systems: A bibliographic perspective." Energy Strategy Reviews 45 (2023): 101017.

[6]. Cioffi, Raffaele, Marta Travaglioni, Giuseppina Piscitelli, Antonella Petrillo, and Fabio De Felice. "Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions." Sustainability 12, no. 2 (2020): 492.

[7]. Ullah, Zaib, Fadi Al-Turjman, Leonardo Mostarda, and Roberto Gagliardi. "Applications of artificial intelligence and machine learning in smart cities." Computer communications 154 (2020): 313-323.

[8]. Winkler, David A. "Role of artificial intelligence and machine learning in nanosafety." Small 16, no. 36 (2020): 2001883.

[9]. Michalski, Ryszard Stanislaw, Jaime Guillermo Carbonell, and Tom M. Mitchell, eds. Machine learning: An artificial intelligence approach. Springer Science & Business Media, 2013.

[10]. Michalski, Ryszard Stanislaw, Jaime Guillermo Carbonell, and Tom M. Mitchell, eds. Machine learning: An artificial intelligence approach. Springer Science & Business Media, 2013.

[11]. Becker, Jan U., David Mayerich, Meghana Padmanabhan, Jonathan Barratt, Angela Ernst, Peter Boor, Pietro A. Cicalese, Chandra Mohan, Hien V. Nguyen, and Badrinath Roysam. "Artificial intelligence and machine learning in nephropathology." Kidney International 98, no. 1 (2020): 65-75.

[12]. Connor, Christopher W. "Artificial intelligence and machine learning in anesthesiology." Anesthesiology 131, no. 6 (2019): 1346.

[13]. Kumar, Indrajeet, Jyoti Rawat, Noor Mohd, and Shahnawaz Husain. "Opportunities of artificial intelligence and machine learning in the food industry." Journal of Food Quality 2021, no. 1 (2021): 4535567.

[14]. Guo, Kai, Zhenze Yang, Chi-Hua Yu, and Markus J. Buehler. "Artificial intelligence and machine learning in design of mechanical materials." Materials Horizons 8, no. 4 (2021): 1153-1172.

[15]. Patel, Veer, and Manan Shah. "Artificial intelligence and machine learning in drug discovery and development." Intelligent Medicine 2, no. 3 (2022): 134-140.

[16]. Koh, Dow-Mu, Nickolas Papanikolaou, Ulrich Bick, Rowland Illing, Charles E. Kahn Jr, Jayshree Kalpathi-Cramer, Celso Matos et al. "Artificial intelligence and machine learning in cancer imaging." Communications Medicine 2, no. 1 (2022): 133.

